

Pre-processing in DNA microarray experiments

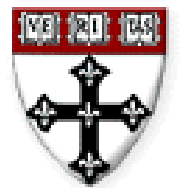
Sandrine Dudoit, Robert Gentleman,
Rafael Irizarry, and Yee Hwa Yang

Bioconductor short course

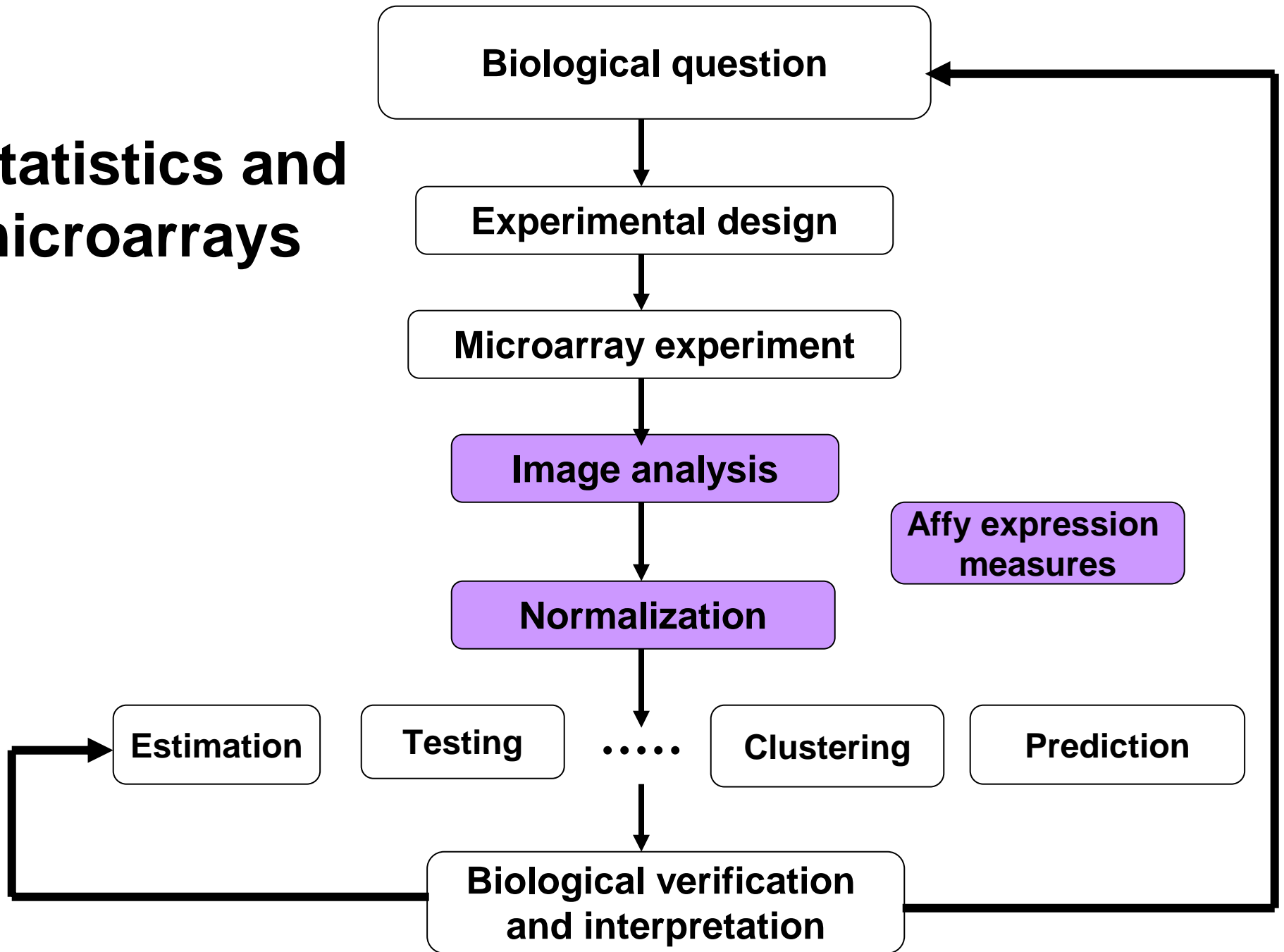
Summer 2002



© Copyright 2002, all rights reserved



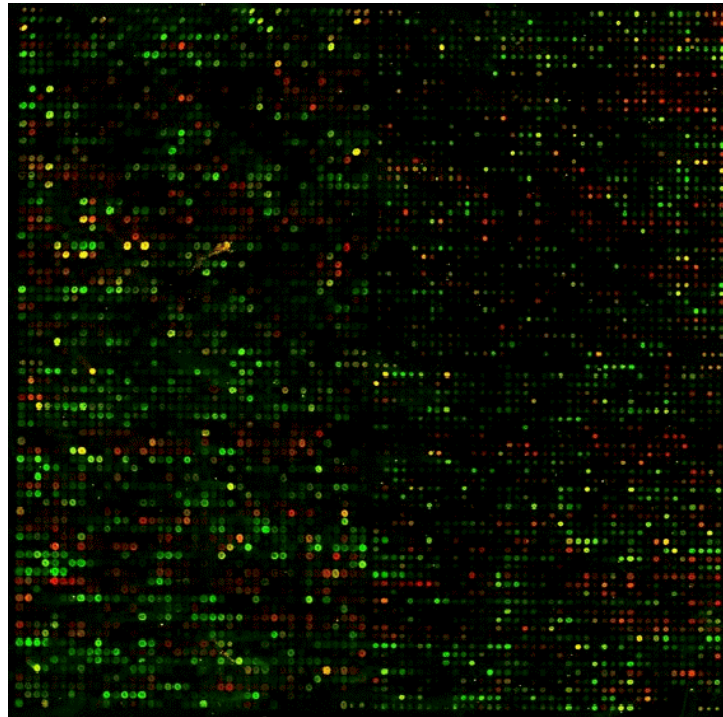
Statistics and microarrays



Outline

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

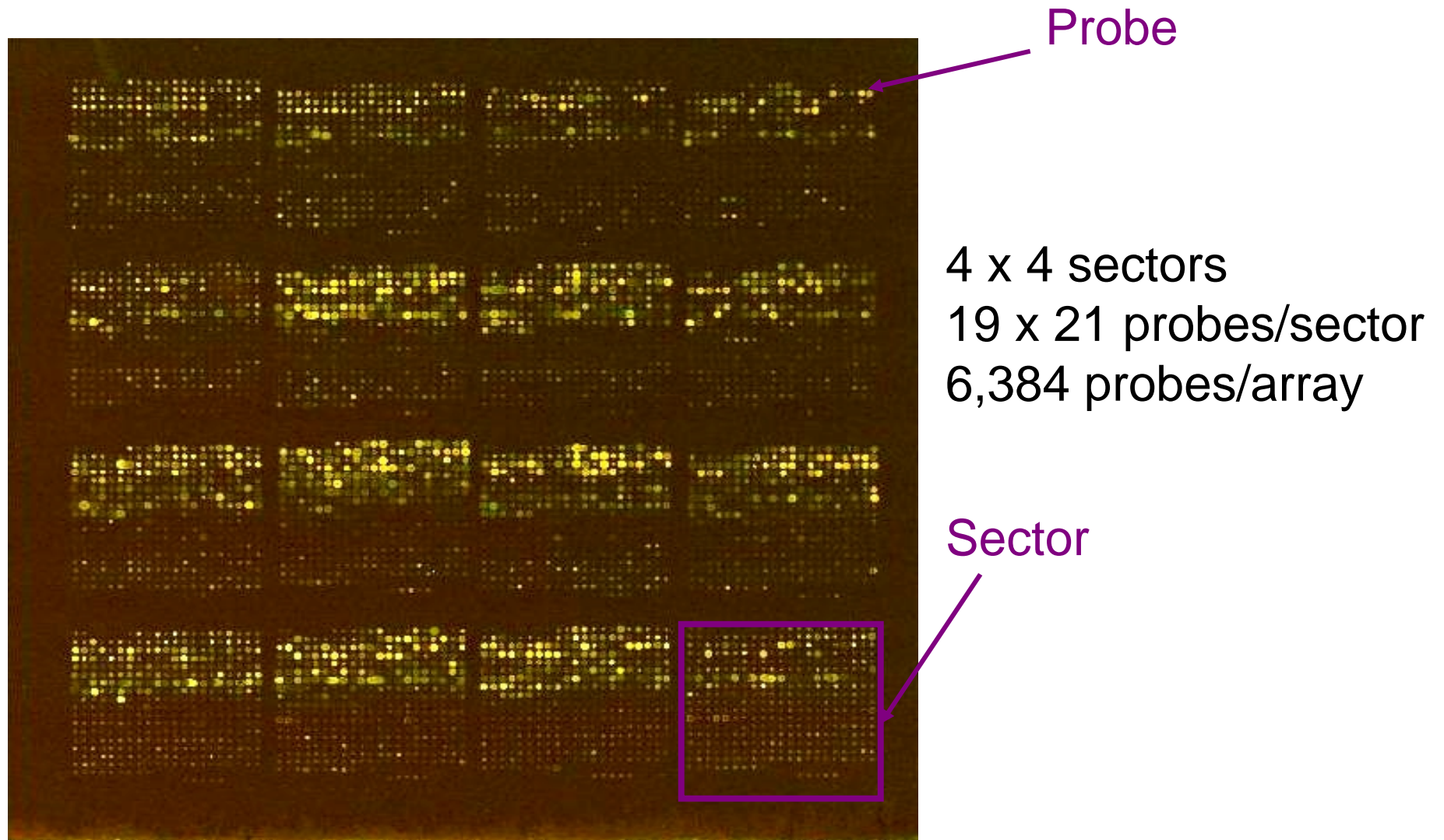
cDNA microarrays



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



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

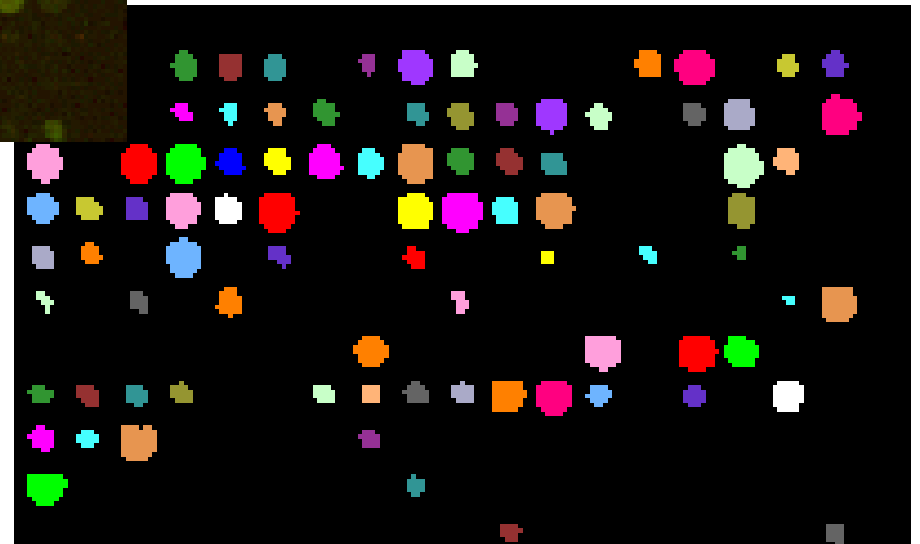
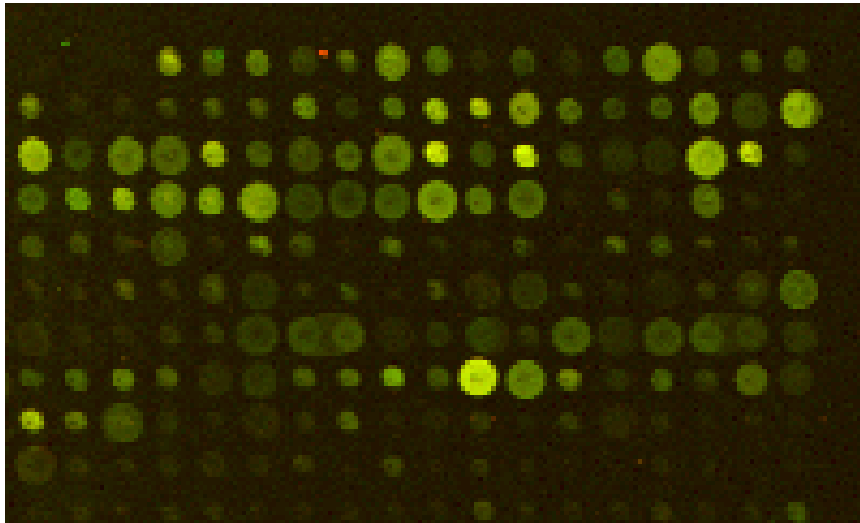
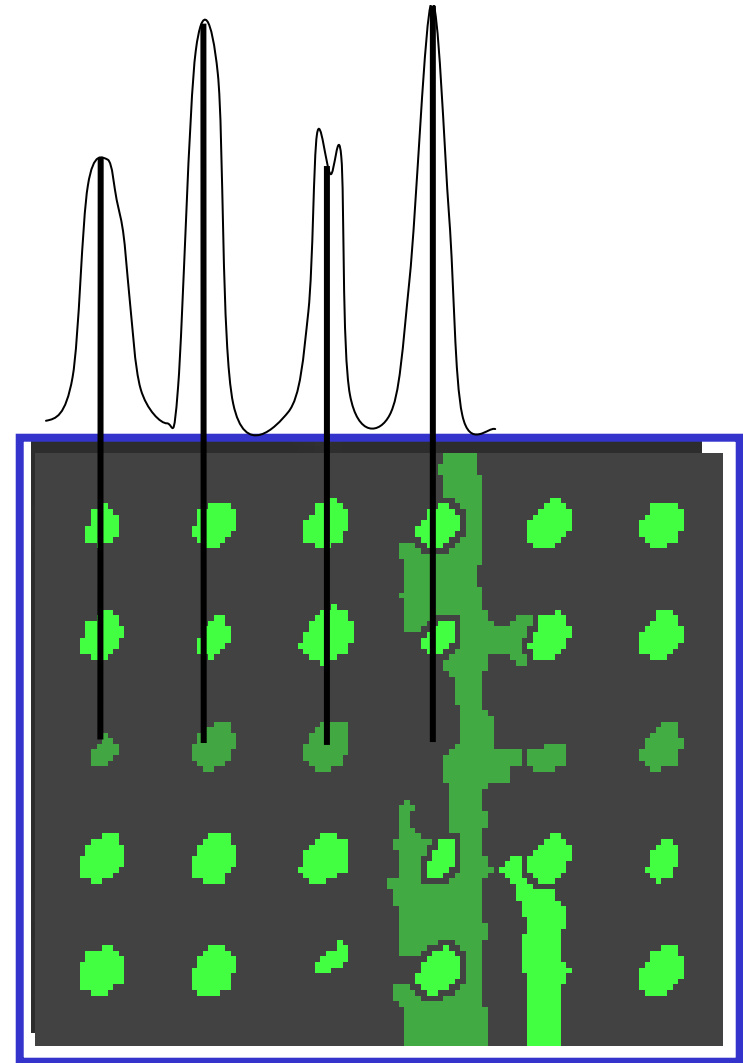


Image analysis

- 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.

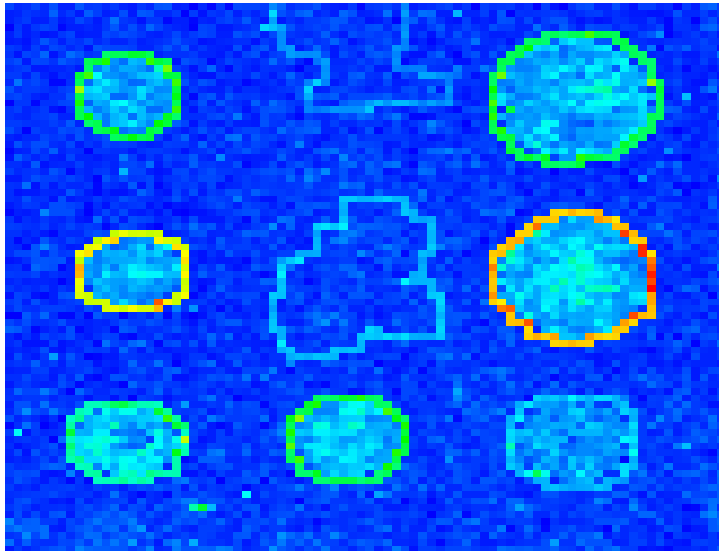
Image analysis

- 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.

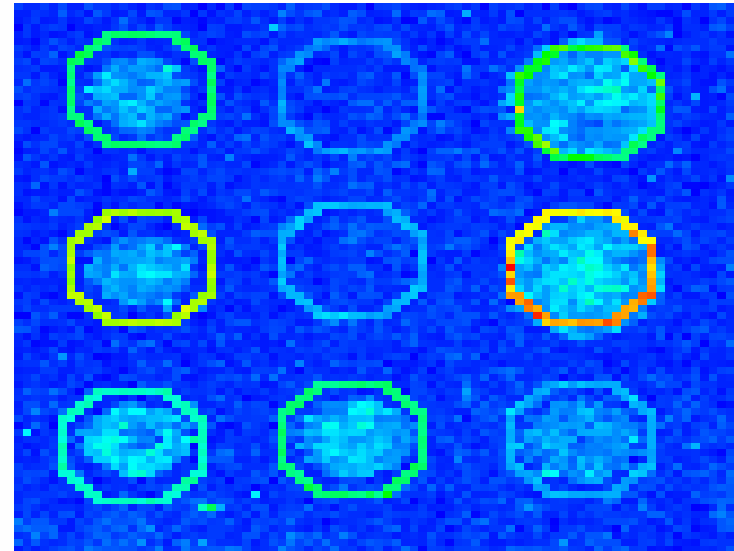


→ **R** and **G** for each spot on the array.

Segmentation



Adaptive segmentation, SRG



Fixed circle segmentation

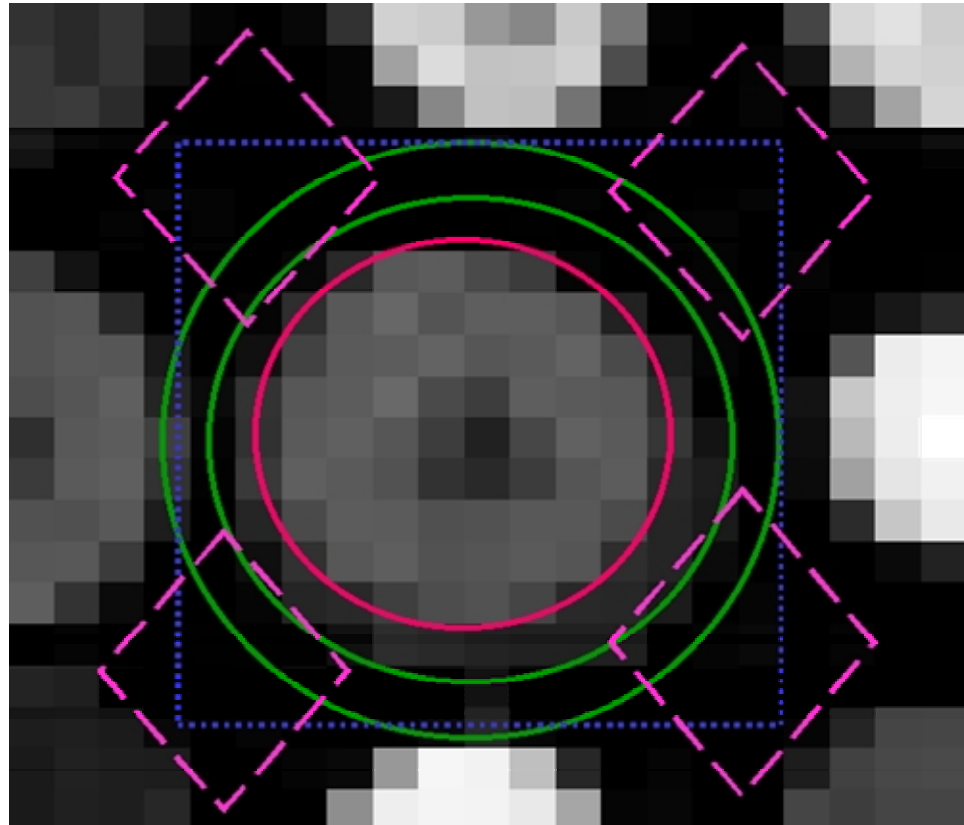
Spots usually vary in size and shape.

Seeded region growing

- **Adaptive** segmentation method.
- Requires the input of **seeds**, either individual pixels or groups of pixels, which control the formation of the regions into which the image will be segmented.
Here, based on fitted foreground and background **grids** from the addressing step.
- The decision to add a pixel to a region is based on the absolute gray-level difference of that pixel's intensity and the average of the pixel values in the neighboring region.
- Done on combined red and green images.
- Ref. Adams & Bischof (1994)

Local background

- GenePix
- QuantArray
- ScanAnalyze

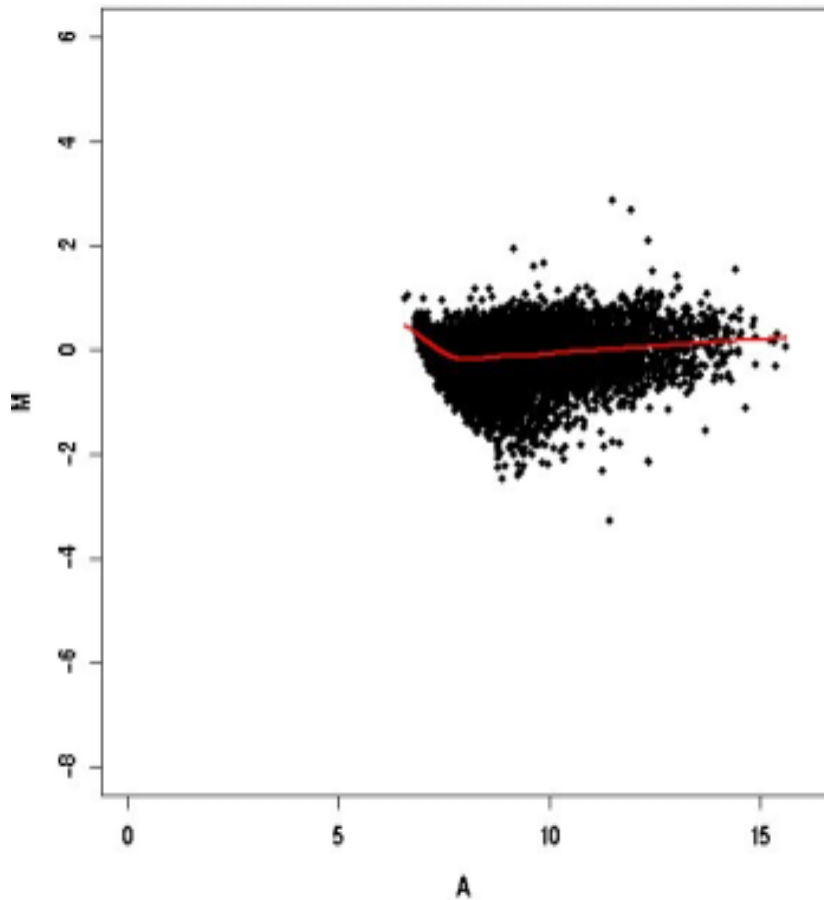


Morphological opening

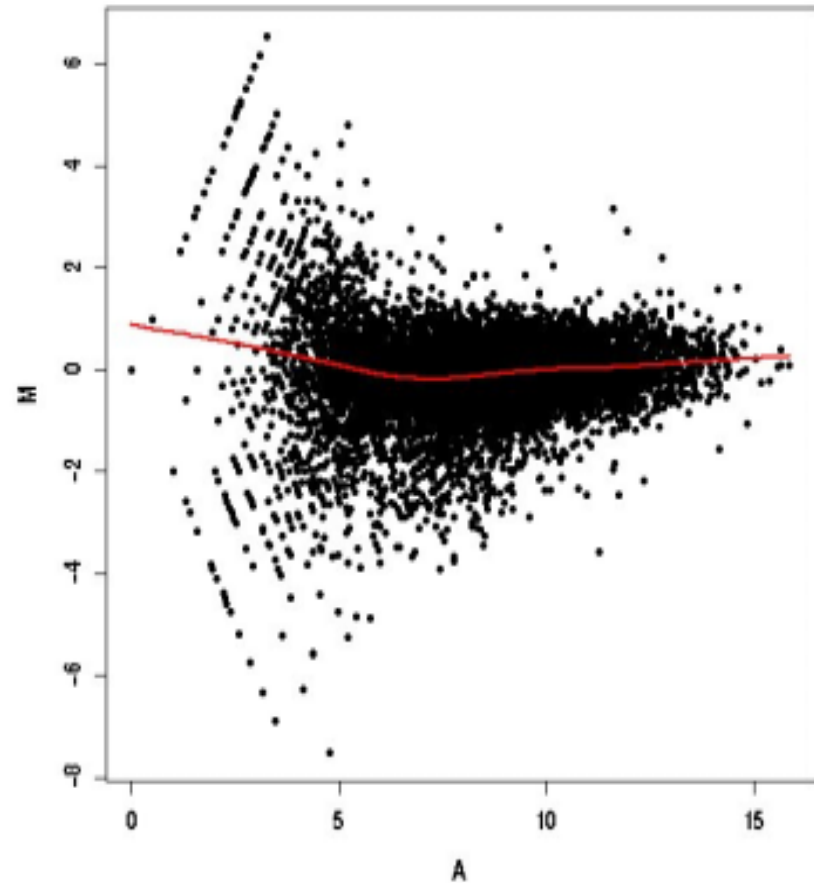
- The image is probed with a **structuring element**, here, a square with side length about twice the spot-to-spot distance.
- **Erosion** (**Dilation**): the eroded (dilated) value at a pixel x is the **minimum** (**maximum**) value of the image in the window defined by the structuring element when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

Background matters

Morphological opening



Local background



$$M = \log_2 R - \log_2 G \quad \text{vs.} \quad A = (\log_2 R + \log_2 G)/2$$

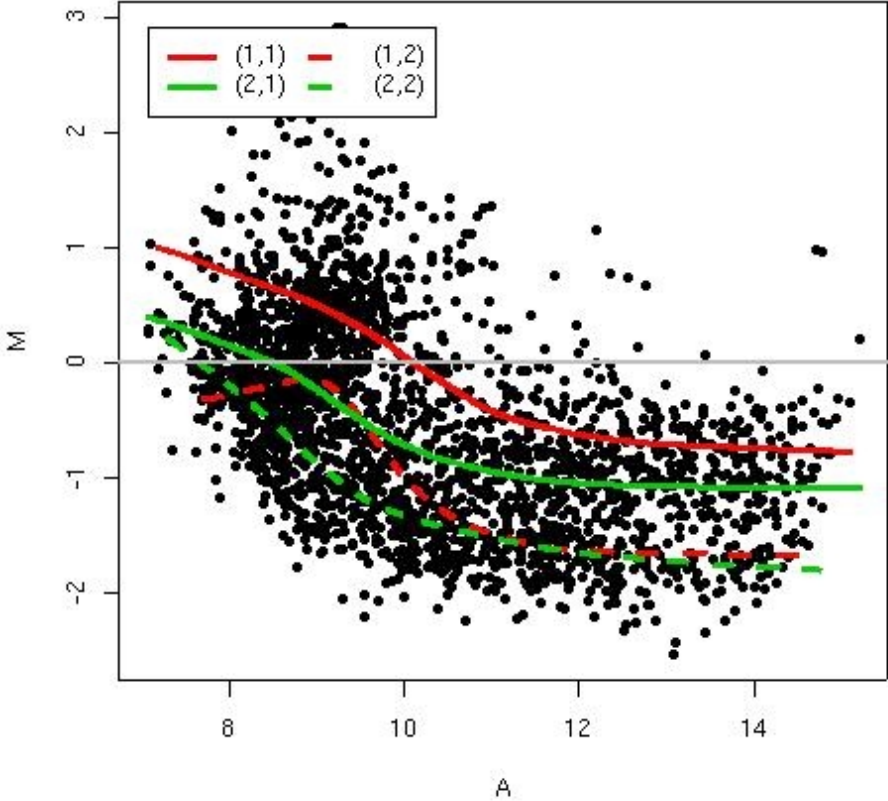
Quality measures

- **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?

Spot image analysis software

- Software package **Spot**, built on the **R** language and environment for statistical computing and graphics.
- 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.

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

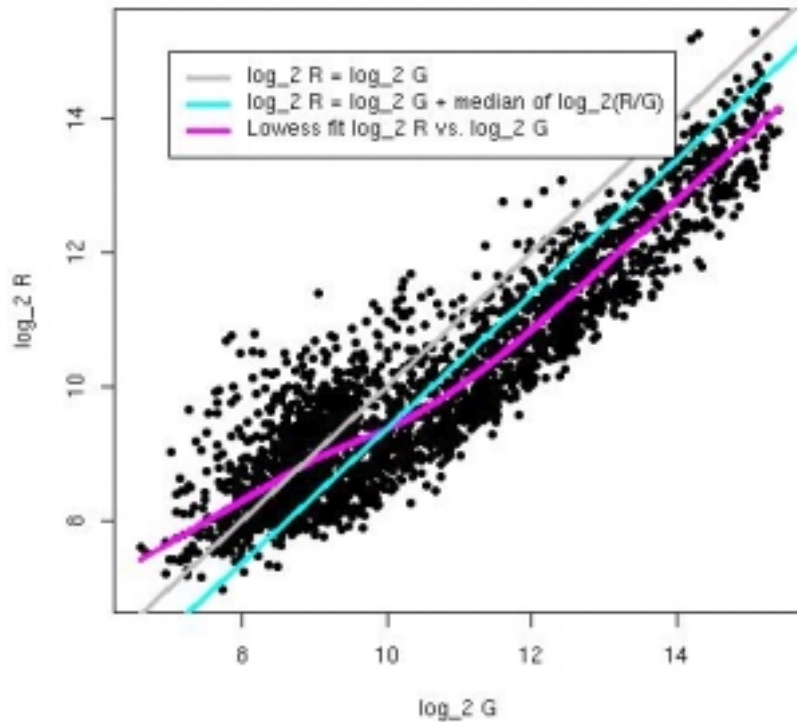
- The need for normalization can be seen most clearly in **self-self hybridizations**, where the same mRNA sample is labeled with the Cy3 and Cy5 dyes.
- The imbalance in the red and green intensities is usually **not constant** across the spots within and between arrays, and can vary according to overall spot intensity, location, plate origin, etc.
- These factors should be considered in the normalization.

Single-slide data display

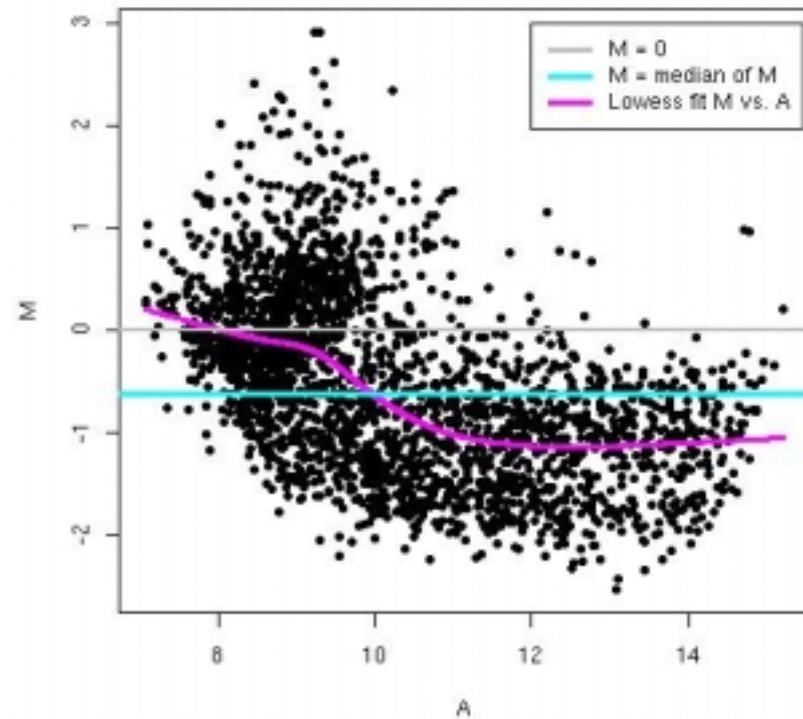
- Usually: R vs. G
 $\log_2 R$ vs. $\log_2 G$.
- Preferred
 $M = \log_2 R - \log_2 G$
vs. $A = (\log_2 R + \log_2 G)/2$.
- An MA-plot amounts to a 45° counterclockwise rotation of a $\log_2 R$ vs. $\log_2 G$ plot followed by scaling.

Self-self hybridization

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



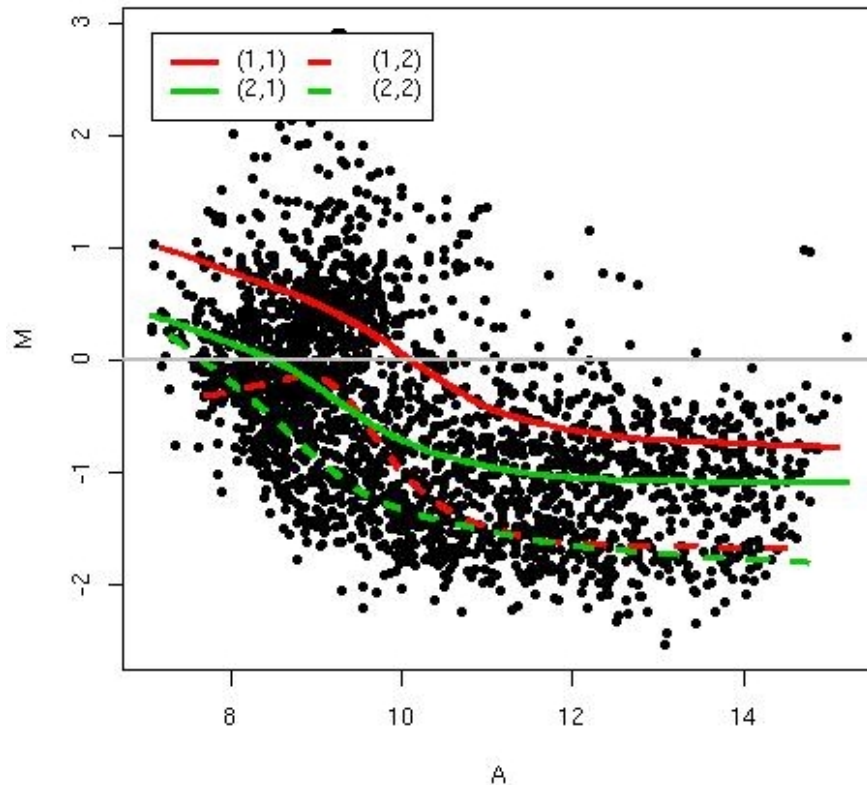
M vs. A



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

Self-self hybridization

M vs. A



Robust local regression
within sectors
(print-tip-groups)
of intensity log-ratio M
on average log-intensity
A.

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

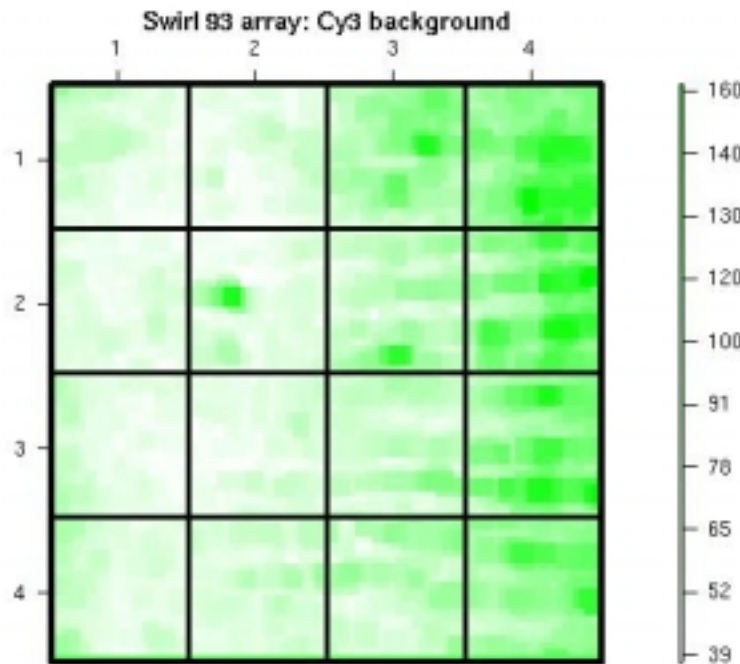
Swirl zebrafish experiment

- **Goal.** Identify genes with altered expression in Swirl mutants compared to wild-type zebrafish.
- 2 sets of dye-swap experiments (n=4).
- Arrays:
 - 8,448 probes (768 controls);
 - 4 x 4 grid matrix;
 - 22 x 24 spot matrices.
- Data available in Bioconductor package **marrayInput**.

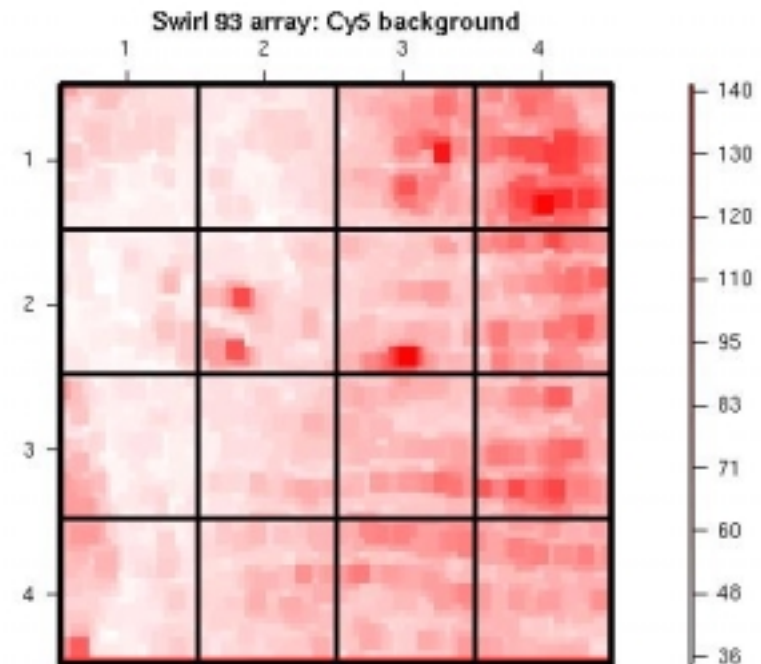
Diagnostic plots

- **Diagnosics 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



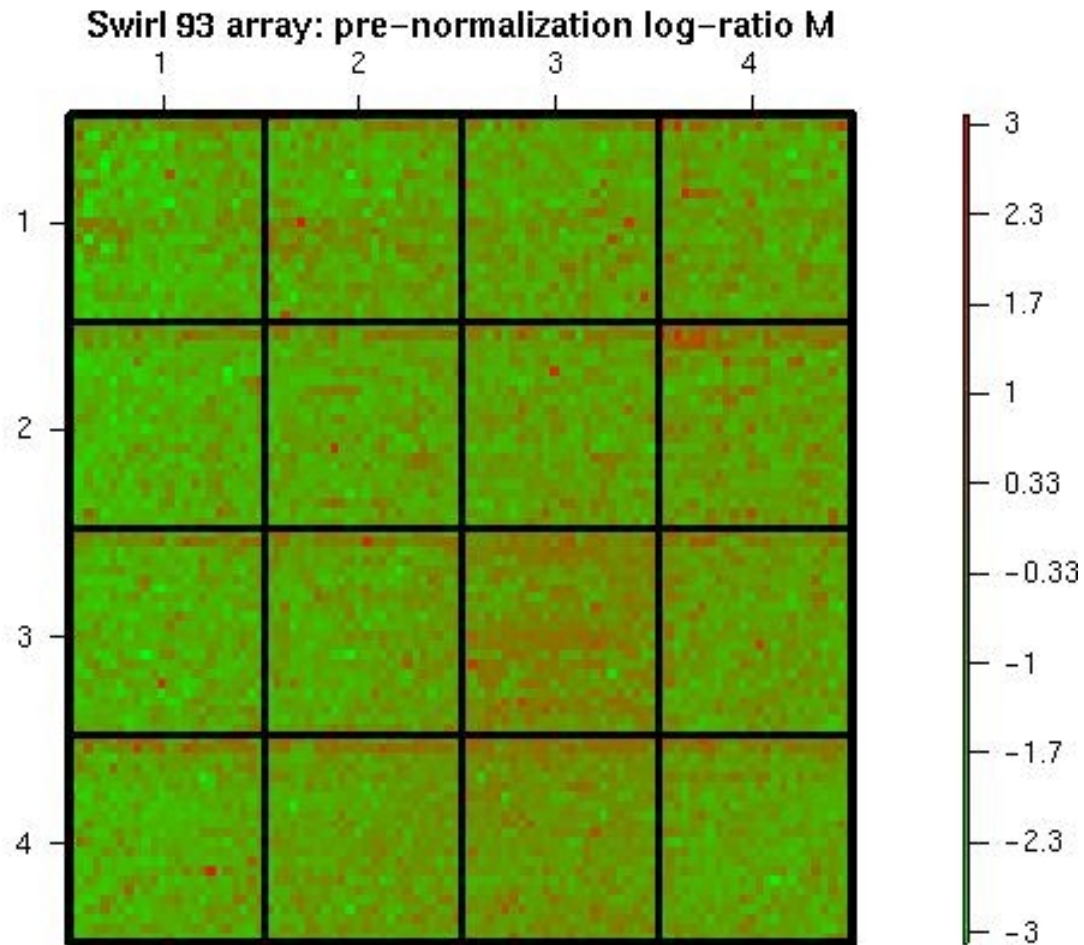
Cy3 background intensity



Cy5 background intensity

2D spatial images

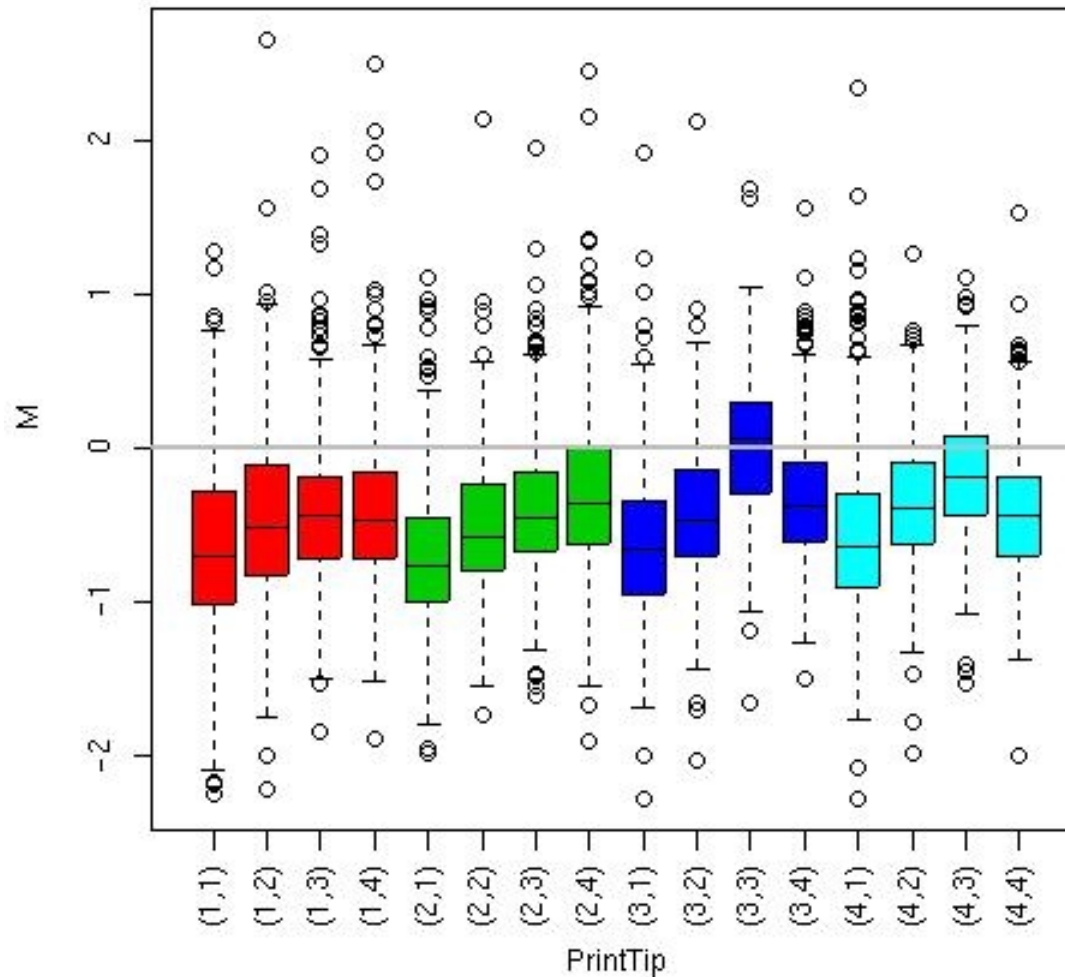
Intensity
log-ratio, M



Boxplots by print-tip-group

Swirl 93 array: pre-normalization log-ratio M

Intensity
log-ratio, M

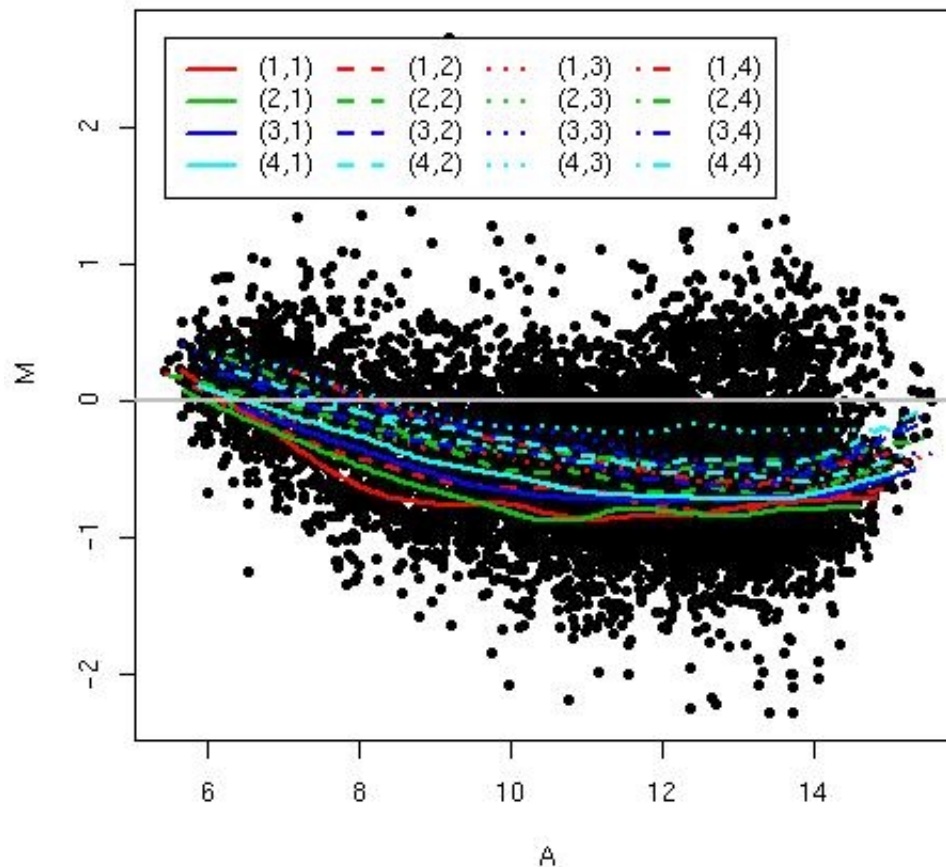


MA-plot by print-tip-group

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

Swirl 93 array: pre-normalization log-ratio M

Intensity
log-ratio, M



Average
log-intensity, A

Location normalization

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

- **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.

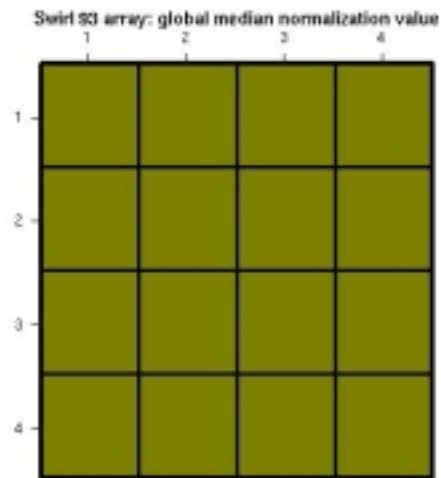
- Regression method: e.g. lowess or loess (Cleveland, 1979; Cleveland & Devlin, 1988).

Location normalization

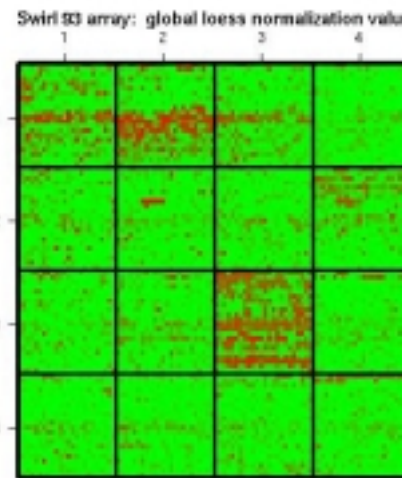
- **Intensity-dependent normalization.**
Regression of M on A (*global loess*).
- **Intensity and sector-dependent normalization.**
Same as above, for each sector separately (*within-print-tip-group loess*).
- **2D spatial normalization.**
Regression of M on 2D-coordinates.
- Other variables: time of printing, plate, etc.
- **Composite normalization.** Weighted average of several normalization functions.

2D images of L values

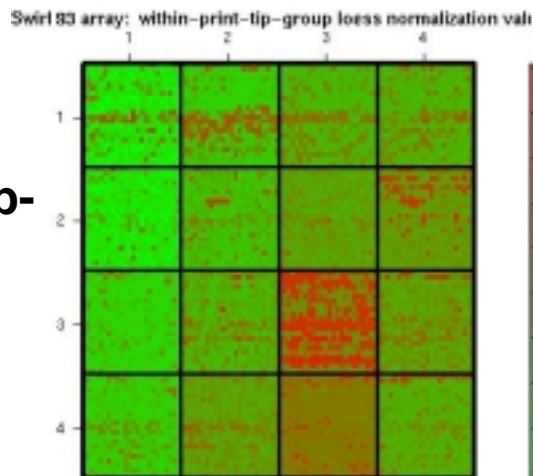
Global median normalization



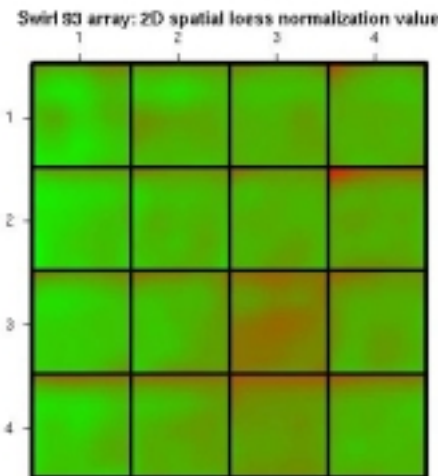
Global loess normalization



Within-print-tip-group loess normalization

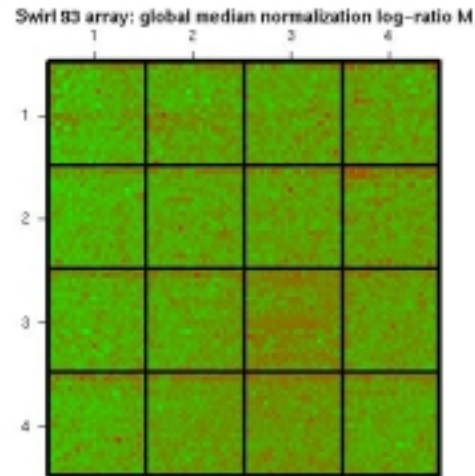


2D spatial normalization

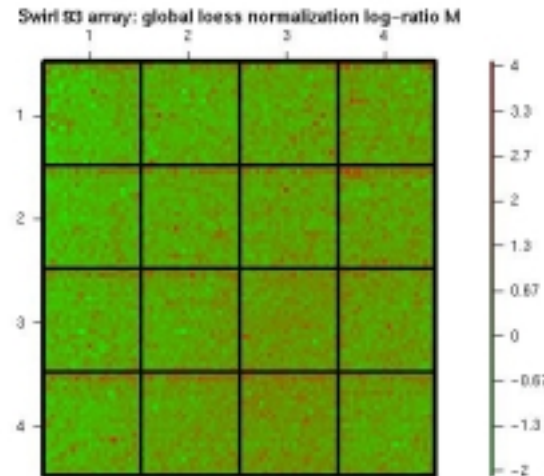


2D images of normalized M-L

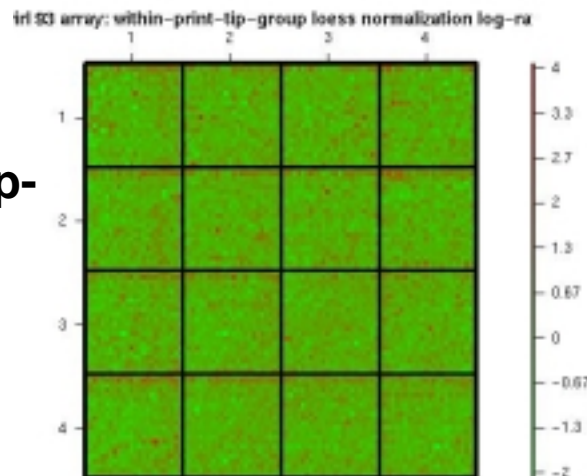
Global median normalization



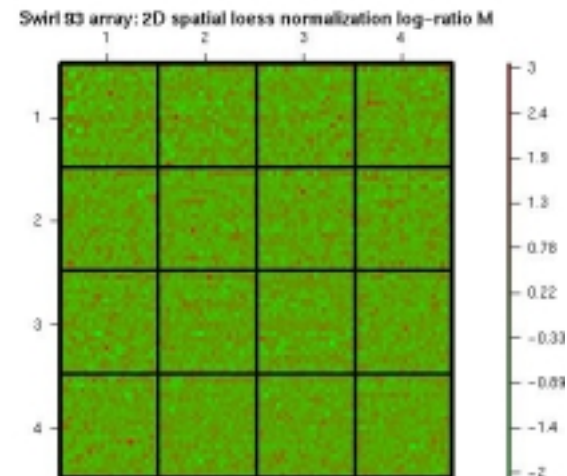
Global loess normalization



Within-print-tip-group loess normalization

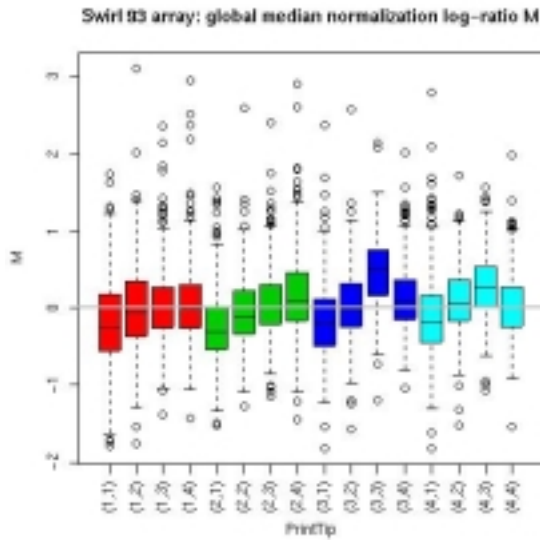


2D spatial normalization

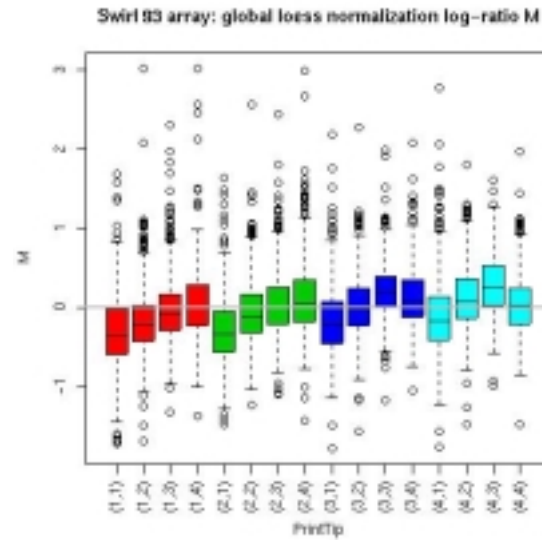


Boxplots of normalized M-L

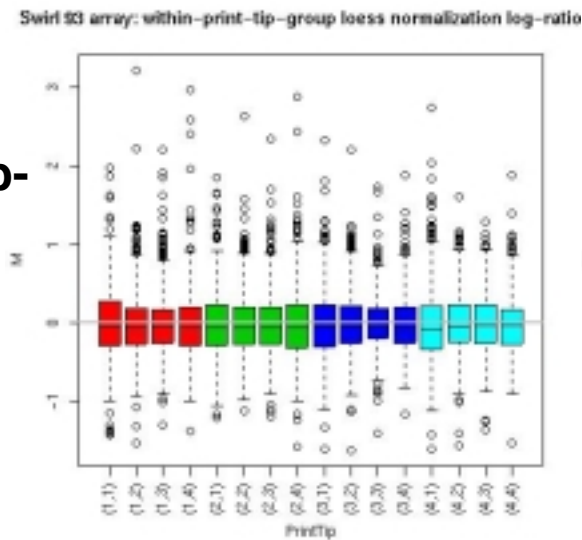
Global median normalization



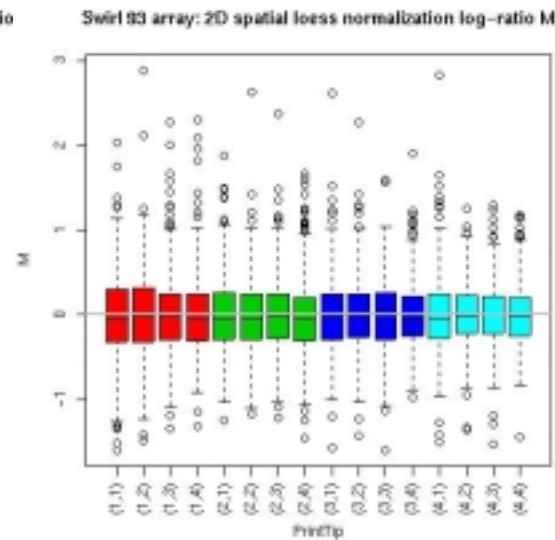
Global loess normalization



Within-print-tip-group loess normalization

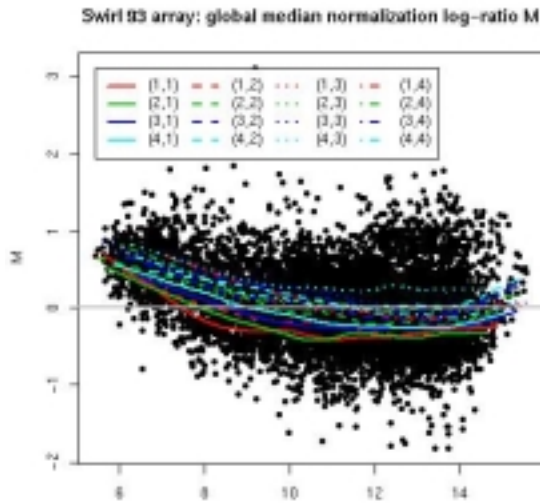


2D spatial normalization

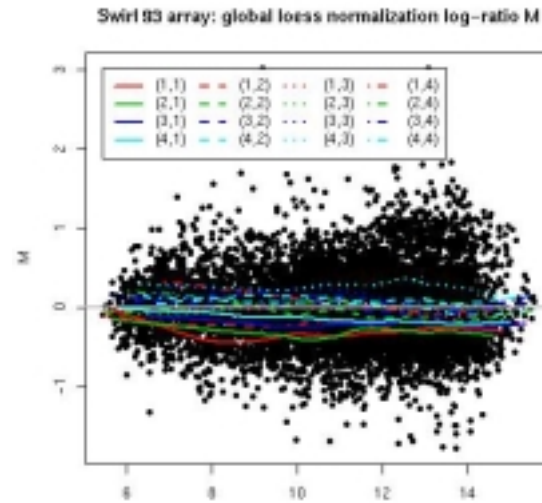


MA-plots of normalized M-L

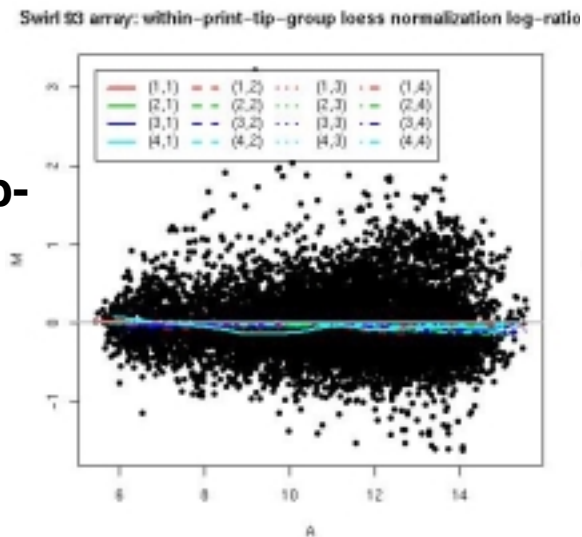
Global median normalization



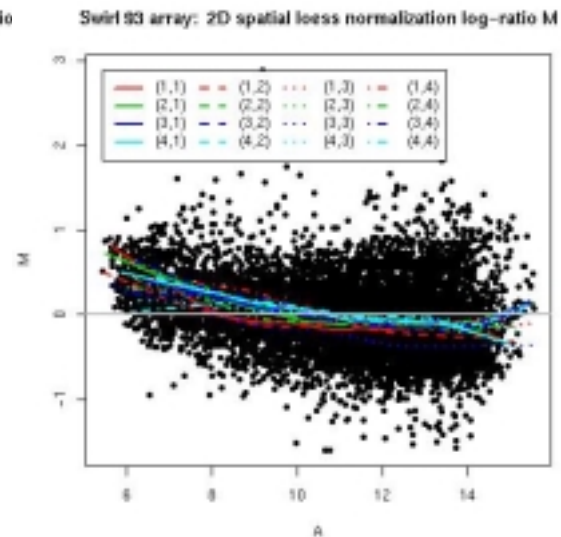
Global loess normalization



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

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 μ_{ij} and M_{ij} , resp. Robust estimate of μ_{ij} is $\hat{a}_i = \frac{MAD_i}{\sqrt[3]{\prod_{i=1}^I MAD_i}}$ and i indexes print-tip-group and j spots.

where MAD_i is MAD of M_{ij} in print-tip-group

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

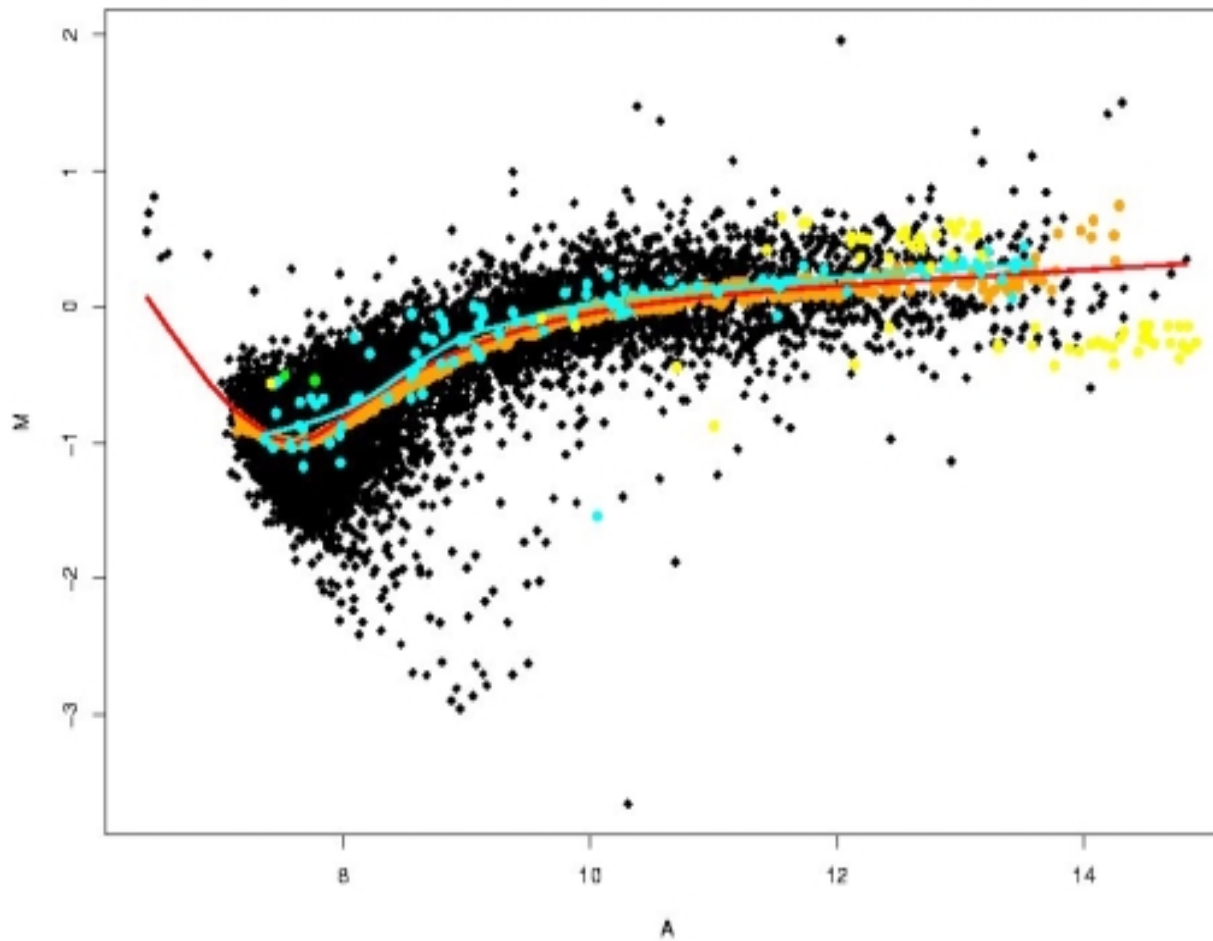
- **Microarray Sample Pool, MSP**: Control sample for normalization, in particular, when it is not safe to assume most genes are equally expressed in both channels.
- MSP: **pooled** all 18,816 ESTs from RIKEN release 1 cDNA mouse library.
- Six-step **dilution series** of the MSP.
- MSP samples were 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.

Microarray sample pool



MSP

Rank invariant

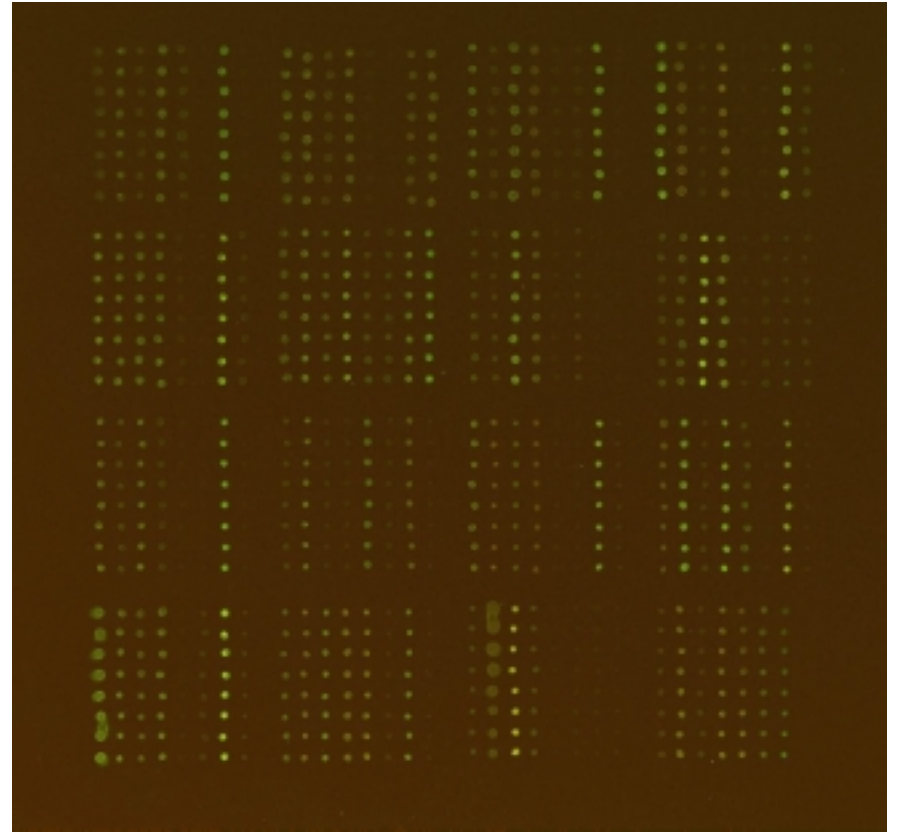
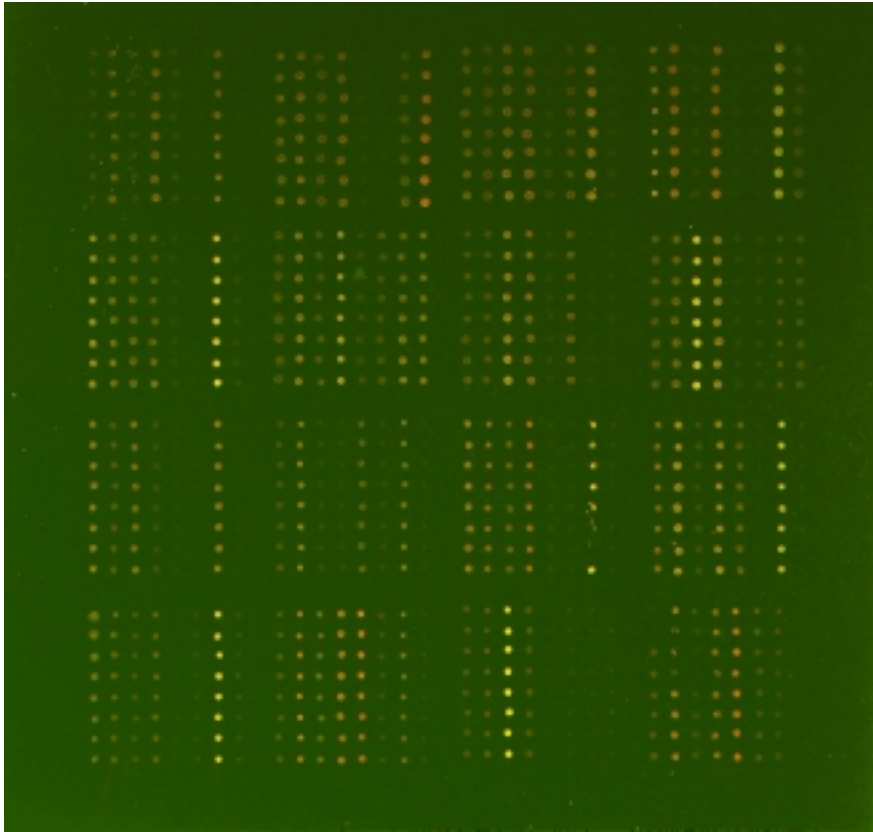
Housekeeping

Tubulin, GAPDH

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:

$$M - M'$$

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

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

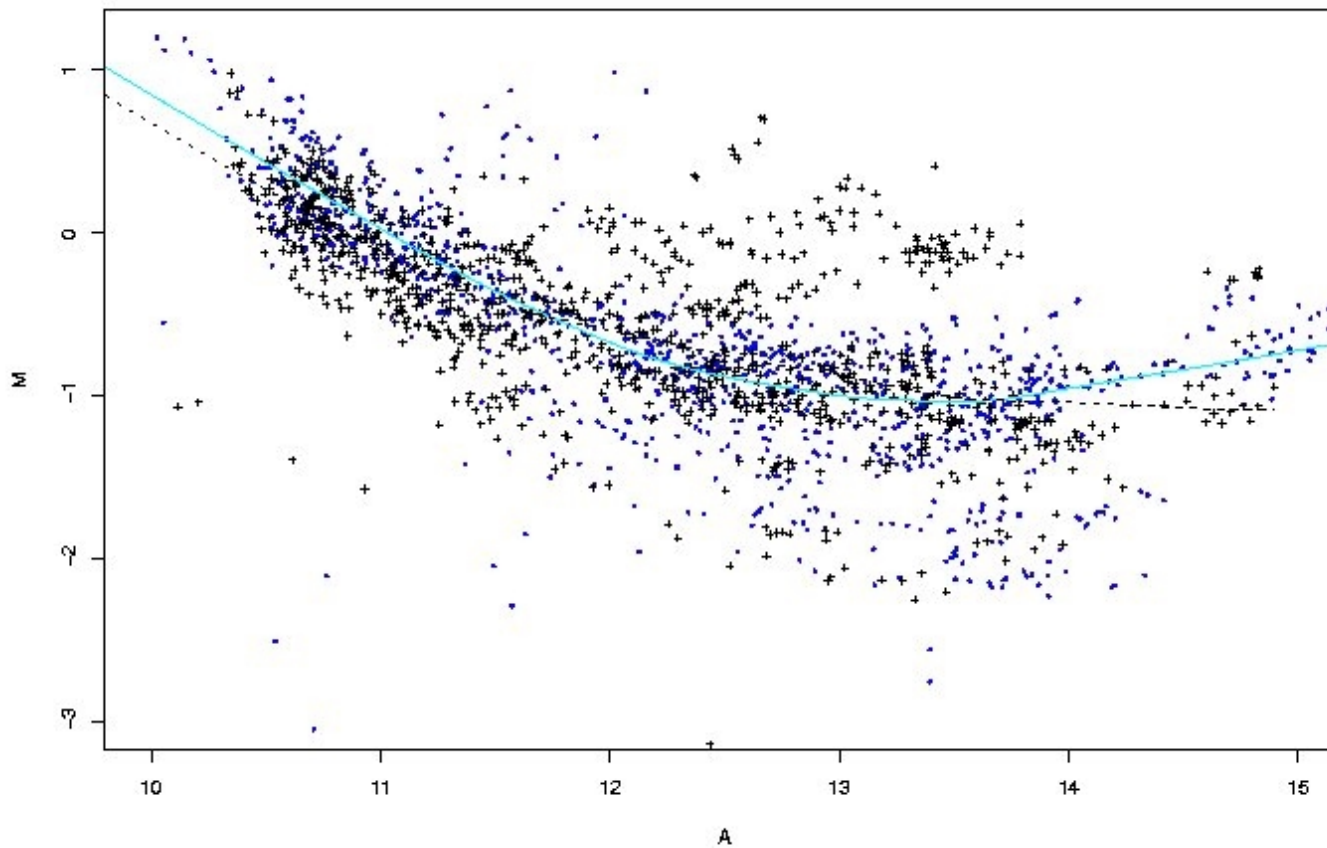
$$\approx [\log_2 (RG'/GR')] / 2$$

provided $L = L'$.

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

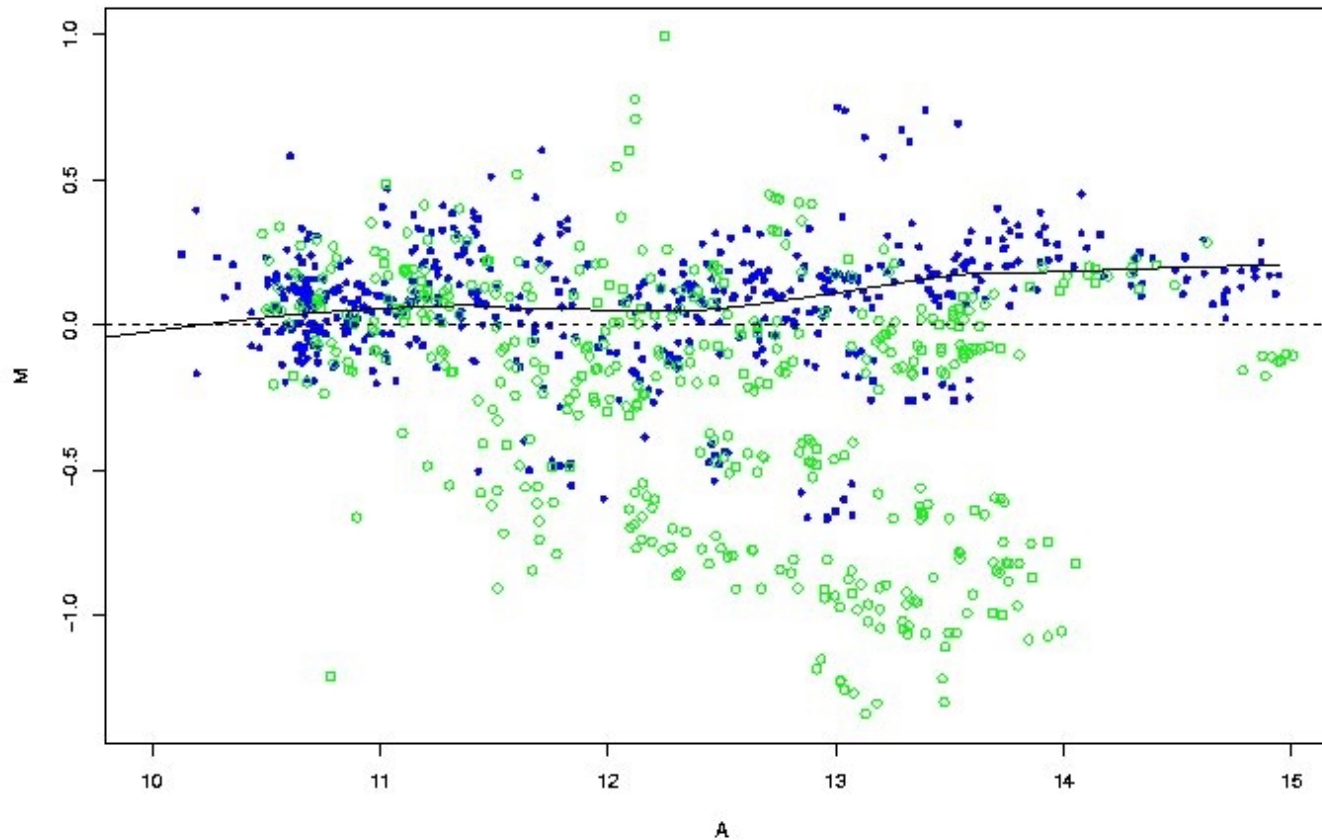
Checking the assumption

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.

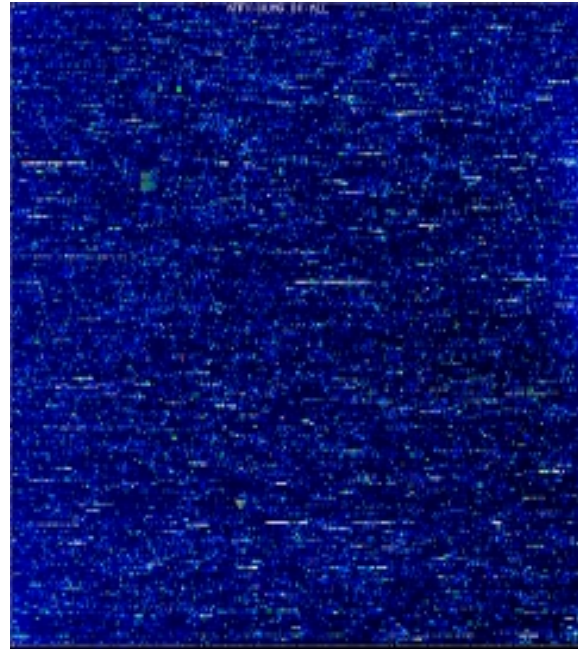
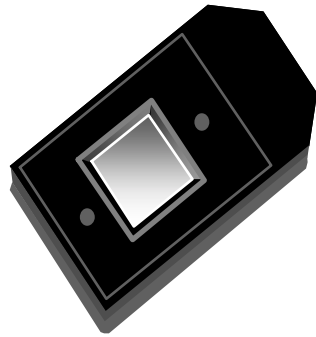
Case 2. Many genes are expected to change.

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

R software for normalization cDNA microarrays

- Bioconductor R packages
 - **marrayClasses**:
 - class definitions for microarray data objects;
 - basic methods for manipulation of microarray objects.
 - **marrayInput**:
 - reading in intensity data and textual data describing probes and targets;
 - automatic generation of microarray data objects;
 - widgets for point & click interface.
 - **marrayPlots**: diagnostic plots.
 - **marrayNorm**: robust adaptive location and scale normalization procedures.

Oligonucleotide chips



Affymetrix files

- Main software from Affymetrix company
MAS - MicroArray Suite, now version 5.
- **DAT** file: Image file, $\sim 10^7$ pixels, ~ 50 MB.
- **CEL** file: Cell intensity file, probe level PM and MM values.
- **CDF** file: Chip Description File. Describes which probes go in which probe sets (genes, gene fragments, ESTs).

Image analysis

- Raw data, **DAT image files** → **CEL files**
- Each probe cell: 10x10 pixels.
- **Gridding**: estimate location of probe cell centers.
- **Signal**:
 - Remove outer 36 pixels → 8x8 pixels.
 - The probe cell signal, PM or MM, is the 75th percentile of the 8x8 pixel values.
- **Background**: Average of the lowest 2% probe cell values is taken as the background value and subtracted.
- Compute also quality measures.

Data and notation

- PM_{ijg} , MM_{ijg} = Intensity for perfect match and mismatch probe in cell j for gene g in chip i .
 - $i = 1, \dots, n$ -- from one to hundreds of chips,
 - $j = 1, \dots, J$ -- usually 16 or 20 probe pairs,
 - $g = 1, \dots, G$ -- between 8,000 and 20,000 probe sets.
- Task: summarize for each probe set the probe level data, i.e., 20 PM and MM pairs, into a single **expression measure**.
- Expression measures may then be compared within or between chips for detecting differential expression.

Expression measures

MAS 4.0

- GeneChip[®] MAS 4.0 software uses **AvDiff**

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

where A is a set of “suitable” pairs, e.g. pairs with $d_j = PM_j - MM_j$ within 3 SDs of the average of $d_{(2)}, \dots, d_{(J-1)}$.

- Log-ratio version is also used: average of $\log(PM/MM)$.

Expression measures

MAS 5.0

- GeneChip[®] MAS 5.0 software uses **Signal**
 $signal = \text{Tukey Biweight}\{\log(PM_j - MM_j^*)\}$
with MM^* a new version of MM that is never larger than PM.
- If $MM < PM$, $MM^* = MM$.
- If $MM \geq PM$,
 - $SB = \text{Tukey Biweight}(\log(PM) - \log(MM))$
(log-ratio).
 - $\log(MM^*) = \log(PM) - \log(\max(SB, +ve))$.
- Tukey Biweight: $B(x) = (1 - (x/c)^2)^2$ if $|x| < c$, 0 ow.

Expression measures

Li & Wong

- Li & Wong (2001) fit a model for each probe set, i.e., gene

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \propto N(0, \sigma^2)$$

where

- θ_i : **model based expression index** (MBEI),
- ϕ_j : probe sensitivity index.
- Maximum likelihood estimate of MBEI is used as expression measure for the gene in chip i .
- Need at least 10 or 20 chips.
- Current version works with PMs only.

Expression measures

- Most expression measures are based on **PM-MM**, with the intention of correcting for non-specific binding and background noise.
- Problems:
 - MMs are PMs for some genes,
 - removing the middle base does not make a difference for some probes .
- Why not simply average PM or log PM? Not good enough, still need to adjust for background.
- Also need to normalize.

Expression measures

RMA

Irizarry et al. (2002).

1. Estimate **background** BG and use only background-corrected PM: $\log_2(\text{PM}-\text{BG})$.
2. Probe level **normalization** of $\log_2(\text{PM}-\text{BG})$ for suitable set of chips.
3. **Robust Multi-chip Analysis, RMA**, of $\log_2(\text{PM}-\text{BG})$.

RMA background, I

Simple background estimation

- Estimate $\log_2(\text{BG})$ as the mode of the $\log_2(\text{MM})$ distribution for a given chip (kernel density estimate).
- Quick fix when $\text{PM} \leq \text{BG}$: use half of the minimum of $\log_2(\text{PM}-\text{BG})$ for $\text{PM} > \text{BG}$ over all chips and probes.

RMA background, II

More refined background estimation

- Model observed PM as the sum of a signal intensity SG and a background intensity BG

$$PM = SG + BG,$$

where it is assumed that SG is *Exponential* (α), BG is *Normal* (μ, σ^2), and SG and BG are independent.

- Background adjusted PM values are then **$E(SG|PM)$** .

Quantile normalization

- **Probe level quantile normalization** (Bolstad et al., 2002).
- Co-normalize probe level intensities, e.g. PM-BG or just PM or MM, for n chips by averaging each quantile across chips.
- Assumption: same probe level intensity distribution across chips.
- No need to choose a baseline or work in a pairwise manner.
- Deals with non-linearity.

Curve-fitting normalization

- Astrand (2001). Generalization of M vs. A robust local regression normalization for cDNA arrays.
- For n chips, regress orthonormal contrasts of probe level statistics on the average of the statistics across chips.

RMA expression measures, I

Simple measure

$$\text{RMA} = \frac{1}{|A|} \sum_{j \in A} \log_2(PM_j - BG_j)$$

with A a set of “suitable” pairs.

RMA expression measures, II

- **Robust regression method** to estimate expression measure and SE from PM-BG values.
- Assume additive model

$$\log_2(PM_{ij} - BG) = a_i + b_j + \varepsilon_{ij}$$

- Estimate RMA = a_i for chip i using robust method, such as **median polish** (fit iteratively, successively removing row and column medians, and accumulating the terms, until the process stabilizes).
- Fine with $n=2$ or more chips.

Conclusions

- Don't use MM.
- “Background correct” PM. Even global background improves on probe-specific MM.
- Take logs: probe effect is additive on log scale.
- PMs need to be normalized (e.g. quantile normalization).
- RMA is arguably the best summary in terms of bias, variance, and model fit. Comparison study in Irizarry et al. (2002).

R software for pre-processing of Affymetrix data

- Bioconductor R package **affy**.
- Background estimation.
- Probe-level normalization: quantile, curve-fitting.
- Expression measures: AvDiff, Signal, Li & Wong (2001), RMA.
- Two main functions: **ReadAffy**, **express**.

Combining data across slides

Data on G genes for n hybridizations

→ $G \times n$ genes-by-arrays data matrix

		Arrays					...
		Array1	Array2	Array3	Array4	Array5	
Genes	Gene1	0.46	0.30	0.80	1.51	0.90	...
	Gene2	-0.10	0.49	0.24	0.06	0.46	...
	Gene3	0.15	0.74	0.04	0.10	0.20	...
	Gene4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	Gene5	-0.06	1.06	1.35	1.09	-1.09	...

$M = \log_2(\text{Red intensity} / \text{Green intensity})$
expression measure, e.g, RMA

Combining data across slides

... but columns have **structure**

How can we design experiments and combine data across slides to provide accurate estimates of the effects of interest?

Experimental design
Regression analysis

