# CSE 5559 – Translational Bioinformatics Week 5

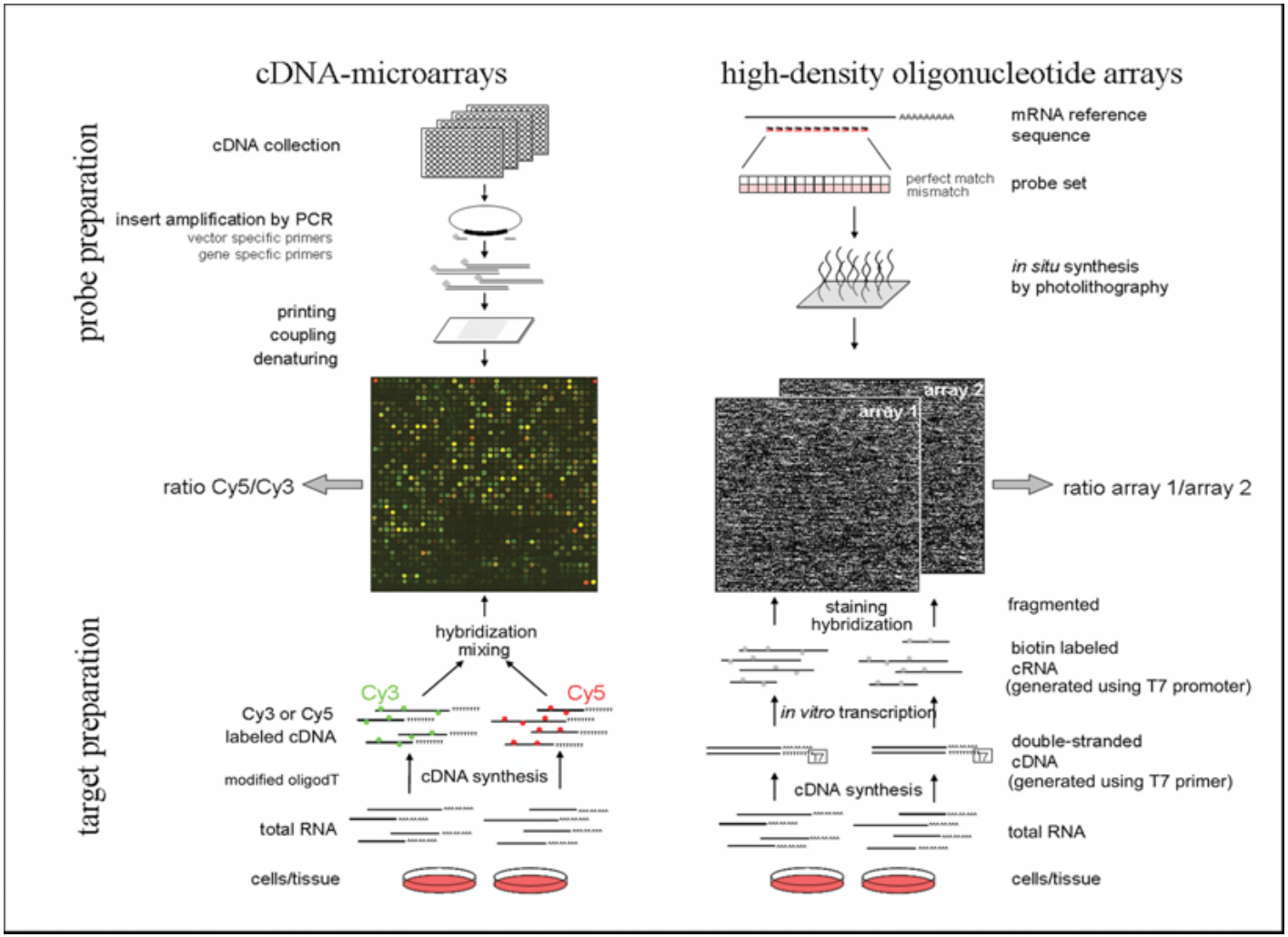
Wexner Medical Center

# Two major technologies

- Affymetrix-like arrays
  – **single channel** (background-green, foreground-red)

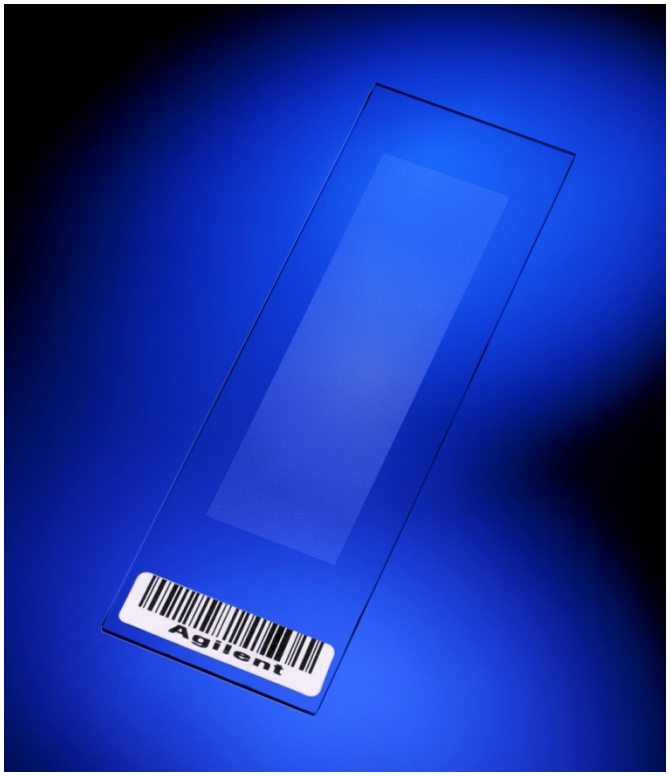
- cDNA arrays
  - – **two channel** (red, green, yellow)

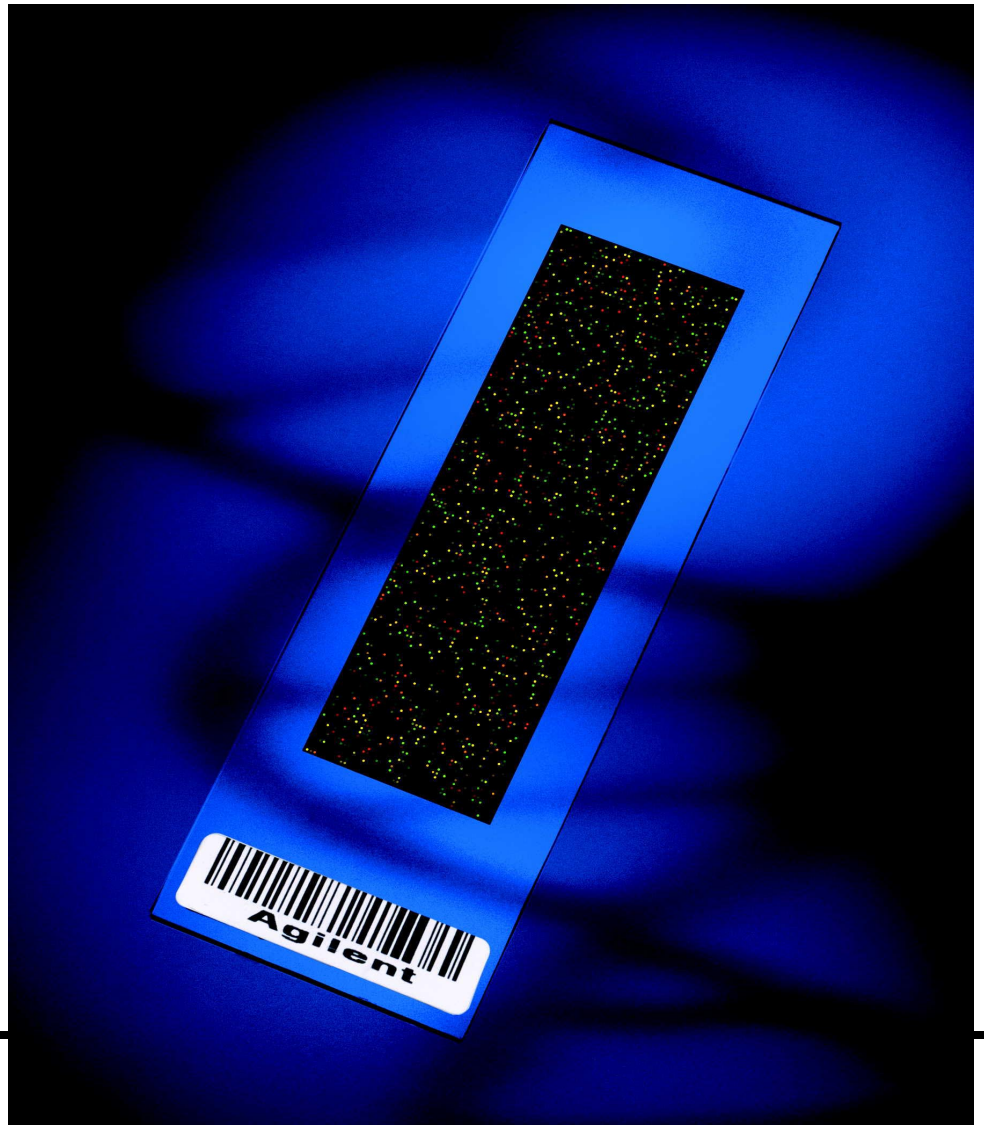
# Both


cDNA-microarrays
high-density oligonucleotide arrays
probe preparation
target preparation
cDNA collection
insert amplification by PCR
vector specific primers
gene specific primers
printing
coupling
denaturing
ratio Cy5/Cy3
hybridization
mixing
Cy3
Cy5
Cy3 or Cy5
labeled cDNA
modified oligodT
cDNA synthesis
total RNA
cells/tissue
mRNA reference
sequence
perfect match
mismatch
probe set
in situ synthesis
by photolithography
array 2
array 1
ratio array 1/array 2
staining
hybridization
fragmented
biotin labeled
cRNA
(generated using T7 promoter)
in vitro transcription
double-stranded
cDNA
(generated using T7 primer)
cDNA synthesis
total RNA
cells/tissue

# Reading Material – Chapter 3, Text

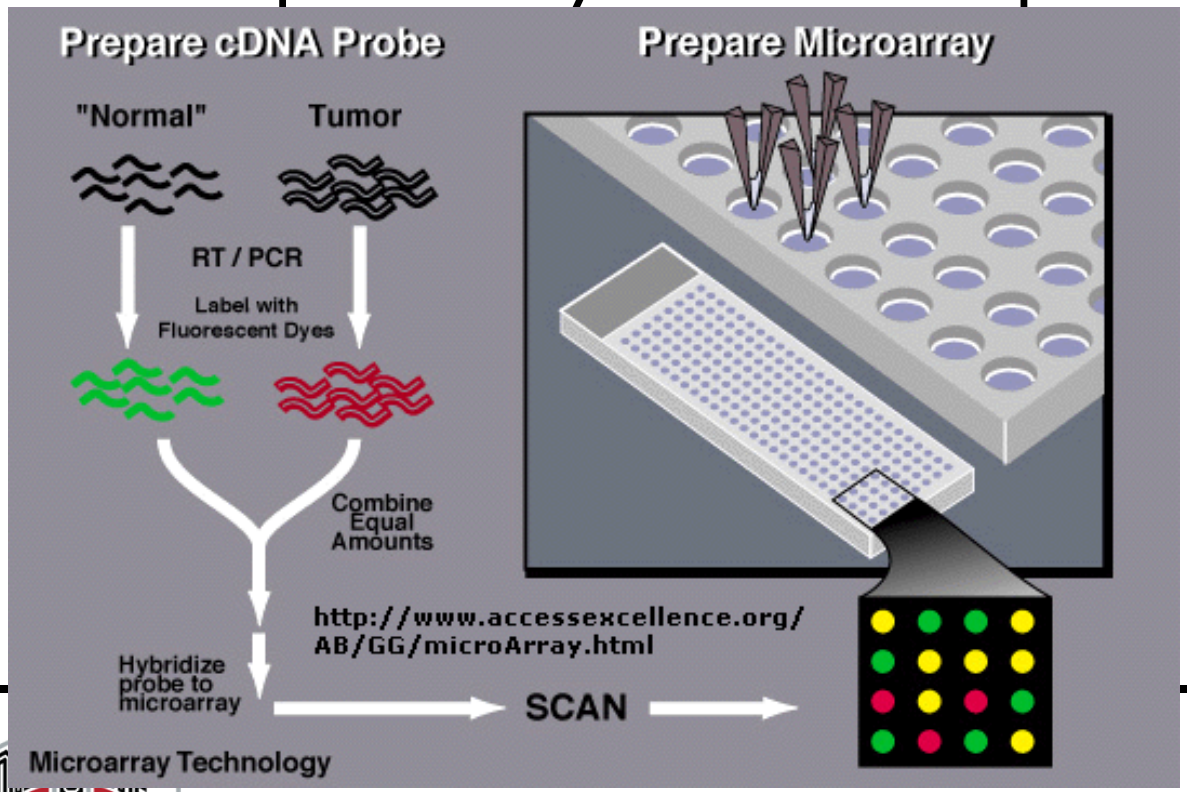
# Two-color CDNA

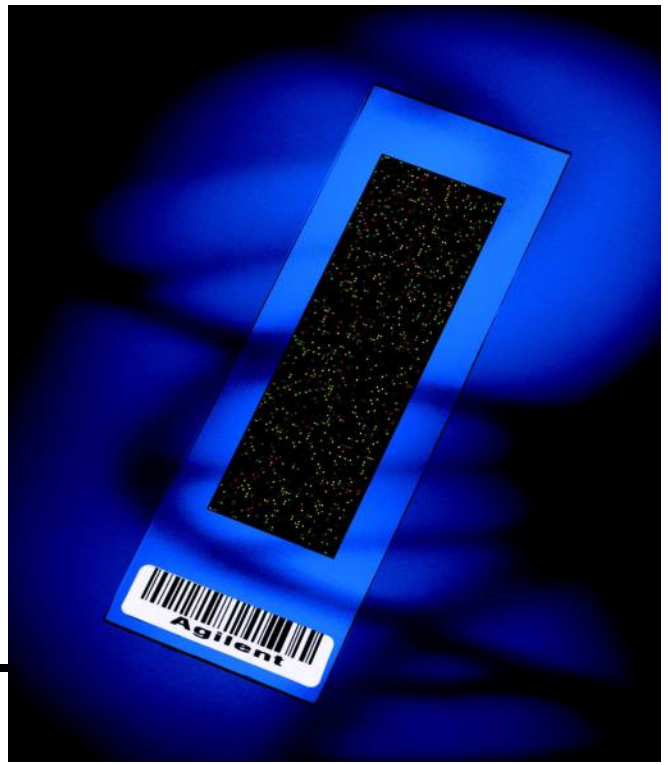


# Two-channel microarray

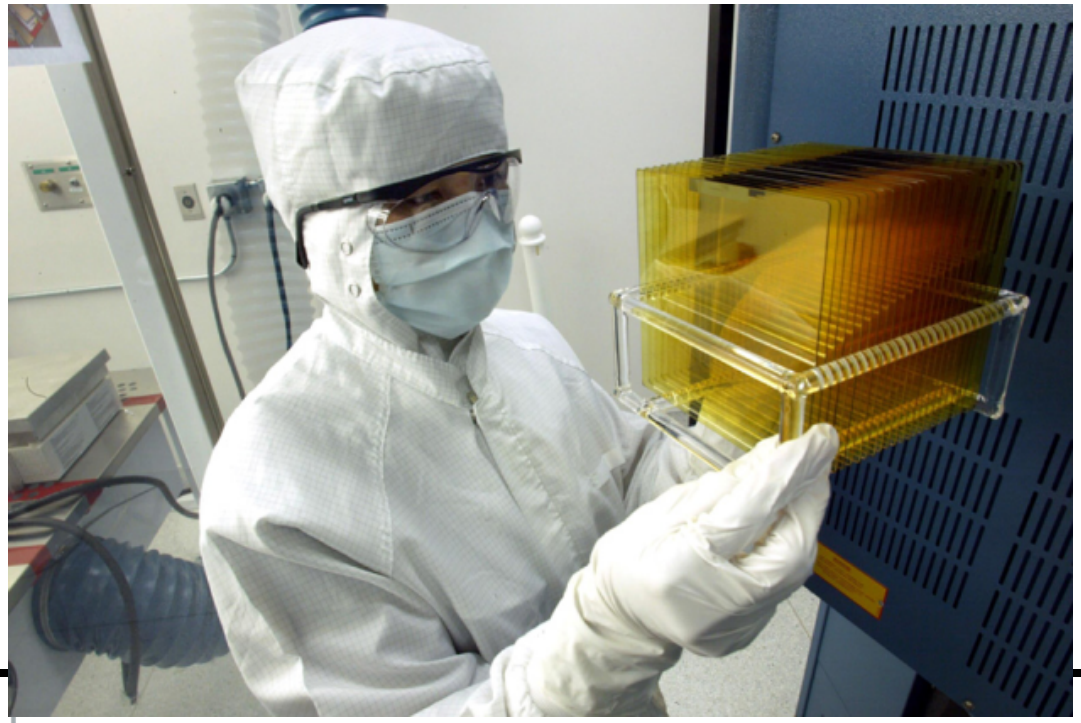
- Printed microarrays
- Long probe oligonucleotides (80-100) long are "printed" on the glass chip
- Comparative hybridization experiment



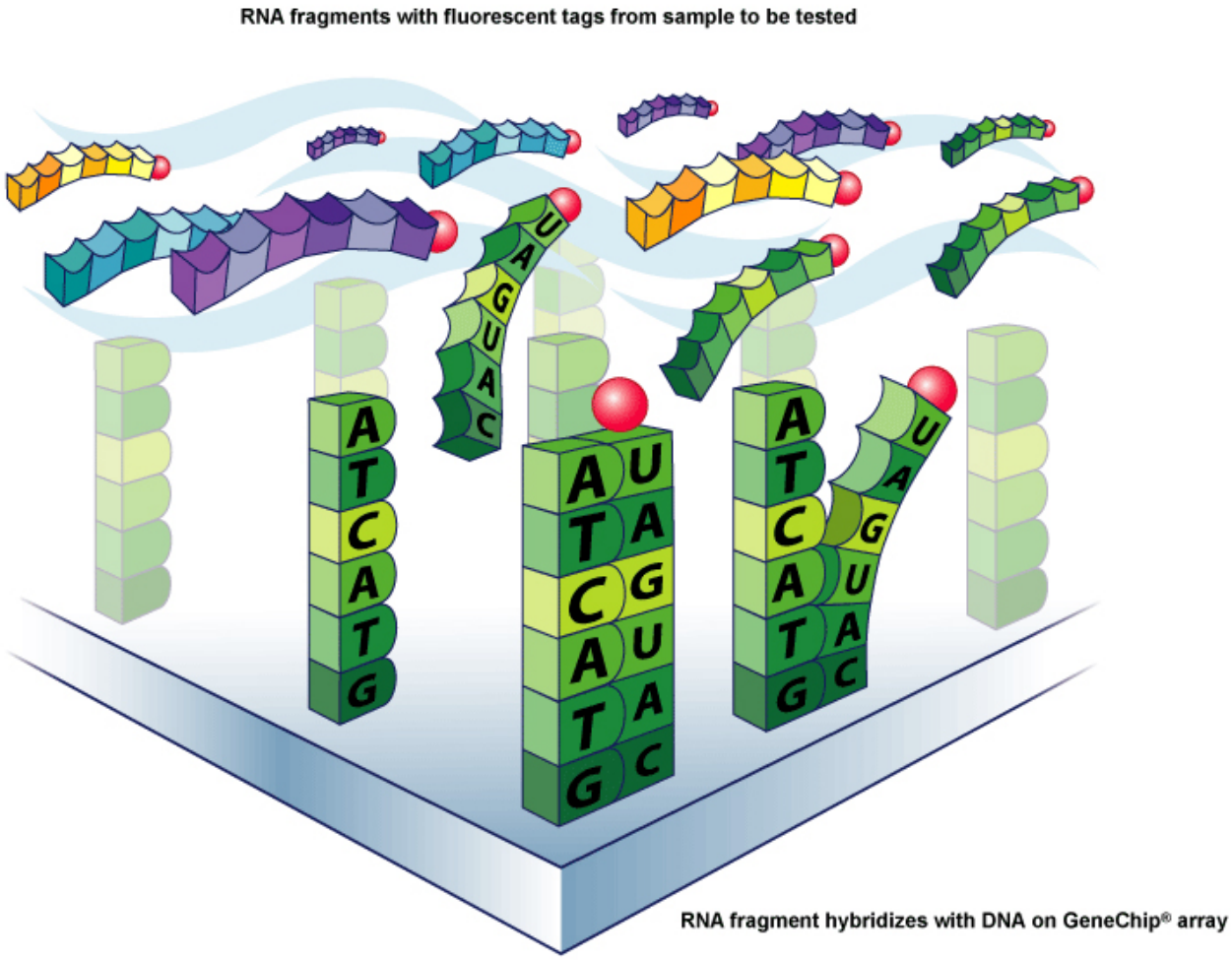



# Affymetrix GeneChip

- silicon chip
- oligonucleiotide probes lithographically synthesized on the array



# Matching


RNA fragments with fluorescent tags from sample to be tested
RNA fragment hybridizes with DNA on GeneChip® array

# Photolithography

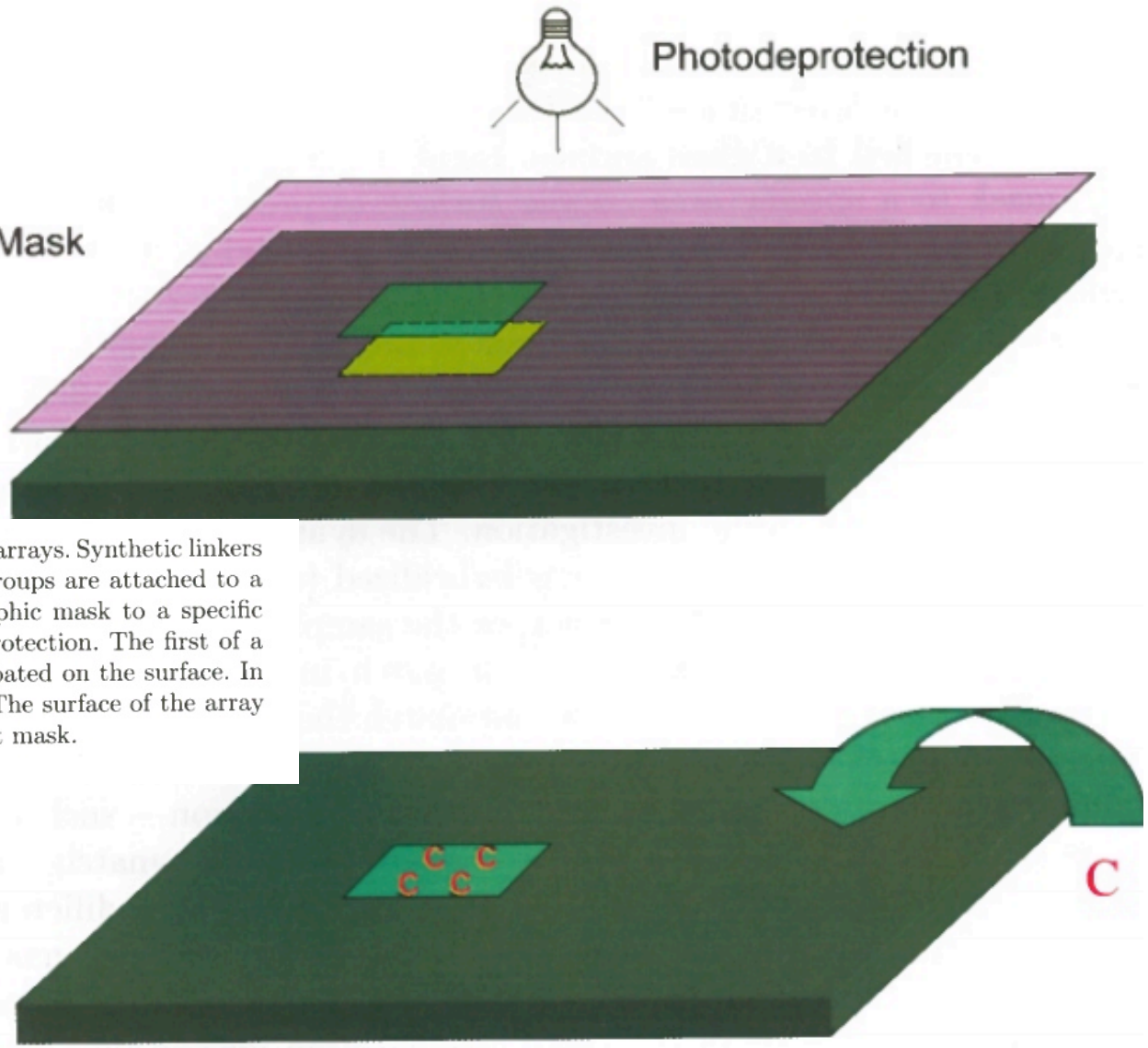
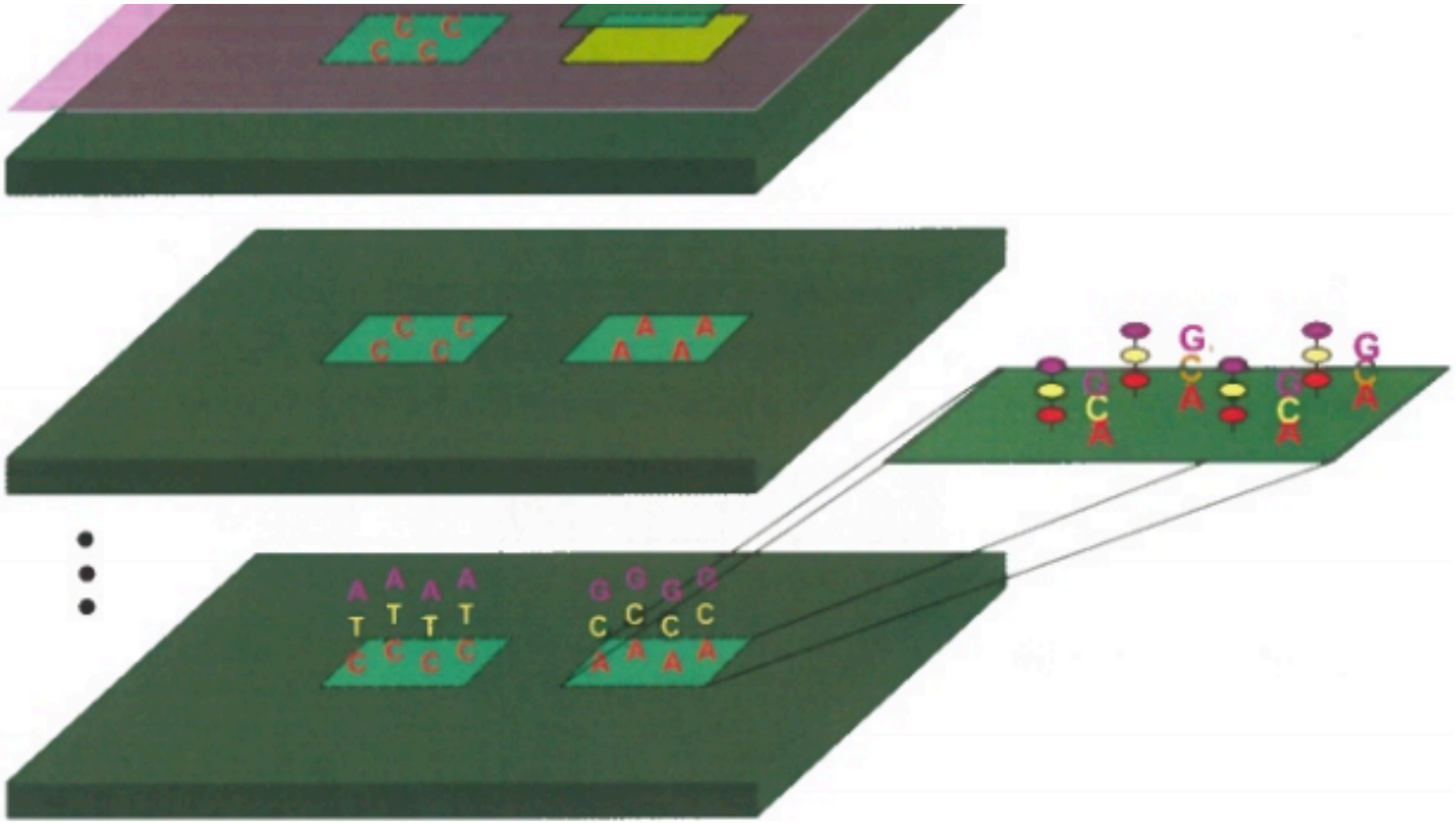


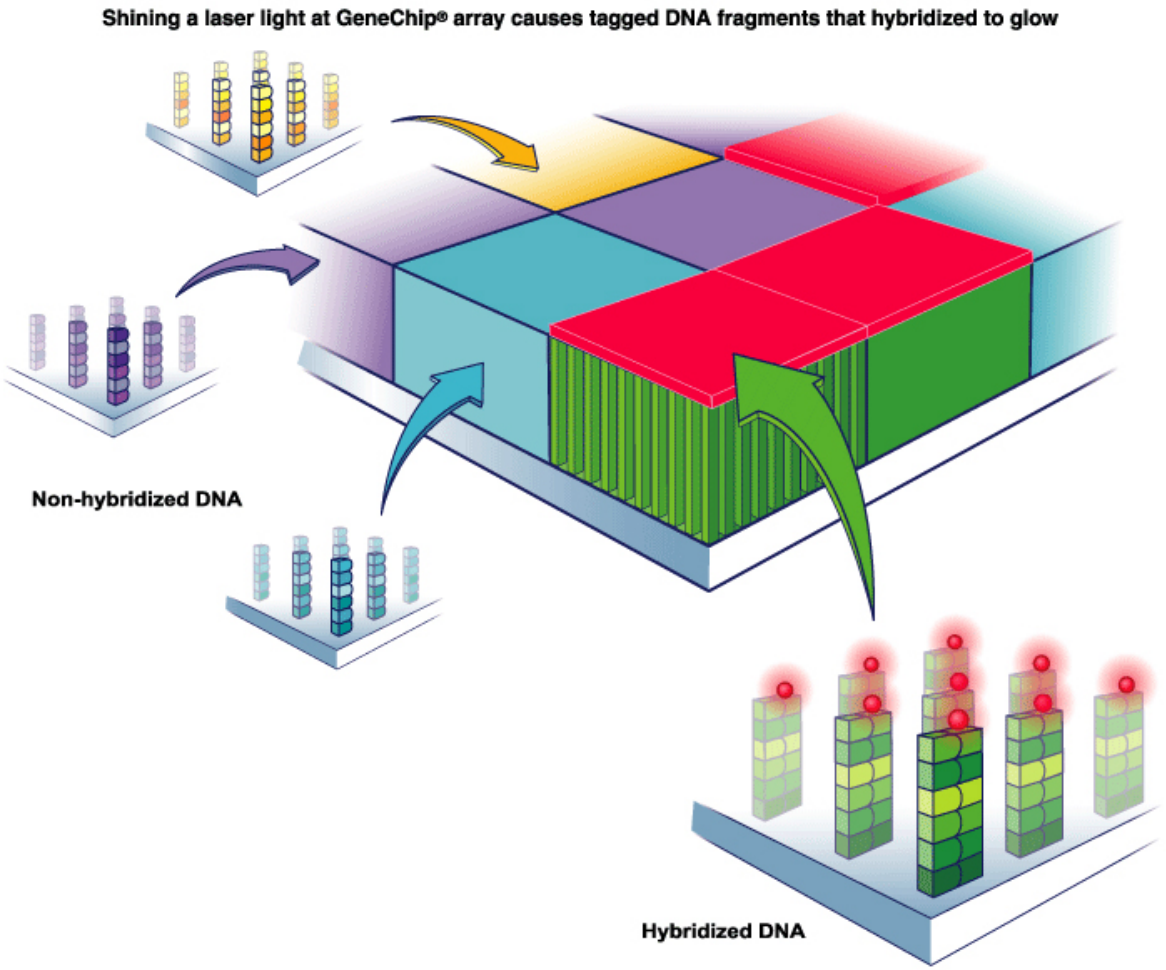

**FIGURE 3.9**: Photolithographic fabrication of microarrays. Synthetic linkers modified with photochemical removable protecting groups are attached to a glass surface. Light is shed through a photolithographic mask to a specific area on the surface to produce a localized photodeprotection. The first of a series of hydroxyl-protected deoxynucleosides is incubated on the surface. In this example, it is the protected deoxynucleoside C. The surface of the array is protected again, and the array is ready for the next mask.

# Photolithography

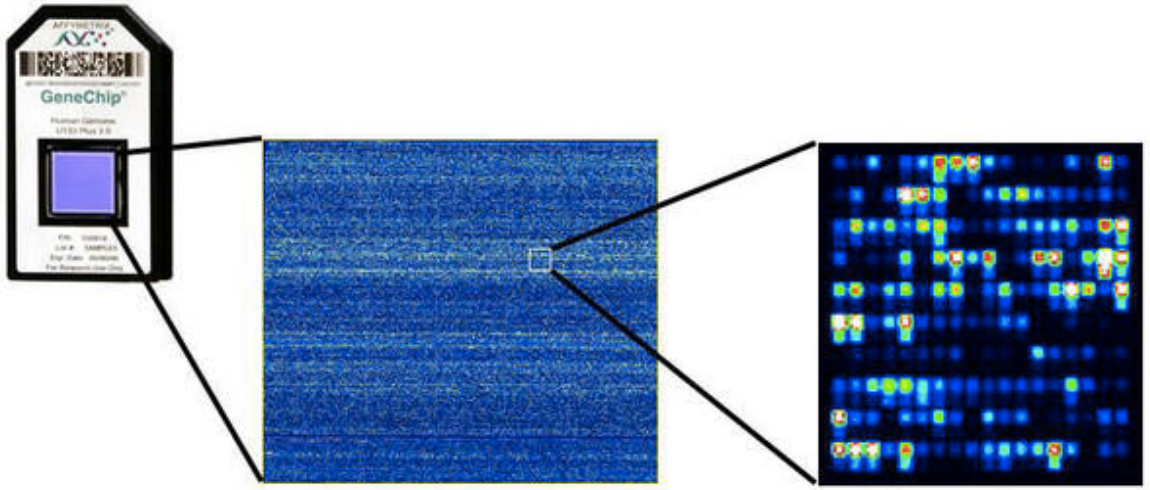
**FIGURE 3.10**: Photolithographic fabrications of microarrays. The sec
mask is applied and light is used to deprotect the areas that are desig
to receive the next nucleoside (A). The fabrication process would gener
require 4 masking steps for each element of the probes. Several steps la
each area has its own sequence as designed.



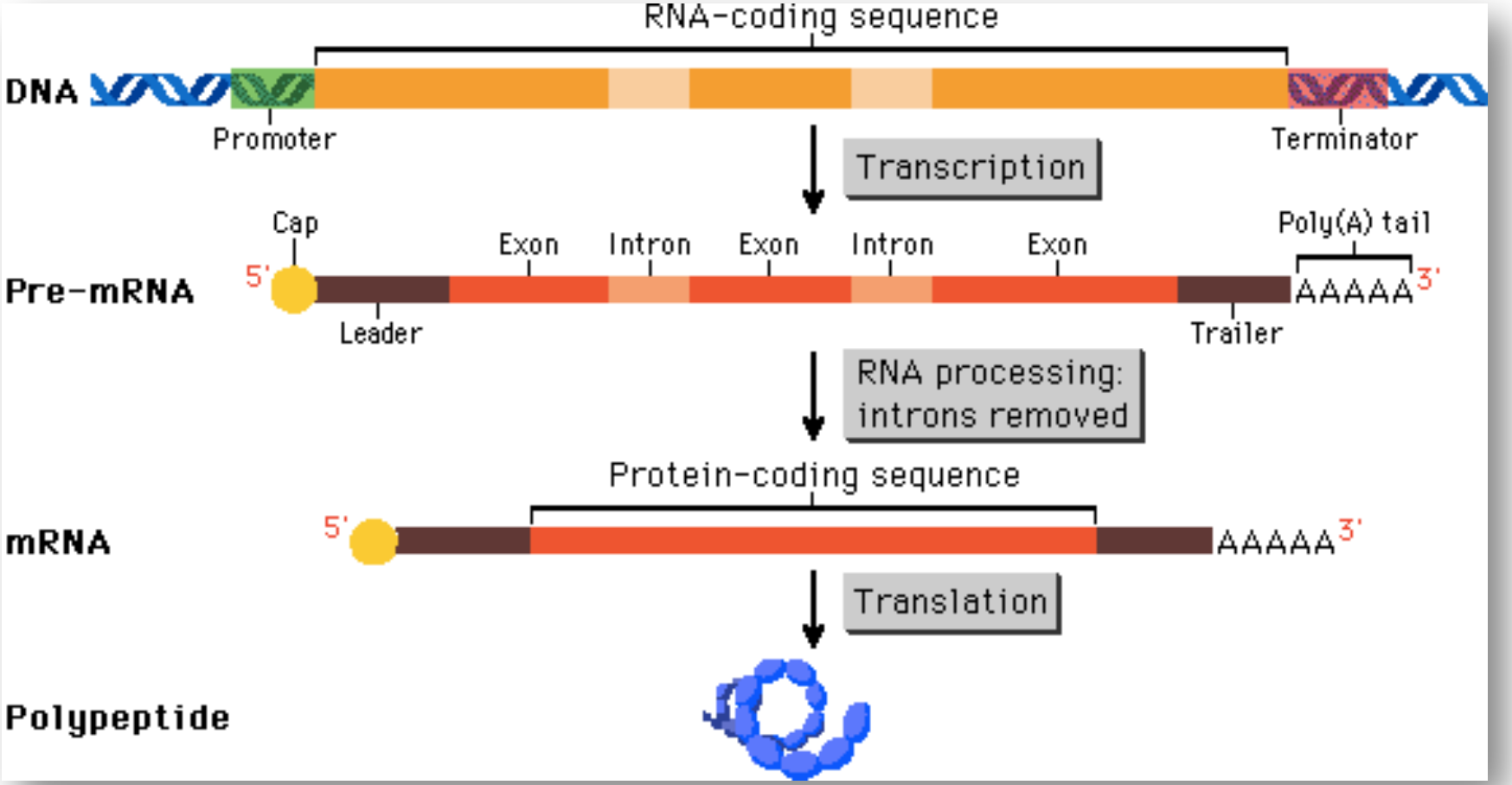
# Fluorescence



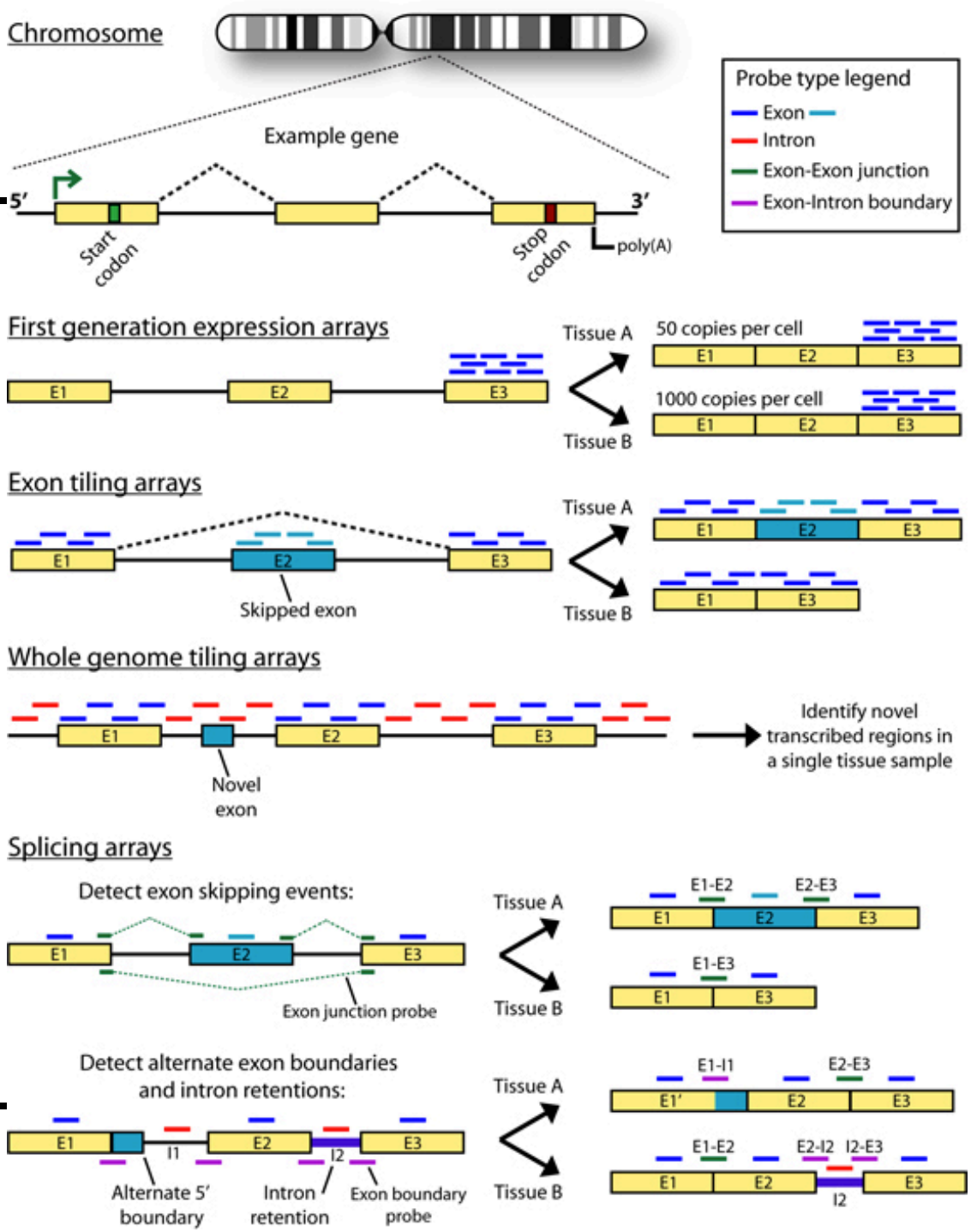
# Image



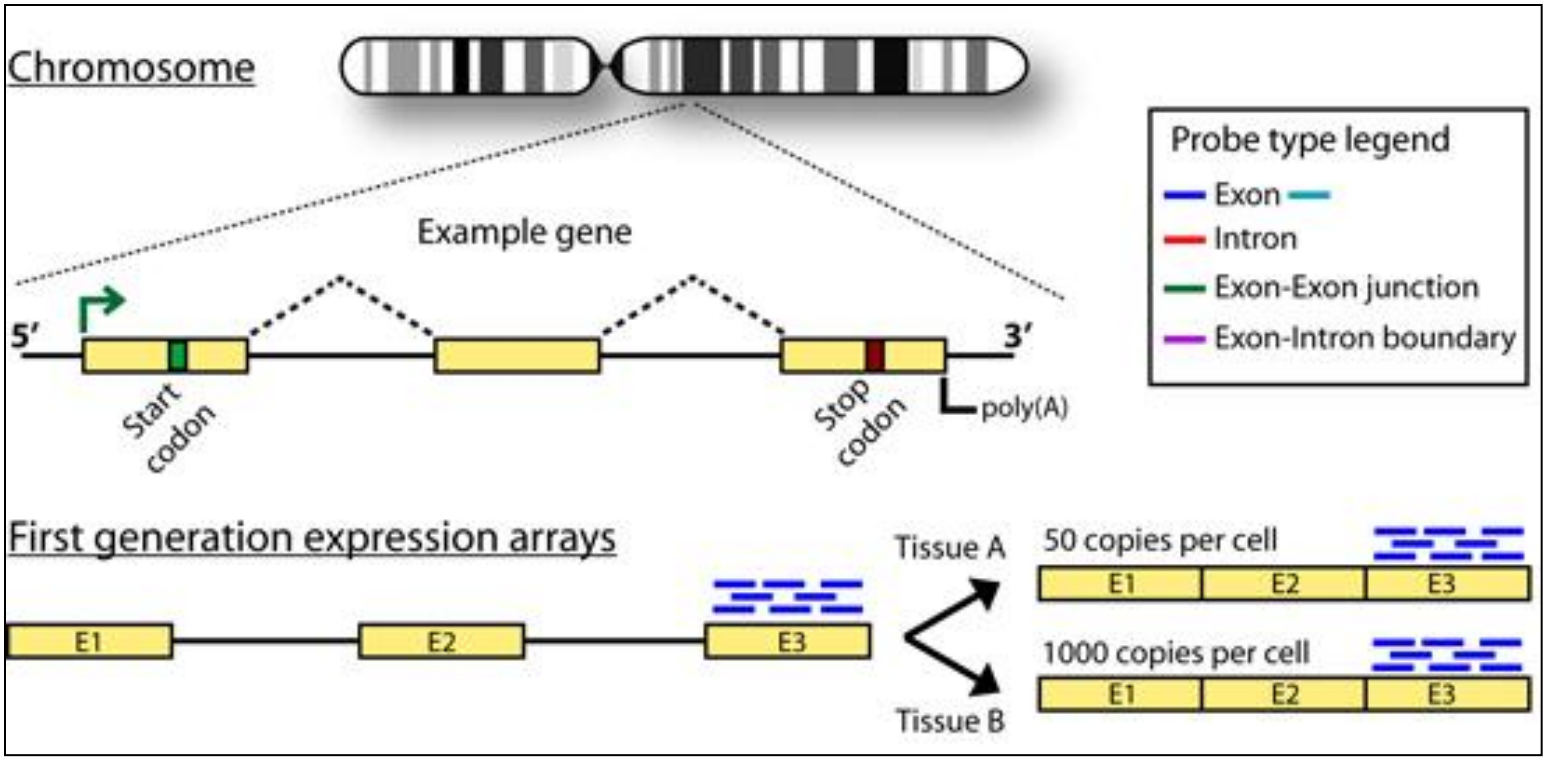
# Recap


RNA-coding sequence
DNA
Promoter
Terminator
Transcription
Cap
Exon
Intron
Exon
Intron
Exon
Poly(A) tail
Pre-mRNA
5'
AAAAA
3'
Leader
Trailer
RNA processing:
introns removed
Protein-coding sequence
mRNA
5'
AAAAA
3'
Translation
Polypeptide

# Probes

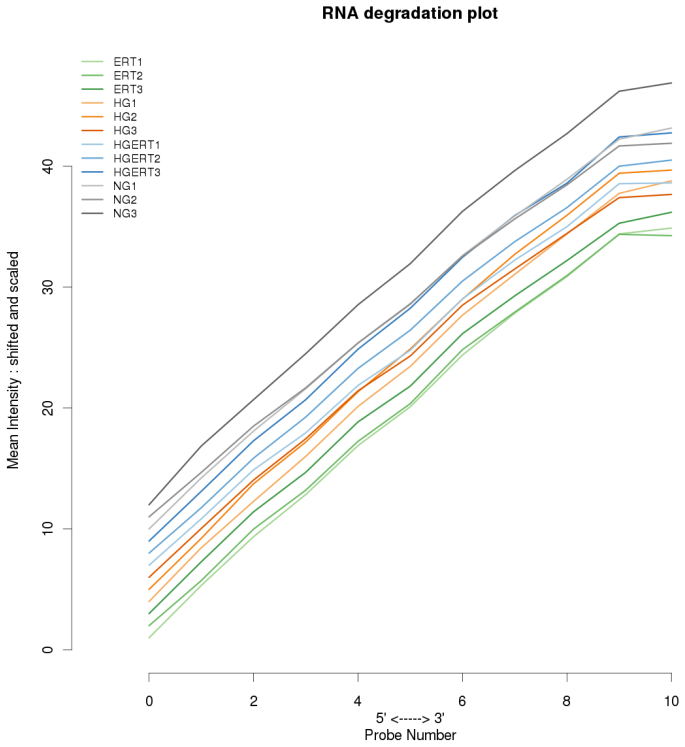

Chromosome
Example gene
5′
3′
Start codon
Stop codon
poly(A)
Probe type legend
Exon
Intron
Exon-Exon junction
Exon-Intron boundary
First generation expression arrays
E1
E2
E3
Tissue A
Tissue B
50 copies per cell
1000 copies per cell
Exon tiling arrays
Skipped exon
Whole genome tiling arrays
Novel exon
Identify novel transcribed regions in a single tissue sample
Splicing arrays
Detect exon skipping events:
Exon junction probe
E1-E2
E2-E3
E1-E3
Detect alternate exon boundaries and intron retentions:
I1
I2
Alternate 5′ boundary
Intron retention
Exon boundary probe
E1-I1
E1′
E1-E2
E2-I2
I2-E3

# Probe Selection



# Probe selection

- ✓ 3’ bias – why ? Think degradation
- ✓ Multiple probes for one region
- ✓ G-C content - Think binding




**http://www.bcgsc.ca/people/malachig/htdocs/alexa_platform/alexa_arrays/intro.htm**

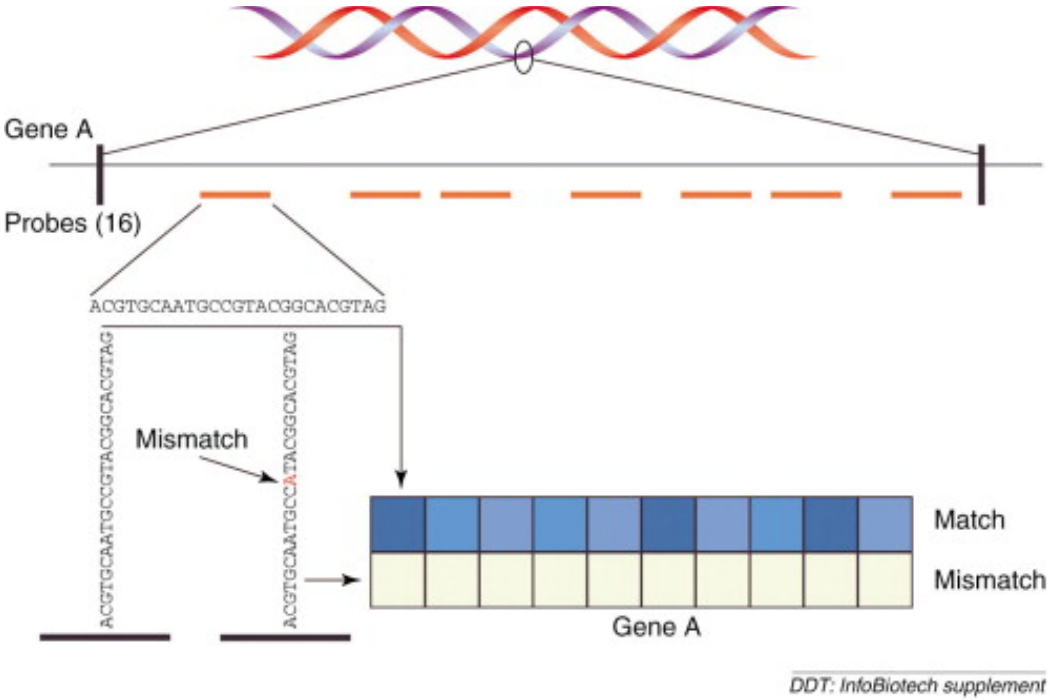
# Worth Visiting

http://www.arrayanalysis.org/

# Probe Design

- Fabrication expense and frequency of error increases with probe length - 25 oligonucleotide probes are employed
- Problem: cross hybridization
- Solution: introduce mismatched probe with one position (central) different with the matched probe. The difference gives a more accurate reading.

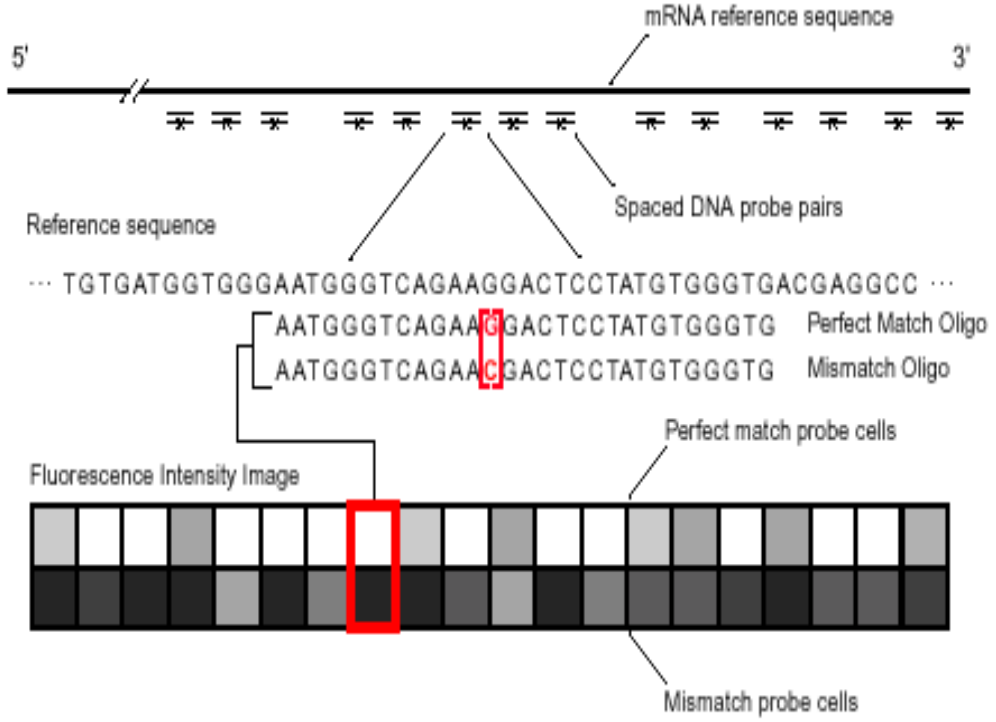
# Probe Design




The principle of the Affymetrix technology. The probes correspond to short oligonucleotide sequences thought to be representative for a given gene. Each oligonucleotide sequence is represented by two probes: one with the exact sequence of the chosen fragment of the gene (perfect match or PM), and one with a mismatch nucleotide in the middle of the fragment (mismatch or MM). The hybridization occurs at a critical temperature – a single mismatch can totally inhibit hybridization. For each gene, the value that is usually taken as representative for the expression level of the gene is the average difference between PM and MM.

# Affymetrix Genechips

- Each gene represented by 11-20 '*probe pairs*'
- Probe pairs are 3' biased.
- '*Probe Pair*' consists of Perfect Match (PM) and MisMatch (MM) probes.
- MM has altered middle (13th) base. Designed to measure non-specific binding (NSB).



# Actually

Problem: cross hybridization

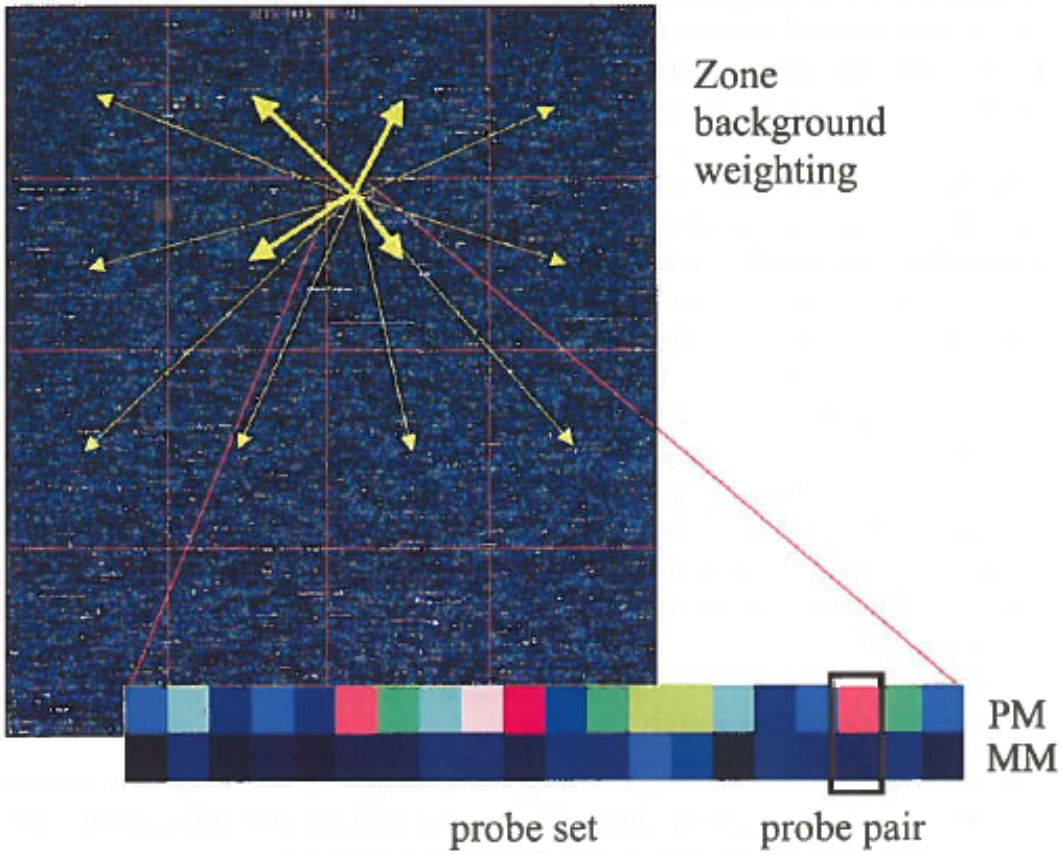
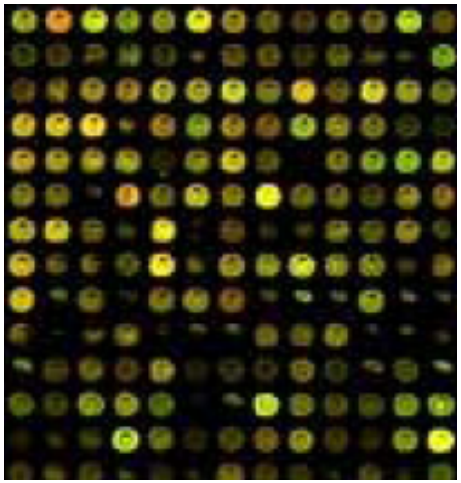



**FIGURE 20.10**: Affymetrix terminology and background zone weighting computation. A gene is represented by a number of perfect match (PM) and mismatch (MM) probe pairs. Each pair is composed of two cells. The background value is computed as a weighted sum of the nearby zone background values. The weight of a zone is inversely proportional to the distance to it. The background value for a given zone is calculated from the cells with the lowest 2% intensities in that zone.

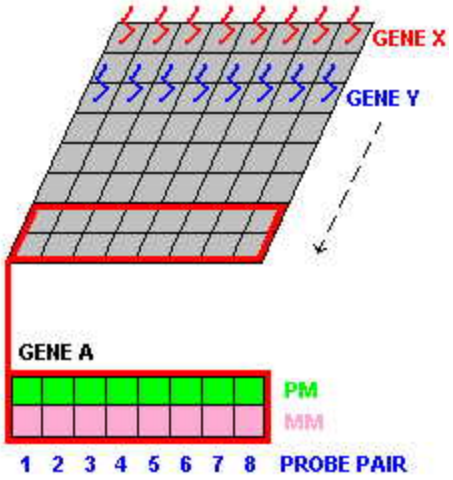
# Reading Data

cDNA array – ratio, log ratio

$$T_i = \frac{R_i}{G_i} \text{ OR } \log \text{ratio} = \log_2 \frac{R_i}{G_i}$$



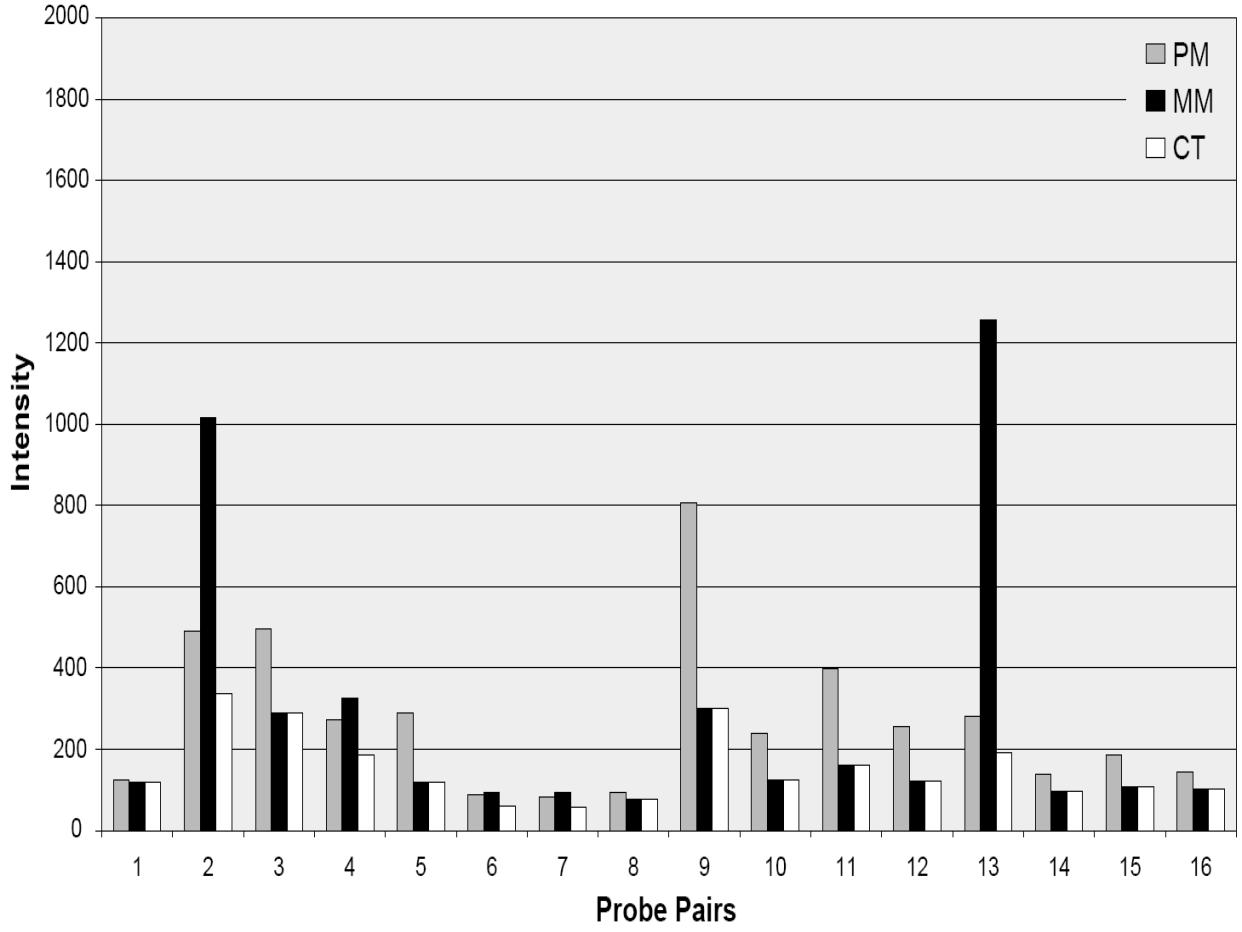
Affymetrix array




$$\text{Difference}_{\text{probe pair}} = PM - MM$$

$$\text{Average Difference}_{\text{probe set}} = \sum_{i=1}^{n} \frac{(PM_i - MM_i)}{n}$$

# Problems with Mis-Match data



# Problems with Mis-Match data

- MM intensity levels are greater than PM intensity levels in ~1/3 of all probes.
- Suggests that MM probes measure actual signal, and not just NSB (non-specific binding).
- Subtracting MM data will result in loss of interesting signal in many probes.
- Several methods have been proposed using only PM data.

# Signal Calculation

- Self-reading – Section 20.4.2, 20.4.3
- Ideal Mismatch: Using MM
- Probe Values
- Detection calls

$$R_i = \frac{PM_i - MM_i}{PM_i + MM_i}$$

$$\text{Present(P)} : p - \text{value} < \alpha_1$$

$$\text{Marginal(M)} : \alpha_1 \le p - \text{value} < \alpha_2$$

$$\text{Absent(A)} : p - \text{value} \ge \alpha_2$$

# The Meaning of A/M/P Calls

| Detection calls exp | Detection calls base | Magnitude of exp/base ratio | Biological meaning | Interesting gene |
|---|---|---|---|---|
| M/A | M/A | meaningless; can be anything | the gene is probably not expressed in either condition | no |
| M/A | P | meaningless; close to zero | gene expressed on the baseline array, not expressed in the experiment | yes |
| P | M/A | meaningless; very large | gene expressed in the experiment, not expressed in the baseline array | yes |
| P | P | meaningful; can be anything | gene expressed on both channels, useful ratio | depends on the ratio |

**TABLE 20.1:** The interplay between the A/M/P calls and the expression values when an experiment array is compared to a baseline array. The only genes that can be safely ignored are those that are absent (A) or only marginally detected (M) in both arrays. All other combinations are or may be interesting from a biological point of view.

# Tiling

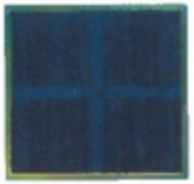
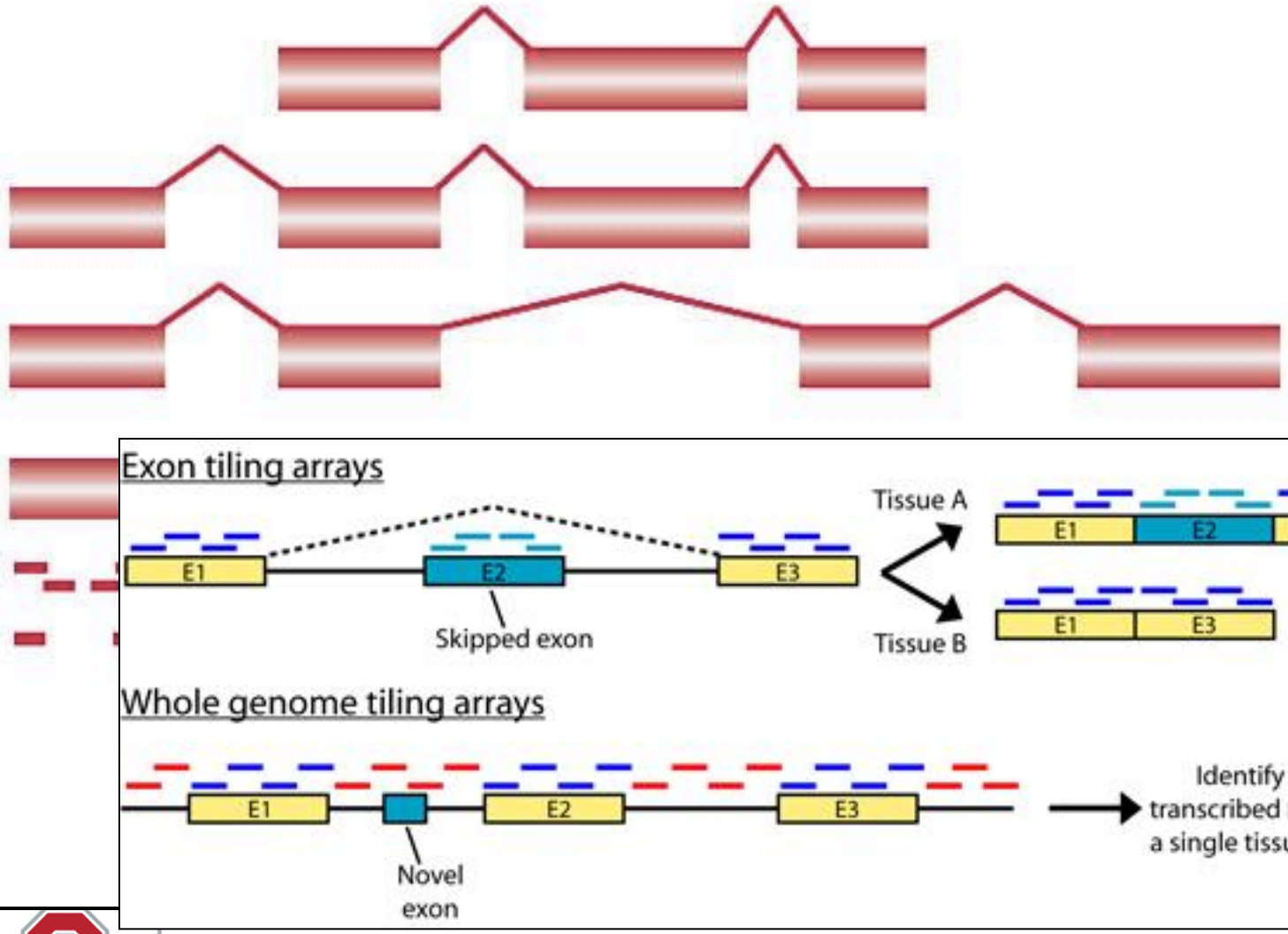
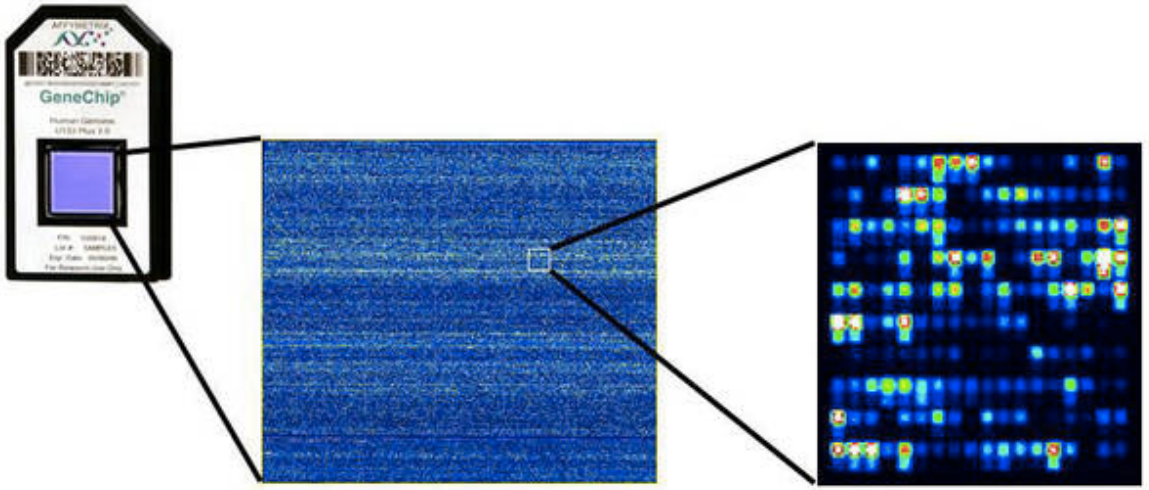


**FIGURE 3.12**: Tiling arrays cover the entire length of the genome after the repetitive elements have been removed. Probes of 25 oligonucleotides are tiled at an average resolution of 35 bps, with an average gap of 10 bps. Some tiling array sets contain both perfect match and mismatch sequences. Others contain only perfect matches.

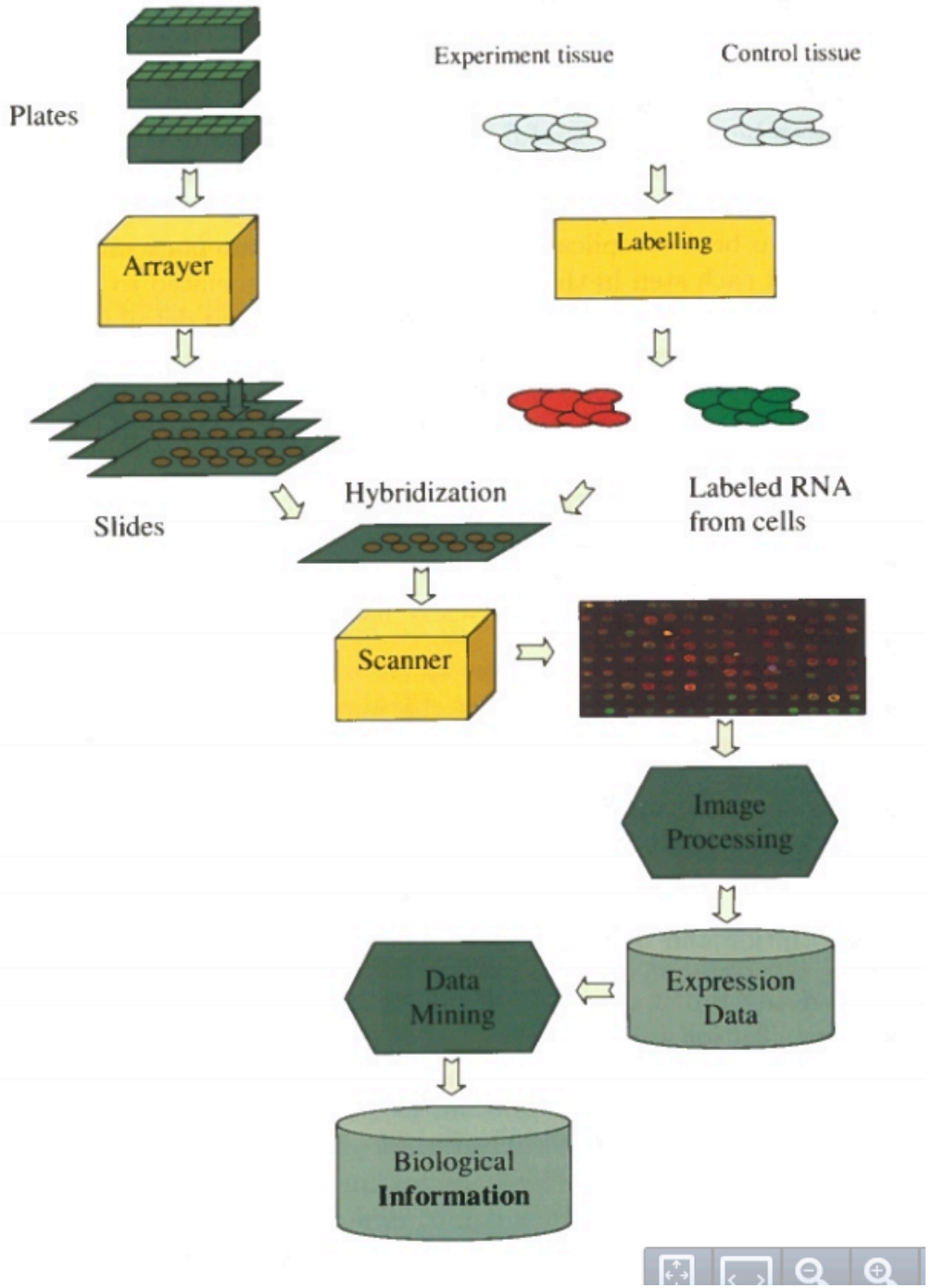
# Tiling


Exon tiling arrays
E1
E2
E3
Skipped exon
Tissue A
Tissue B
E1
E2
E3
E1
E3
Whole genome tiling arrays
E1
E2
E3
Novel exon
Identify novel transcribed regions in a single tissue sample

# Recap - Image



# Pipeline

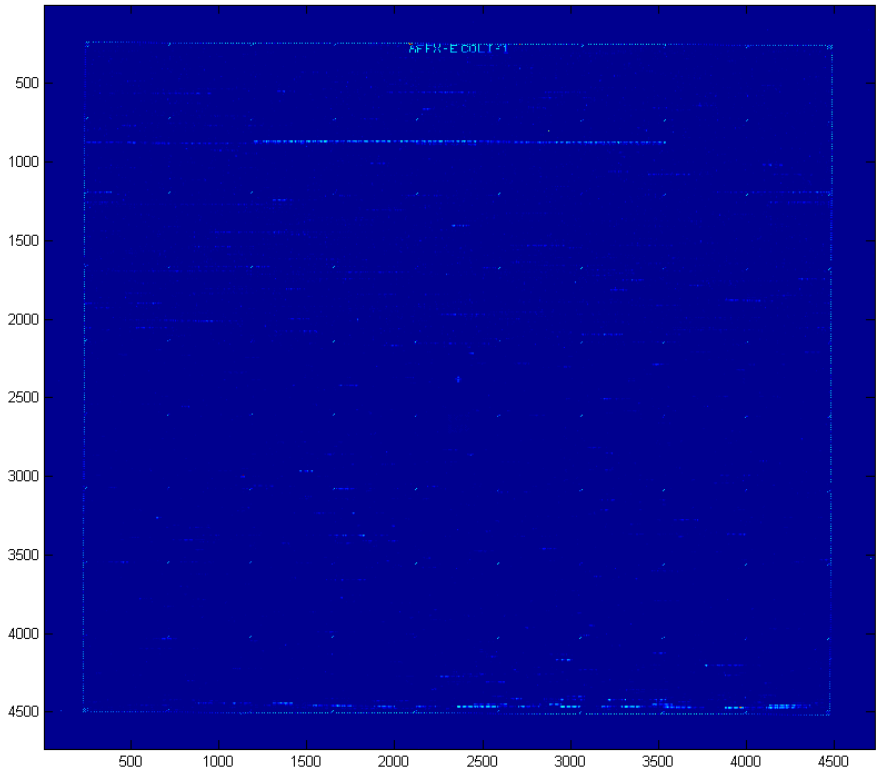

Plates
Arrayer
Slides
Experiment tissue
Control tissue
Labelling
Labeled RNA
from cells
Hybridization
Scanner
Image
Processing
Expression
Data
Data
Mining
Biological
Information

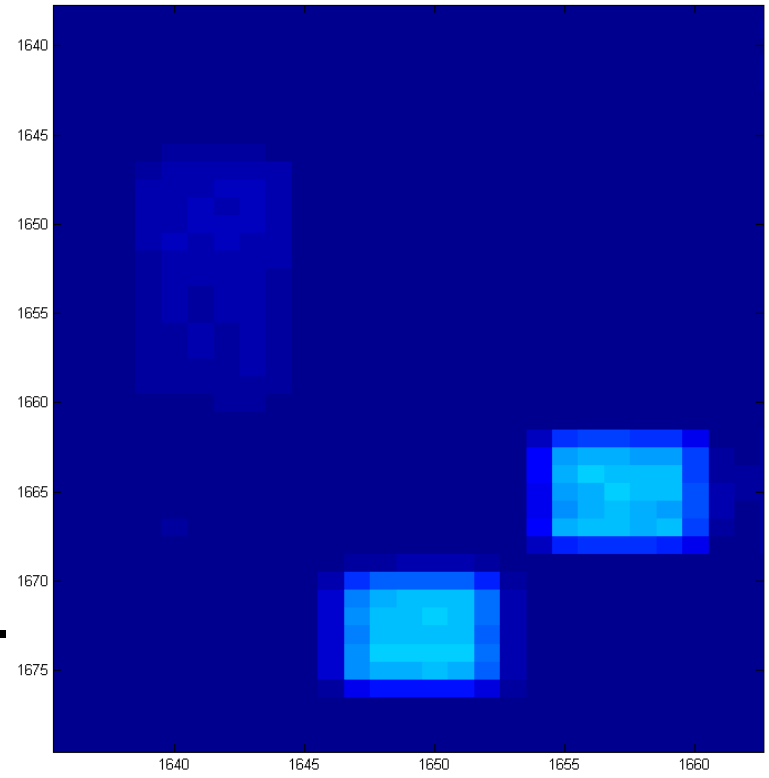
# Genechip Scanning

- RNA sample prepared, labelled and hybridised to chip.
- Chip fluorescently scanned.
- Gives a raw pixelated image - .DAT file.
- Grid used to separate pixels related to individual probes.
- Pixel intensities averaged to give single intensity for each probe - .CEL file.
- Probe level intensities combined for each probe set to give single intensity value for each gene.

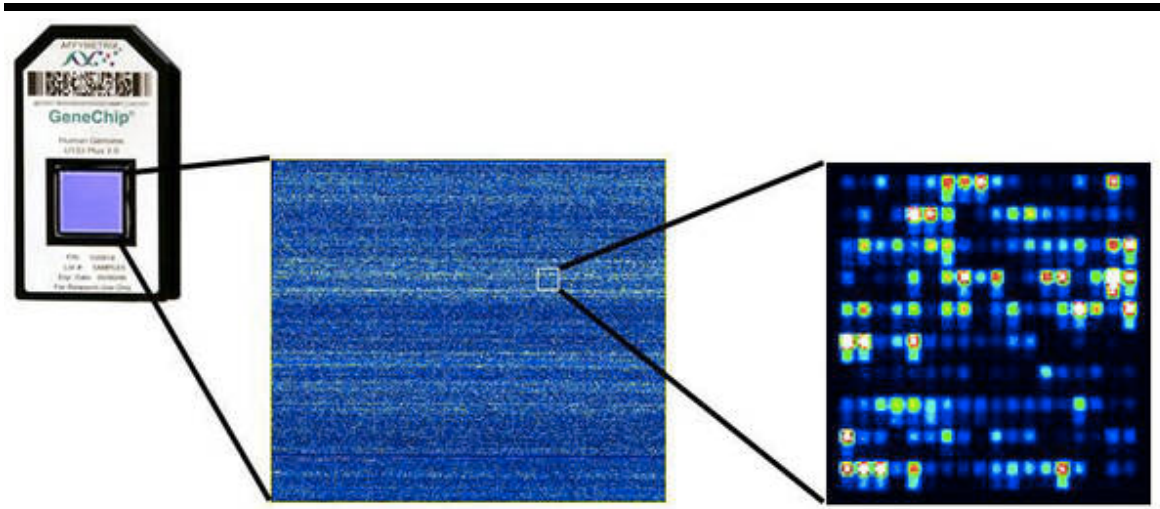
# Affymetrix Data Files

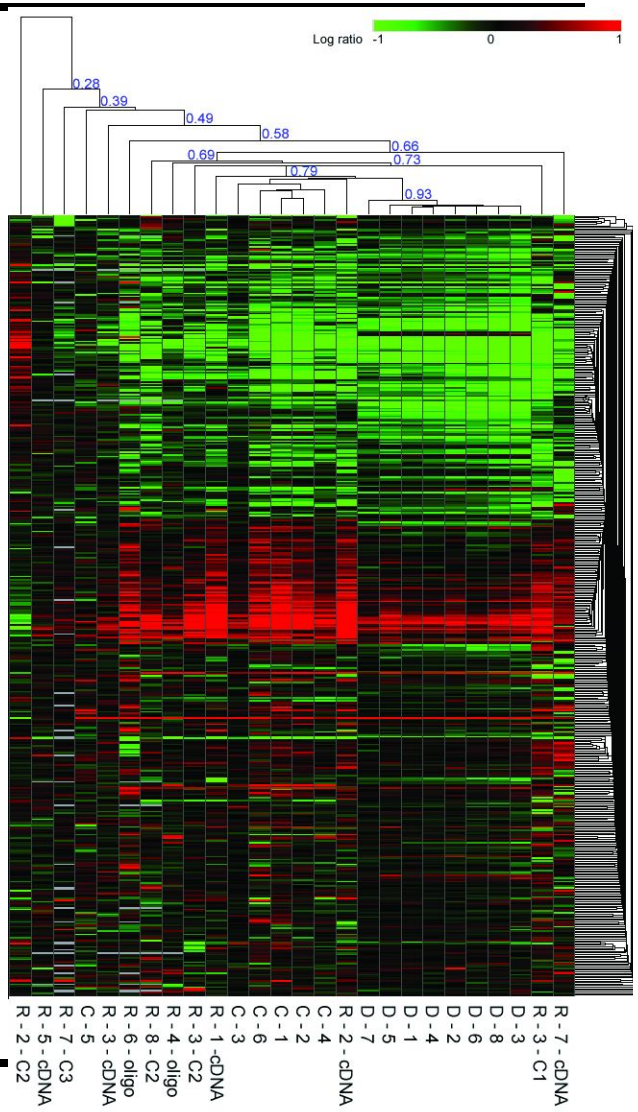
- Image file (.dat file)
- **Probe results file (.cel file)**
- Library file (.cdf, .gin files)
- Results file (.chp file)



# Microarray Data Analysis

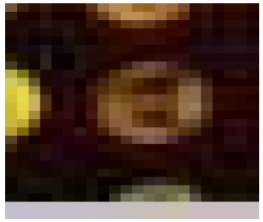



Supplementary Figure 1: Clustering of laboratory/platform combinations using log ratio values of commone genes

# Problems ?

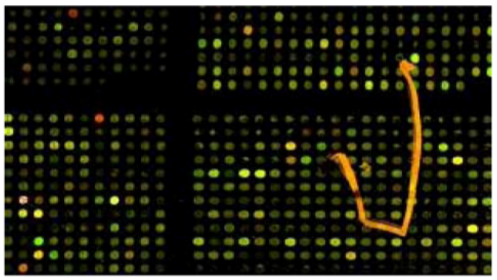
Poorly defined borders



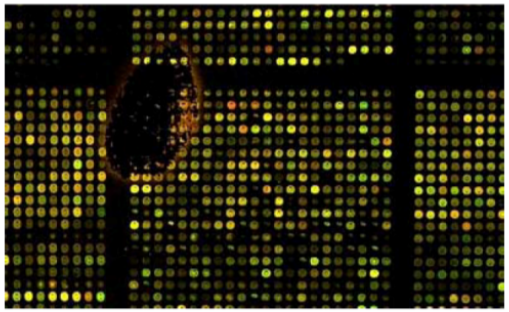
Large holes

Saturated spot

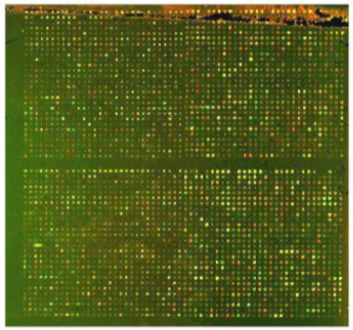
Dust specs



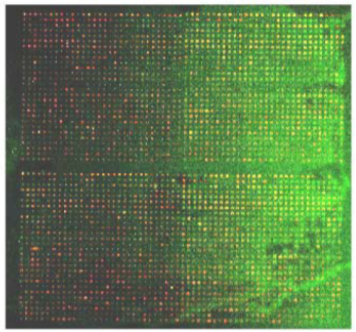
Fiber or scratch?



Bubble



Edge effect

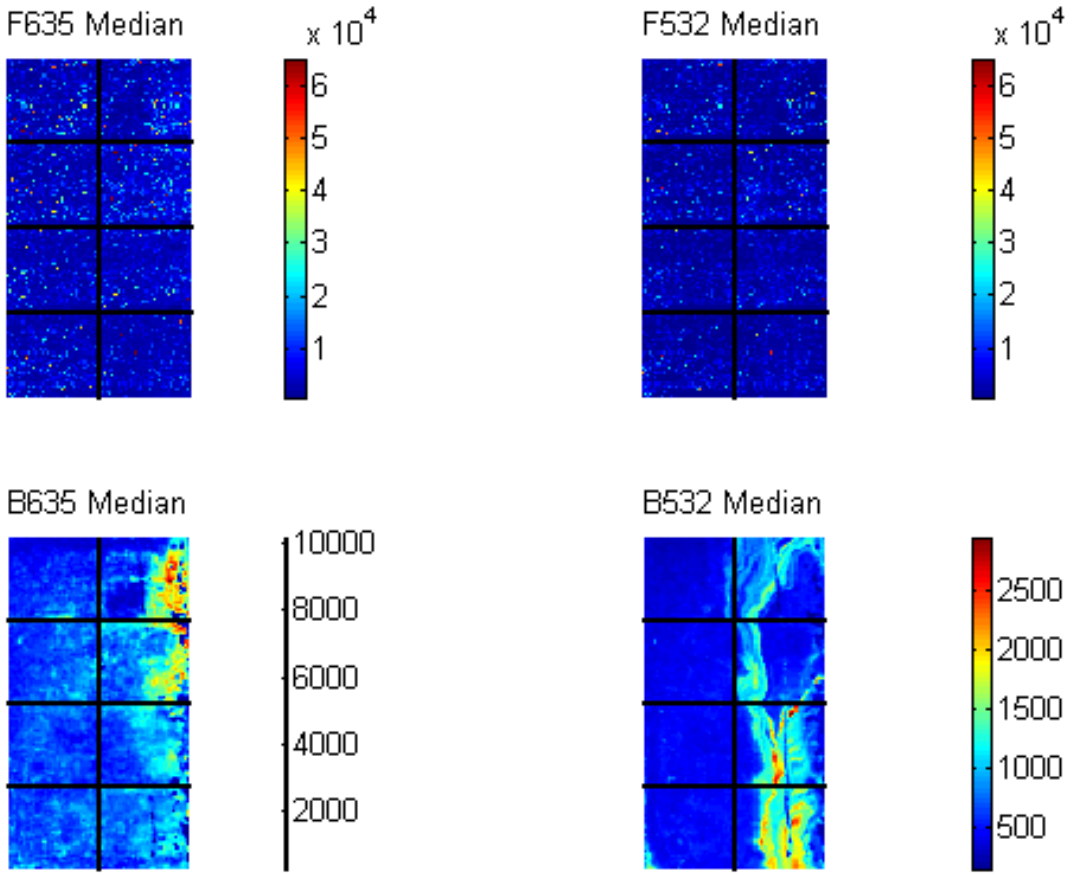


Background haze

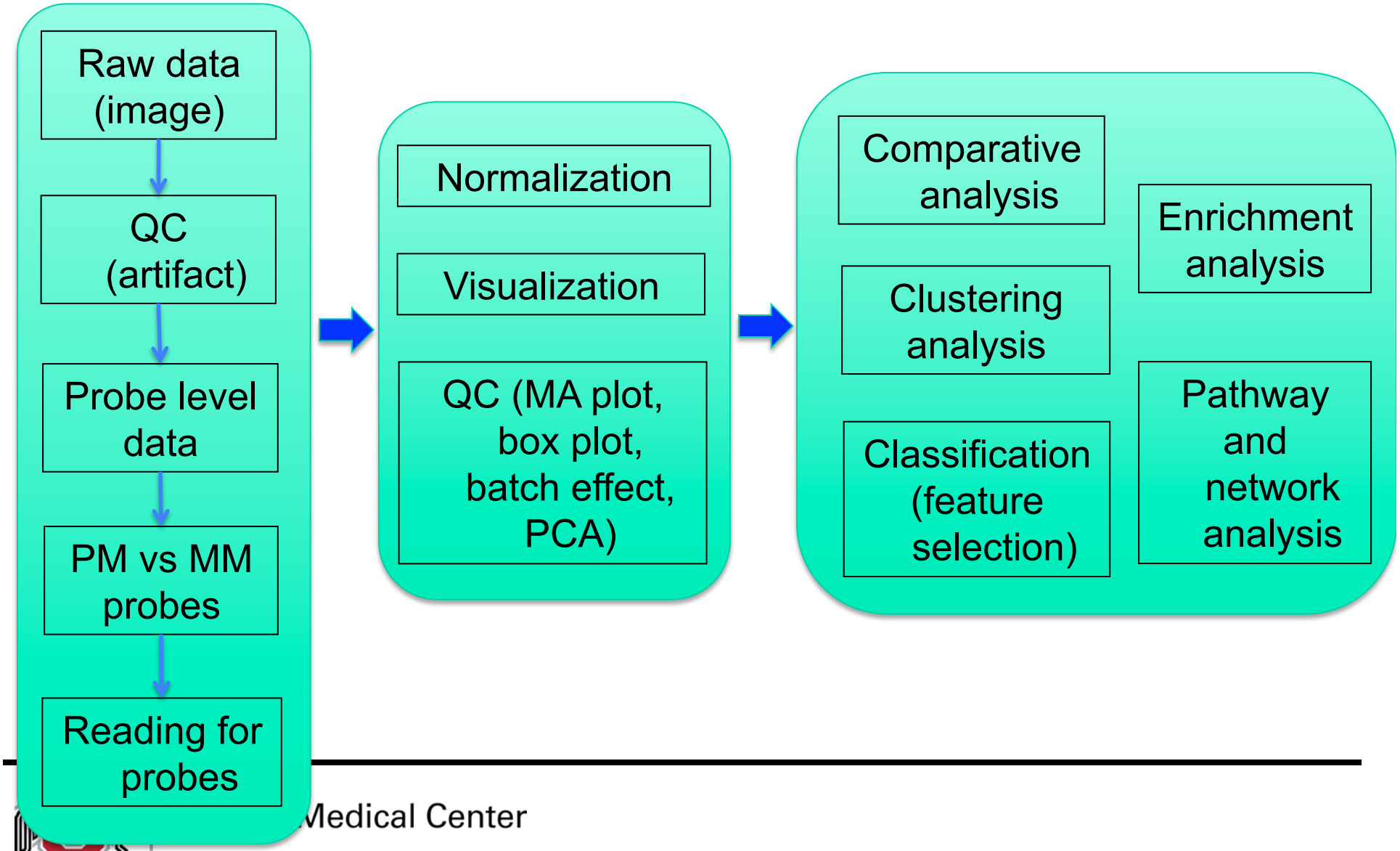
**(McShane, NCI)**

# Spatial Images of Microarrays

- Data for the same brain voxel but for the untreated control mouse
- Background levels are much higher than those for the Parkinson's disease model mouse
- There appears to be something non random affecting the background of the green channel of this slide



# Microarray Data Analysis



# Preprocessing/Data Analysis

Chapter 8, 9, 10 – Background in Statistics

Chapter 20 – Data Preprocessing and Normalization

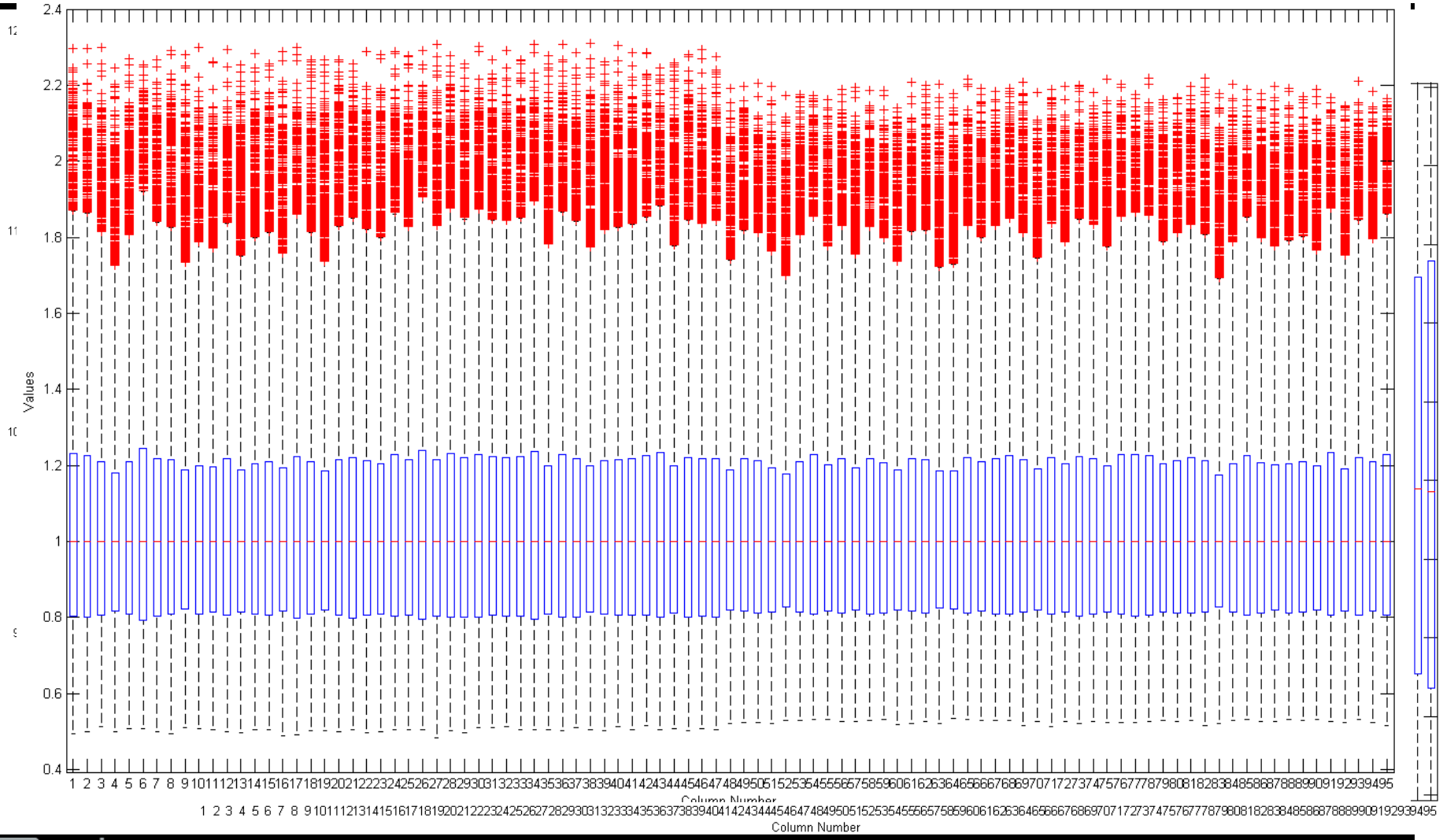
Chapter 11, 12, 13, 14, 16–Classical/Linear Approaches

# Visualization

- Chapter 17

# Why Normalize ?

# Array Normalization



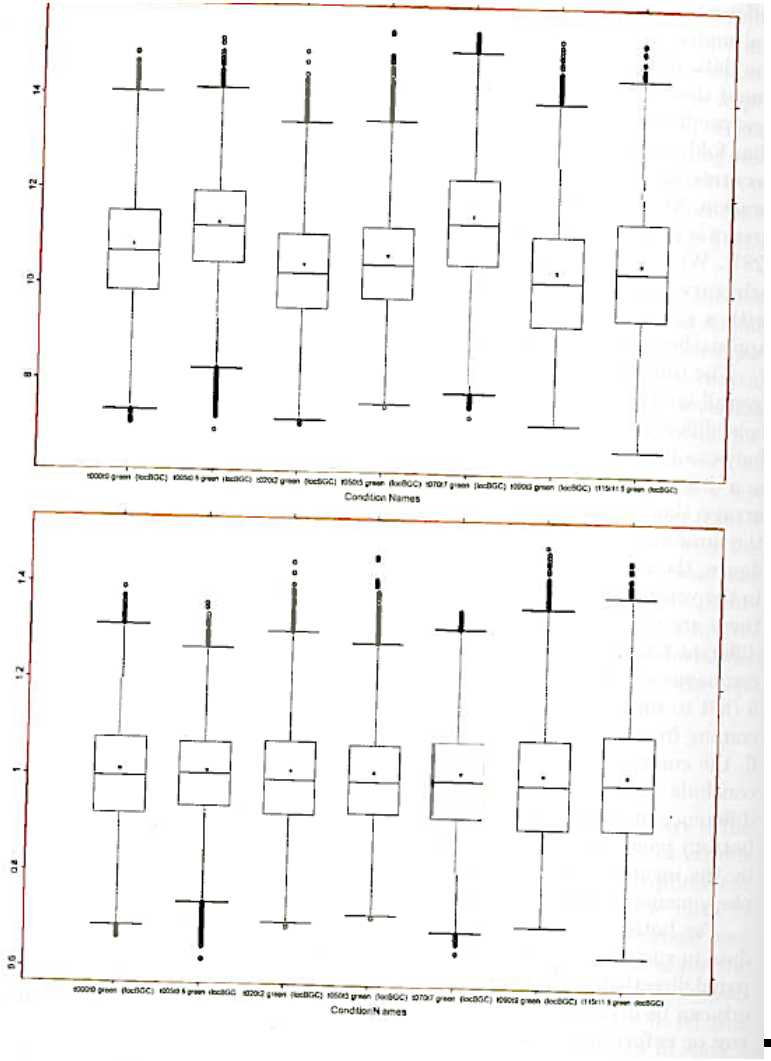
# Goal of Normalization

- Arbitrary comparison of gene expressions in various condition and/or tissues
- Goal of normalization – data is independent of experiment and technology
- Ideal measurement – number of mRNA copies per cell

# Confounding Reasons

+ Intensity imbalance between RNA samples
+ Affects all genes
+ Not due to biology of samples
+ Reasons include
  - difference in the settings of the photodetector voltage
  - imbalance in total amount of RNA in each sample
  - difference in uptaking of the dyes, etc.
  - difference in sites !

# Array Normalization

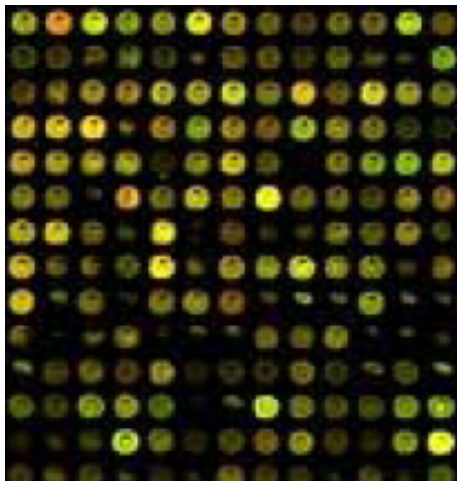


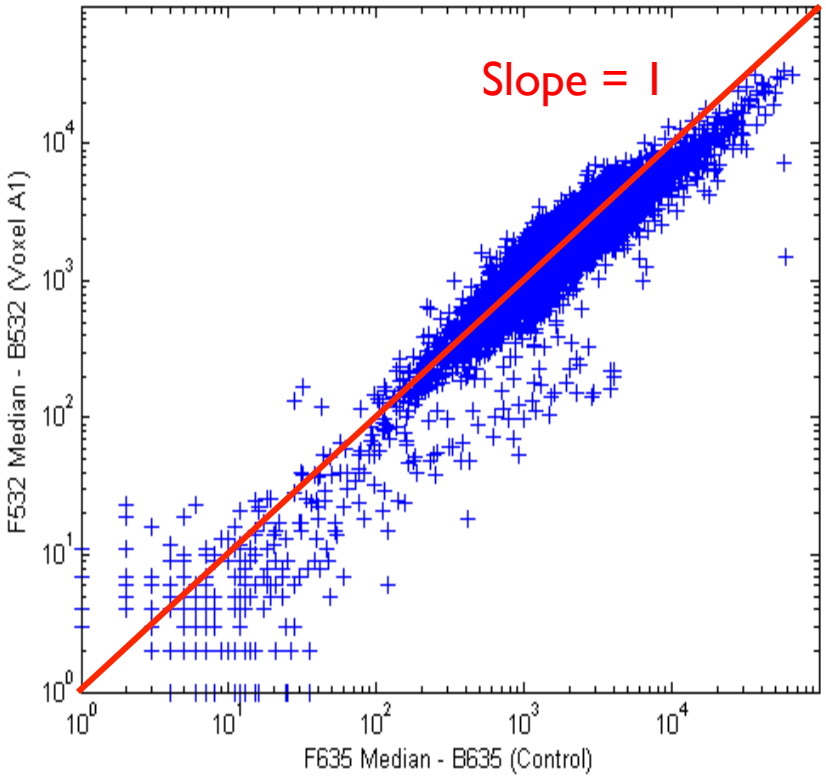
# Consider This

+ A measure of the actual expression levels

+ Differences between median foreground and median background for the red and green channels:

"F635 Median - B635"
"F532 Median - B532"

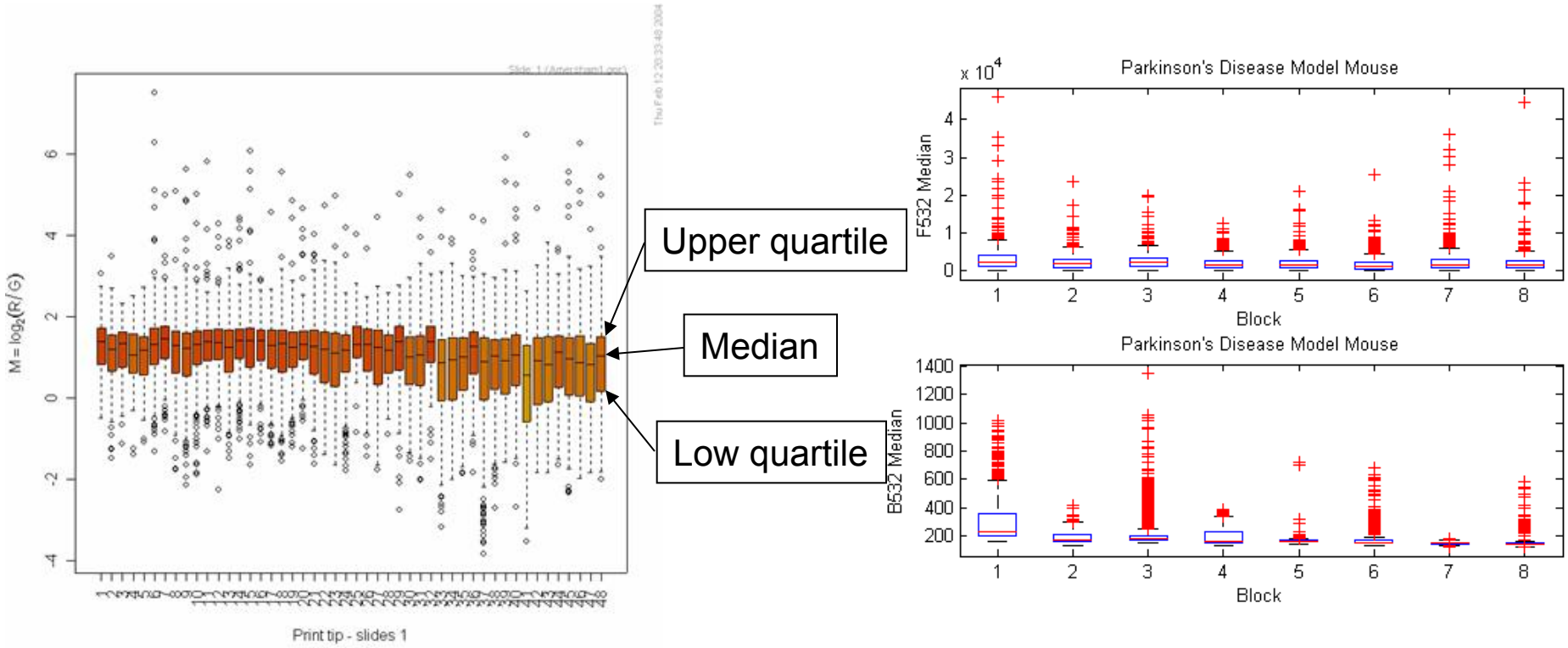




# The Task

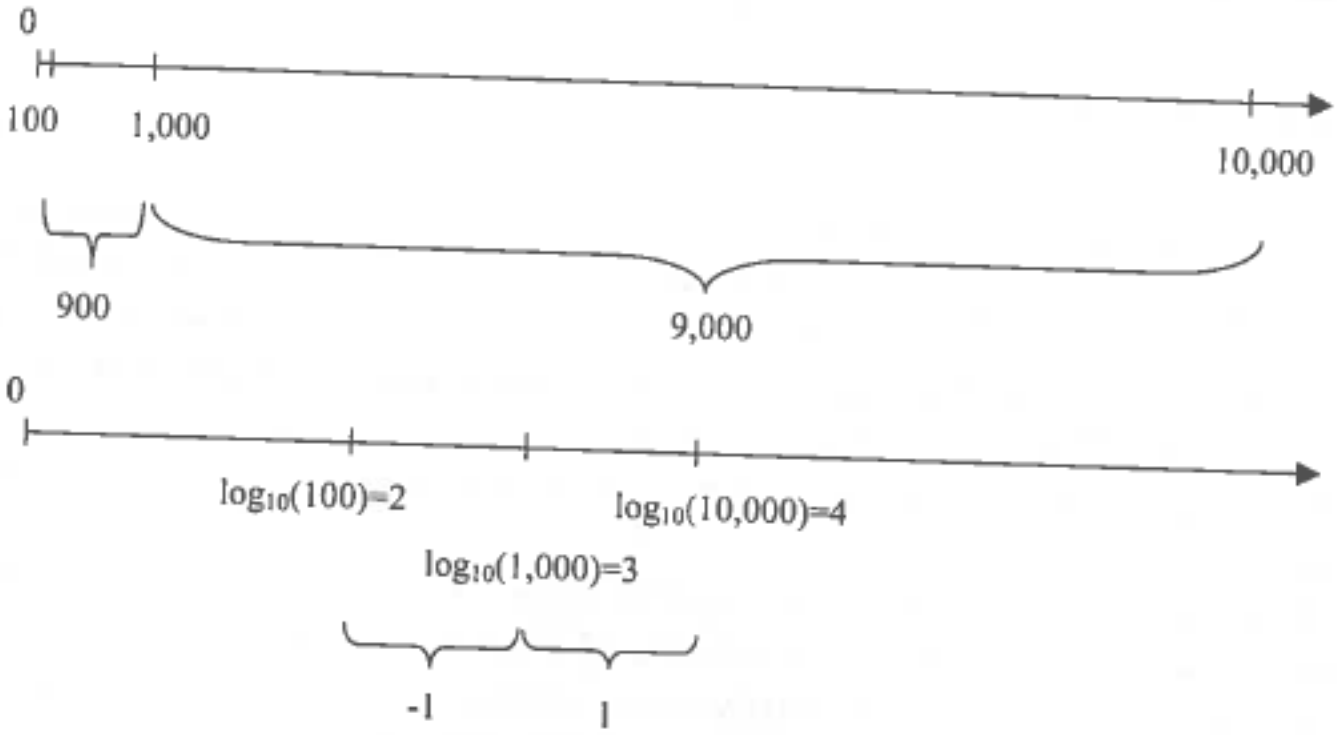
Adjust expression values of all genes so that the ones that are not really differentially expressed have similar values across the array(s)

# The MV Box Plot

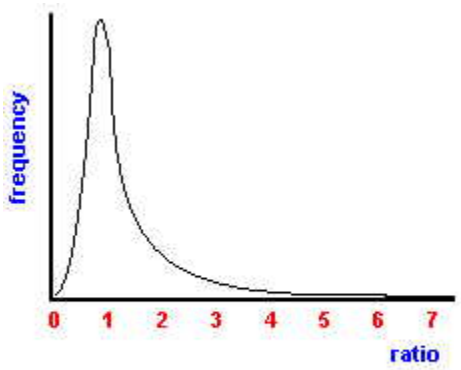


# General Preprocessing Techniques

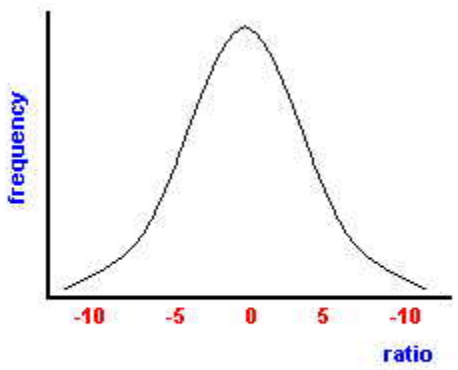
# Log Transform – Fold Changes



# Log Transform - Distribution

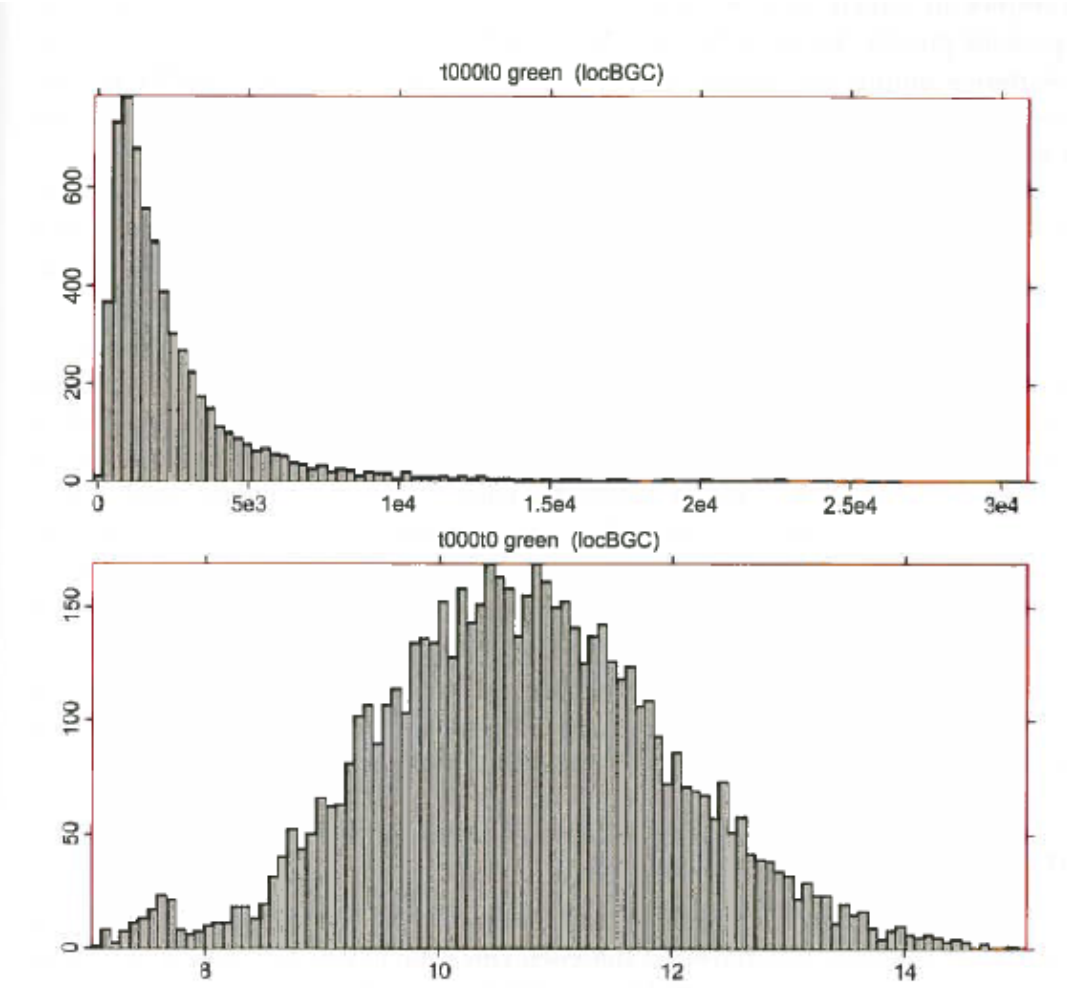



NORMAL RATIO




LOG RATIO

# Log Transform - Distribution



# Platforms and Slides

+ Do not mix platforms

+ Affymetrix - Intra-slide normalization

# Normalization in R

# Use

- library(affy)
- CELfiles <- list.celfiles()

- Rawdata <- ReadAffy(filenames=CELfiles)
- Intensity(Rawdata) # extract intensity

#Visualize

# M-A Plot

- par(mfrow=c(2,3))
- Image(rawdata,transfo=log)
- Maplot(rawdata, pairs=T,plot.method="smoothScatter",cex =0.9,…)

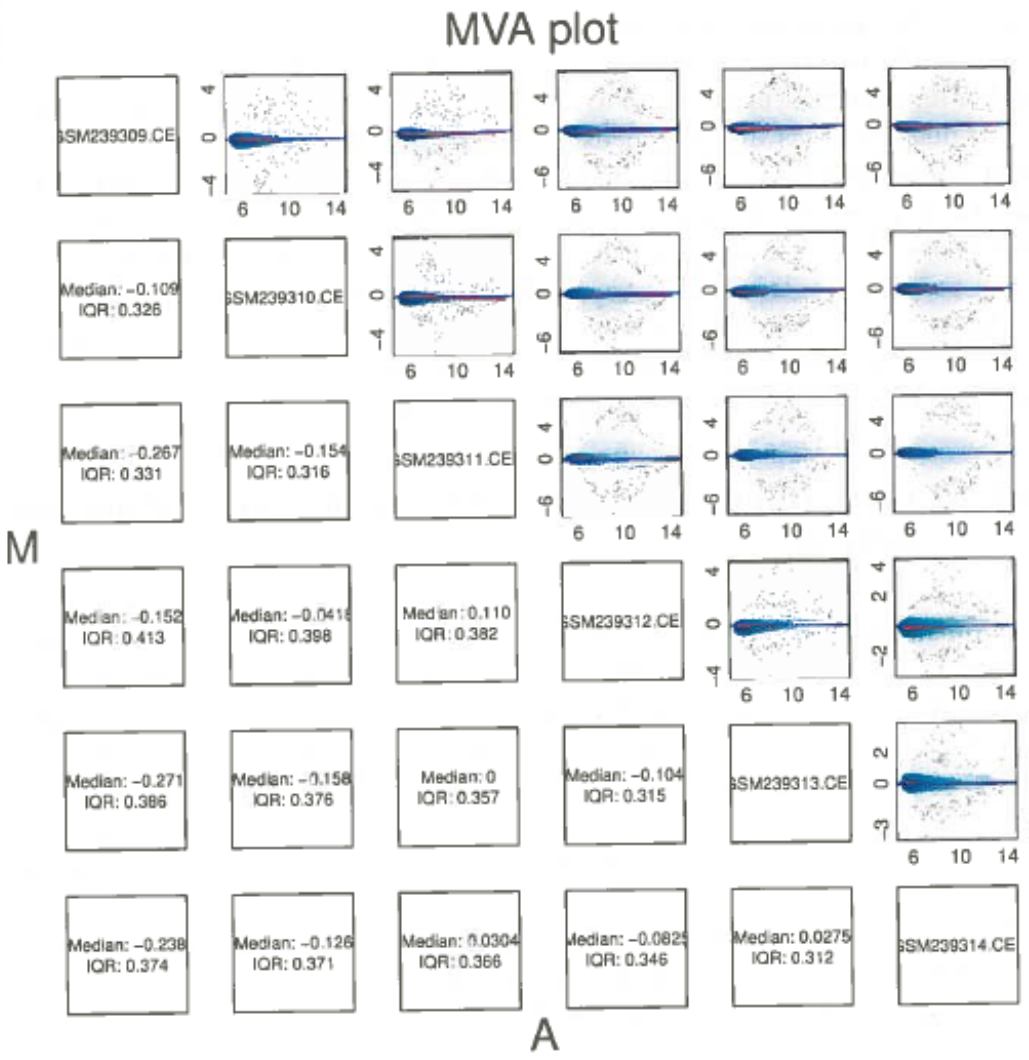



**FIGURE 20.14**: MA-plots of the Affymetrix arrays in this sample data set. The smoothed blue color density representations are derived from kernel density estimations for each array. Data points located in low-density area are depicted as black dots. The red line is the fitted Loess curve. The two summary statistics, median and IQR (interquantile range), for M-values are also shown.

# Background Adjustment

```
> data.bg<-bg.correct(rawdata, method="rma")
> data.qtl<-normalize(data.bg, method="quantiles")
> data.scl<-normalize(data.bg, method="constant")
> data.dChip<-normalize(data.bg, method="invariantset")
> usr.col<-rep(c("red", "blue"), each=3)
> usr.line<-rep(1:3, 2)

> hist(data.bg, lty=usr.line, col=usr.col, main="background
+ corrected")
> legend("topright", sampleNames(rawdata), lty=usr.line,
+ col=usr.col)

> hist(data.scl, lty=usr.line, col=usr.col, main="scaling")

> legend("topright", sampleNames(rawdata), lty=usr.line,
+ col=usr.col)

> hist(data.qtl, lty=usr.line, col=usr.col, main="quantiles")
> legend("topright", sampleNames(rawdata), lty=usr.line,
+ col=usr.col)

> hist(data.dChip,lty=usr.line,col=usr.col,main="invariantset")
> legend("topright", sampleNames(rawdata), lty=usr.line,
+ col=usr.col)
```

# Compare

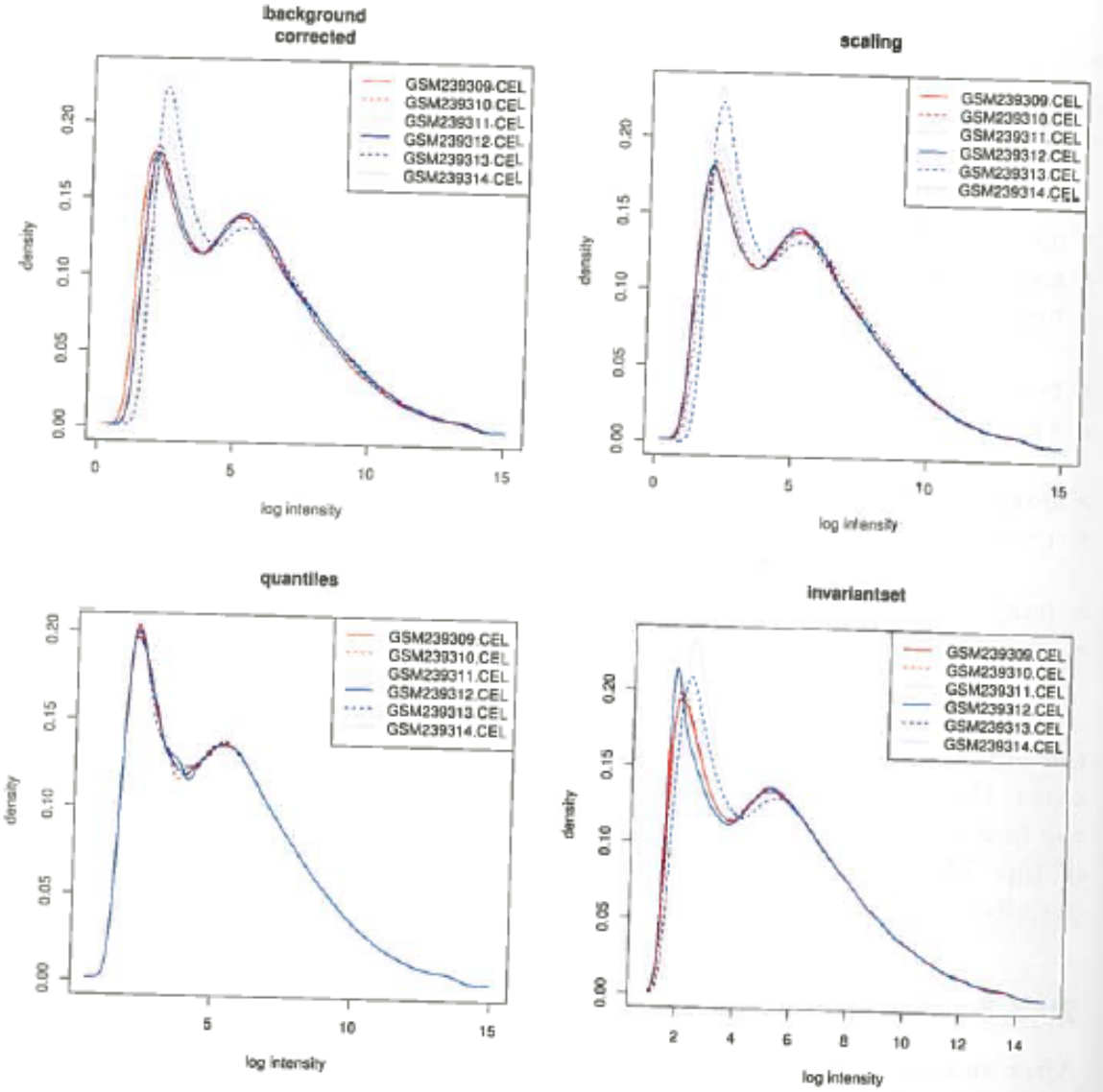



**FIGURE 20.15**: A comparison of three normalization methods. The figure shows the probe intensity distributions of the background corrected data (upper left), and the same data after three different normalization methods: scaling (upper right), quantiles (bottom left), and invariant set (bottom right).

# Background Adjustment

```
> boxplot(data.bg, col=usr.col, las=3, names=
+ sampleNames(rawdata),ylim=c(4,8),cex.axis=0.6,main=
+ "background corrected")

> boxplot(data.scl, col=usr.col, las=3, names=sampleNames
+ (rawdata), ylim=c(4,8), cex.axis=0.6, main="scaling")

> boxplot(data.qtl, col=usr.col, las=3, names=sampleNames
+ (rawdata), ylim=c(4,8), cex.axis=0.6, main="quantiles")

> boxplot(data.dChip, col=usr.col, las=3, names=sampleNames
+ (rawdata), ylim=c(4,8), cex.axis=0.6, main="invariantset")
```

# Compare

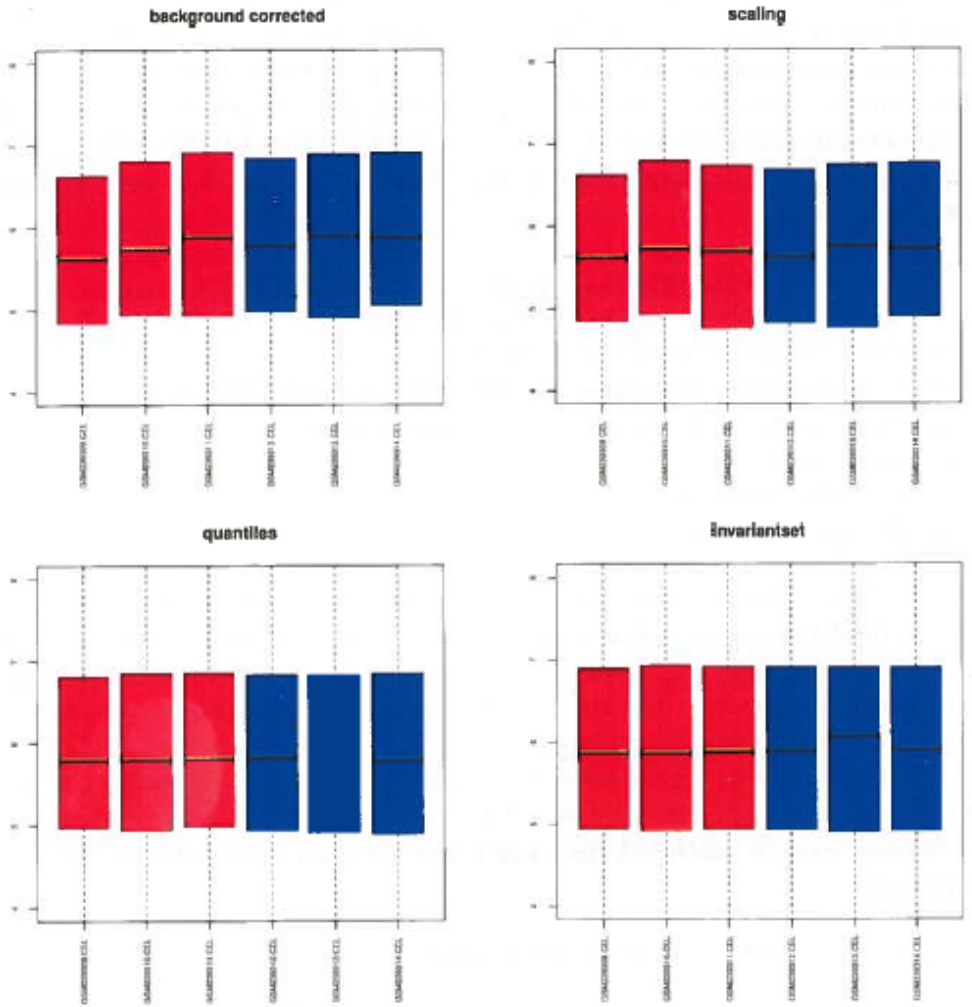



**FIGURE 20.16**: A comparison of three normalization methods. The figure shows the box plot of probe intensities of the background corrected data (upper left) and after three different normalization methods: scaling (upper right), quantiles (bottom left), and invariant set (bottom right). Note the almost perfect alignment of the distributions provided by the quantile normalization, and the very good alignment provided by the invariant set method used by dChip. The red and blue colors show the class membership of that particular sample.

# Array Normalization

- ✓ Which genes to use for normalization
- ✓ Which methods to use

# Which Genes ?

+ Housekeeping genes

+ Genes involved in activities of cell maintenance and survival

+ Not in cell function & proliferation

+ These genes will be similarly expressed in all samples

# Affymetrix & House Genes

+ Difficult to identify – need to be confirmed

+ Affymetrix GeneChip provides a set of house keeping genes based on a large set of tests on different tissues

+ Found to have low variability in these samples (but still no guarantee).

# Spiked Controls

- ✓ Genes that are not usually found in the samples – both control and test sample
- ✓ E.g., yeast gene in human tissue samples.

# The Simple Route

+ Use all adequately expressed genes for normalization

+ **Assumption - majority of genes on array are housekeeping genes and proportion of over expressed genes is similar to under expressed genes**

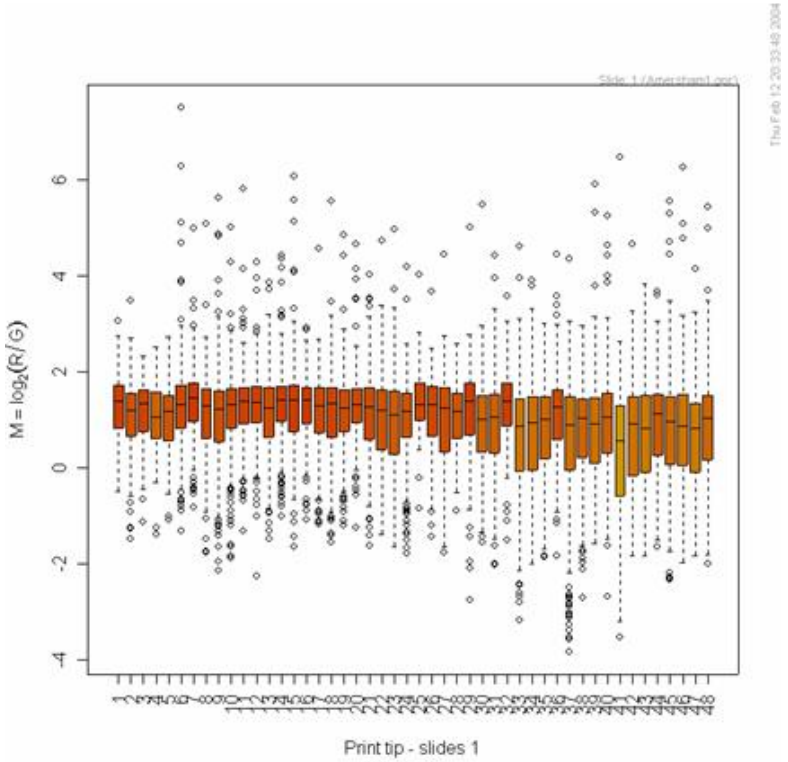
+ If the genes on chip are specially selected, then this method will not work.

# Methods

# Array Normalization

- Scaling/Non-linear
  - Using baseline array or reference
  - Baseline – Median of median intensities
  - Linear Regression
  - Standard Affymetrix method
- Complete - Combine information from all arrays
  - Cyclic Loess
  - Quantile Normalization

# First Examine



# Scaling - Dividing by Array Mean

- z-transform
- Substitute mean by median/mode/percentile
- Remove upper/lower 10% and divide by mean
- cDNA arrays – adjusts overall intensity problems

# Subtracting Mean

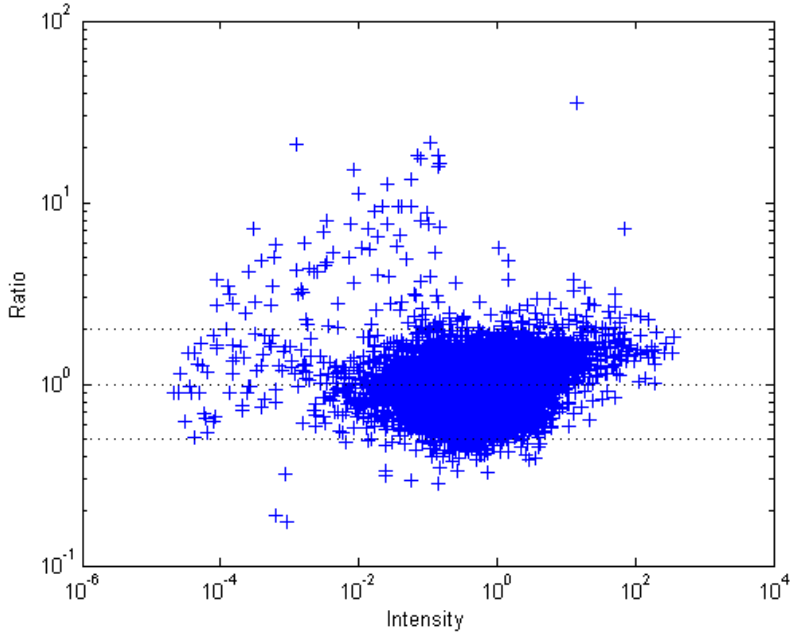
$$x_{in} = \log\left(\frac{x_i}{\bar{x}_i}\right)$$

$$x_{in} = \log\left(x_i\right) - \frac{\sum \log\left(x_i\right)}{n} = \log\left(\frac{x_i}{\bar{x}_{ig}}\right)$$

# Iterative linear Regression

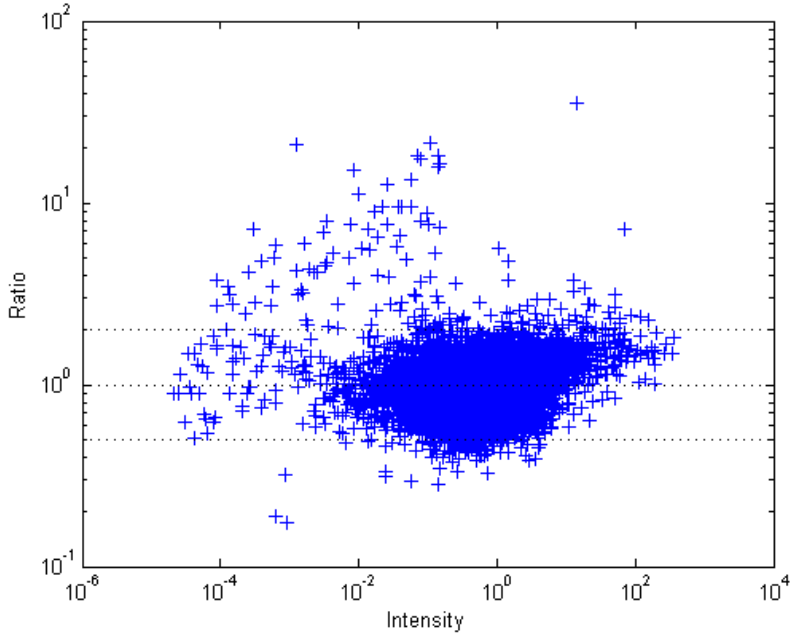
Repeat regression by discarding points: error > 2σ

$$X_i^{norm} = k * X_i$$
$$c = log(k)$$
$$M_i^{norm} = log(X_i^{norm}) = c + M_i$$



# Line (global) normalization

- Simplest but most consistent
- Move the median to zero –
  - slope 1 in scatter plot
  - changes the intersection



# Using Control Spots & Genes

# Affymetrix MicroArray Suite (MAS) v4.0

- Uses MM probes to correct for NSB.
- MAS4.0 used simple Average Difference method:

$$AvDiff = \frac{1}{|A|} \sum_{j \in A} \left( PM_j - MM_j \right)$$

- A is the subset of probes where $d_j = \left( PM_j - MM_j \right)$ is within 3 SDs of the average of $d_{(2)}, \ldots, d_{(J-1)}$
- Excludes outliers, but not a robust averaging method.

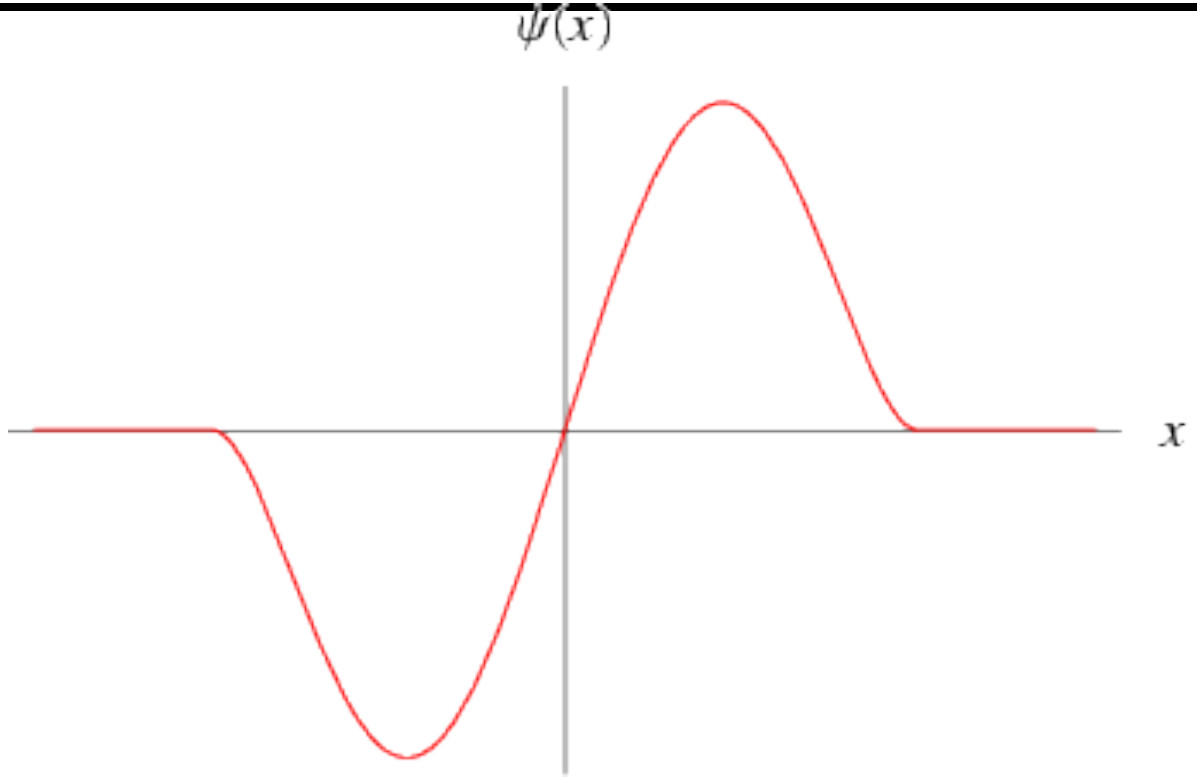
# Affymetrix MicroArray Suite (MAS) v5.0

- Current method employed by Affymetrix.
- Weighted mean using one-step Tukey Biweight Estimate:

$$signal = \log^{-1}\left(Tukey\ Biweight\left(\log\left(PM_j - CT_j\right)\right)\right)$$

- $CT_j$ (contrast tau) is a quantity derived from $MM_j$ never larger than $PM_j$.
- Weights each probe intensity based on it's distance from the mean.
- Robust average (insensitive to small changes from any assumptions made).

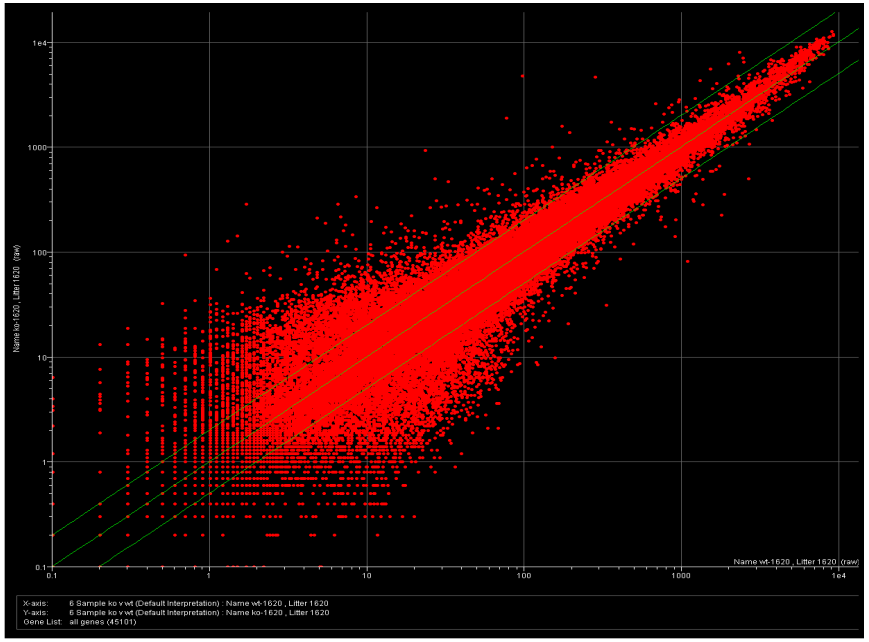
# Tukey Biweight




$$\psi(x) = \begin{cases} x\left(1 - \frac{x^2}{c^2}\right)^2 & \text{for } |x| < c \\ 0 & \text{for } |x| > c \end{cases}$$

# Problems with MAS5.0

- Loss of probe-level information
- Background estimate may cause noise at low intensity levels due to subtraction of MM data.

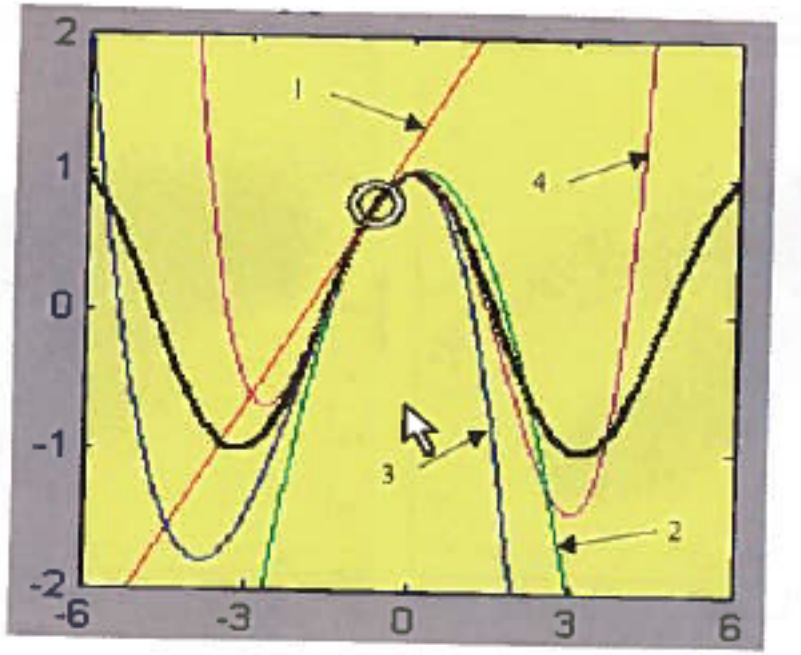


# LOWESS/LOESS

- Locally weighted Polynomial Regression
- Over-fitting vs. under-fitting
- LOWESS-1 – degree 1, LOWESS-2 - degree 2
- Divide domain into windows of size *w*
- And fit

$$w(x) = \begin{cases} (1 - |x|^3), |x| < 1 \\ 0, |x| \geq 1 \end{cases}$$

# Effects of LOWESS



# Effects of LOWESS

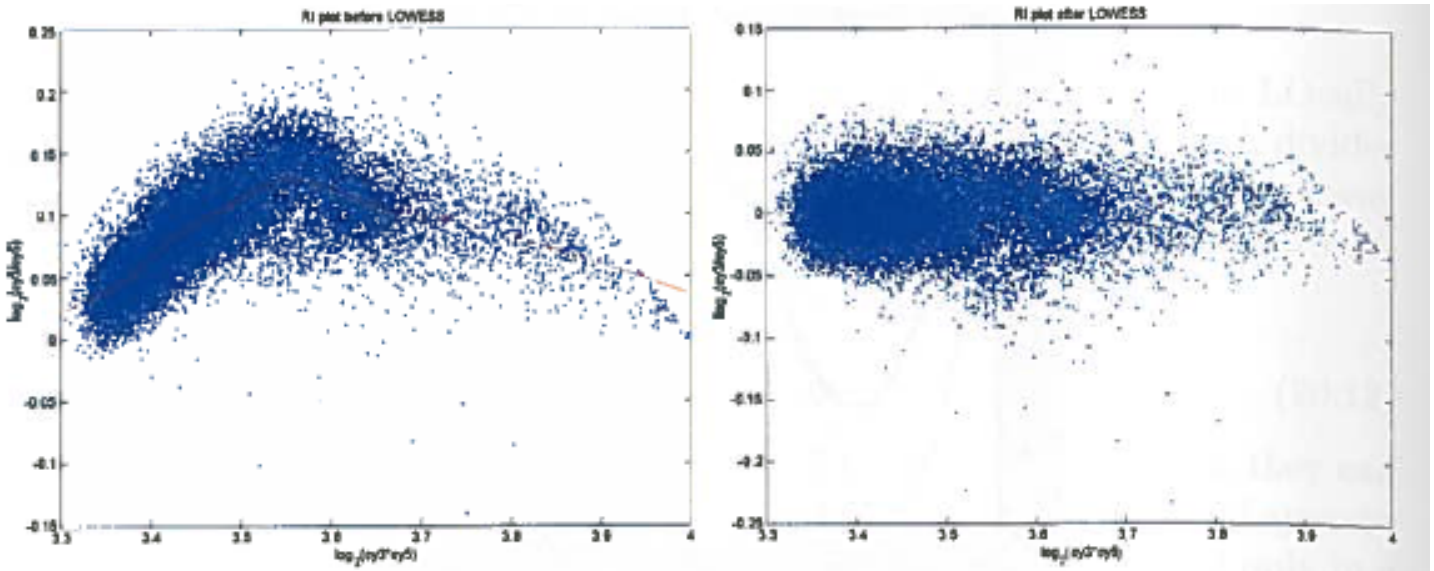


**FIGURE 20.7**: LOWESS normalization. A ratio-intensity plot before (left panel) and after (right panel) the LOWESS correction. The horizontal axis represents $\log_2(cy3)+\log_2(cy5)$ which is a quantity proportional to the overall signal intensity. The vertical axis represents the log-ratio of the two samples $\log_2(cy3/cy5)$. The left panel also shows the nonlinear regression curve.

# Effects of LOWESS

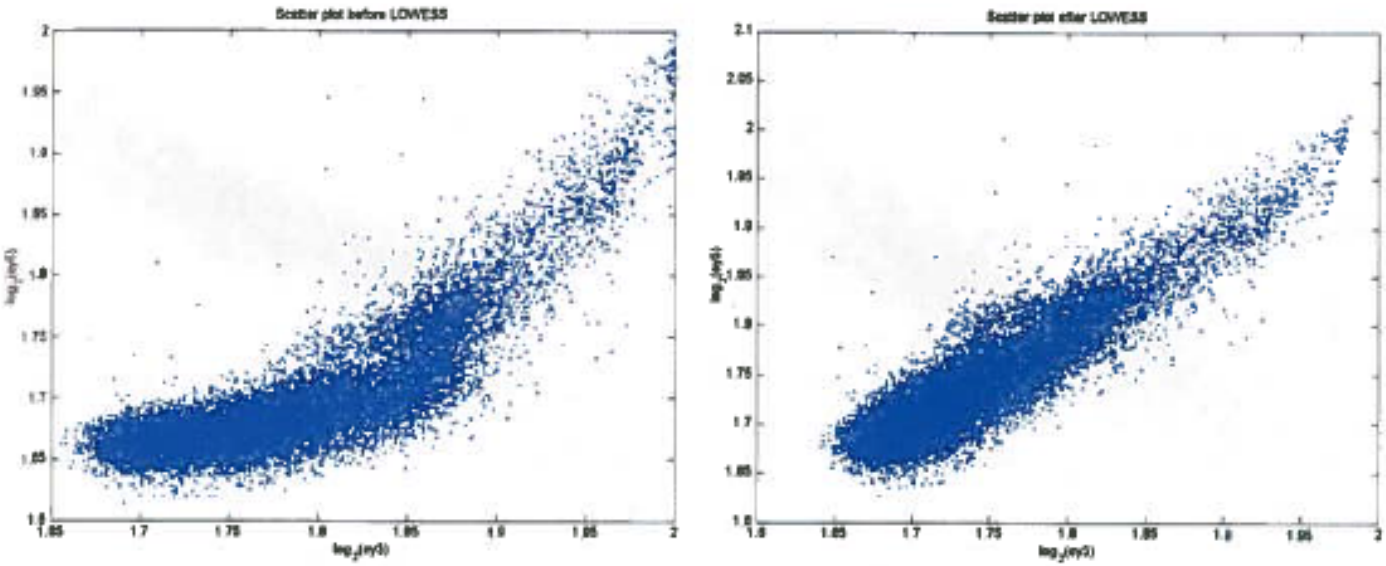



**FIGURE 20.8**: LOWESS normalization. A scatter plot before (left panel) and after (right panel) the LOWESS correction. The horizontal axis represents $\log_2(cy3)$ and the vertical axis represents $\log_2(cy5)$.

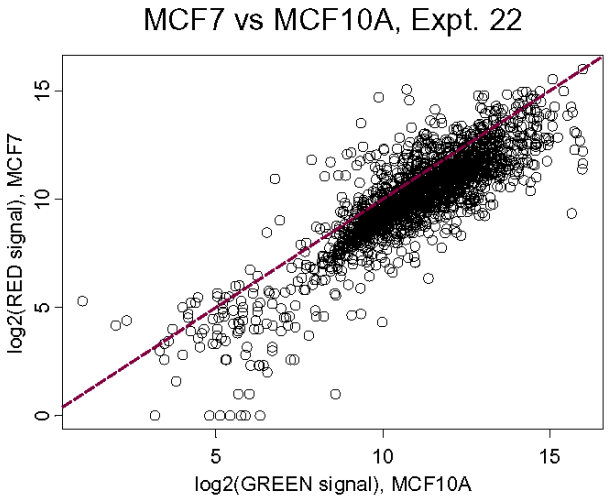
# Operation

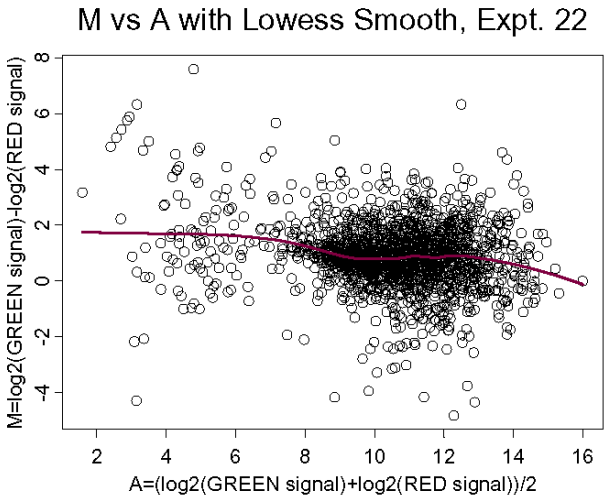
- For n points $\frac{d+1}{n} < q < 1$

- In each window, $n \times q$ points will be chosen
- Small q – less smooth but close to function
- Large q – more smooth

# Use for LOWESS

Overall magnitude of  spot intensity has impact on relative intensity between channels.

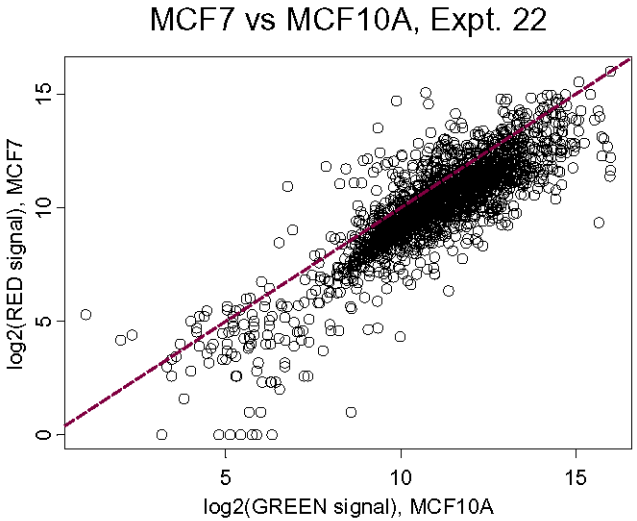


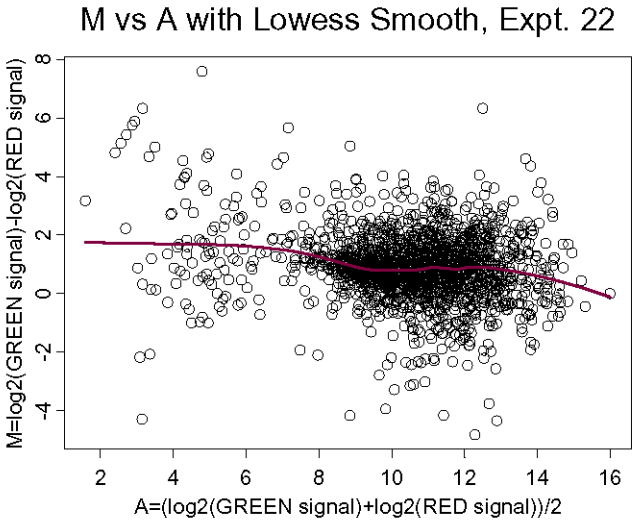




(McShane, NCI)

# Use for LOWESS

- Nonlinear
- Gene-by-gene, could introduce bias
- Use only when there is a compelling reason




(McShane, NCI)

# Advantages

- No need to specify type of function
- Specify degree of polynomial and smoothing factor q
- Least square is well studied

# Disadvantages

- No model or function is generated
- LOWESS will need to be applied for every data; cannot transfer models
- Very expensive
- Suspetible to noise and outliers

# Cyclic LOESS – M&A

- Arrays *i, j*
- Probe *k*
- Apply to pairs of arrays
- M – fold change, A log average
- Do LOESS on M-A plot
- Repeat till threshold adjustment reached

$$M_k = \log_2\left(\frac{x_{ki}}{x_{kj}}\right)$$

$$A_k = \frac{1}{2}\log_2\left(x_{ki} \cdot x_{kj}\right)$$

$$M_k' = M_k - \hat{M}_k$$

$$x_{ki} = 2^{A_k + \frac{M_k'}{2}}$$

$$x_{ki} = 2^{A_k - \frac{M_k'}{2}}$$

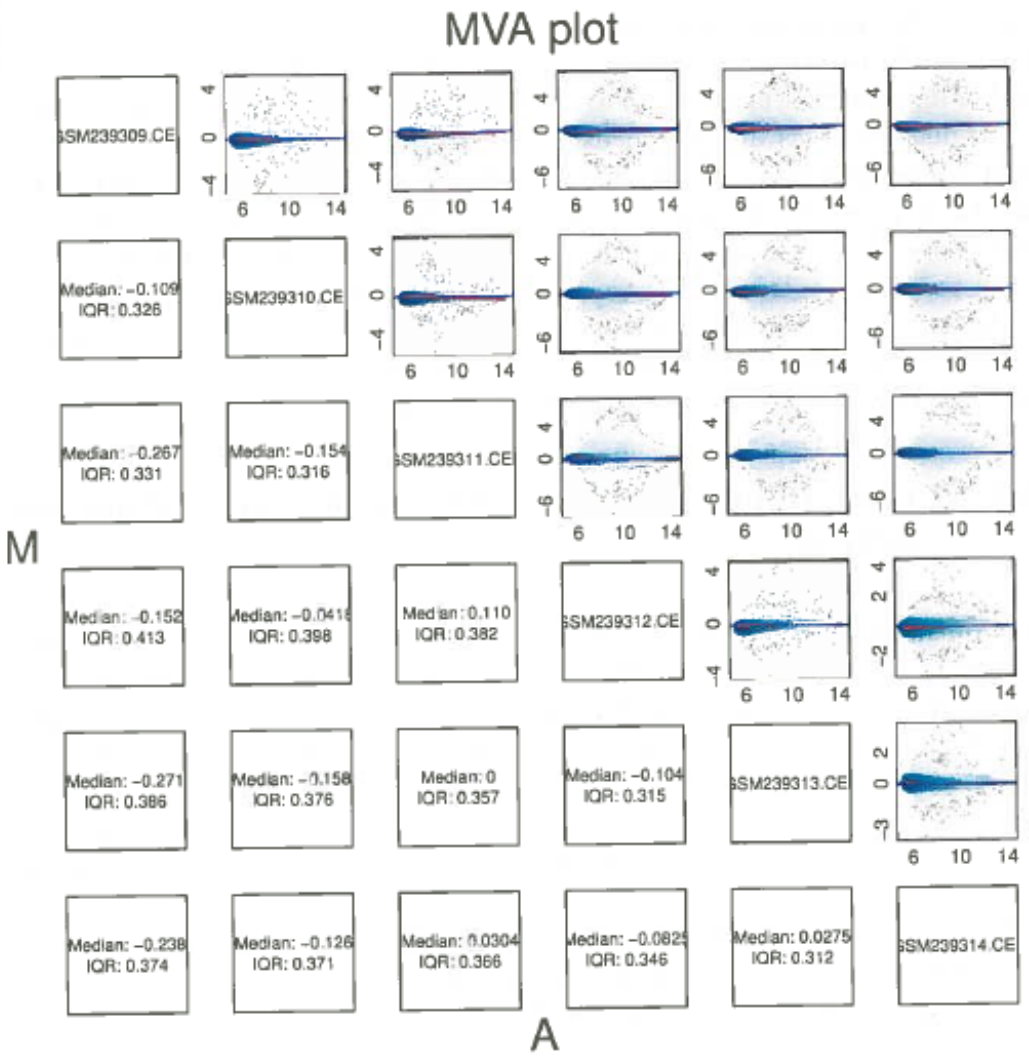
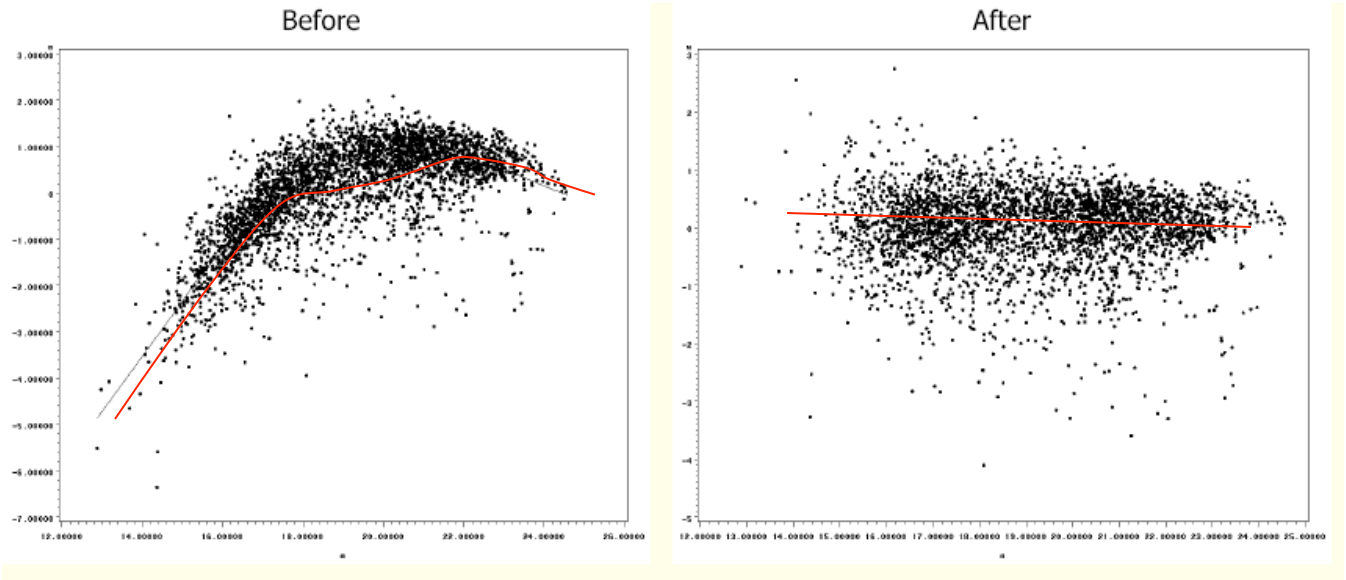



**FIGURE 20.14**: MA-plots of the Affymetrix arrays in this sample data set. The smoothed blue color density representations are derived from kernel density estimations for each array. Data points located in low-density area are depicted as black dots. The red line is the fitted Loess curve. The two summary statistics, median and IQR (interquantile range), for M-values are also shown.

# Application

"Straighten" Lowess fit line (MA plot) to horizontal & move to zero




$$X_i^{norm} = k(A) * X_i$$

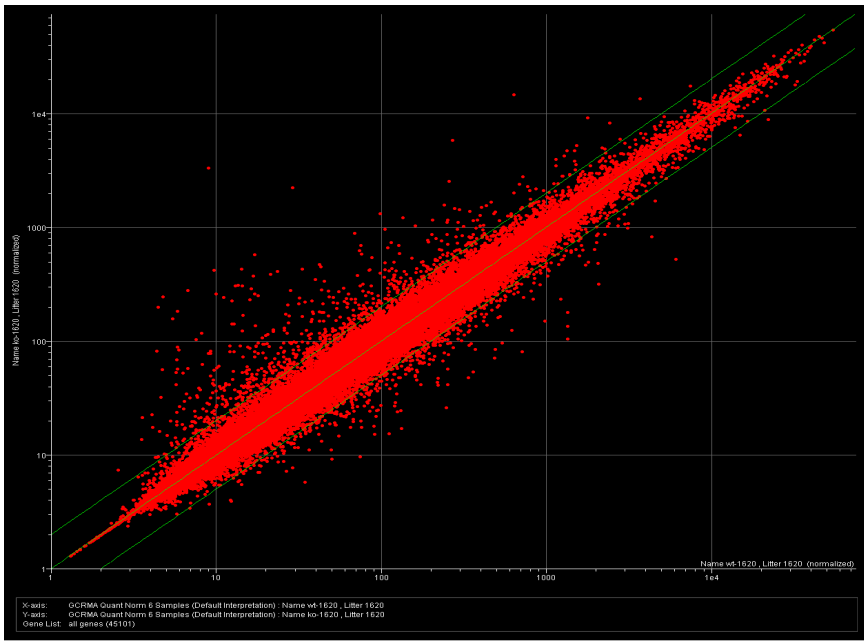
$$c(A) = log(k(A))$$

$$M_i^{norm} = log(X_i^{norm}) = c(A) + M_i$$

# Robust Multiarray Average (RMA)

# The Need

- Subtraction of MM data corrects for NSB, but introduces noise.
- Want a method that gives positive intensity values.
- Normalising at probe level avoids the loss of information.



# Background Correction

- PM data is combination of background and signal.

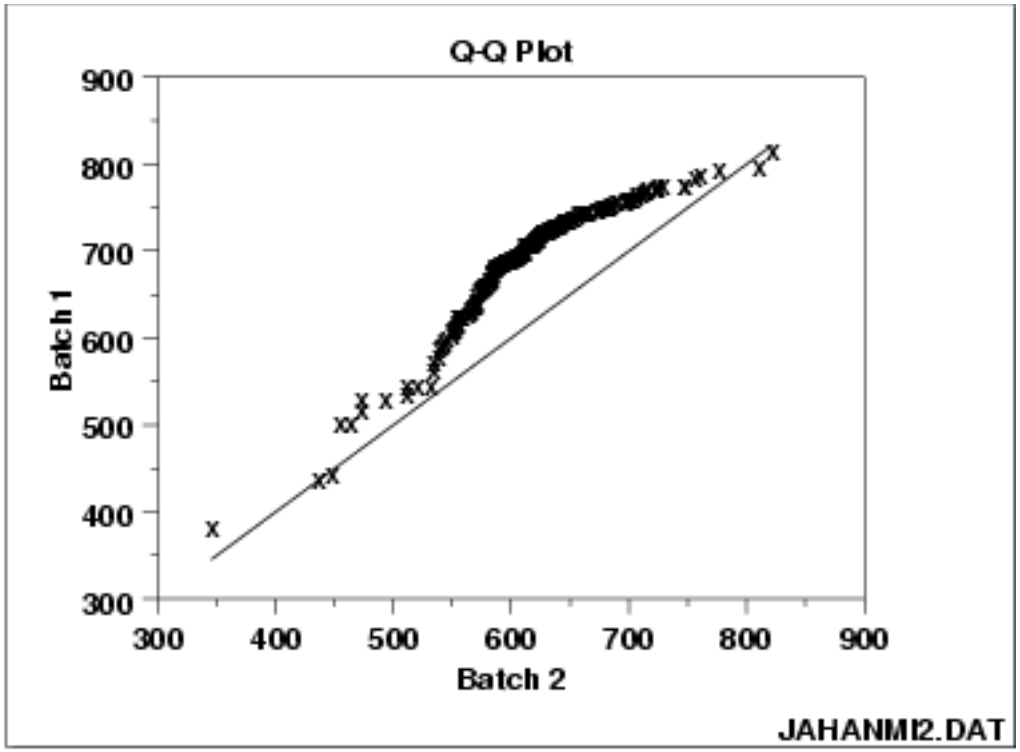
$$B\left(PM_{ijn}\right) \equiv \mathrm{E}\left(S_{ijn} \middle| PM_{ijn}\right)$$

- Assume strictly positive distribution for signal. Then background corrected signal is also positively distributed.
- Background correction performed on each array separately.

# Normalization

- '*Quantile Normalization*' used to correct for array biases.
- Compares expression levels between arrays for various quantiles.
- Can view this on quantile-quantile plot.
- Protects against outliers.

# Q-Q plot




These 2 batches do not appear to have come from populations with a common distribution.

The batch 1 values are significantly higher than the corresponding batch 2 values.

The differences are increasing middle of range. Then the values for the 2 batches get closer again.

# Robust Multiarray Average (RMA)

- Linear model.
- Uses background corrected, normalised, log transformed probe intensities: $Y_{ijn} = T(Pm_{ij})$

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}$$

$\mu_{in}$ = Log scale expression level (RMA measure).

$\alpha_{jn}$ = Probe affinity affect.

$\varepsilon_{ijn}$ = Independent identically distributed error term (with mean 0).

- Not all Probes are used

# Comparison

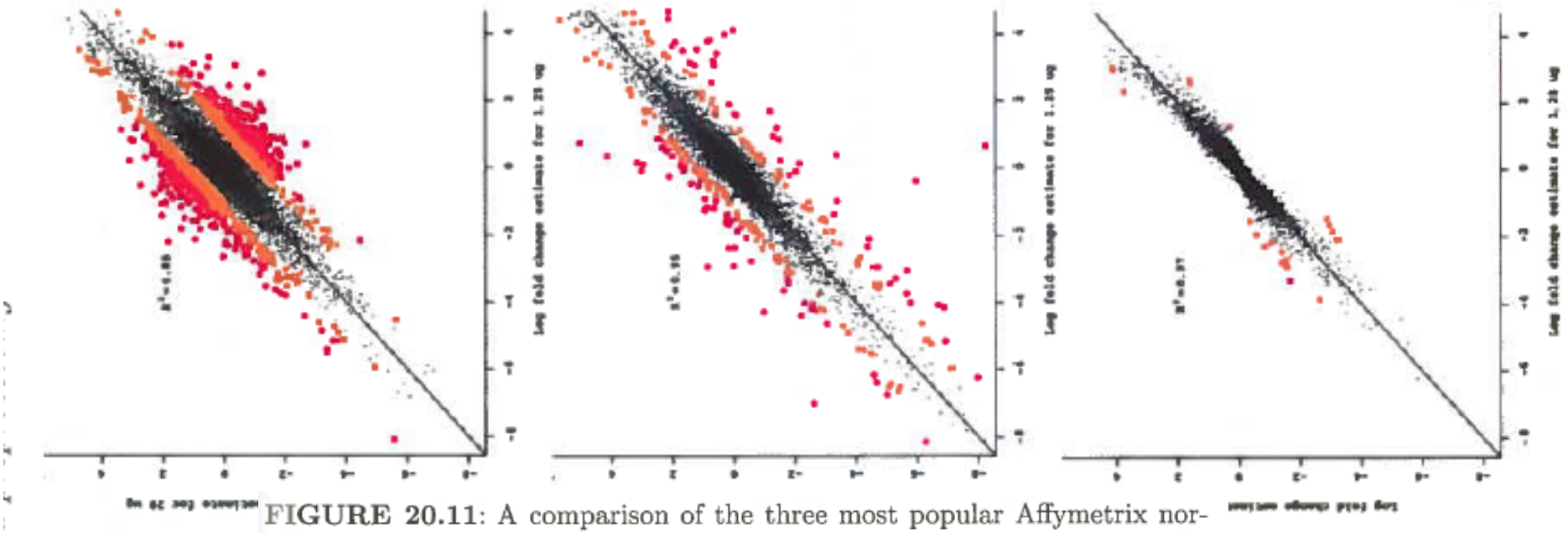


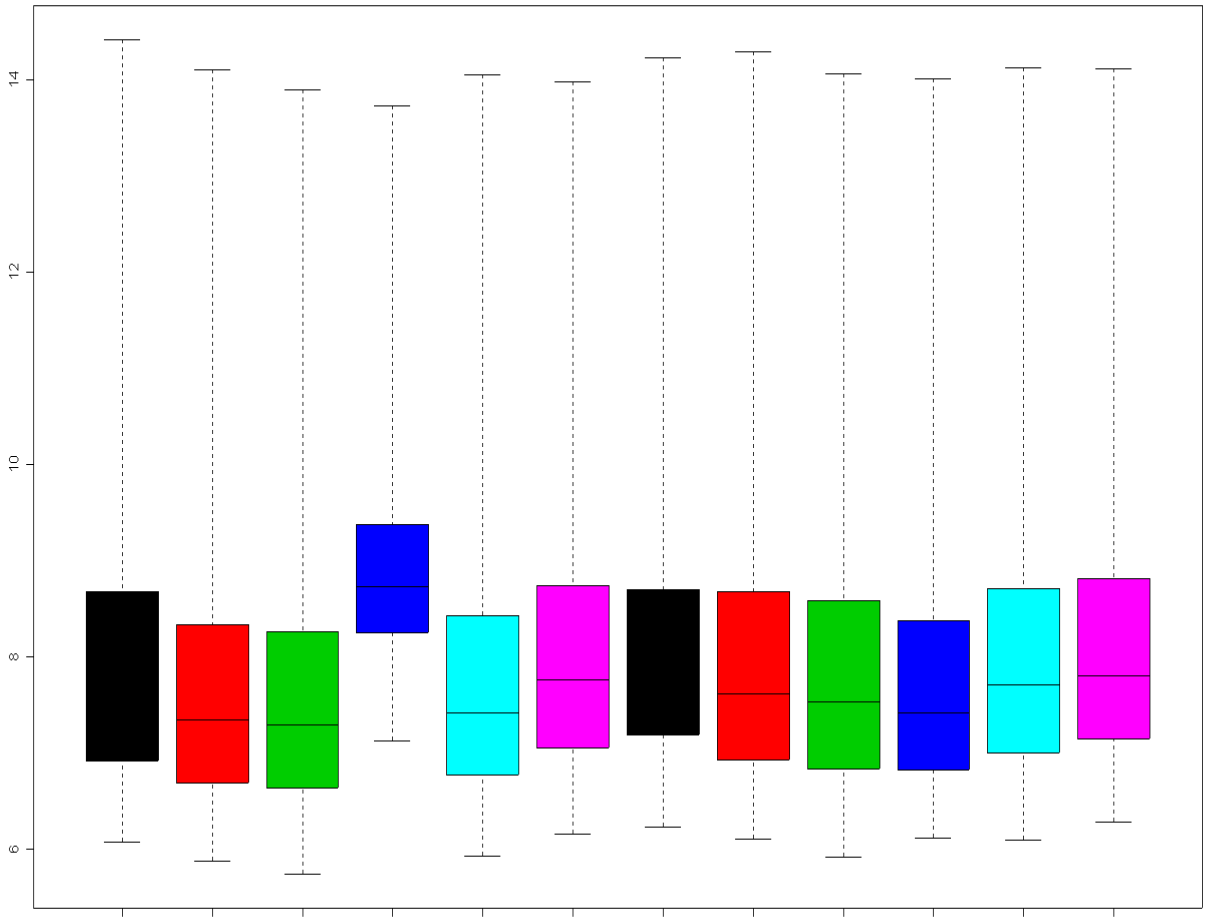
**FIGURE 20.11**: A comparison of the three most popular Affymetrix normalization approaches. Each panel shows a scatter plot of the normalized values obtained after hybridizing two different amounts from the same sample. Since these are different amounts from the same biological sample, there should be no differentially expressed (DE) genes. The results obtained with the Affymetrix MAS 5 package (section 20.4), dChip (section 20.5.2), and RMA (section 20.5.3) are shown in panels A, B, and C, respectively. Genes reported as having two- to three-fold changes are shown in yellow. Genes reported as having fold changes larger than three fold are shown in red. Figure reprinted with permission from [235].

# Summarization

- Combine intensity values from the probes in the probe set to get a single intensity value for each gene.
- Uses '*Median Polishing*' .
- Each chip normalised to its median.
- Each gene normalised to its median.
- Repeated until medians converge.
- Maximum of 5 iterations to prevent infinite loops.

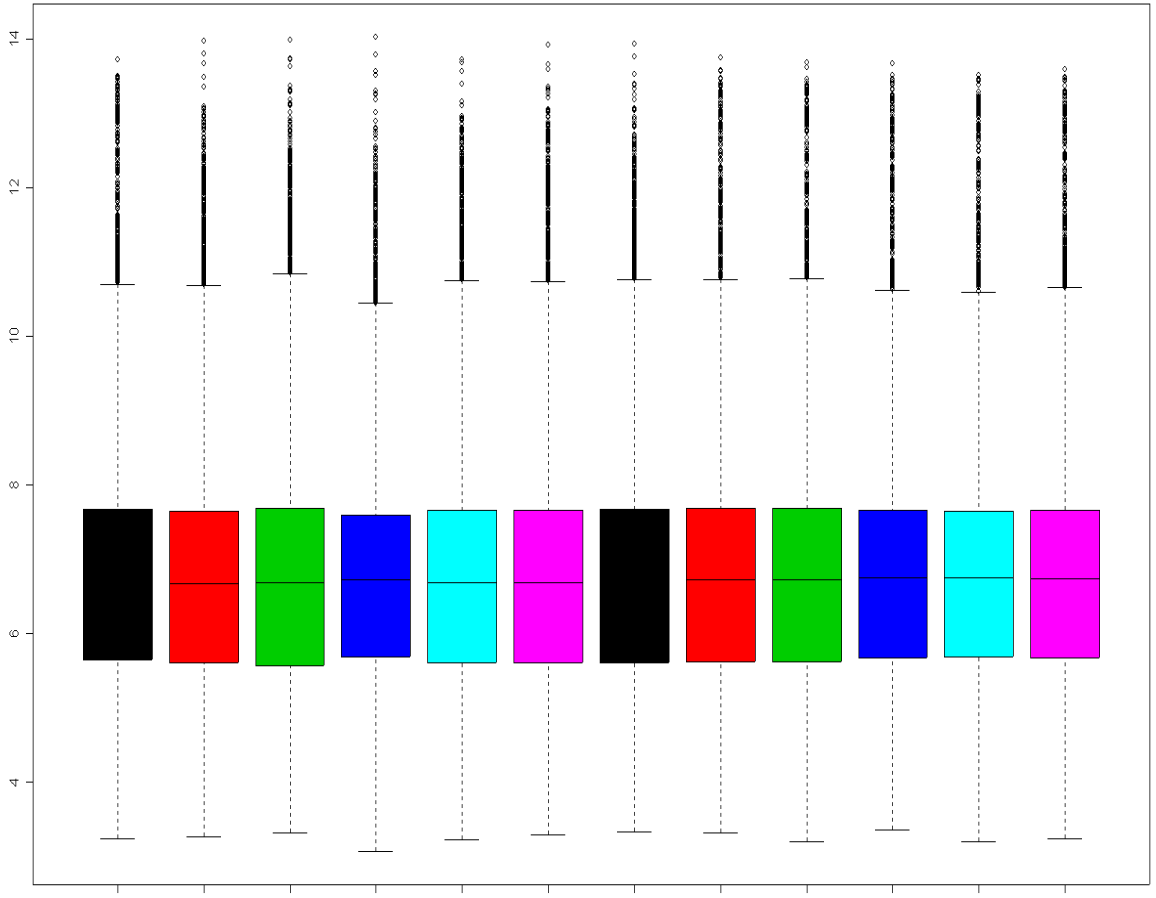
# Effect

## Pre-Normalisation



# Effect

## Post-Normalisation



# GC-RMA

- Corrects for background noise as well as NSB.
- Probe affinity calculated using position dependant base effects:

$$\alpha = \sum_{k=1}^{25} \sum_{j \in \{A,T,G,C\}} \mu_{j,k} 1_{b_k=j} \text{ with } \mu_{j,k} = \sum_{l=0}^{3} \beta_{j,l} k^l,$$

- MM data adjusted based on probe affinity, then subtracted from PM.
- Does not loose MM data.

# Advantages of RMA/GC-RMA

- Gives less false positives than MAS5.0.
- See less variance at lower expression levels than MAS5.0.
- Provides more consistent fold change estimates.
- Exclusion of MM data in RMA reduces noise, but loses information.
- Inclusion of adjusted MM data in GC-RMA reduces noise, and retains MM data.

# Disadvantages

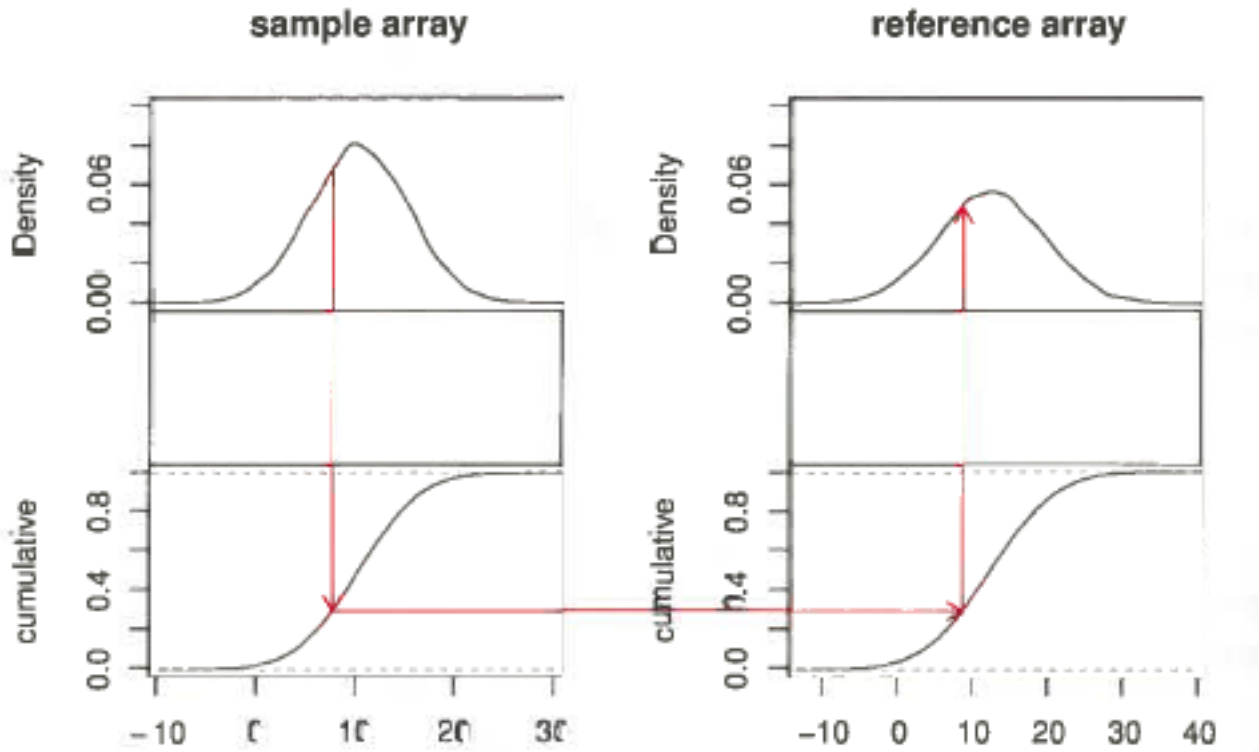
- May hide real changes, especially at low expression levels (false negatives).
- Makes quality control after normalisation difficult.
- Normalisation assumes equal distribution which may hide biological changes.

# Conclusions

- RMA is more precise than MAS5.0, but may result in false negatives at low expression levels.
- Useful for fold change analysis, but not for studying statistical significance. Makes quality control difficult.
- Ideal solution – Use standard MAS5.0 techniques for quality control. Then go back and perform probe level normalisation on quality controlled genes.

# The Best for The Last

# Quantile Normalization

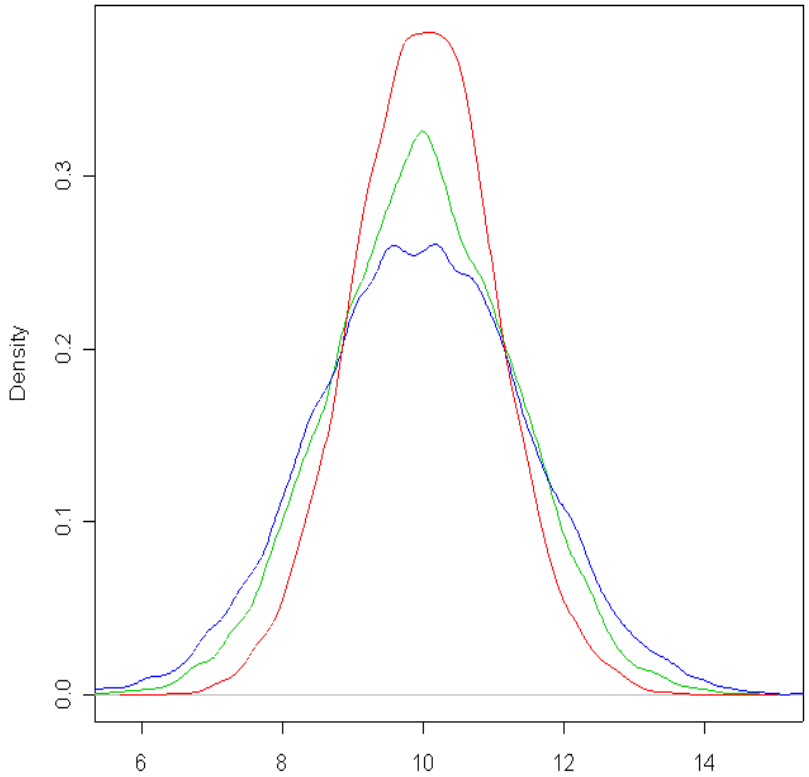


# Steps

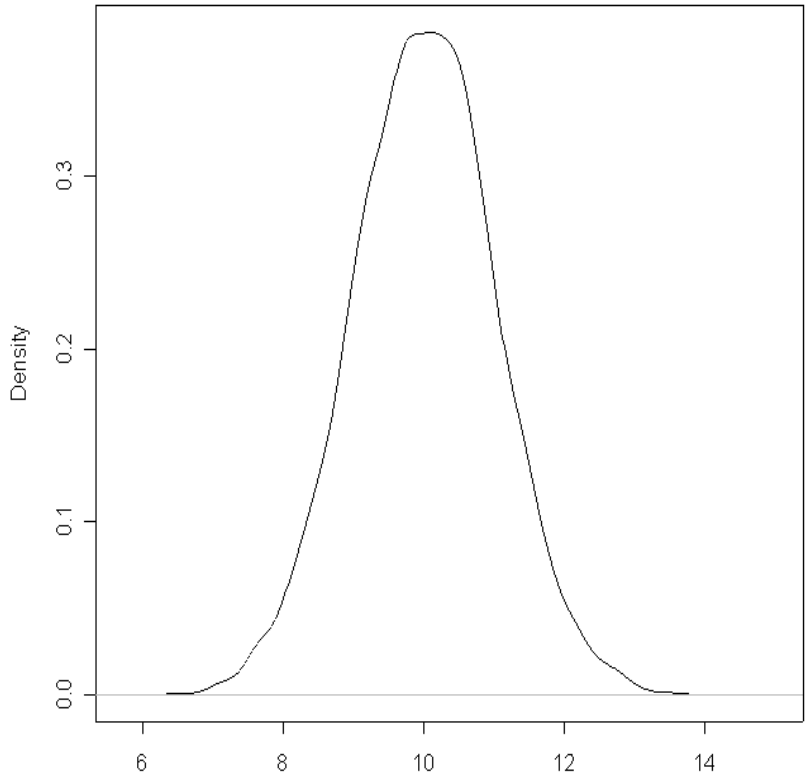
1. Given a set of $n$ arrays with $p$ probes, construct a matrix $X$ with $p$ rows and $n$ columns, in which each column is an array of raw intensities
2. Sort each column to construct the matrix $X_{sort}$
3. Calculate the mean across each row of $X_{sort}$ and assign this mean to each element in the row to get $X'_{sort}$
4. Construct $X_{normalized}$ by rearranging each column of $X'_{sort}$ in the ordering of the original $X$.

The idea of the quantile normalization is illustrated in Fig. 20.12. In

# Quantile Normalization




After Lowess normalization




After quantile normalization

# Compare

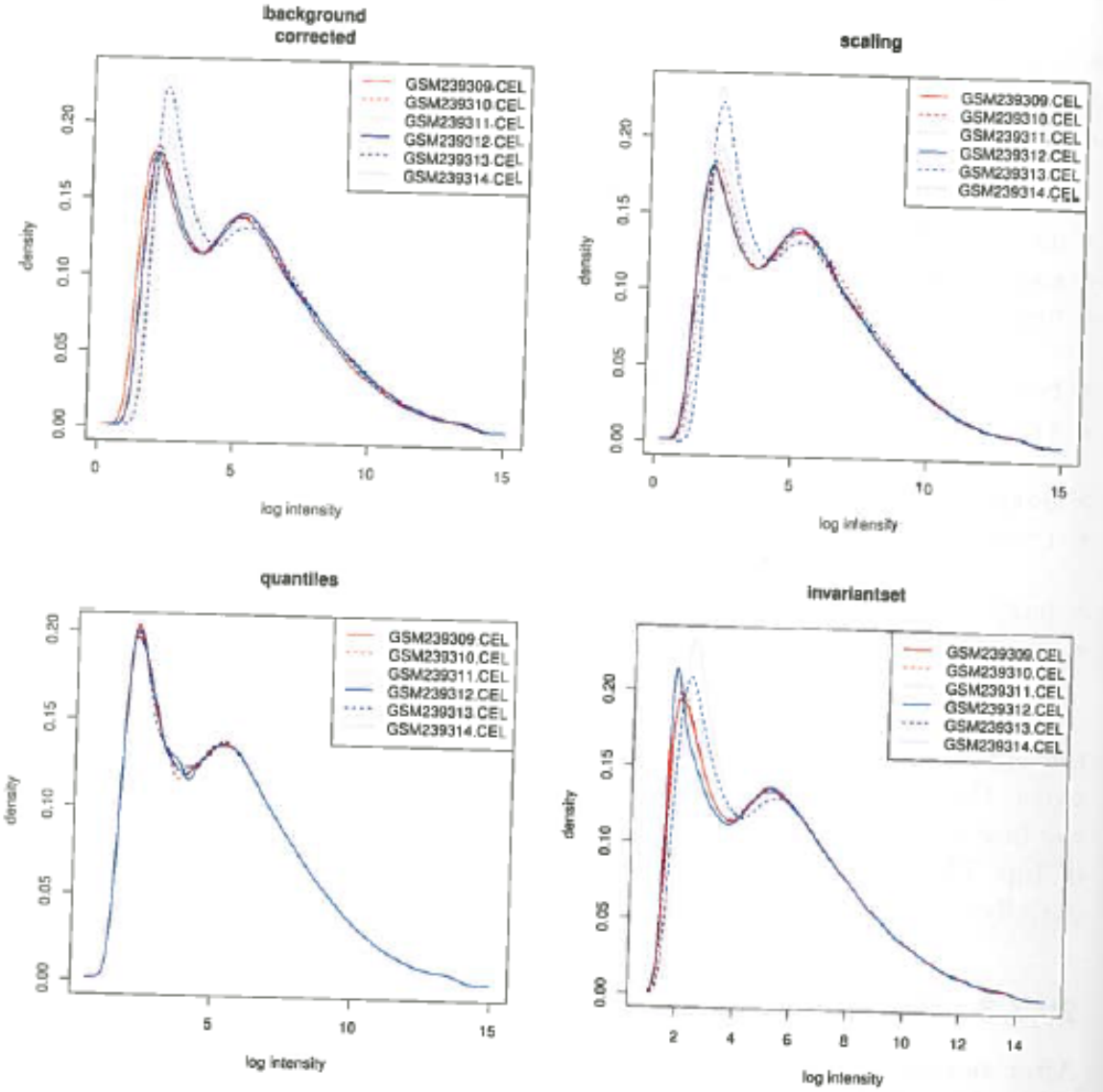



**FIGURE 20.15**: A comparison of three normalization methods. The figure shows the probe intensity distributions of the background corrected data (upper left), and the same data after three different normalization methods: scaling (upper right), quantiles (bottom left), and invariant set (bottom right).

# Compare

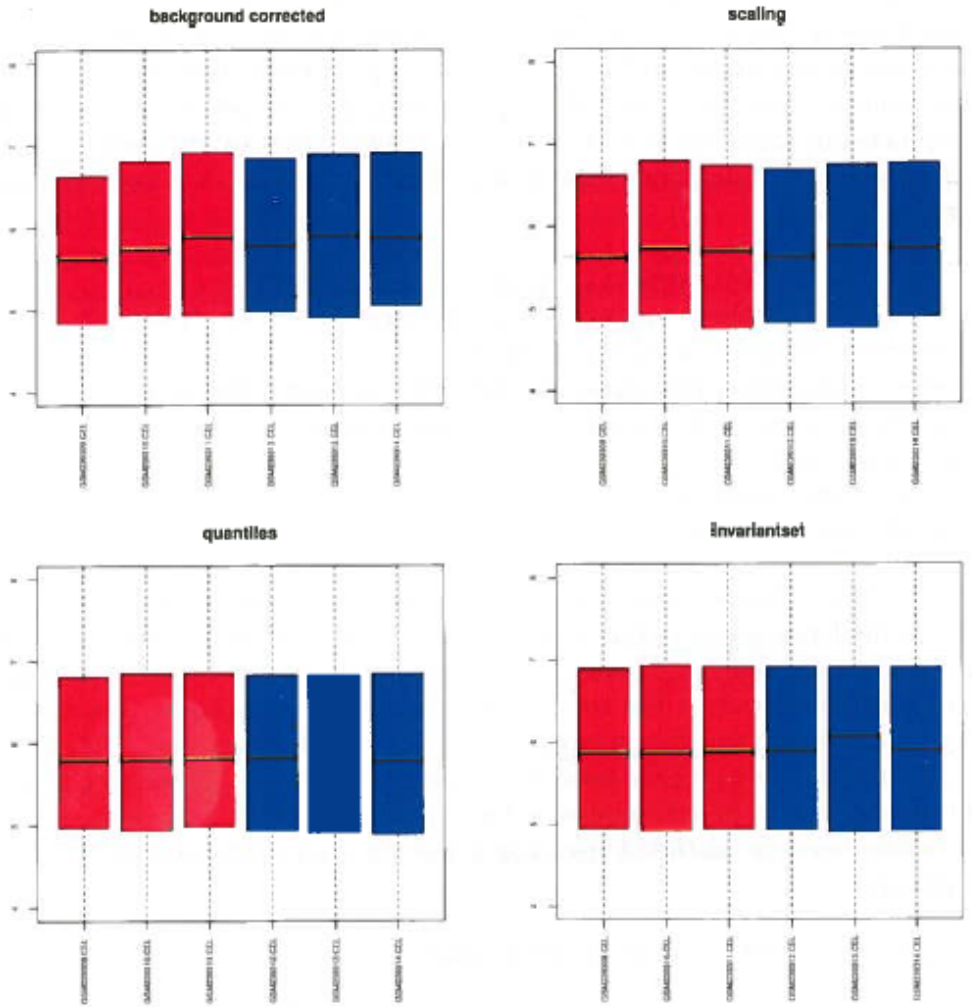



**FIGURE 20.16**: A comparison of three normalization methods. The figure shows the box plot of probe intensities of the background corrected data (upper left) and after three different normalization methods: scaling (upper right), quantiles (bottom left), and invariant set (bottom right). Note the almost perfect alignment of the distributions provided by the quantile normalization, and the very good alignment provided by the invariant set method used by dChip. The red and blue colors show the class membership of that particular sample.

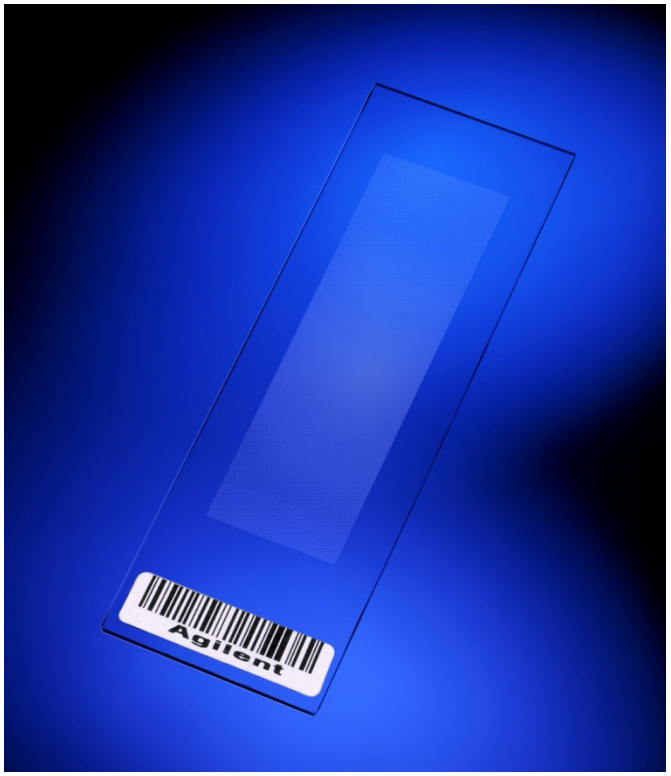
# In Closing

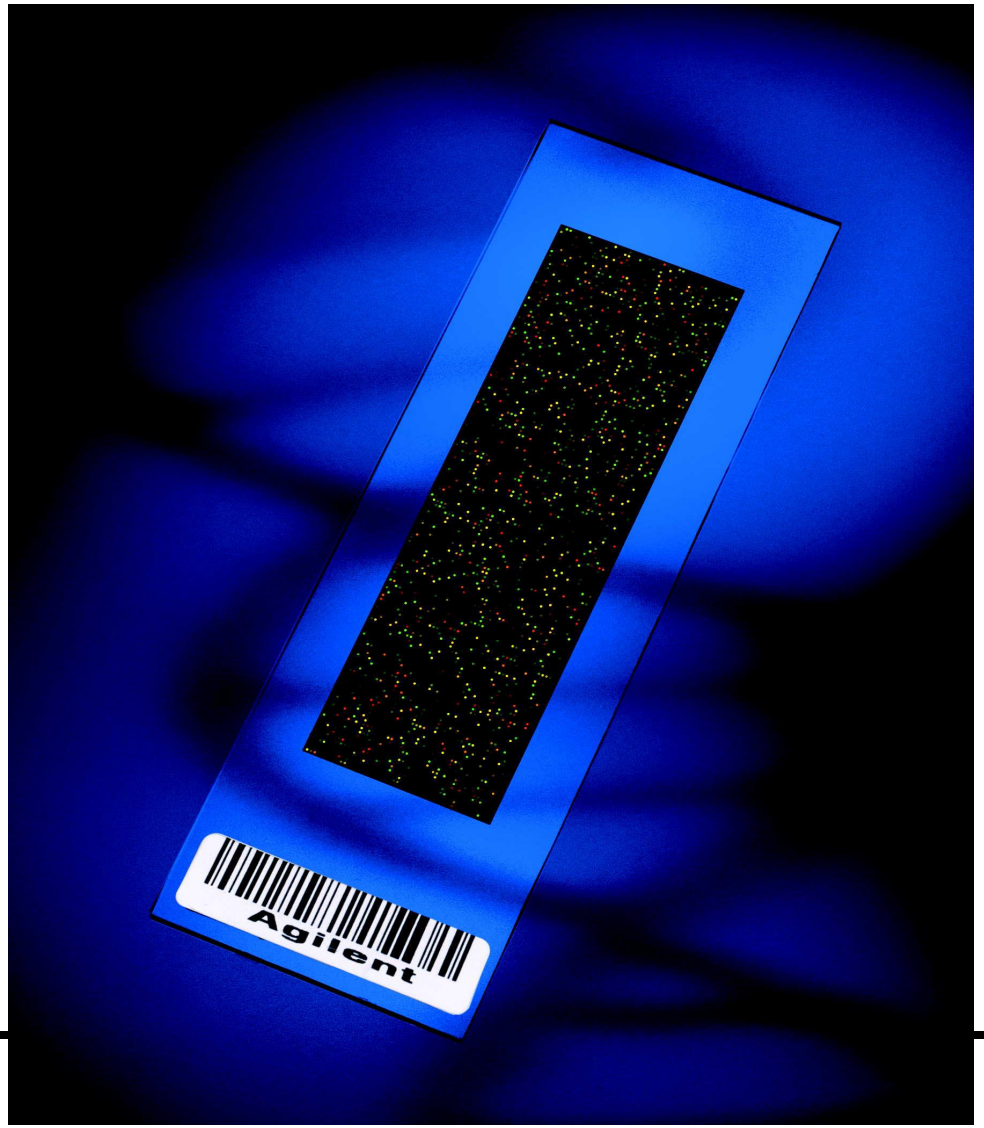
# Overall Comparison

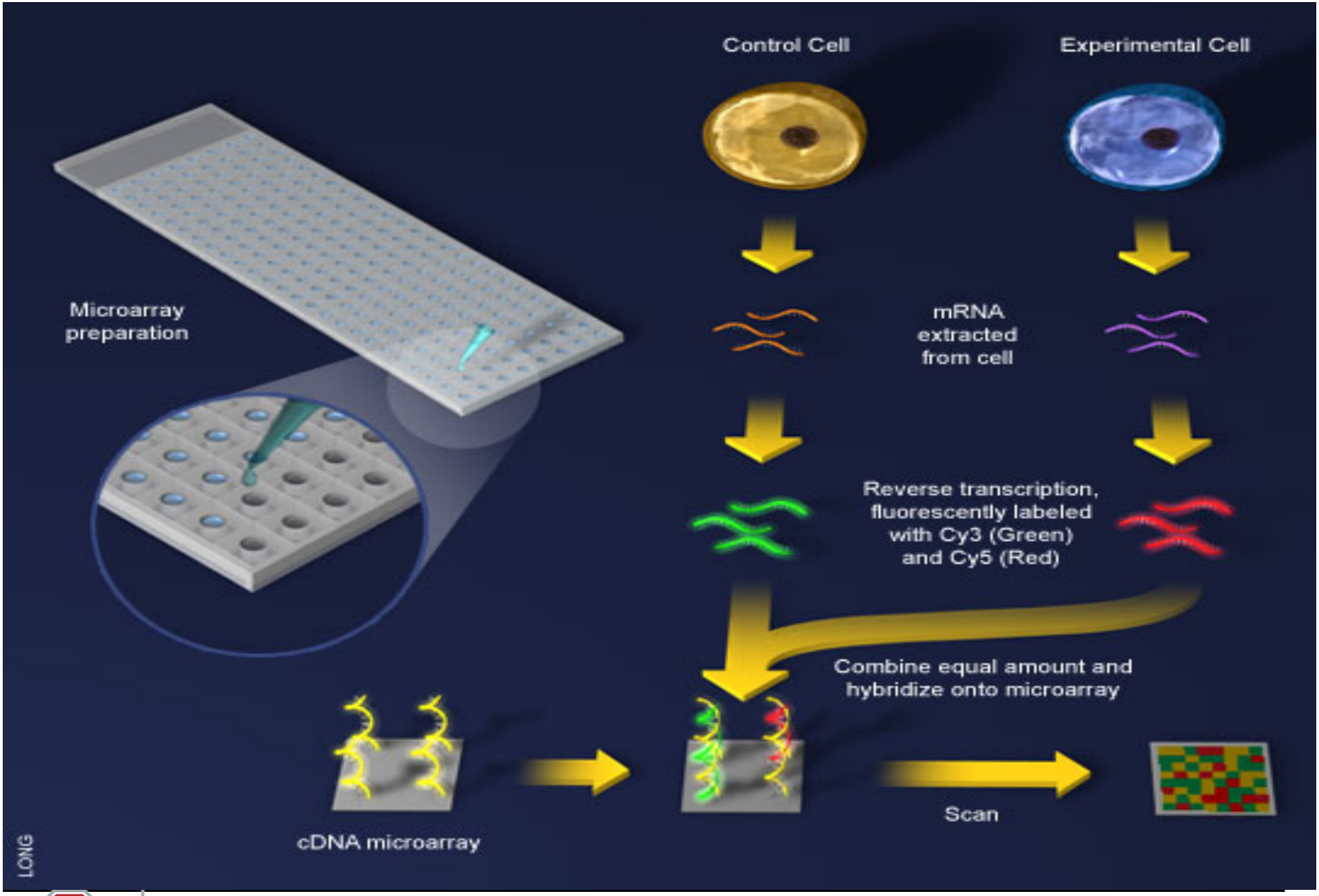
- Linear (global) – the chips have equal median (or mean) intensity
- Intensity-based (Lowess) – the chips have equal medians (means) at all intensity values
- Quantile – the chips have identical intensity distribution
- Quantile is the "best" in term of normalizing the data to desired distribution, however it also changes the gene expression level individually

# The End

# Two-color CDNA

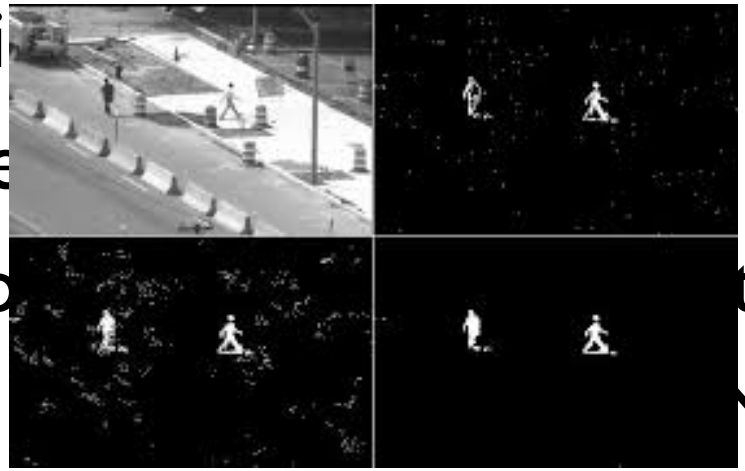


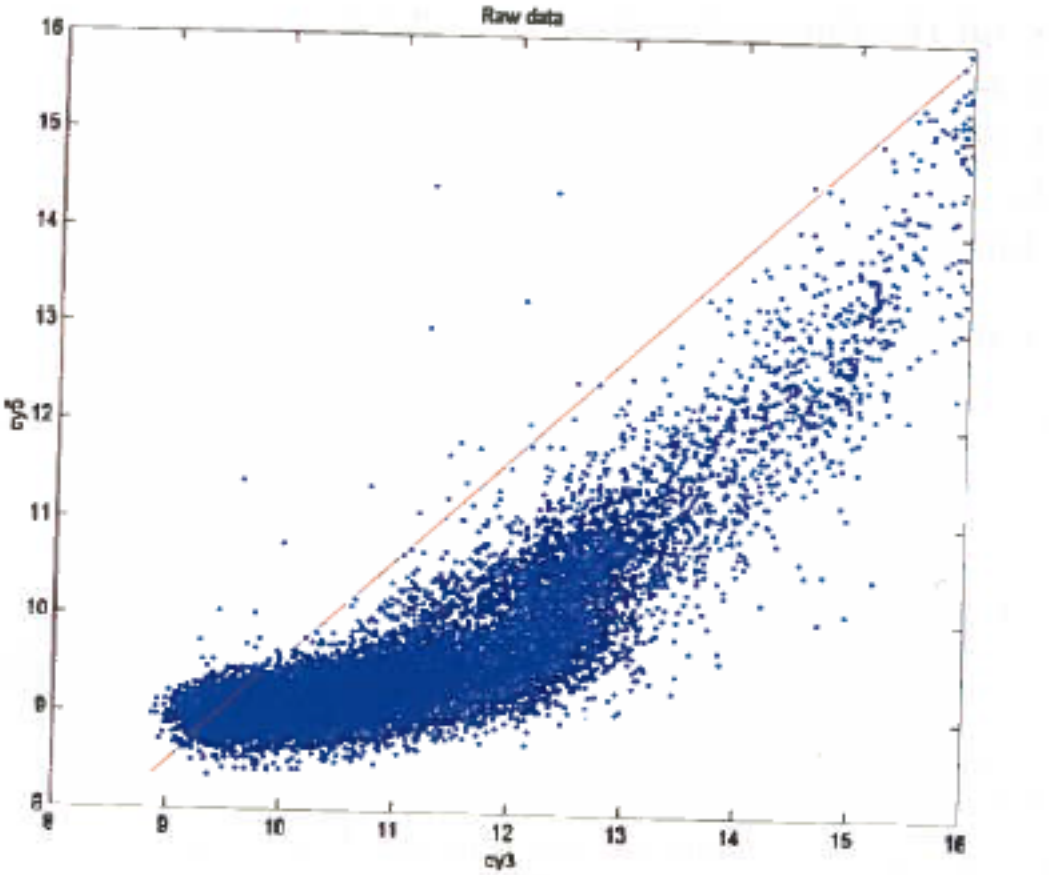

Control Cell
Experimental Cell
Microarray preparation
mRNA extracted from cell
Reverse transcription, fluorescently labeled with Cy3 (Green) and Cy5 (Red)
Combine equal amount and hybridize onto microarray
Scan
cDNA microarray
LONG

# Background Correction

- Local correcti
- Sub-grid corre
- Using blank sp ted)
- Using control NA)



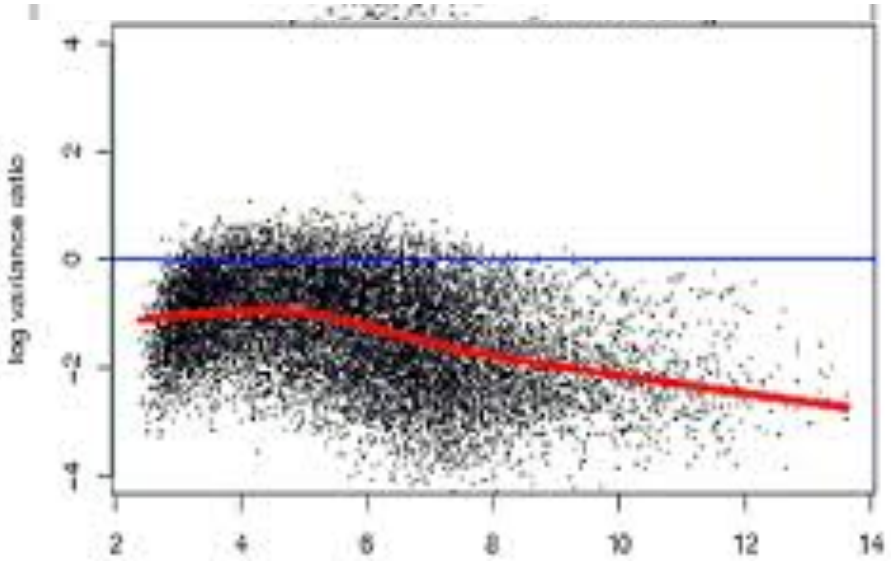
# Color Normalization



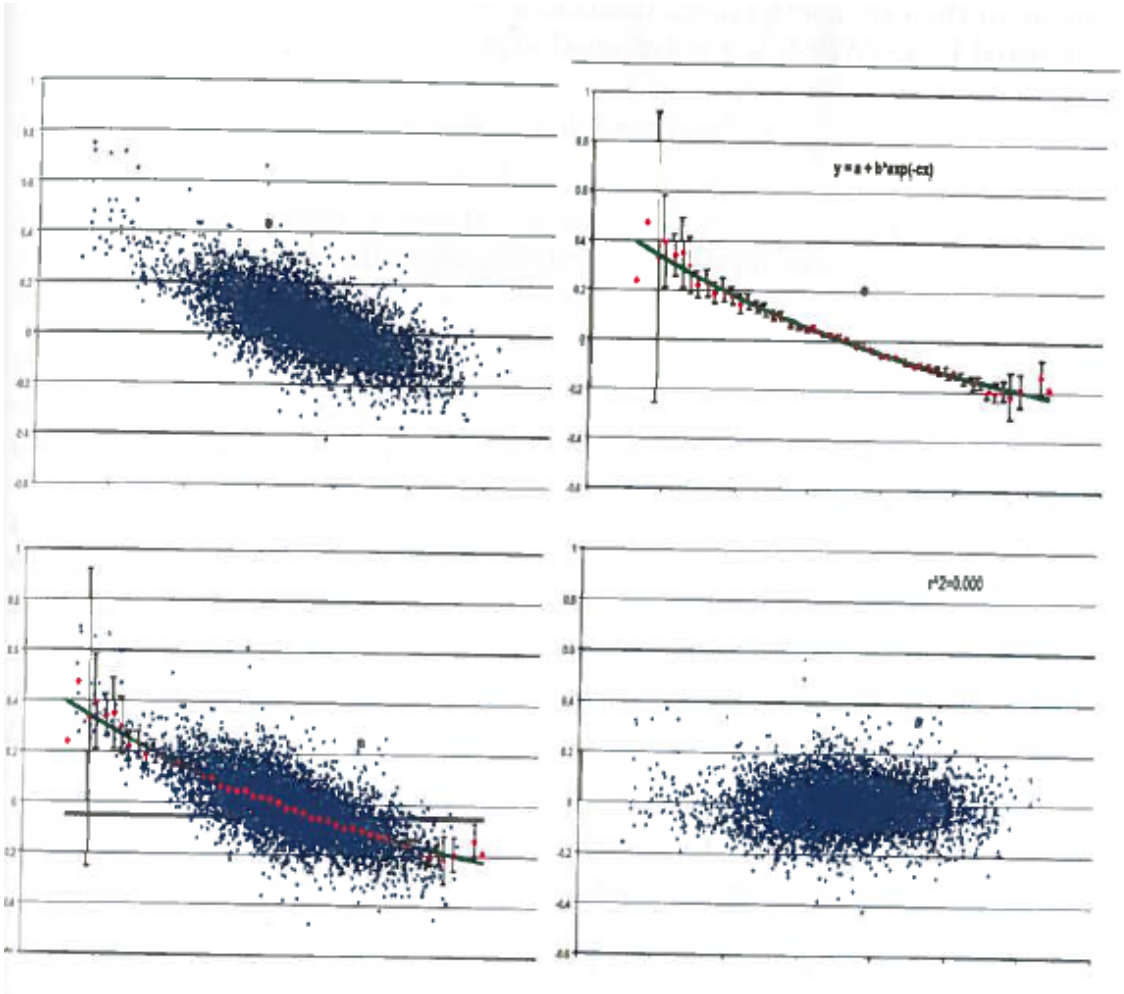
# Color Normalization

- Cy3 dye is more efficient
- mRNA for Cy3 dye is abundant

# The Curse of the Bannana



# Curve Fitting and Correction



# cDNA Variation (Table 3.3)

| Factor | Comments |
|---|---|
| mRNA preparation | Tissues, kits, and procedures vary |
| transcription | Inherent variation in the reactions, type of enzymes used |
| Labeling | Depends on the type of labeling and procedures as well as age of labels |
| Amplification (PCR protocol) | PCR is difficult to quantify |
| Pin geometry variations | Different surfaces and properties due to production random errors |
| Target volume | Fluctuates stochastically even for the same pin |
| Target fixation | The fraction of target cDNA that is chemically linked to the slide surface from the droplet is unknown |
| Hybridization parameters | Influenced by many factors, such as temperature of the laboratory, time, buffering conditions, and others |
| Slide inhomogeneities | Slide production parameters, batch-to-batch variations |
| Nonspecific hybridization | cDNA hybridizes to background or to sequences that are not their exact complement |
| Gain setting (PMT) | Shifts the distribution of the pixel intensities |
| Dynamic range limitations | Variability at low end or saturation at the high end |
| Image alignment | Images of the same array at various wavelengths corresponding to different channels are not aligned; different pixels are considered for the same spot |
| Grid placement | Center of the spot is not located properly |
| Nonspecific background | Erroneous elevation of the average intensity of the background |
| Spot shape | Irregular spots are hard to segment from background |
| Segmentation | Bright contaminants can seem like signal (e.g., dust) |
| Spot quantification | Pixel mean, median, area, etc. |

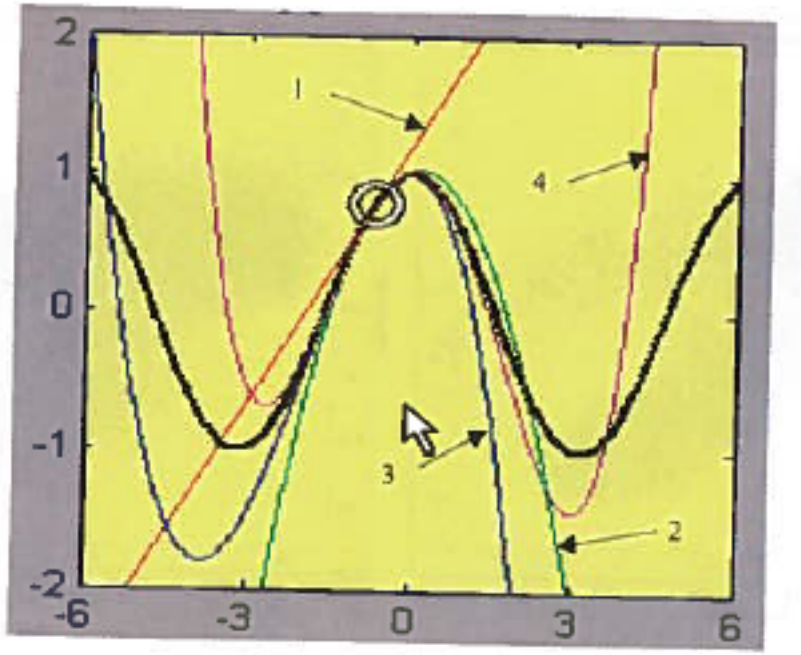
**TABLE 3.3:** Sources of fluctuations in a typical cDNA microarray experiment.

# LOWESS/LOESS

- Locally weighted Polynomial Regression
- Over-fitting vs. under-fitting
- LOWESS-1 – degree 1, LOWESS-2 - degree 2
- Divide domain into windows of size *w*
- And fit

$$w(x) = \begin{cases} (1 - |x|^3), |x| < 1 \\ 0, |x| \geq 1 \end{cases}$$

# Effects of LOWESS



# Effects of LOWESS

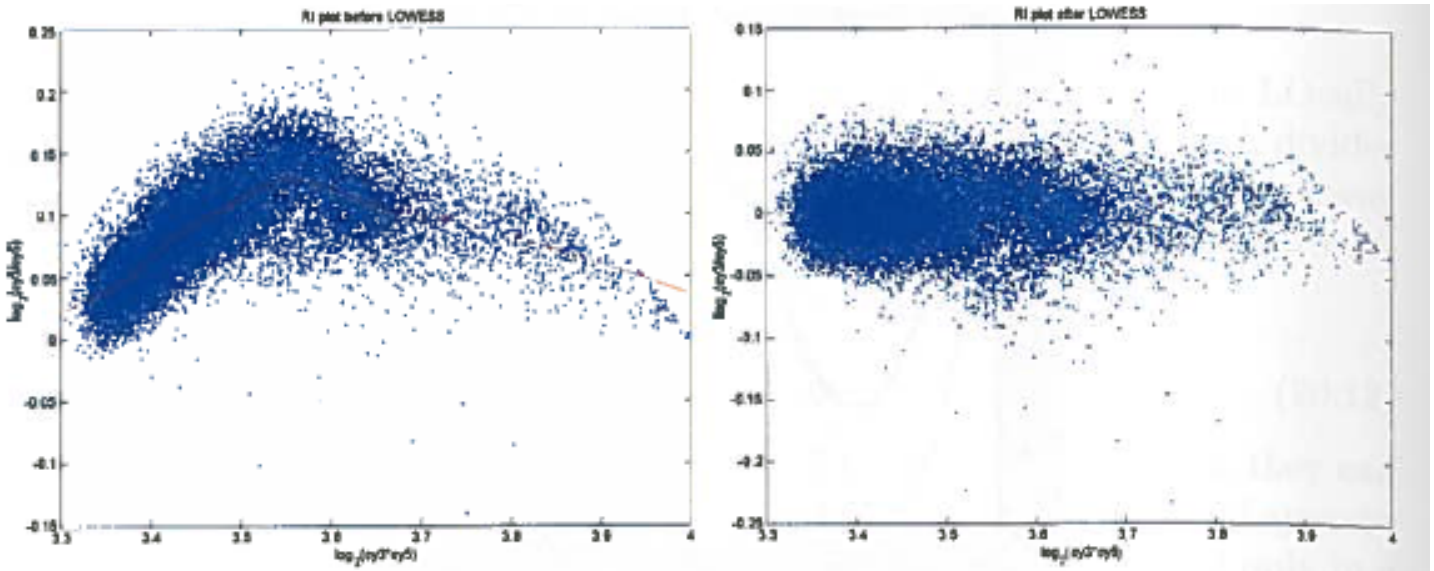


**FIGURE 20.7**: LOWESS normalization. A ratio-intensity plot before (left panel) and after (right panel) the LOWESS correction. The horizontal axis represents $\log_2(cy3)+\log_2(cy5)$ which is a quantity proportional to the overall signal intensity. The vertical axis represents the log-ratio of the two samples $\log_2(cy3/cy5)$. The left panel also shows the nonlinear regression curve.

# Effects of LOWESS

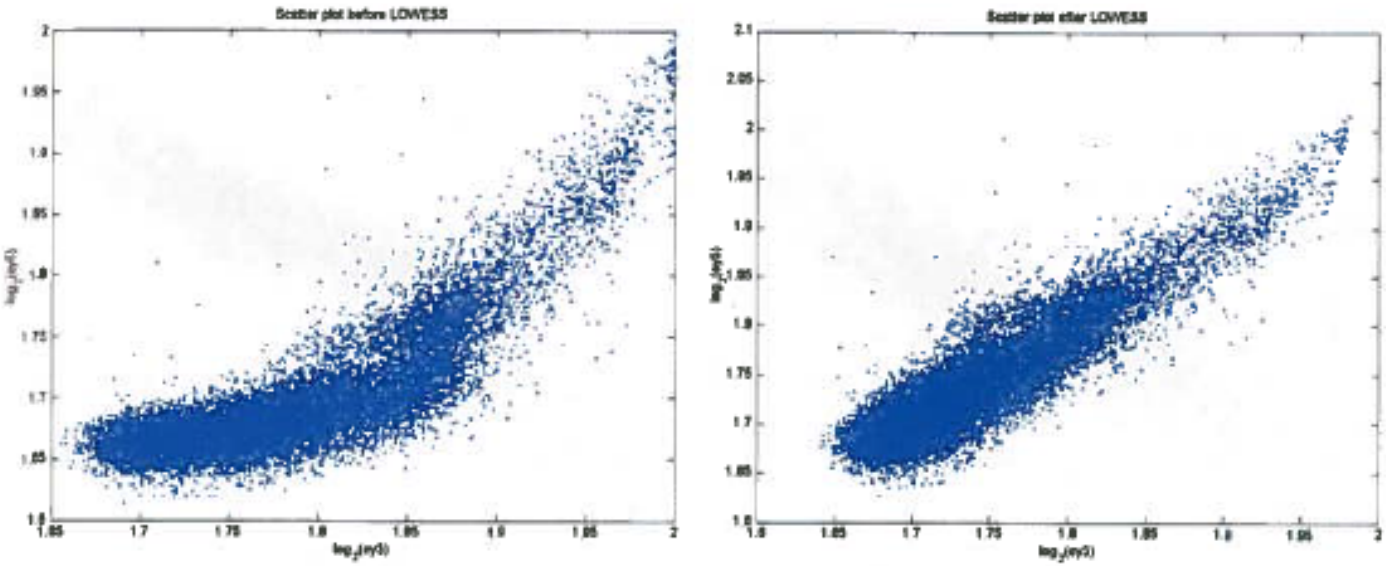



**FIGURE 20.8**: LOWESS normalization. A scatter plot before (left panel) and after (right panel) the LOWESS correction. The horizontal axis represents $\log_2(cy3)$ and the vertical axis represents $\log_2(cy5)$.

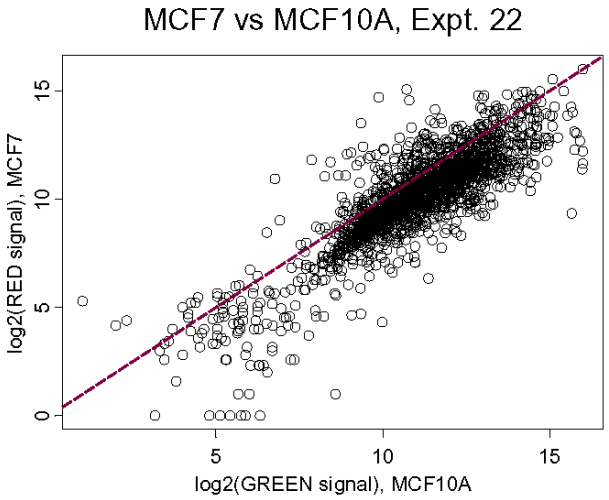
# Operation

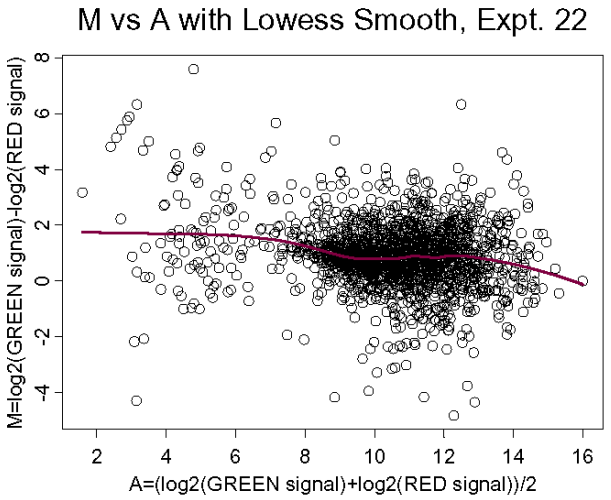
- For n points $\frac{d+1}{n} < q < 1$

- In each window, $n \times q$ points will be chosen
- Small q – less smooth but close to function
- Large q – more smooth

# Use for LOWESS

Overall magnitude of  spot intensity has impact on relative intensity between channels.

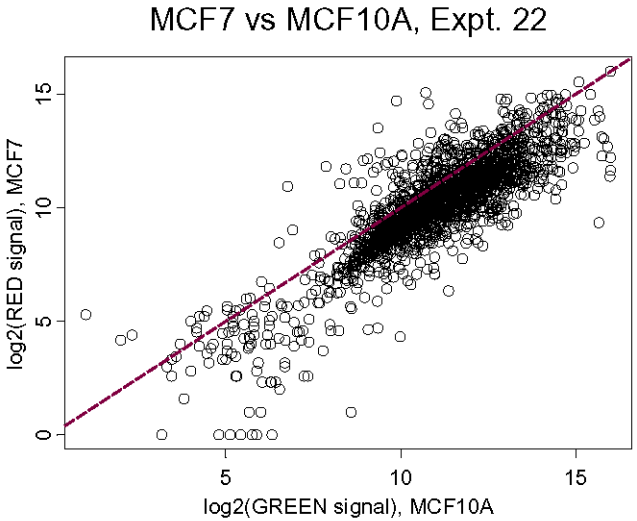


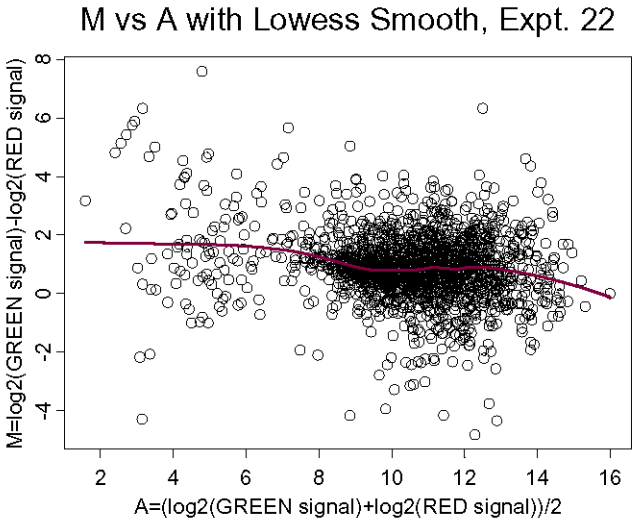




(McShane, NCI)

# Use for LOWESS

- Nonlinear
- Gene-by-gene, could introduce bias
- Use only when there is a compelling reason




(McShane, NCI)

# Advantages

- No need to specify type of function
- Specify degree of polynomial and smoothing factor q
- Least square is well studied

# Disadvantages

- No model or function is generated
- LOWESS will need to be applied for every data; cannot transfer models
- Very expensive
- Suspetible to noise and outliers