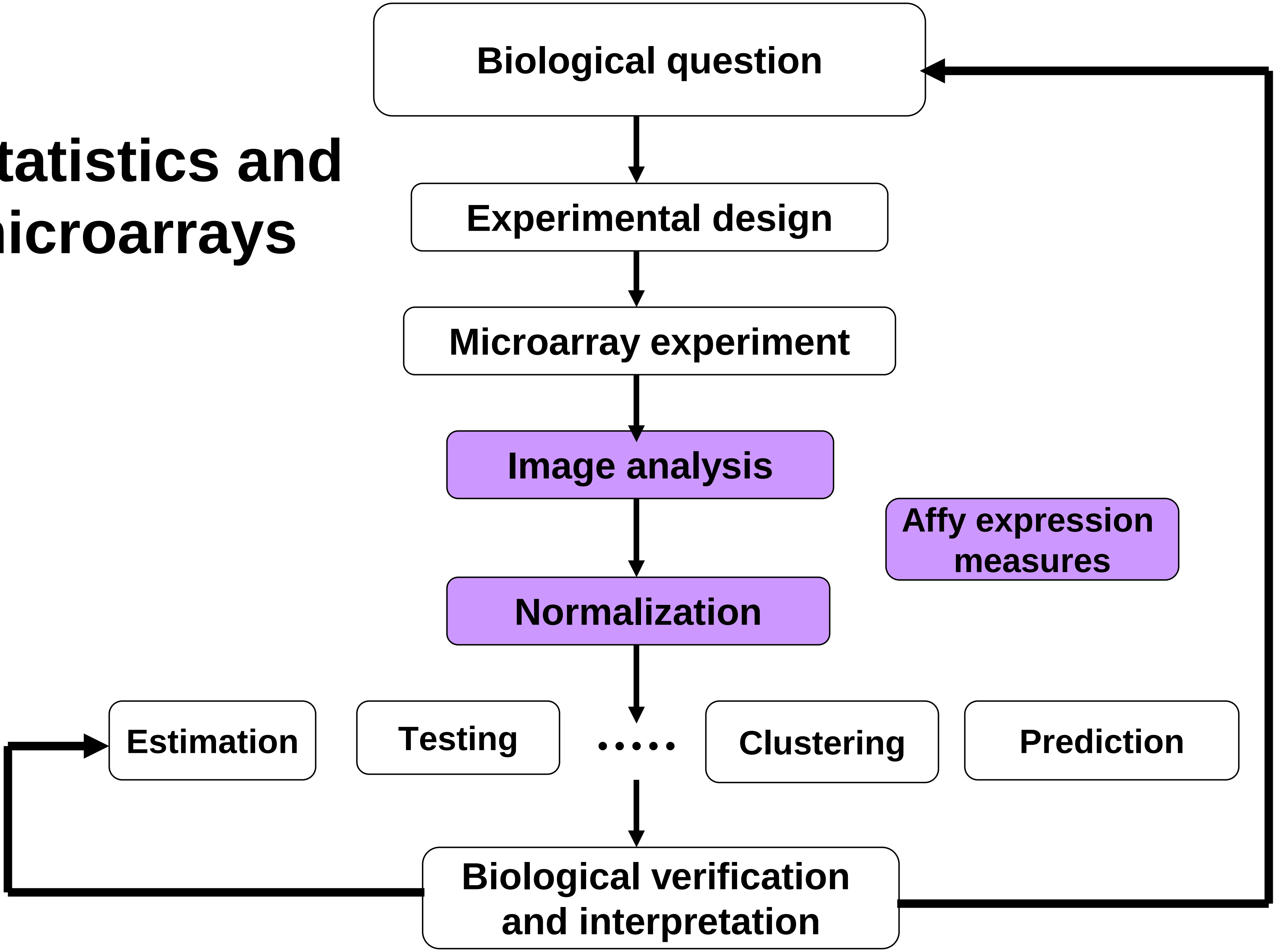
# Pre-processing in cDNA microarray experiments

Sandrine Dudoit, Robert Gentleman,
Rafael Irizarry, and Yee Hwa Yang

# Statistics and microarrays



Biological question

Experimental design

Microarray experiment

Image analysis

Affy expression measures

Normalization

Estimation

Testing

.....

Clustering

Prediction

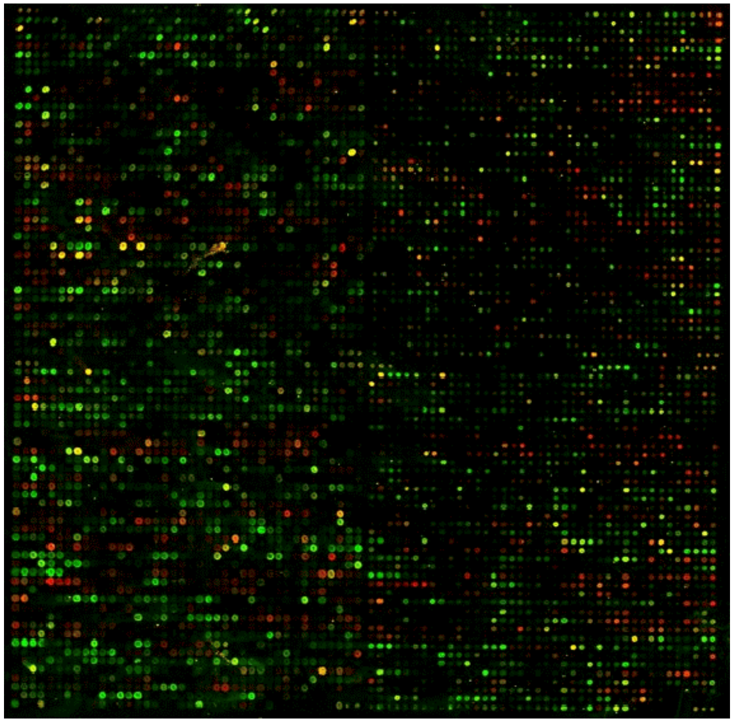
Biological verification and interpretation

# Outline

- cDNA microarrays
  - Image analysis;
  - Normalization.

- Affymetrix oligonucleotide chips
  - Image analysis;
  - Normalization;
  - Expression measures.

# cDNA microarrays



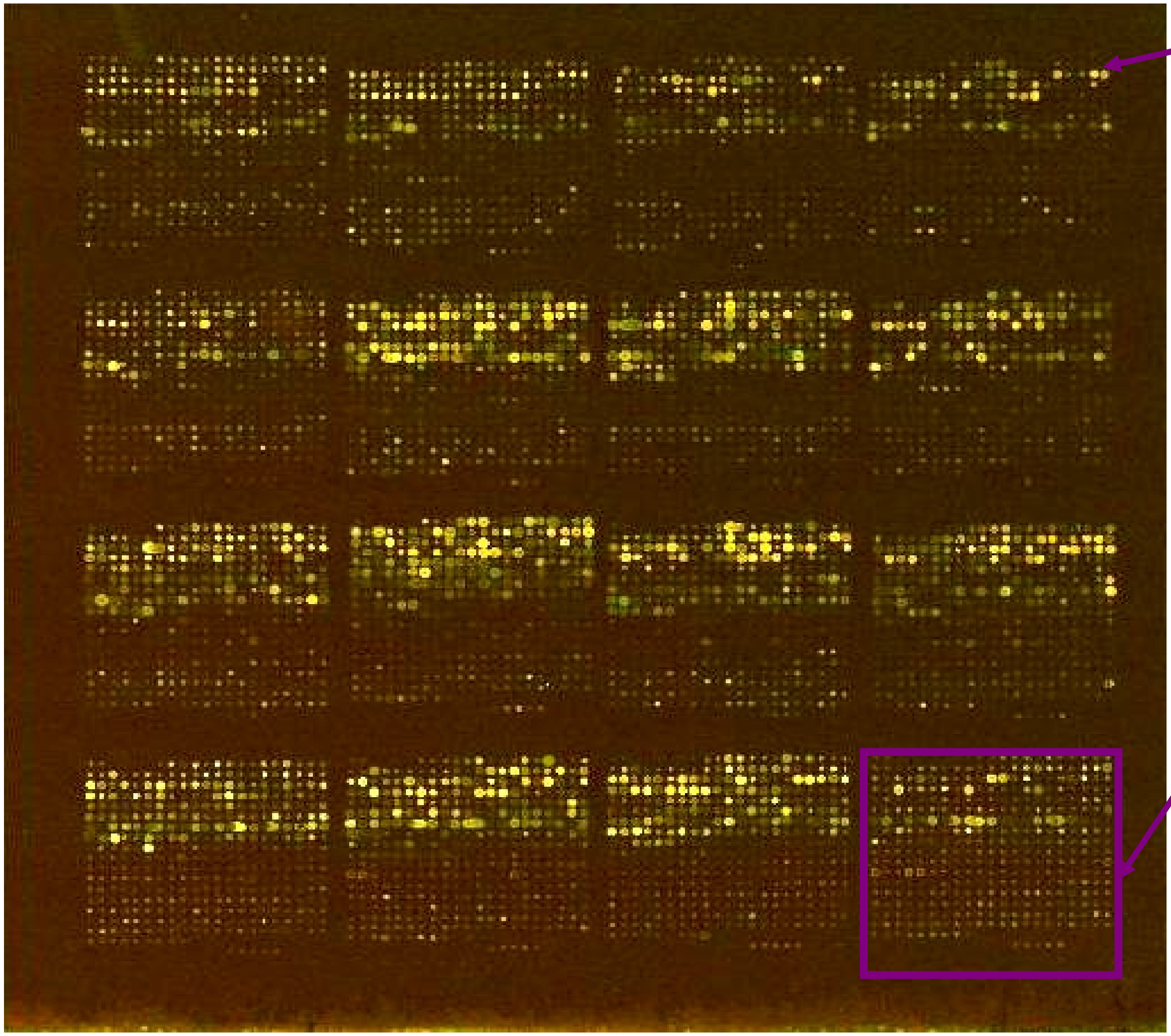
# Terminology

- Target**:** DNA hybridized to the array, mobile substrate.
- Probe**:** DNA spotted on the array,

  aka. spot, immobile substrate.
- Sector**:** collection of spots printed using the same print-tip (or pin),

  aka. print-tip-group, pin-group, spot matrix, grid.
- The terms slide and array are often used to refer to the printed microarray.
- Batch: collection of microarrays with the same probe layout.
- Cy3 = Cyanine 3 = green dye.
- Cy5 = Cyanine 5 = red dye.

# RGB overlay of Cy3 and Cy5 images



Probe

4 x 4 sectors
19 x 21 probes/sector
6,384 probes/array

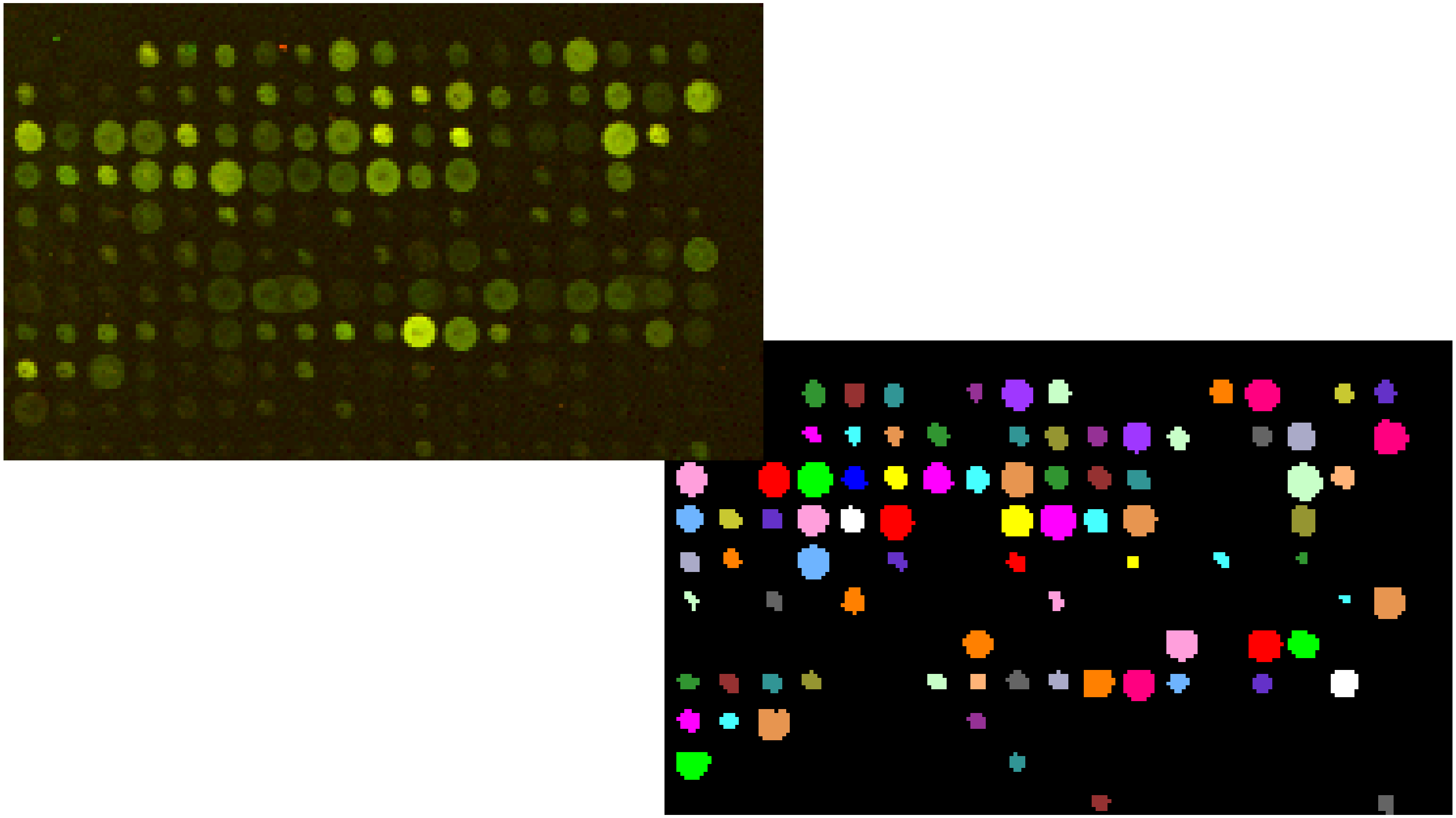
Sector

# Raw data

E.g. Human cDNA arrays

- ~43K spots;
- 16–bit TIFFs: ~ 20Mb per channel;
- ~ 2,000 x 5,500 pixels per image;
- Spot separation: ~ 136um;
- For a “typical” array, the spot area has
  - mean = 43 pixels,
  - med = 32 pixels,
  - SD = 26 pixels.

# Image analysis

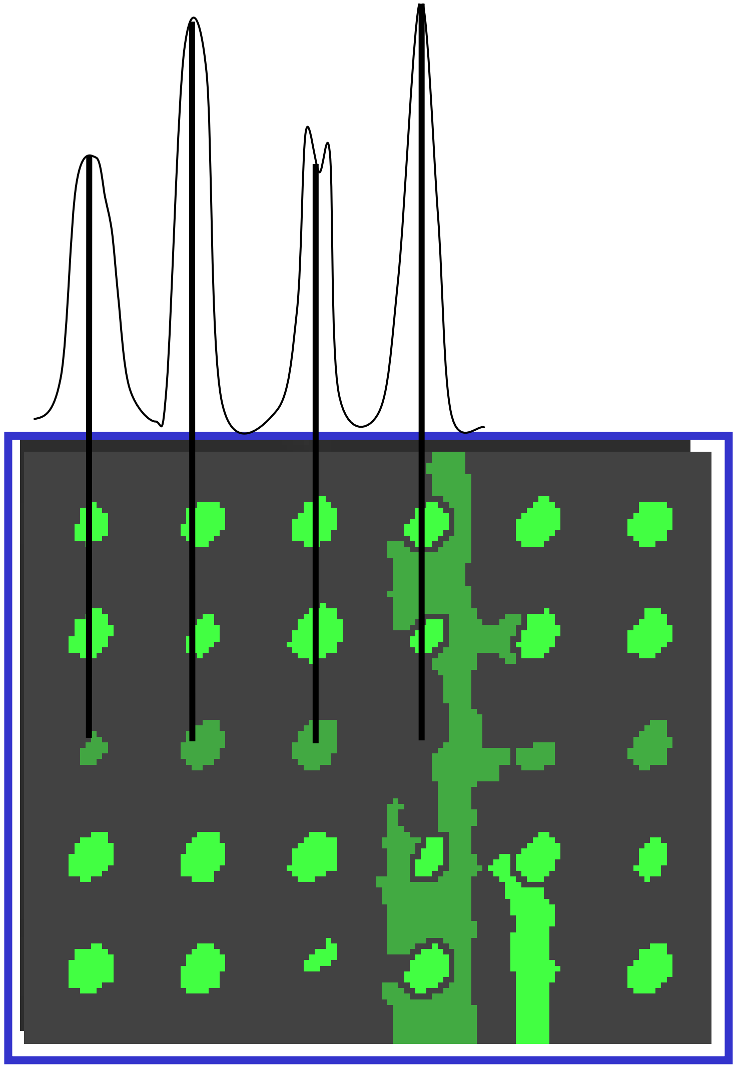


# Image analysis

- The **raw data** from a cDNA microarray experiment consist of pairs of **image files**, 16-bit TIFFs, one for each of the dyes.

- **Image analysis** is required to extract measures of the red and green fluorescence intensities, **R** and **G**, for each spot on the array.

# Image analysis



**1. Addressing.** Estimate location of spot centers.

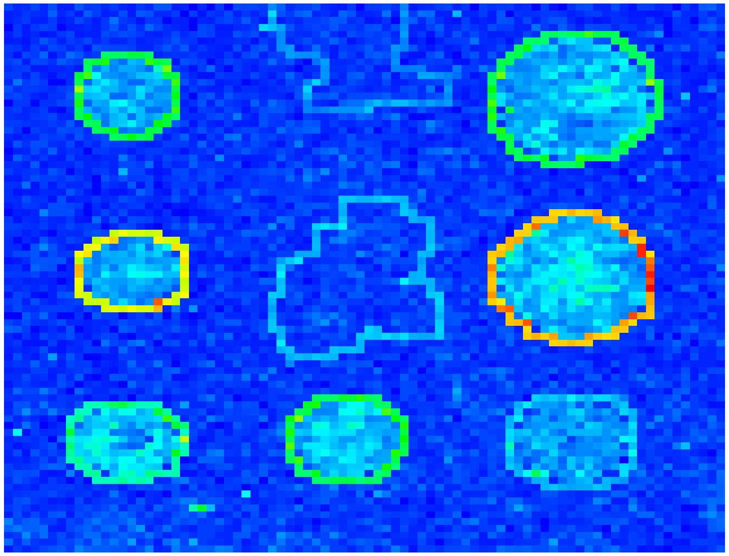
**2. Segmentation.** Classify pixels as foreground (signal) or background.

**3. Information extraction.** For each spot on the array and each dye

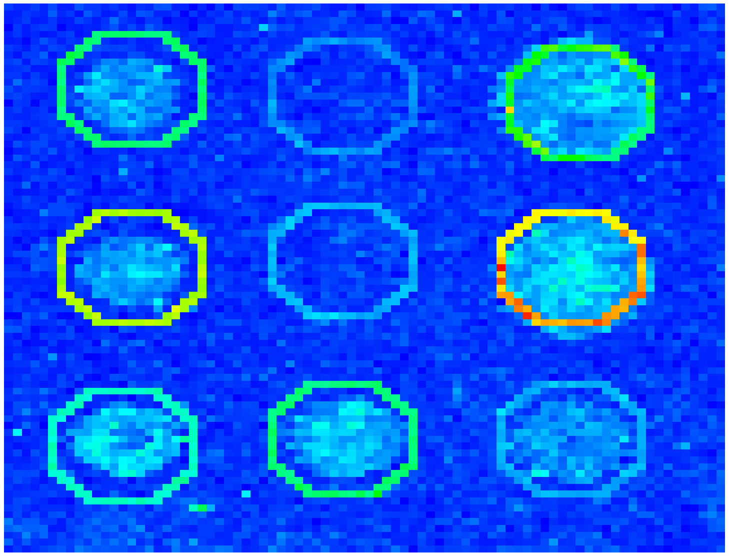
- foreground intensities;
- background intensities;
- quality measures.

→ **R and G for each spot on the array.**

# Segmentation



Adaptive segmentation, SRG



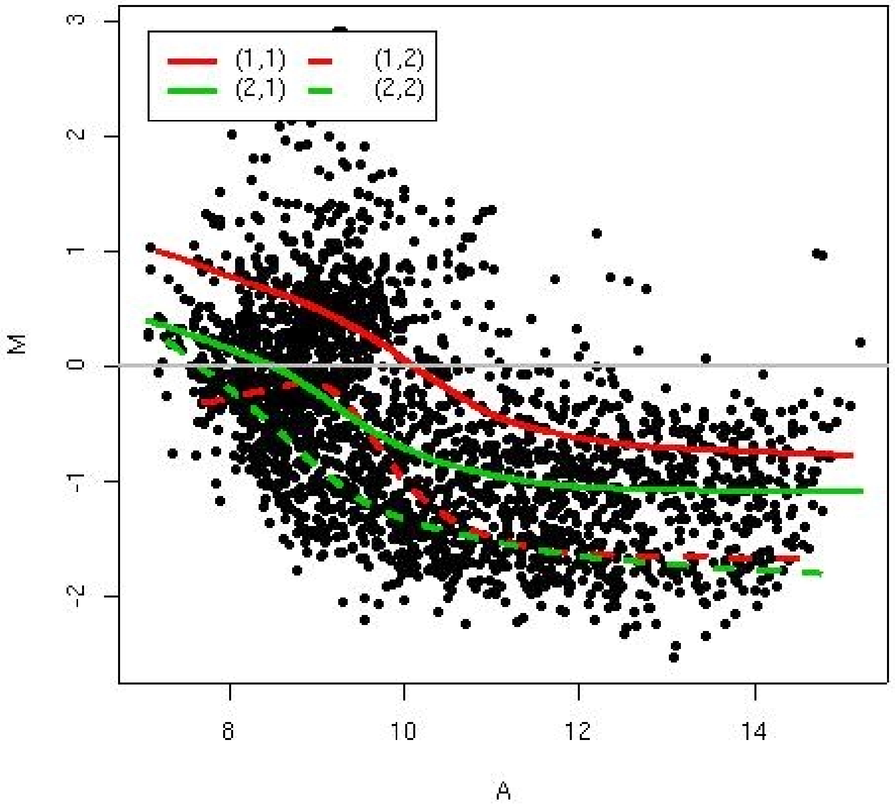
Fixed circle segmentation

**Spots usually vary in size and shape.**

# Quality measures

- **Spot quality**
  - **Brightness:** foreground/background ratio;
  - **Uniformity:** variation in pixel intensities and ratios of intensities within a spot;
  - **Morphology:** area, perimeter, circularity.
- **Slide quality**
  - Percentage of spots with no signal;
  - Range of intensities;
  - Distribution of spot signal area, etc.
- How to use quality measures in subsequent analyses?

# Normalization



# Normalization

- **Purpose.** Identify and remove the effects **of systematic variation** in the measured fluorescence intensities, other than differential expression, for example
  - different labeling efficiencies of the dyes;
  - different amounts of Cy3- and Cy5-labeled mRNA;
  - different scanning parameters;
  - print-tip, spatial, or plate effects, etc.

# Normalization

- Normalization is needed to ensure that differences in intensities are indeed due to differential expression, and not some printing, hybridization, or scanning artifact.

- Normalization is necessary before any analysis which involves within or between slides comparisons of intensities, e.g., clustering, testing.

# Normalization

- 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.

# Single-slide data display

- Usually: R vs. G

  $\log_2 R$ vs. $\log_2 G$.

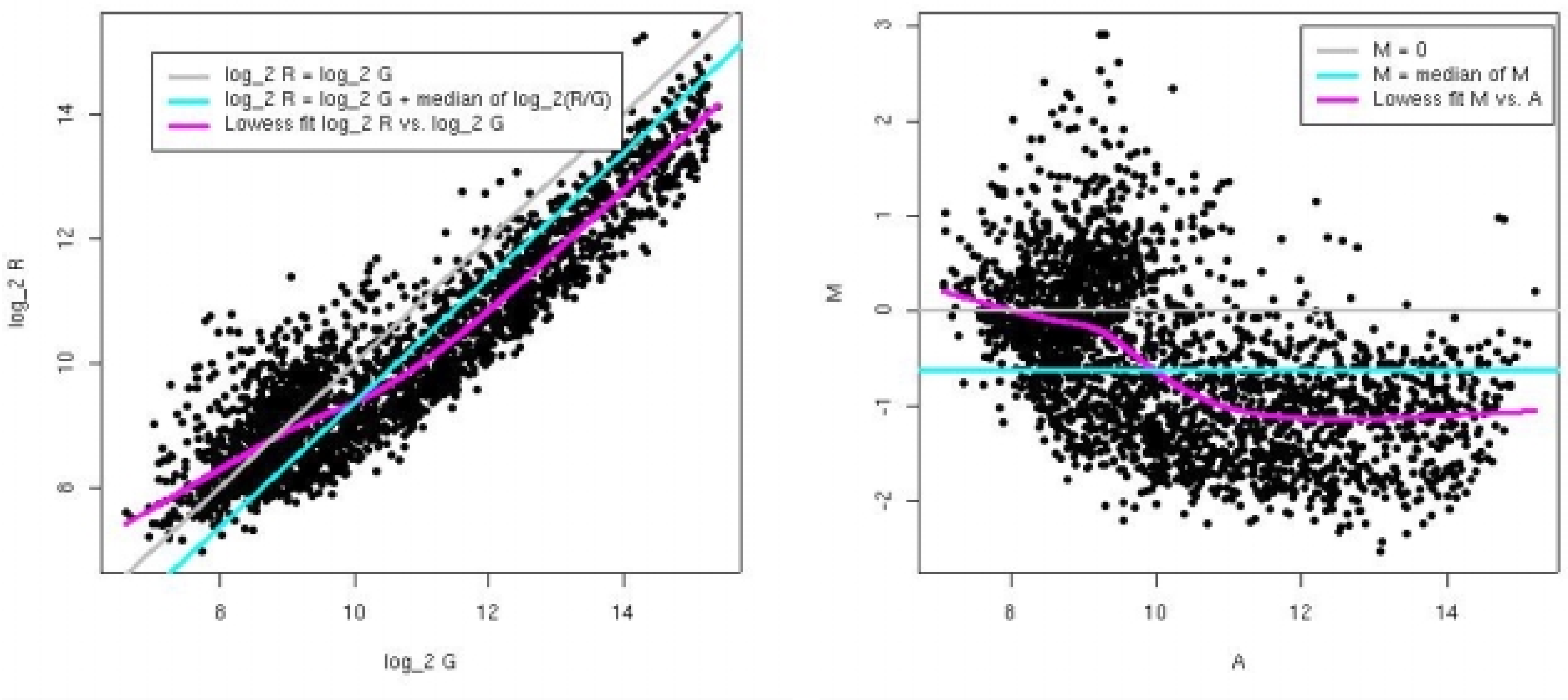
- Preferred

  **$M = \log_2 R - \log_2 G$**

  vs. **$A = (\log_2 R + \log_2 G)/2$**.

- An MA-plot amounts to a 45° counterclockwise rotation of a $\log_2 R$ vs. $\log_2 G$ plot followed by scaling.

# Self-self hybridization



$\log_2$ R vs. $\log_2$ G

M vs. A

$M = \log_2 R - \log_2 G, \quad A = (\log_2 R + \log_2 G)/2$

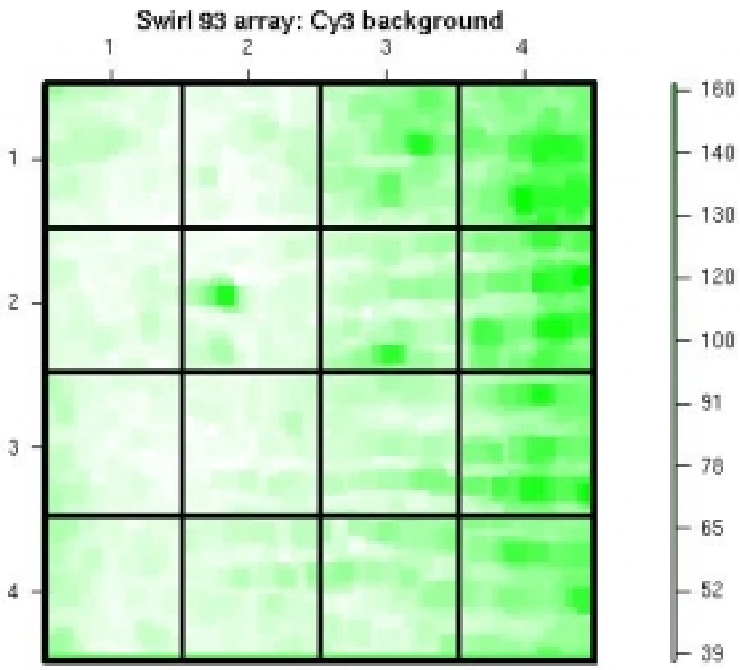
# Diagnostic plots

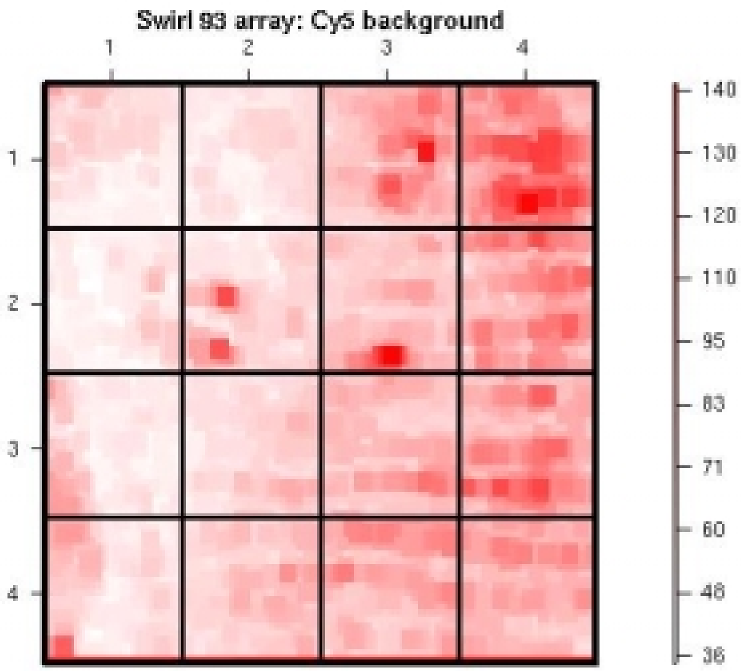
- **Diagnostics plots** of spot statistics

  E.g. red and green log-intensities, intensity log-ratios M, average log-intensities A, spot area.

  - Boxplots;
  - 2D spatial images;
  - Scatter-plots, e.g. MA-plots;
  - Density plots.
- **Stratify** plots according to layout parameters, e.g. print-tip-group, plate.

# 2D spatial images



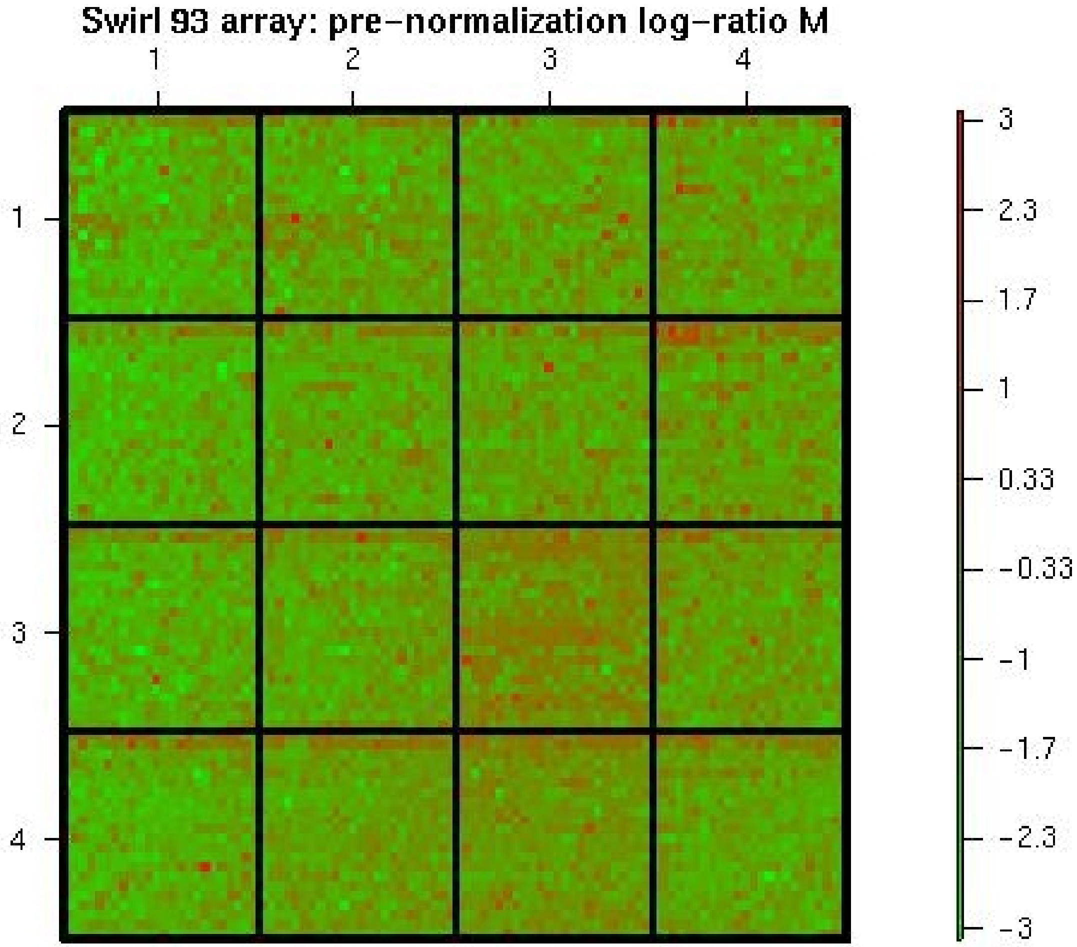


Cy3 background intensity

Cy5 background intensity

# 2D spatial images

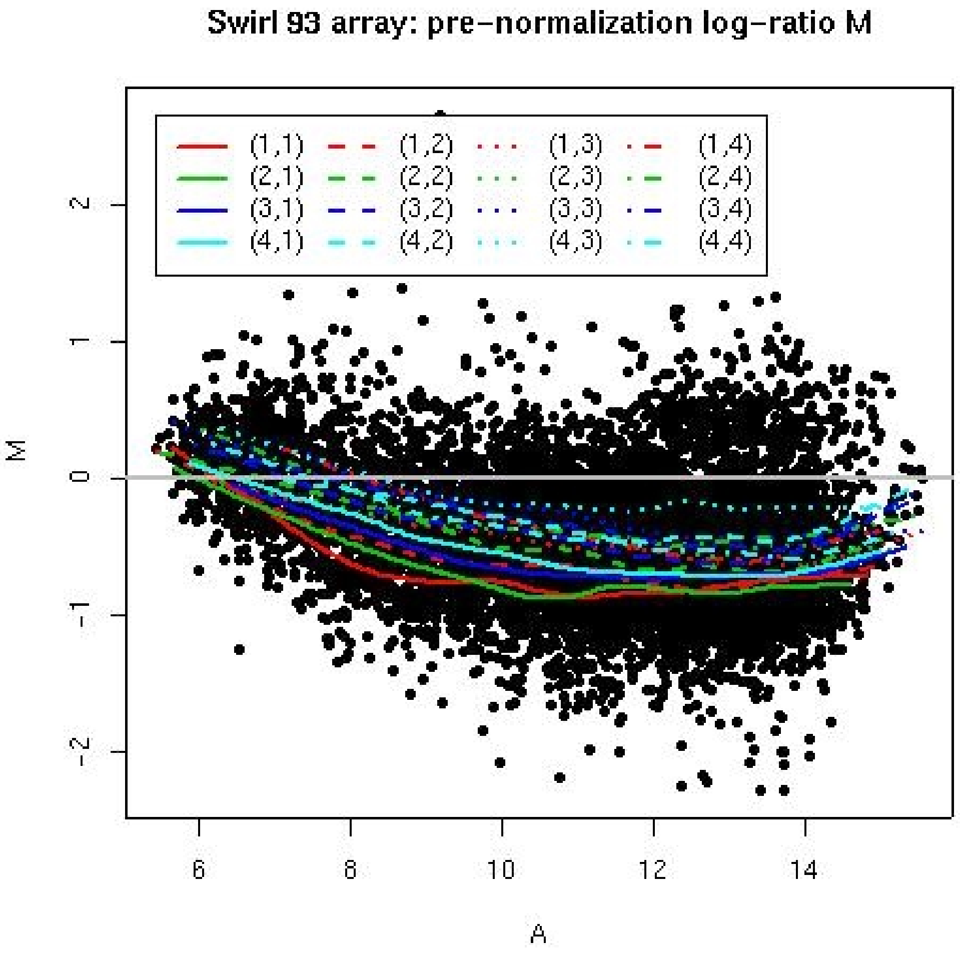
Intensity log-ratio, M



# MA-plot by print-tip-group

$$M = \log_2 R - \log_2 G, \quad A = (\log_2 R + \log_2 G)/2$$

Intensity
log-ratio, M



Average
log-intensity, A

# Location normalization

$$\log_2 R/G \leftarrow \log_2 R/G - L(\text{intensity}, \text{sector}, \ldots)$$

- **Constant normalization.** Normalization function L is **constant** across the spots, e.g. mean or median of the log-ratios M.

- **Adaptive normalization.** Normalization function L depends on a  number of **predictor variables**, such as spot intensity A, sector, plate origin.

# Location normalization

- The normalization function can be obtained by **robust locally weighted regression** of the log-ratios M on predictor variables.

  E.g. regression of M on A within sector.

- Regression method: e.g. lowess or loess (Cleveland, 1979; Cleveland & Devlin, 1988).

# Location normalization

- **Intensity-dependent normalization.**

  Regression of M on A (*global loess*).

- **Intensity and sector-dependent normalization***.*

  Same as above, for each sector separately

  (*within-print-tip-group loess*).

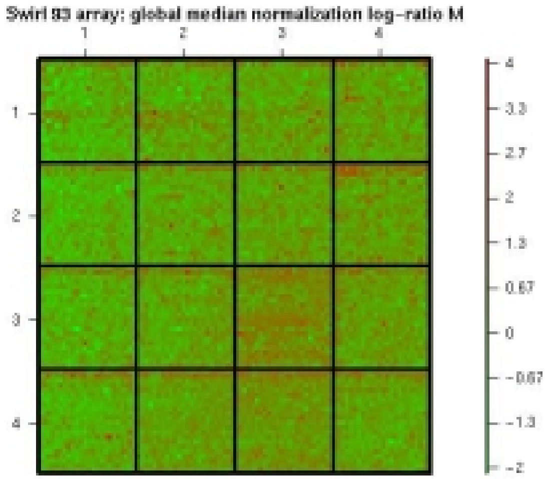
- **2D spatial normalization**.

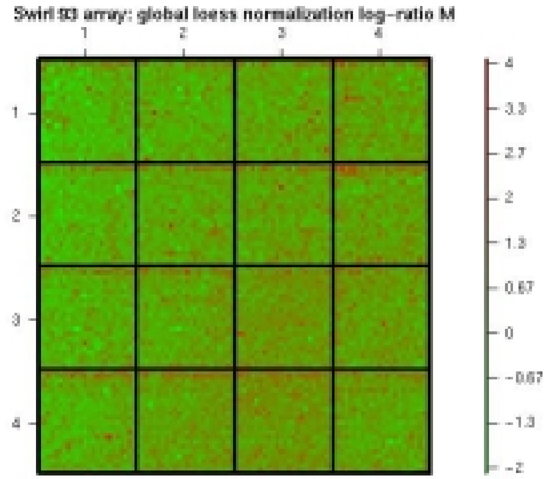
  Regression of M on 2D-coordinates.

- Other variables: time of printing, plate, etc.
- **Composite normalization**. Weighted average of several normalization functions.

# 2D images of normalized M-L

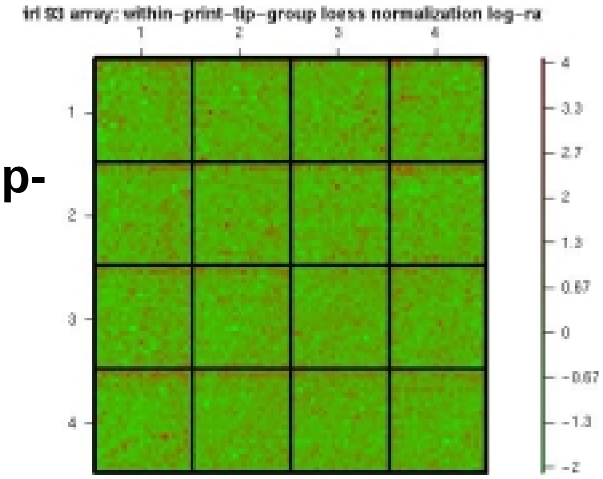
**Global median normalization**

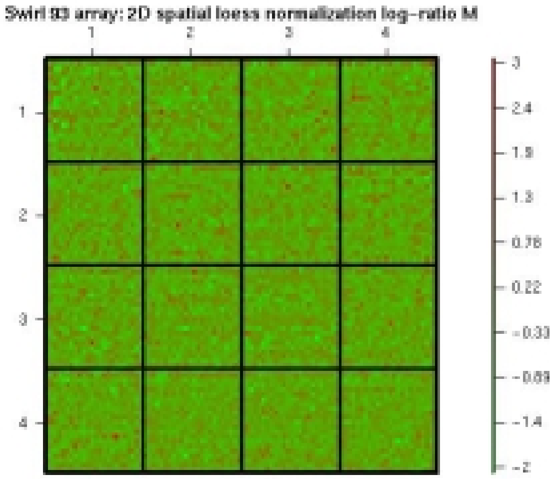




**Global loess normalization**

**Within-print-tip-group loess normalization**



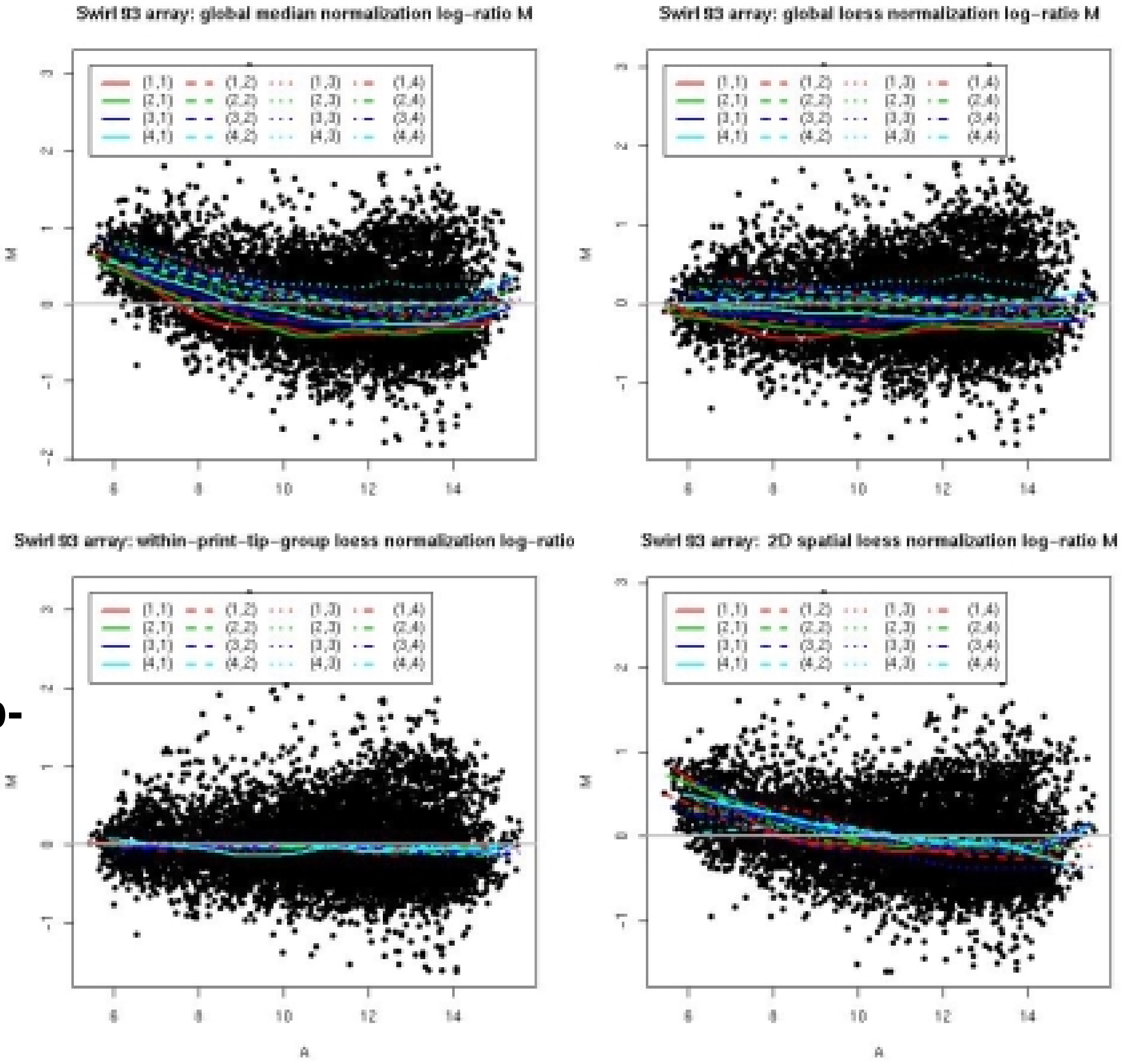


**2D spatial normalization**

# MA-plots of normalized M-L

Global median normalization

Global loess normalization



Within-print-tip-group loess normalization

2D spatial normalization

# Normalization

- Within-slide
  - **Location** normalization - additive on log-scale.
  - **Scale** normalization - multiplicative on log-scale.
  - **Which spots** to use?
- Paired-slides (dye-swap experiments)
  - Self-normalization.
- Between-slides.

# Scale normalization

- The log-ratios M from different sectors, plates, or arrays may exhibit different spreads and some **scale** adjustment may be necessary.

$$\log_2 R/G \leftarrow (\log_2 R/G - L)/S$$

- Can use a robust estimate of scale such as the **median absolute deviation (MAD)**

  MAD = median | M – median(M) |.

# Scale normalization

- For print-tip-group scale normalization, assume all print-tip-groups have the same spread in M.
- Denote ***true*** and ***observed*** log-ratio by $\mu_{ij}$ and $M_{ij}$, resp., i $\mu_{ij}$, and i indexes print- d j spots. Robust estimate of $a_i$ is

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

where *MAD<sub>i</sub>* is MAD of $M_{ij}$ in print-tip-group

# Algorithm Median Absolute Deviation (MAD) scale normalization

**Input**: log intentisity ratios $M_j = \log_2 R_j/G_j$ for the overall genes in a given slide or within a given print-tip-group, $1 \le j \le n$

**Output**: scale normalization factor for a given slide S or a print-tip-group $S_i$

1. m = $\text{median}_j (M_j)$
2. AD = { $ad_j = |m_j - m|$, $1 \le j \le n$ }
3. MAD = $\text{median}_j (ad_j)$
4. Output:

A. Within slide:

$$S = MAD$$

B. Within print-tip-group:

$$S_i = \frac{MAD_i}{\sqrt[I]{\prod_{i=1}^{I} MAD_i}}$$

# Which genes to use?

- **All spots on the array**:
  - Problem when many genes are differentially expressed.

- **Housekeeping genes**: Genes that are thought to be constantly expressed across a wide range of biological samples (e.g. tubulin, GAPDH). Problems:
  - sample specific biases (genes are actually regulated),
  - do not cover intensity range.

# Which genes to use?

- **Genomic DNA titration series**:
  - fine in yeast,
  - but weak signal for higher organisms with high intron/exon ratio (e.g. mouse, human).

- **Rank invariant set** (Schadt et al., 1999; Tseng et al., 2001): genes with same rank in both channels. Problems: set can be small.

# Microarray sample pool

- **Microarray Sample Pool**, **MSP**: Control sample for normalization, in particular, when it is not safe to assume most genes are equally expressed in both channels.
- MSP: **pooled** all 18,816 ESTs from RIKEN release 1 cDNA mouse library.
- Six-step **dilution series** of the MSP.
- MSP samples were spotted in middle of first and last row of each sector.
- Ref. Yang et al. (2002).

# Microarray sample pool

MSP control spots

- provide potential probes for every target sequence;
- are constantly expressed across a wide range of biological samples;
- cover the intensity range;
- are similar to genomic DNA, but without intron sequences → better signal than genomic DNA in organisms with high intron/exon ratio;
- can be used in composite normalization.

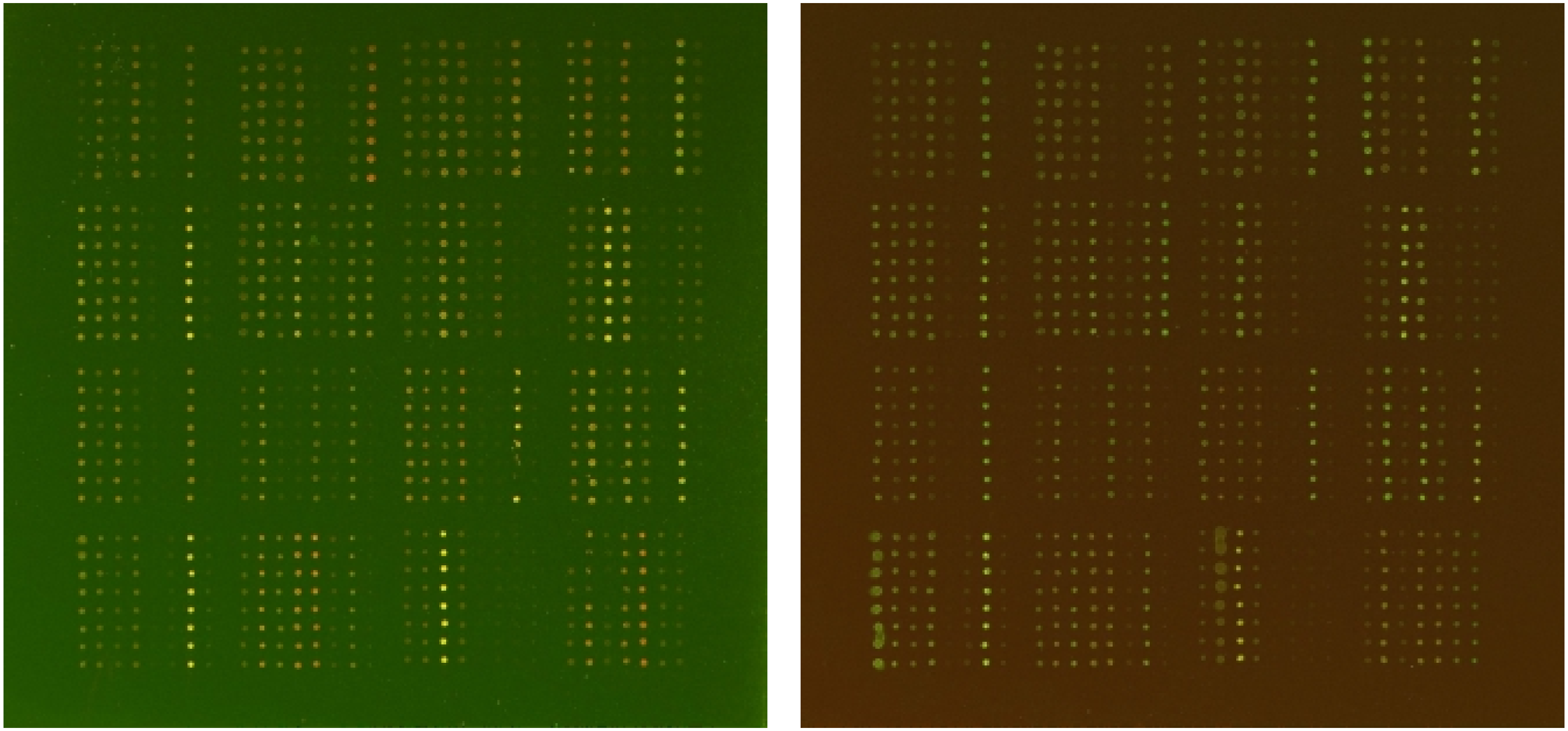
# Dye-swap experiment

- Probes
  - 50 distinct clones thought to be differentially expressed in apo AI knock-out mice compared to inbred C57Bl/6 control mice (largest absolute t-statistics in a previous experiment).
  - 72 other clones.
- Spot each clone 8 times .
- Two hybridizations with dye-swap:

  Slide 1: trt → red, ctl → green.

  Slide 2: trt → green, ctl → red.

# Dye-swap experiment



# Self-normalization

- Slide 1, $M = \log_2 (R/G) - L$
- Slide 2, $M' = \log_2 (R'/G') - L'$

Combine by **subtracting** the normalized log-ratios:

$$M – M'$$

$$= [ (\log_2 (R/G) - L) - (\log_2 (R'/G') - L') ] / 2$$

$$\approx [ \log_2 (R/G) + \log_2 (G'/R') ] / 2$$

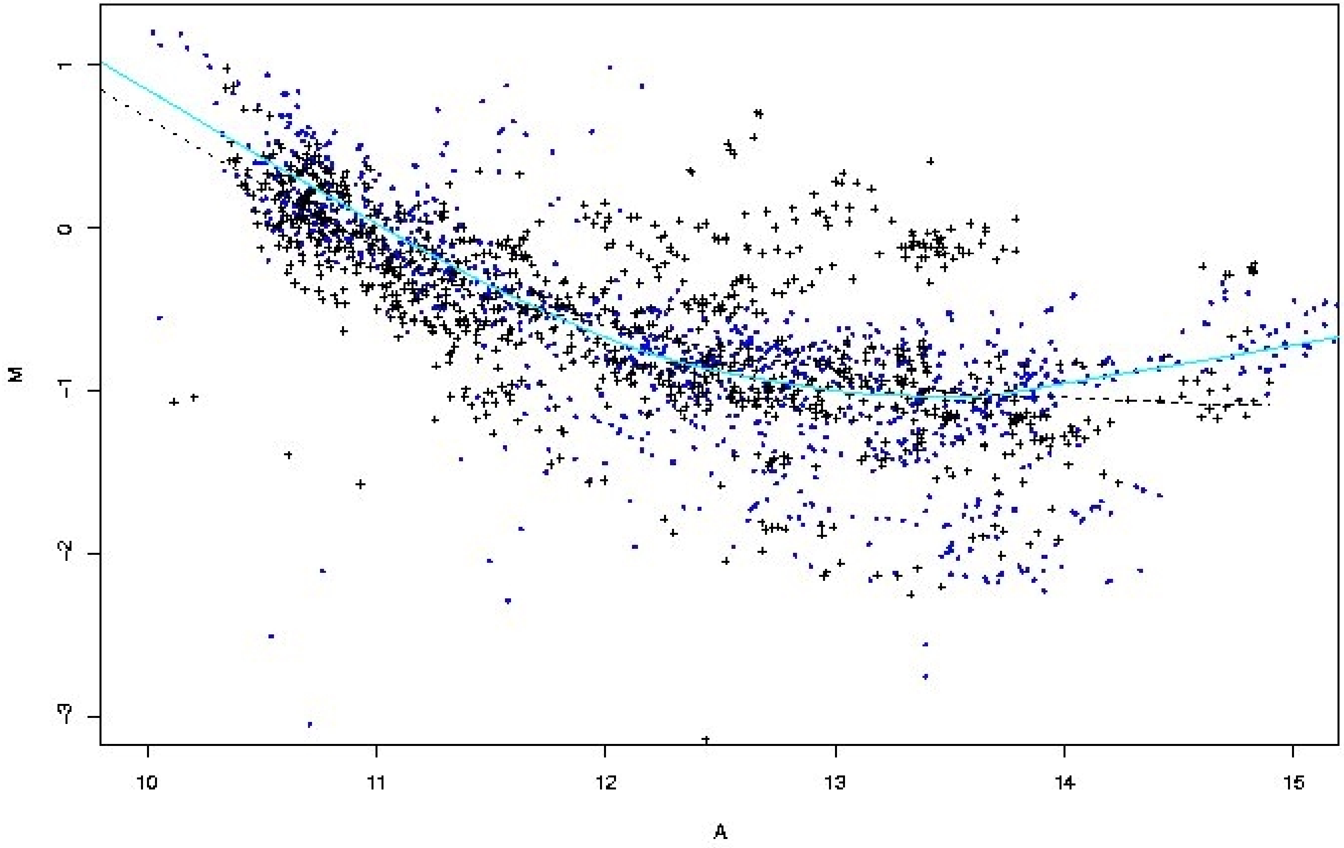
$$\approx [ \log_2 (RG'/GR') ] / 2$$

provided $L= L'$.

***Assumption: the normalization functions are the same for the two slides.***

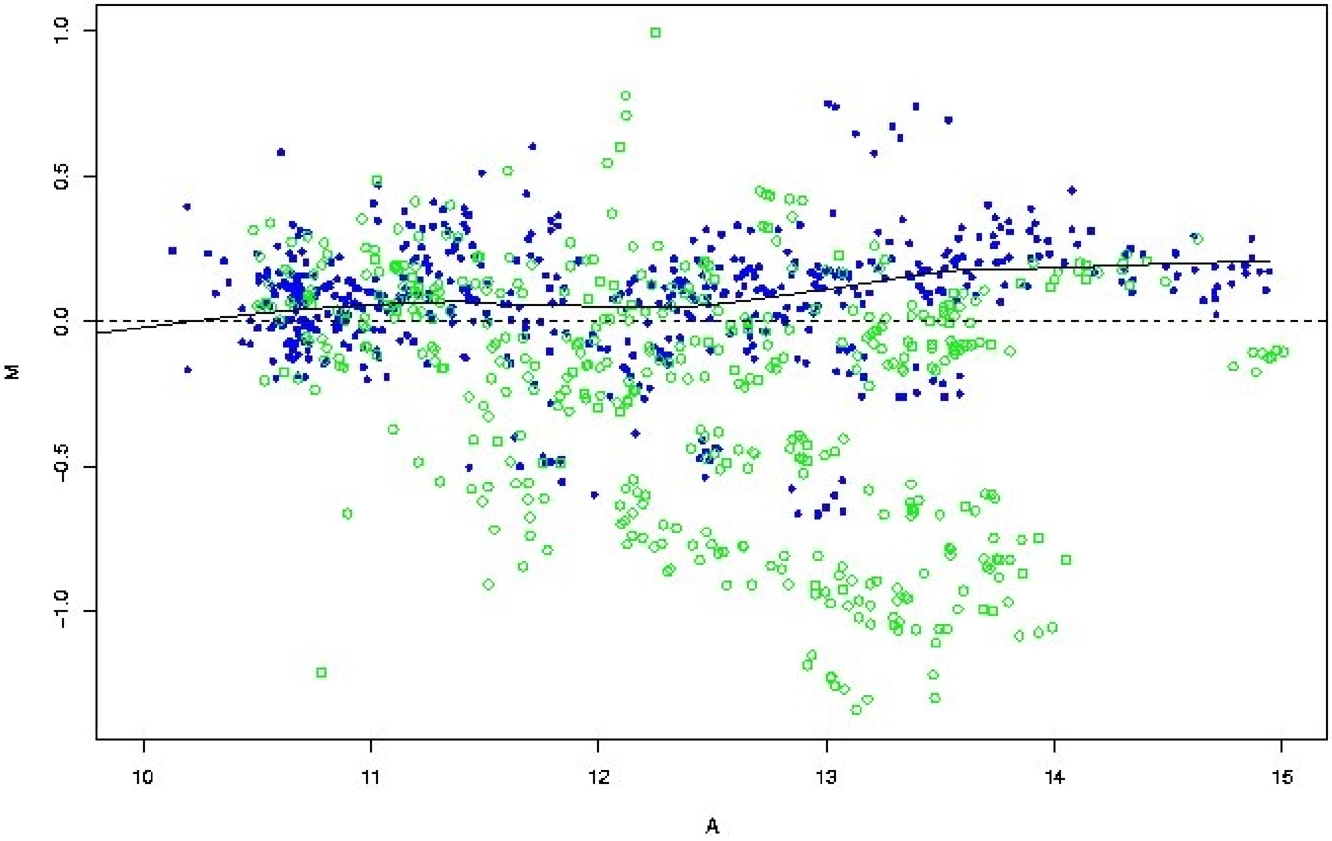
# Checking the assumption

## MA-plot for slides 1 and 2



# Result of self-normalization

## (M - M’)/2 vs. (A + A’)/2



# Summary

Case 1. Only a few genes are expected to change.

Within-slide

- Location: intensity + sector-dependent normalization.
- Scale: for each sector, scale by MAD.

Between-slides

- An extension of within-slide scale normalization.

Case 2. Many genes are expected to change.

- Paired-slides: Self-normalization.
- Use of controls or known information, e.g. MSP.
- Composite normalization.