

# Microarray Data Pre-processing

Ana H. Barragan Lid

# Hybridized Microarray

- Imaged in a microarray scanner
- Scanner produces fluorescence intensity measurements
- Intensities correspond to levels of hybridization
- Fluorescence intensity values are stored as image file = raw data



# What is pre-processing?

- **Convert raw data to useful biological data:**
  - ✓ Image data to intensities values
  - ✓ Quality control
  - ✓ Remove bias  
(Filtering, normalization, transformation)

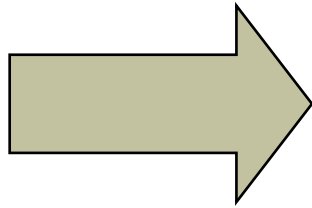
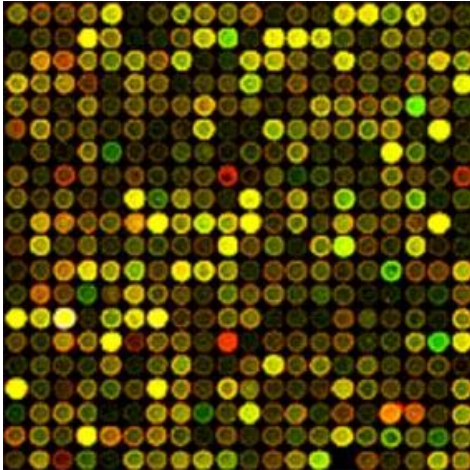
# Why pre-process?

- To avoid using bad data
- To distinguish noise and the actual biological data
- To be able to compare data from multiple arrays

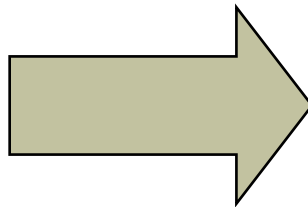
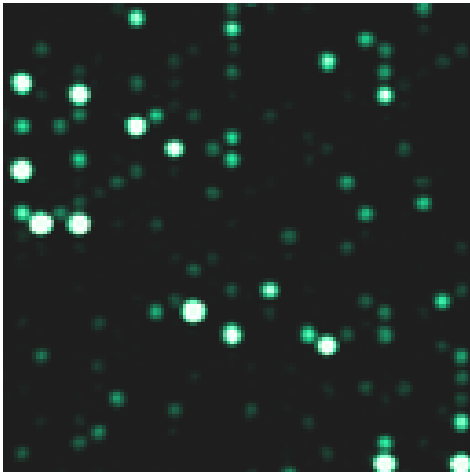
# Pre-processing

- Image Analysis
- Background adjustment
- Filtering
- Normalization
- Quality control

# Image Analysis



	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Gene 1	1235	546	943	263	136	314
Gene 2	1266	32	556	435	687	2718
Gene 3	947	2829	389	3820	2039	1414
Gene 4	392	2398	84	829	4392	512
...	...	...	...	...	...	...



# Image Analysis

Commercial microarrays:

- Specifically design software packages
- Automatically visualize and quality report

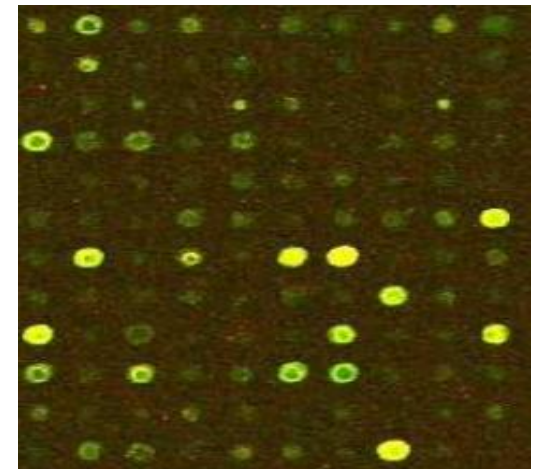
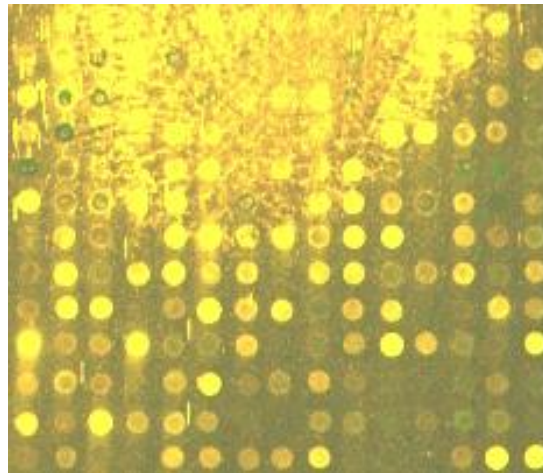
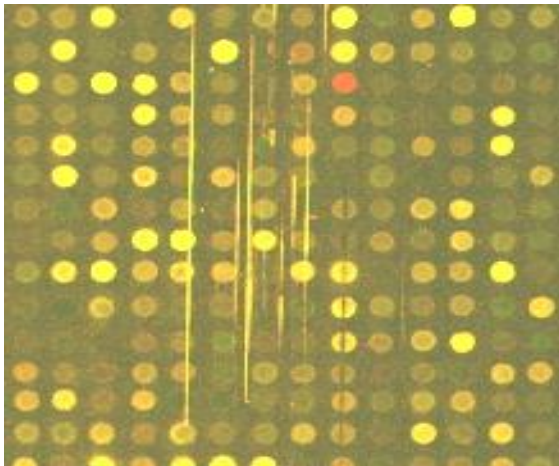
But, commercial arrays are not offered for everything e.g

- ✓ Protein arrays
- ✓ Custom arrays

# Image Analysis

Visual inspection in scanner or platform software

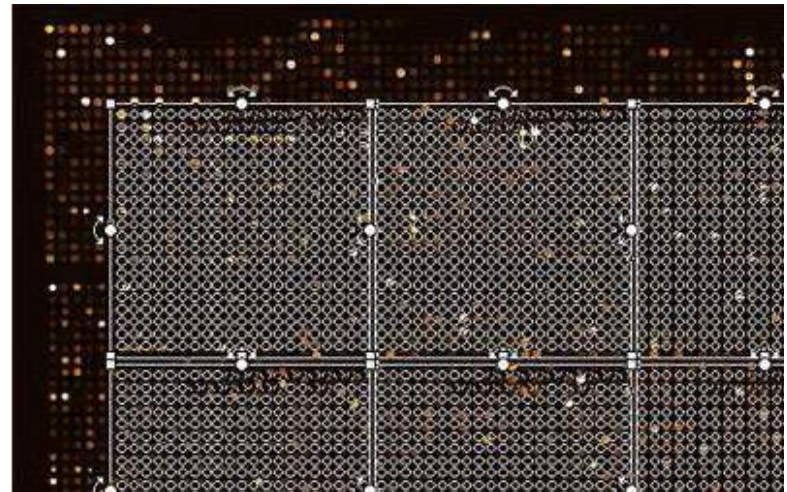
- Look for scratches and shadows
- Washing artifacts
- Manufacture errors
- Odd spots (donut, star shape etc)
- Missing spots



# Image Analysis

Usually automatic from commercial software

- Gridding
- Gene annotation
- Spot segmentation



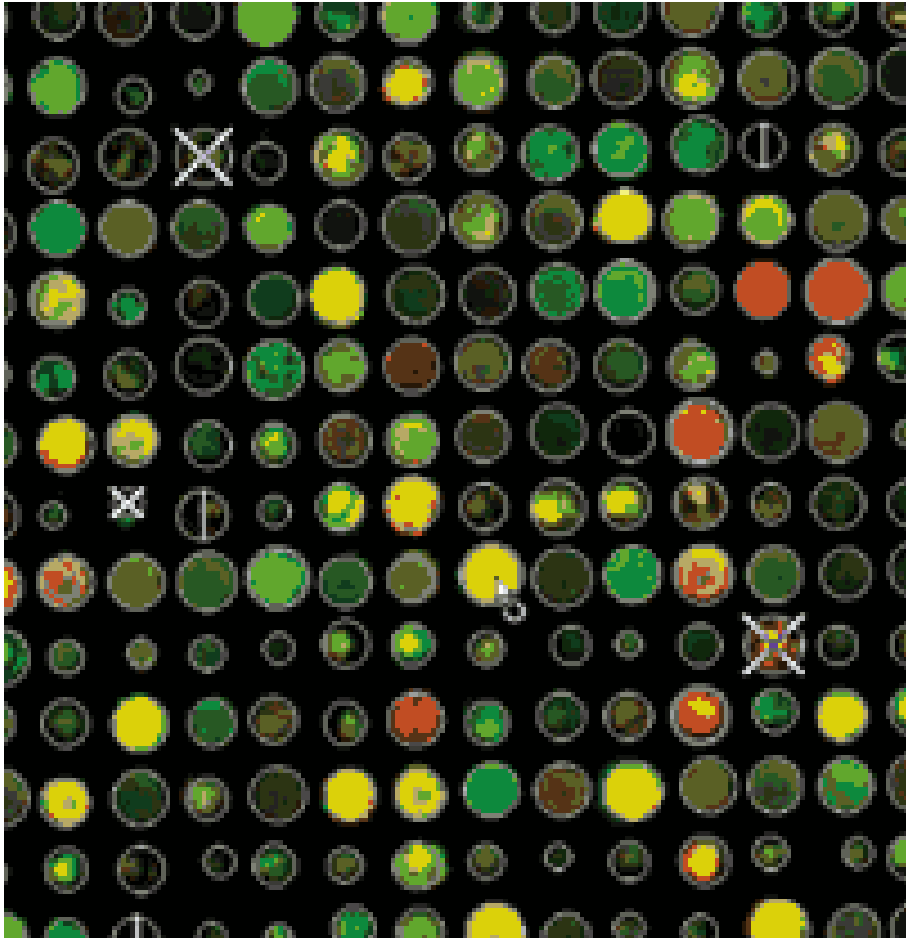
# Image Analysis

## Addressing or gridding

- Assign coordinates/physical position to each spot
- Takes into account small changes caused in array production such as displacement of spots



# Image Analysis



- Flag bad spots
  - ✓ Spot size
  - ✓ Circularity measure
  - ✓ Uniformity
  - ✓ Signal strength
  - ✓ Spot intensity relative to background
  - ✓ Software to extract the information/intensities

# Pre-processing

- Image Analysis
- **Background adjustment**
- Filtering
- Normalization
- Quality control

# Background Adjustment

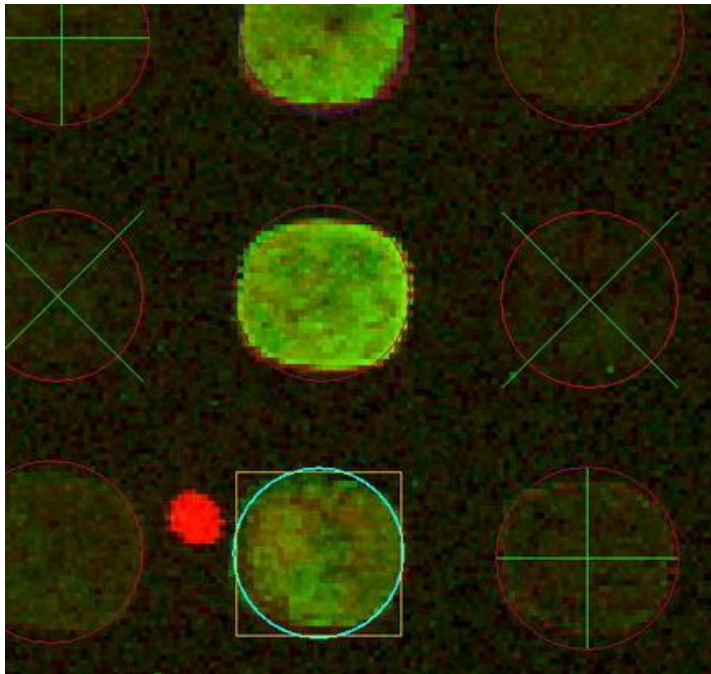
Spot intensity = background + foreground

Surrounding background can include:

- No hybridization
- Non specific hybridization
- Other fluorescent artifacts



# Background adjustment



- Why?
  - ✓ More accurate measure of spot intensity
  - ✓ Reduces bias
- How?
  - ✓ Make background more homogeneous

# Pre-processing

- Image Analysis
- Background adjustment
- **Filtering**
- Normalization
- Quality control



# Filtering

- Remove data that will contribute to noise or bias
- Low intensity, bad quality, empty spots, outliers, control probes



# Filtering

- Filtering criteria
  - ✓ Spot size/shape
  - ✓ Foreground/background intensities
  - ✓ Type of spot
  - ✓ Number of replicas
  - ✓ Variation in replica signal intensities



# Filtering

- Categories of spots to filter
  - ✓ Controls
  - ✓ Saturated
  - ✓ Poor quality
  - ✓ Too weak

# Filtering

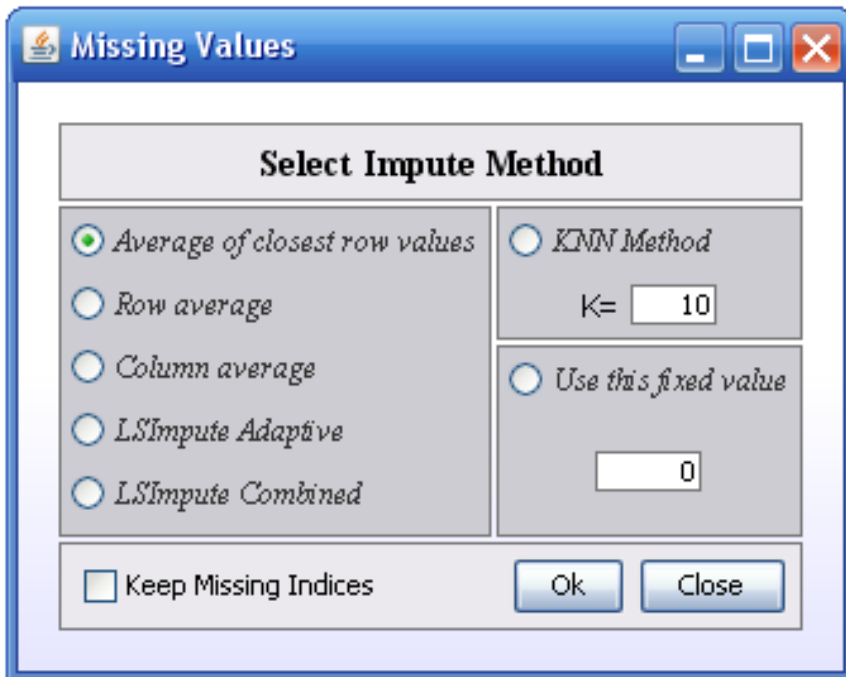
## Missing values

- Removal of bad quality spots may introduce "missing values" for some genes
- Some analysis programs does not tolerate this
- May have to impute missing values
- How?



# Filtering

## Imputing Missing Values



K-nearest neighbor algorithm

- Identifies other genes with expression most similar to the genes of interest (euclidean distance)
- Weighted average of values for those genes is used to estimate the missing values

*KNN-method - Troyanskaya, O, Bioinformatics. 2001 17:520-525.*

# Pre-processing

- Image Analysis
- Background adjustment
- Filtering
- **Normalization**
- Quality control

# Normalization

- Correct for differences not representing true biological variation between samples
- Remove systematic/technical variations in the relative intensities of each channel
- Aims to correct for differences in intensities between samples (same or different slides)

# Normalization assumptions and approaches

- Some genes exhibit constant mRNA levels:
  - Housekeeping genes
- The level of some mRNAs are known:
  - Spike-in controls
- The total of all mRNA remains constant:
  - Global median and mean; Lowess
- The distribution of expression levels is constant
  - quantile

From: WIBR Microarray Course, © Whitehead Institute, November 2004



# Normalization by global mean (total intensity)

- Assumes that some genes are differentially expressed but most are equivalently expressed
- Meaning those genes up- or down-regulated will balance each other out
- The summed intensity values should be equal and where they differ, a constant factor can be calculated to rescale all intensity values

- Multiply/divide all expression values for one color (or array if one-color) by the constant factor calculated to produce a constant mean (or total intensity) for every color/array

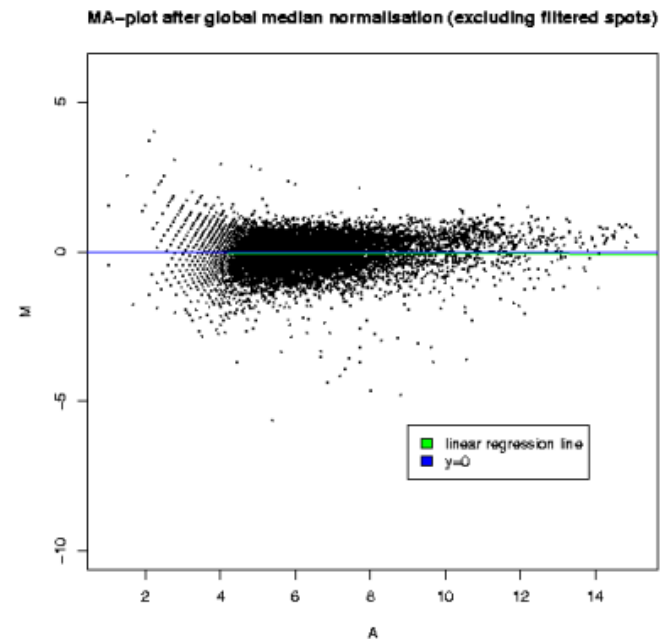
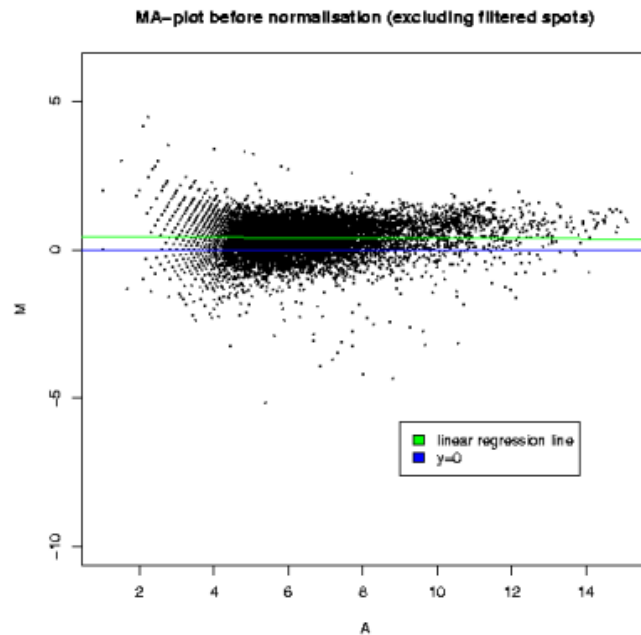
## Example with two one-color arrays

Chip	Mean expr (raw)	Total expr (raw)	Norm. factor	Mean expr (norm)	Total expr (norm)
A	2.000	100,000	<b>0.5</b>	1.000	50,000
B	2.200	110,000	<b>0.4545</b>	1.000	50,000

From: WIBR Microarray Course, © Whitehead Institute, November 2004

# Global median normalization

- Transform all expression values to produce a constant median (instead of mean)
- Linear regression Ratio vs Intensity



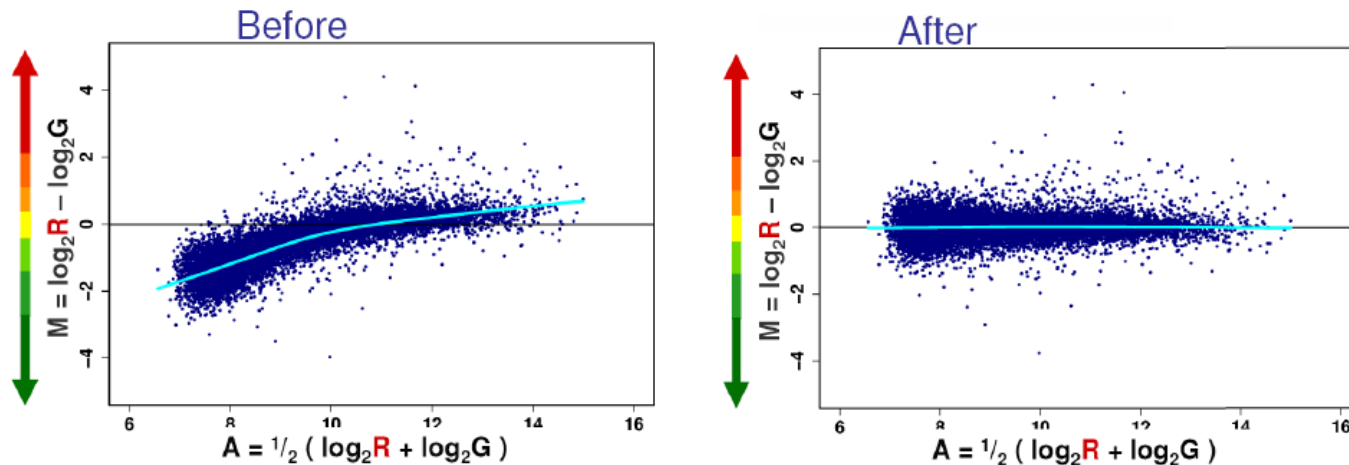
# Lowess

- Non linear regression Ratio vs Intensity
- Used on intensity-depended bias
- As a result, the normalization factor needs to change with spot intensity mean

MA-plot

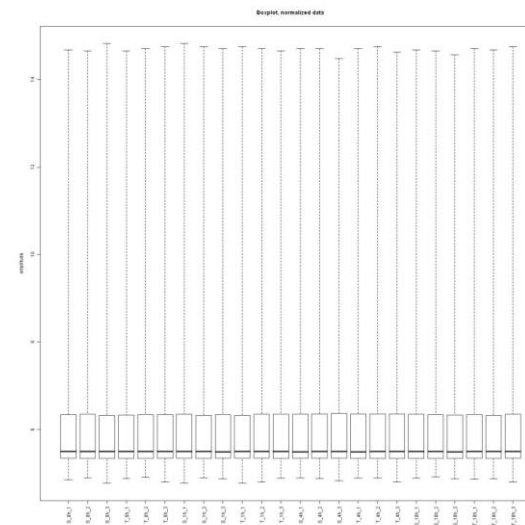
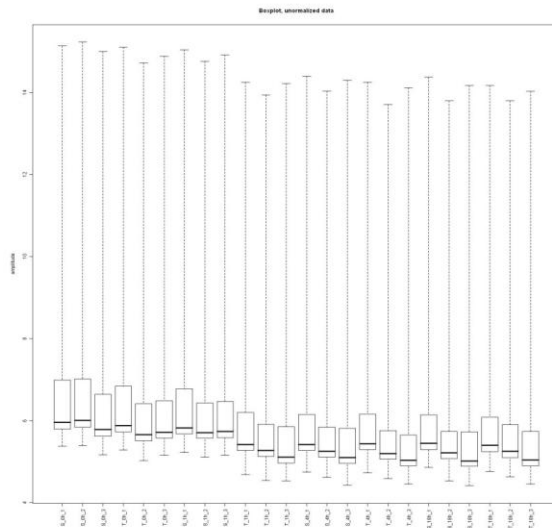
M = Ratio of Red vs green channel or ratio between two different arrays

A = Signal intensity



# Quantile

- Different chips may have the same median or mean but still very different distributions
- Assuming the chips have a common distribution of intensities, they may be transformed to produce similar distributions



# Normalization – between arrays

- The intensity distributions across arrays are assumed to be the same
- This is not always/never true
- Intensity distributions need to be similar for the arrays to be comparable



# Normalization

- Different probes / spots can be involved in the normalization process
  - ✓ Based on all the genes on the array
  - ✓ Based on controls
- Which algorithm
  - ✓ Technology
  - ✓ The shape of the data distribution
  - ✓ Always look at the data before and after normalization



# Quality control

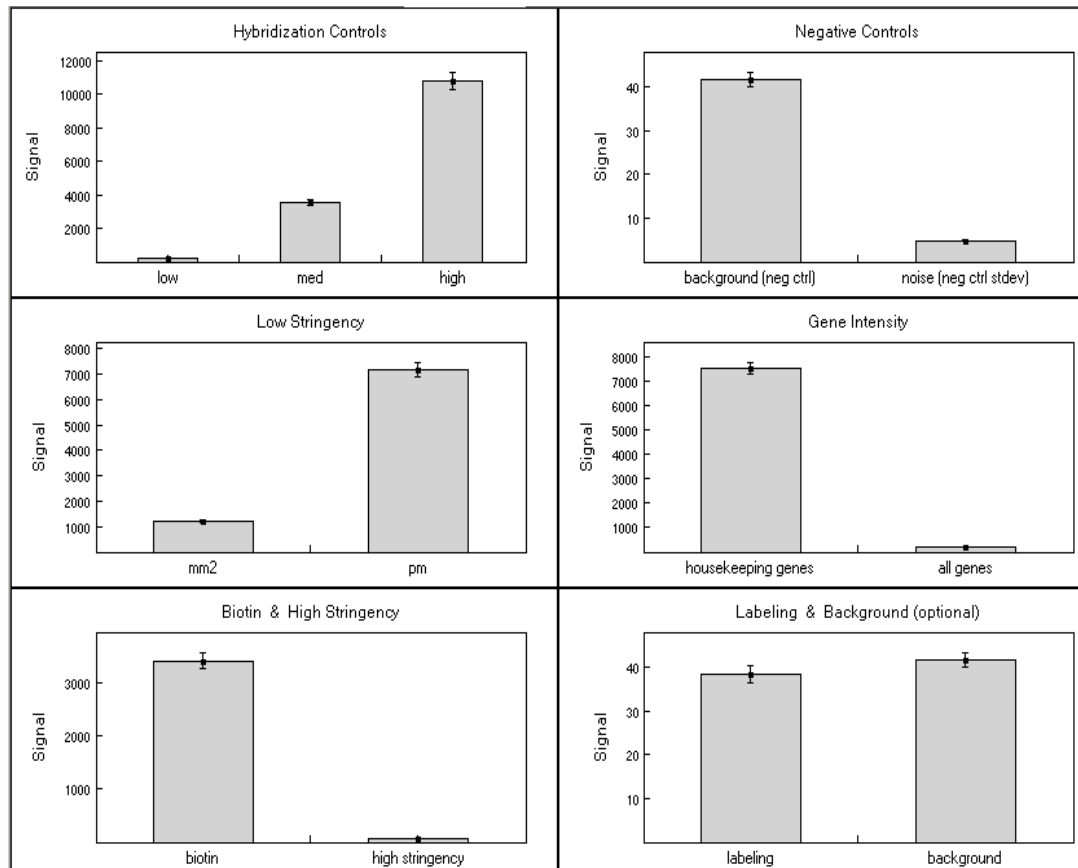
- Many steps influence data:
  - ✓ Sampling
  - ✓ Extraction
  - ✓ Labeling (sample dependent control)
  - ✓ Hybridization (sample **in**dependent control)
  - ✓ Scanning (sample **in**dependent control)
  - ✓ Extraction of data



# Different levels of quality control

- Array level
  - ✓ Assess each spot and surroundings
    - Foreground and background
    - Control spots
    - Flags
    - Plot
- Experiment level
  - ✓ Comparing all arrays to identify outliers and batch effects

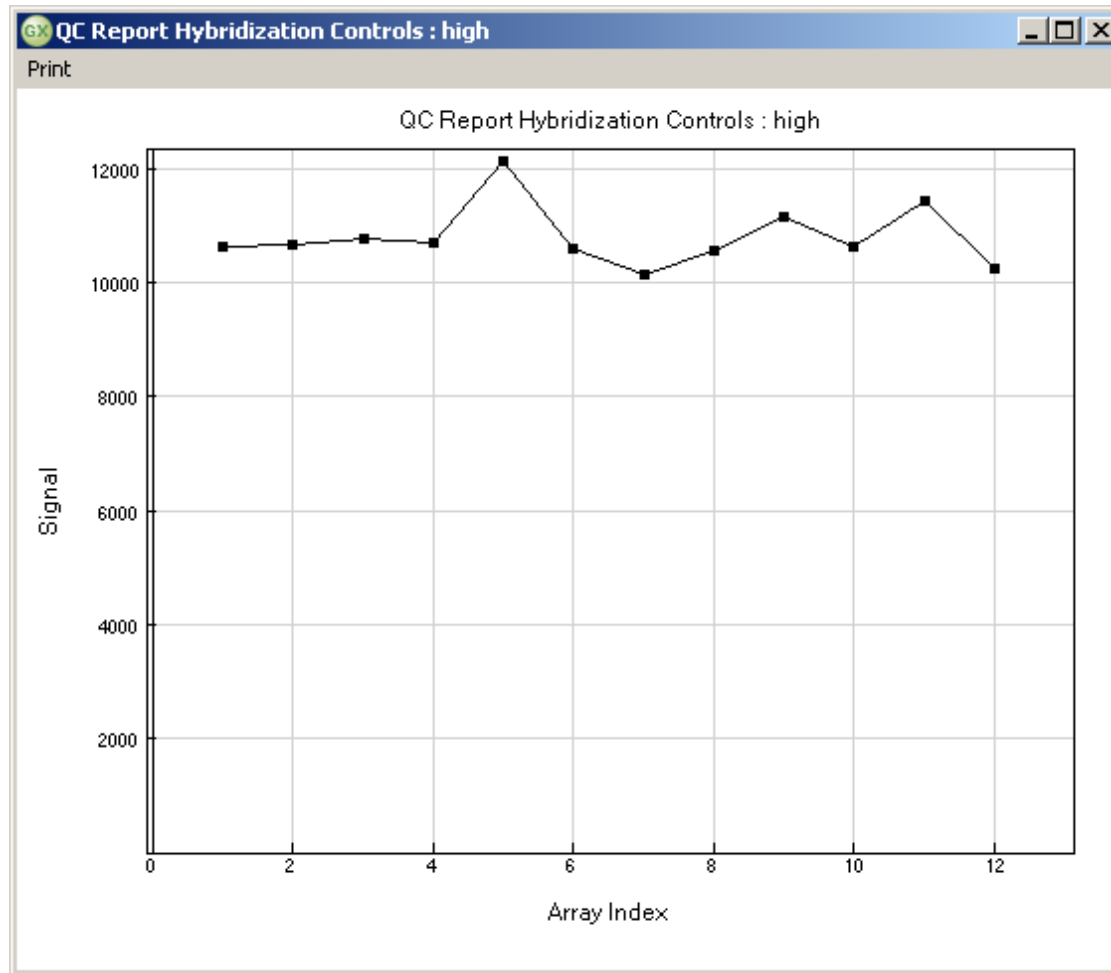
# Illumina, GenomeStudio



Sample independent controls

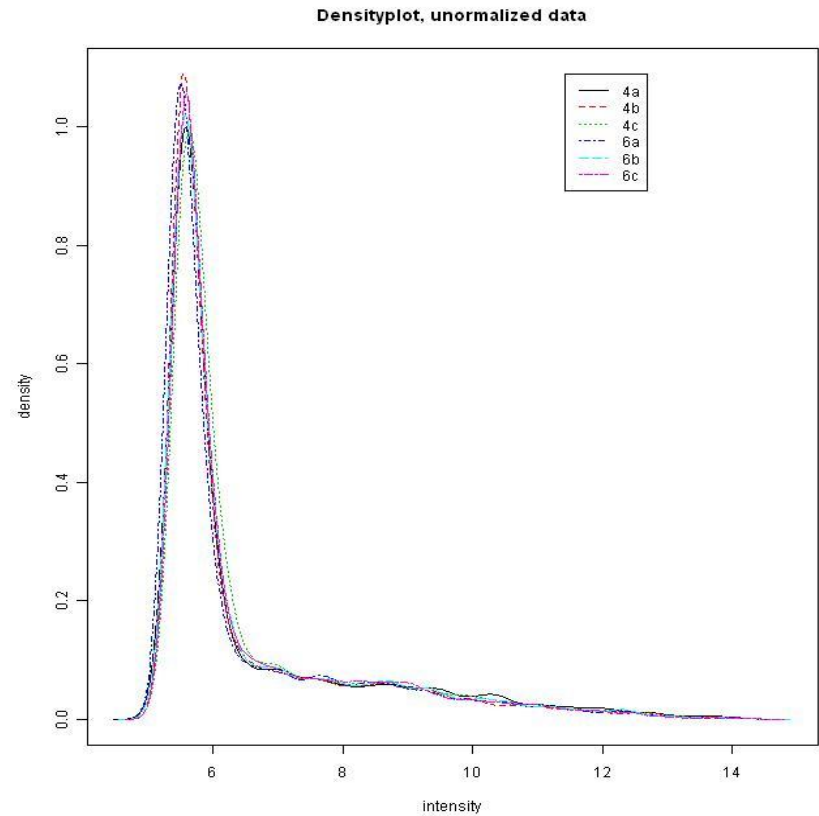
Sample dependent controls

# Illumina, GenomeStudio



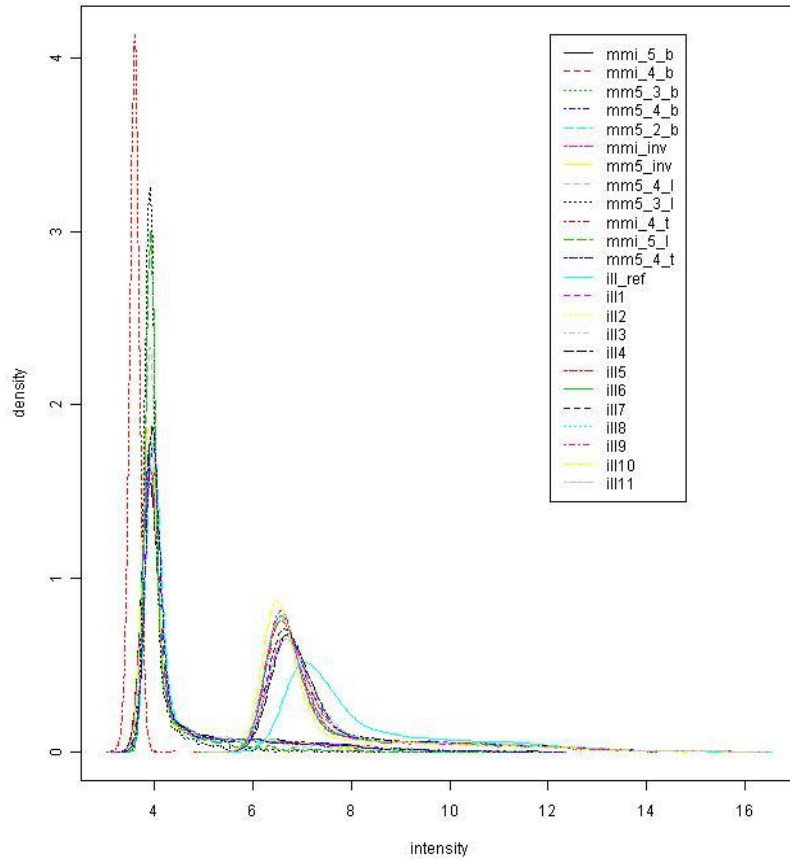
# Density plot

- Histogram/density plot
  - ✓ Distribution of the intensity for each array

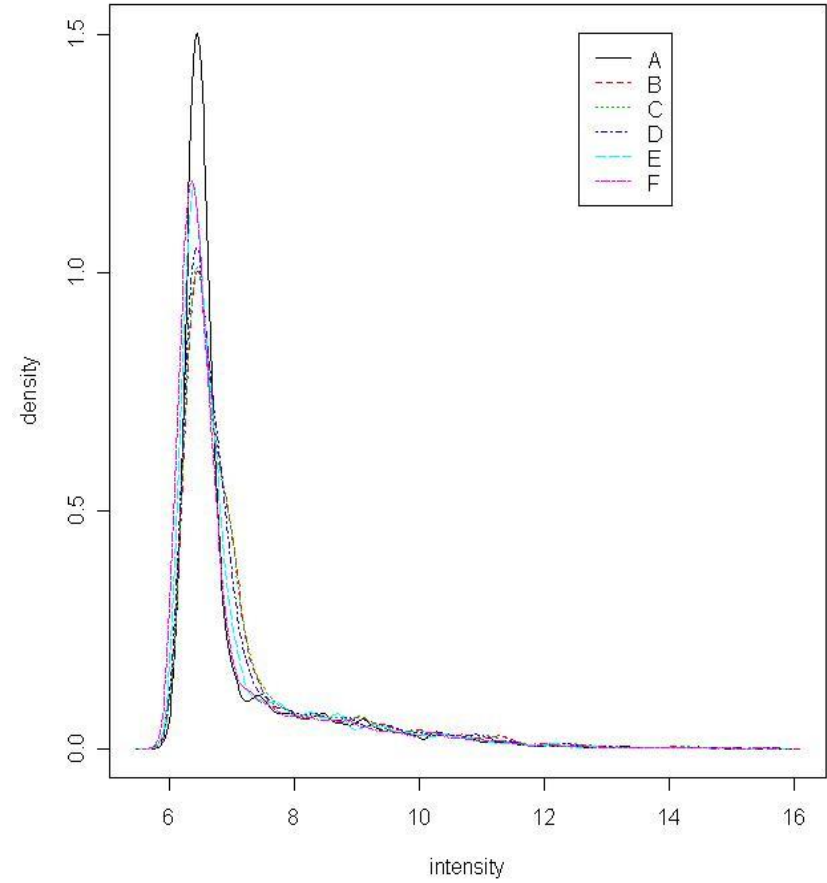


# Density plot

Densityplot, unnormalized data

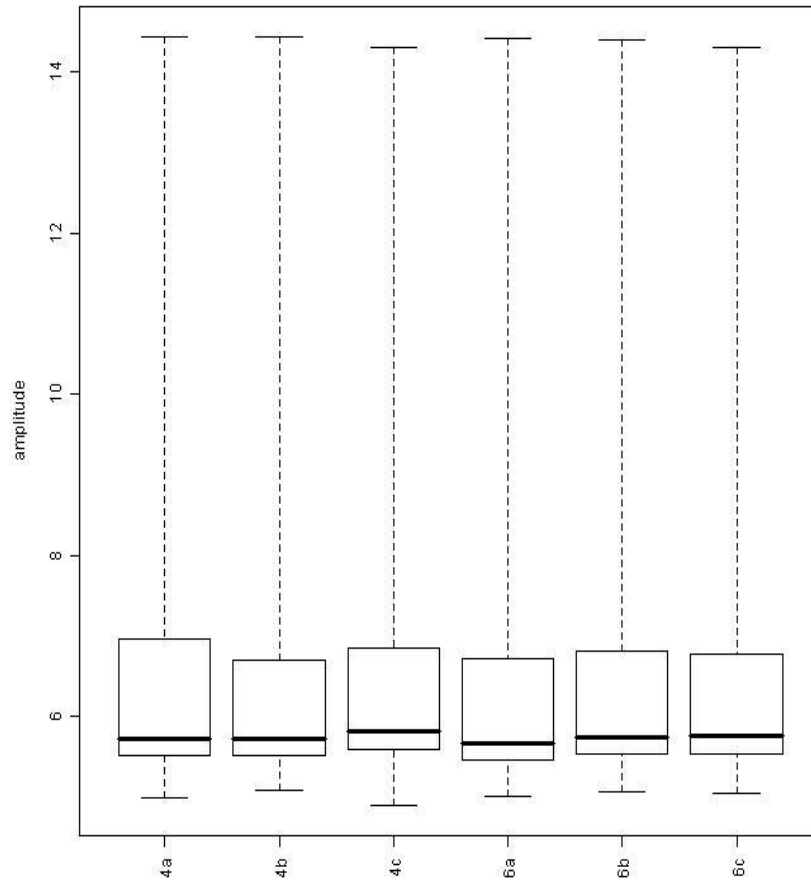


Densityplot, unnormalized data

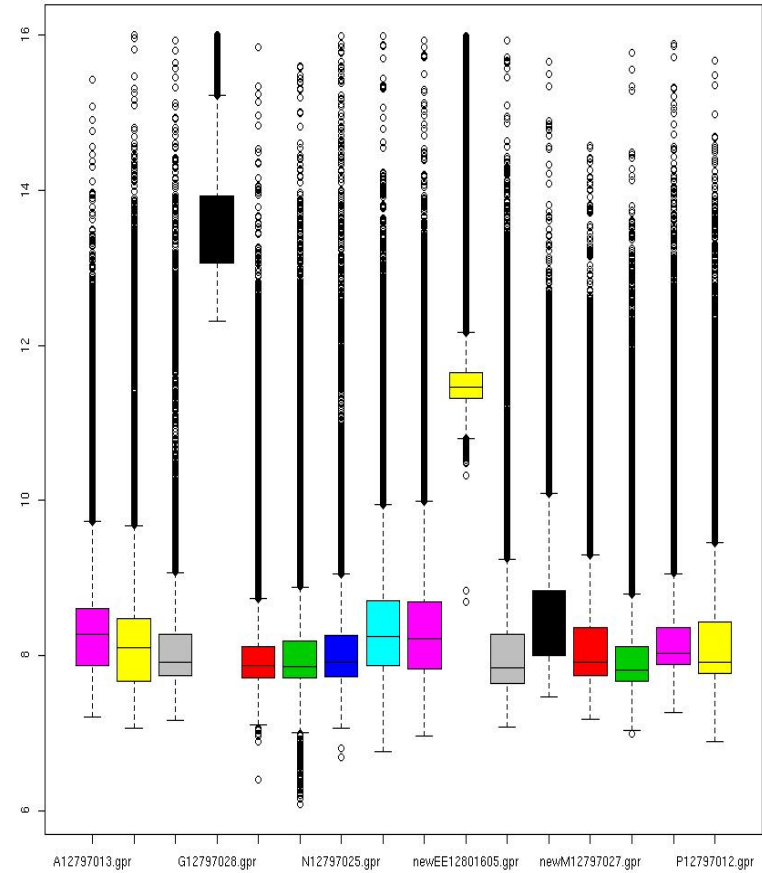


# Box plot

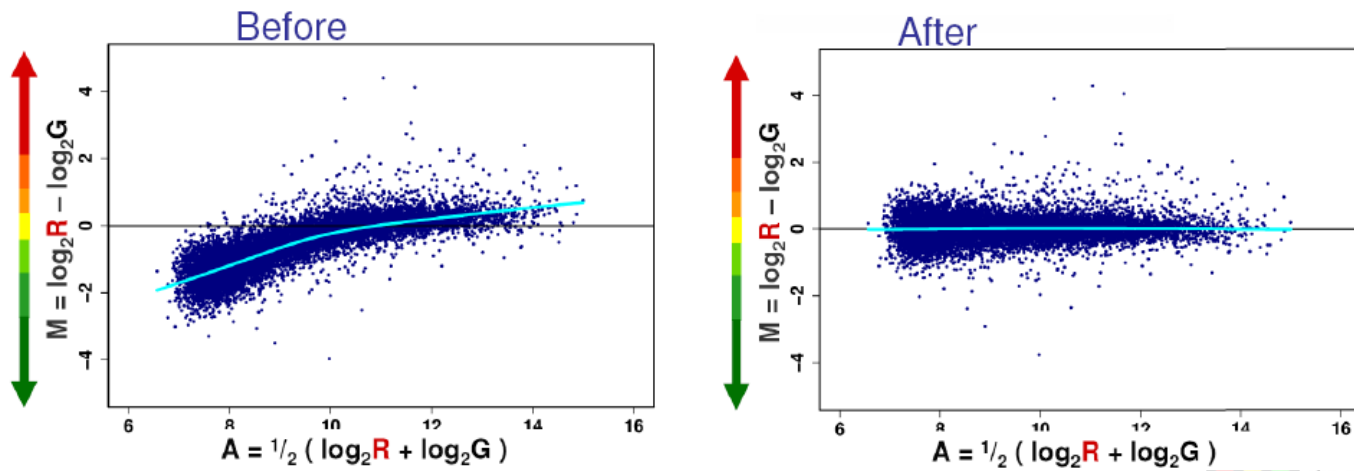
Boxplot, unnormalized data



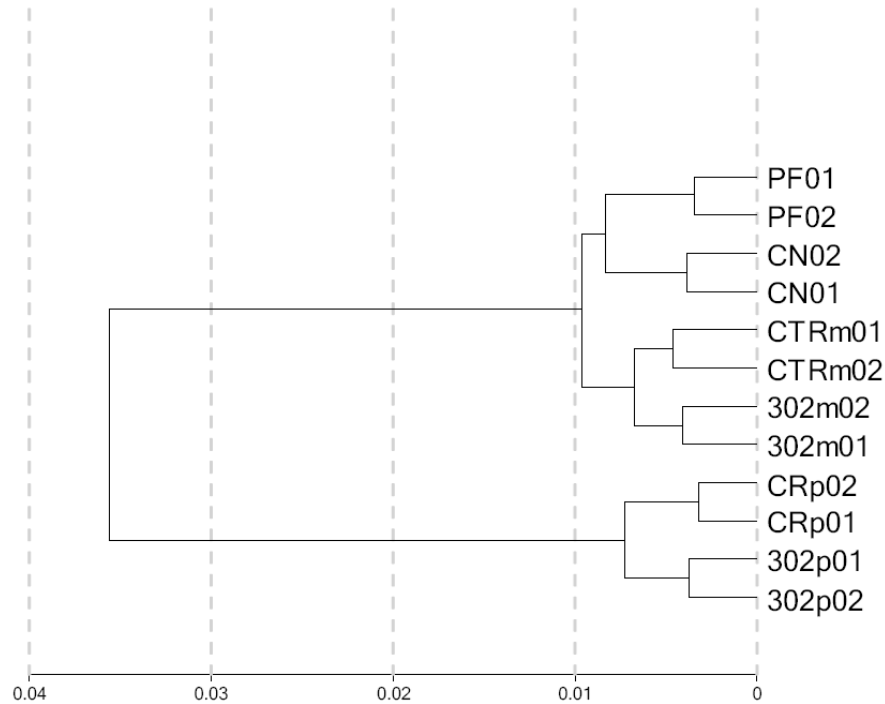
Raw Green Foreground (log2)



# Scatter plot



# Clustering



The similarity between the expression profiles is determined by cluster analysis and shown as a dendrogram. The length of the line connecting two samples indicates the similarity between the samples (short line: highly similar pattern). Here, the clustering is used as a final quality check on the data.

# Pre-processing

- Different ways/order
- Differences between technologies
- Be modest

# Summary

- A certain amount of pre-processing is needed
- But do not over pre-process
- Different technologies, different people, different implementations, different ways
- Read and understand what you are doing



Thank you!