cDNA microarray experiments: pre-processing and experimental design

Statistics and Genomics - Lecture 1, Part II

Department of Biostatistics
Harvard School of Public Health

January 23-25, 2002

Outline of lecture 1, Part II

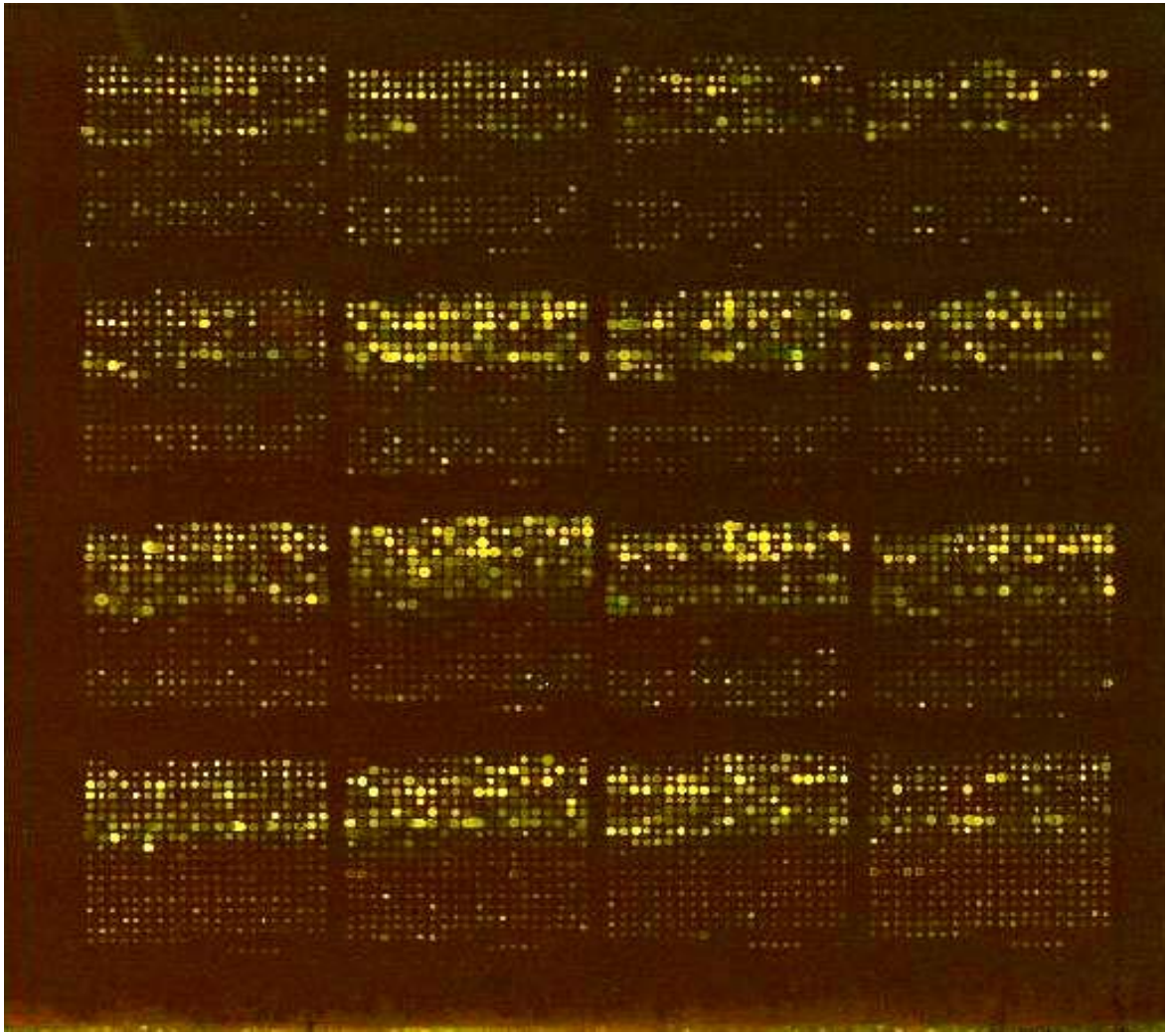
cDNA microarrays

- Pre-processing: Image analysis;
- Pre-processing: Normalization;
- Experimental design.

Terminology

- **Probe:** DNA spotted on the array, aka. spot, immobile substrate.
- **Target:** DNA hybridized to the array, mobile substrate.
- **Sector:** collection of spots printed using the same print-tip (or pin),

  aka. **print-tip-group**, **pin-group, spot matrix**, **grid**.
- **Batch:** collection of slides with the same probe layout.
- The terms **slide** or **array** are often used to refer to the printed microarray.



4 x 4 sectors
399 probes/sector
6,384 probes/array

Image analysis

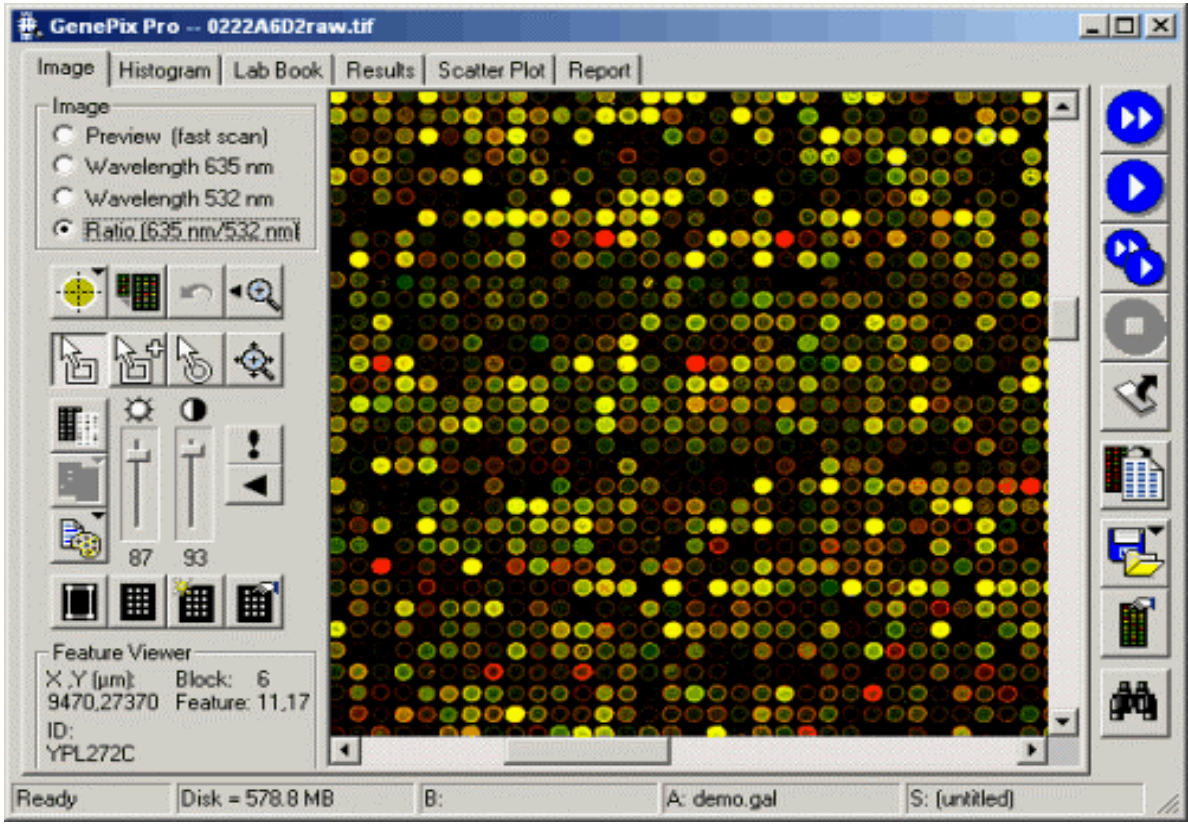
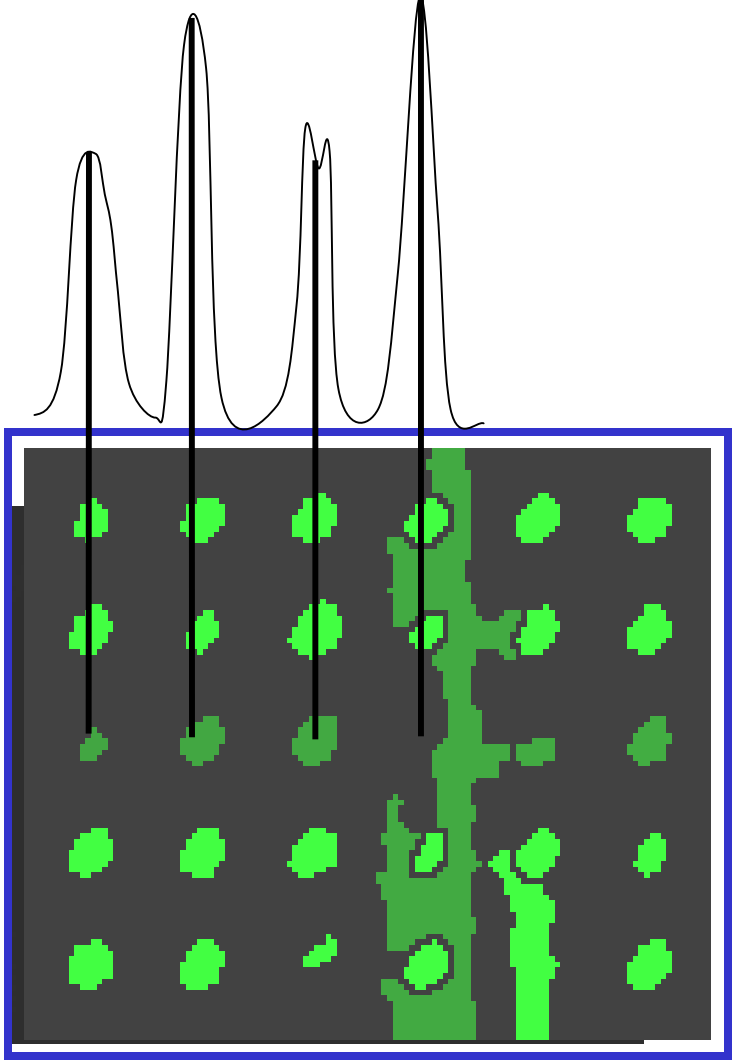


Image analysis

- 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 for each spot on the array.

Image analysis



1. Addressing. Estimate location of spot centers.

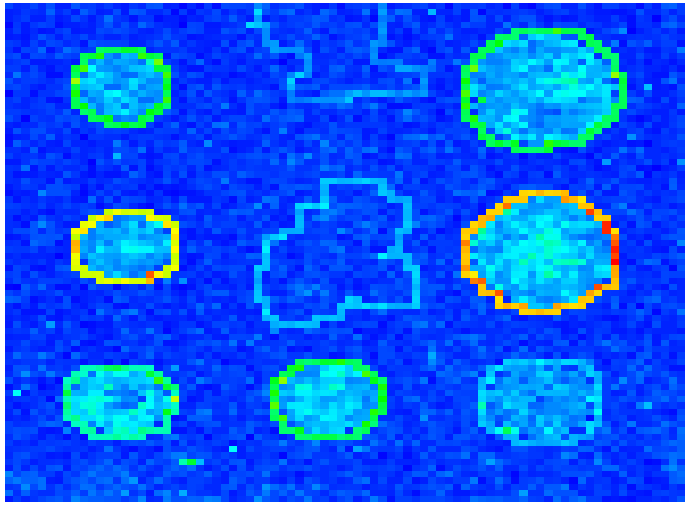
2. Segmentation. Classify pixels as foreground (signal) or background.

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

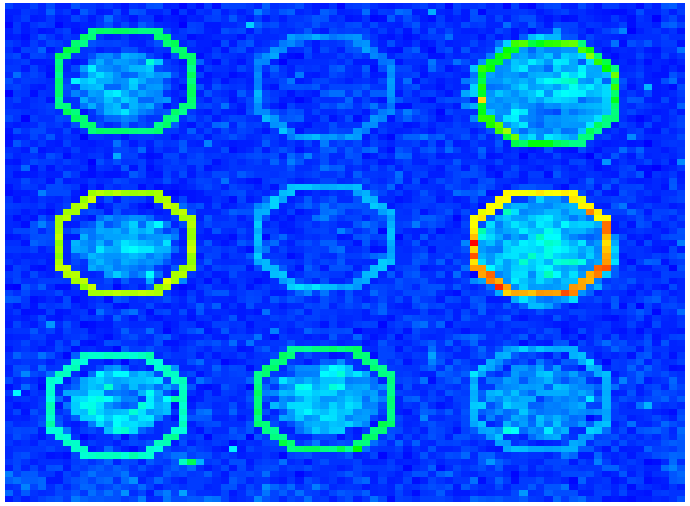
- signal intensities;
- background intensities;
- quality measures.

→ **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.

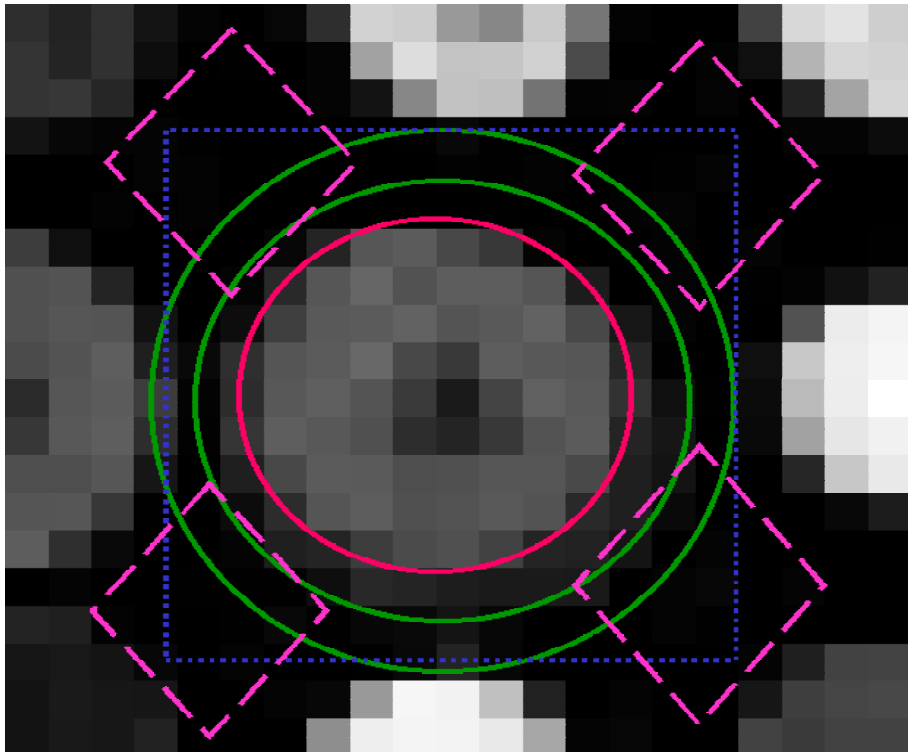
  Here, based on fitted foreground and background grids from the addressing step.
- The decision to add a pixel to a region is based on the absolute gray-level difference of that pixel's intensity and the average of the pixel values in the neighboring region.
- Done on combined red and green images.

Local background

---- GenePix

---- QuantArray

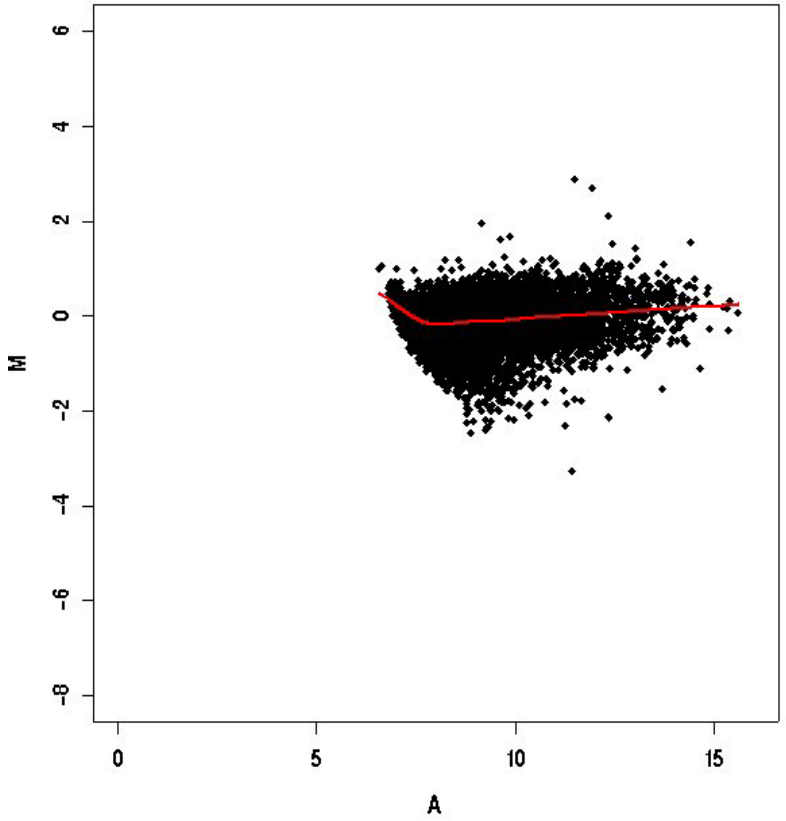
---- ScanAnalyze



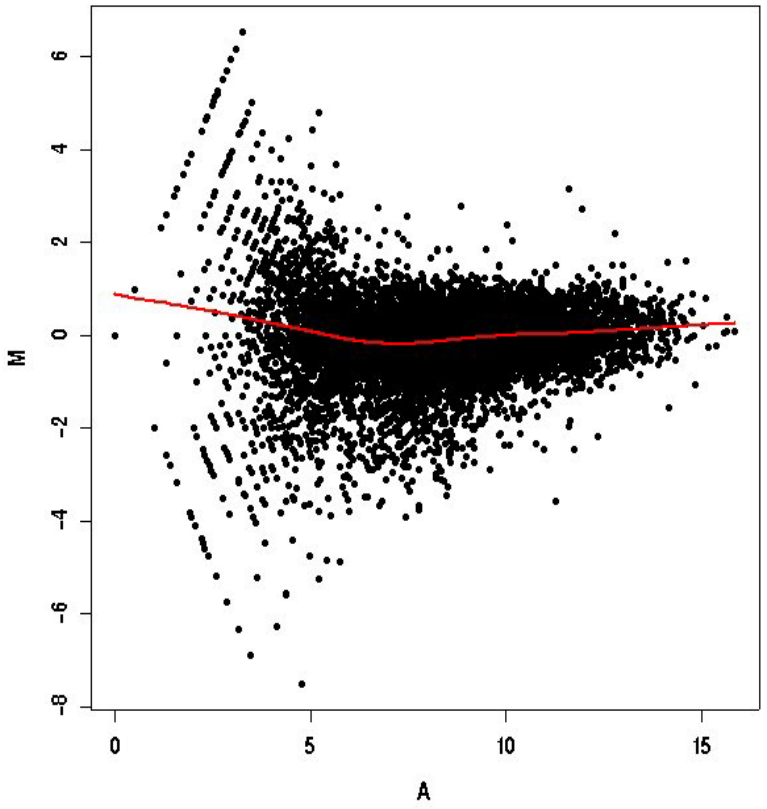
Morphological opening

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

Background matters



Morphological opening



Local background

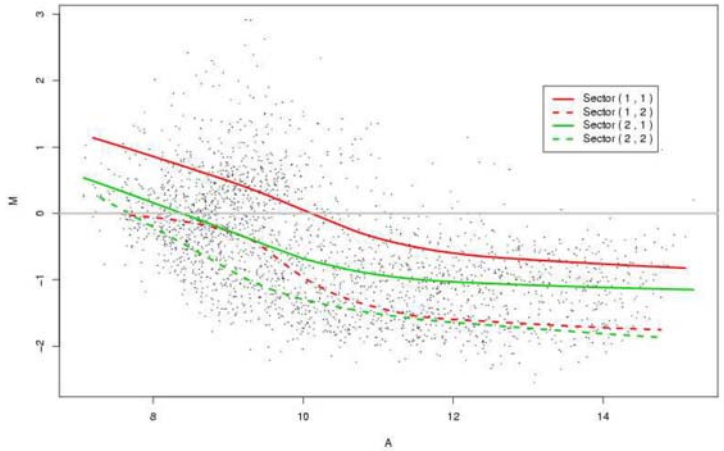
Quality measures

- Spot quality
  - **Brightness:** foreground/background ratio;
  - **Uniformity:** variation in pixel intensities and ratios of intensities;
  - **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

- 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

- Identify and remove sources 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.
- Necessary for within and between slides comparisons of expression levels.

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

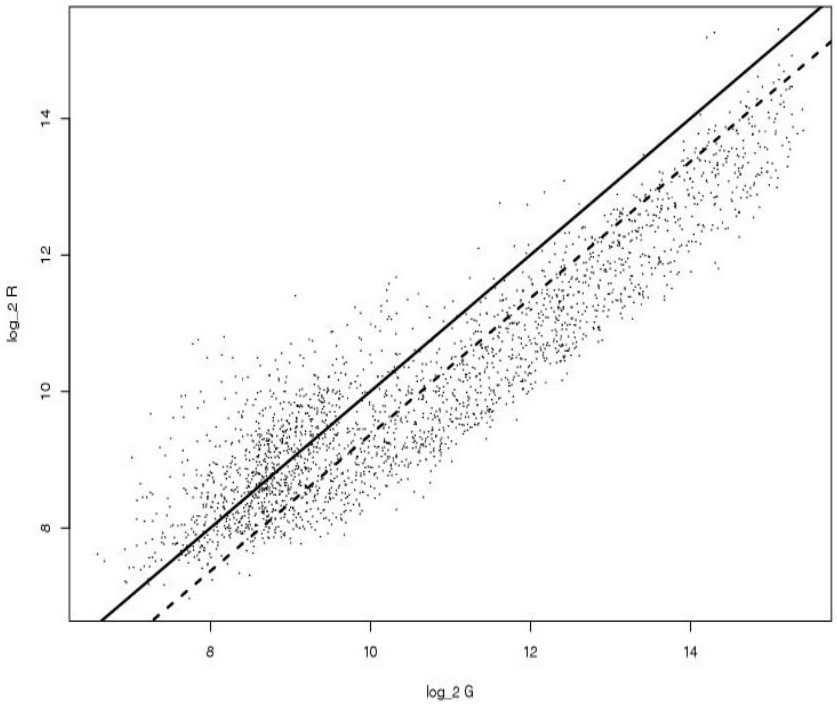
  $M = \log_2 R - \log_2 G$

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

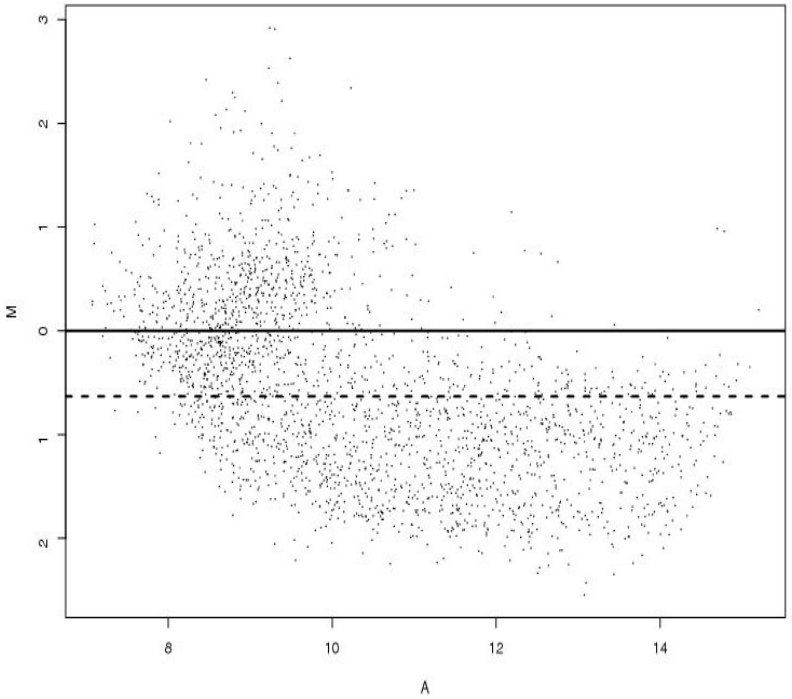
- An MA-plot amounts to a 45° counterclockwise rotation of a $\log_2 R$ vs. $\log_2 G$ plot followed by scaling.

Self-self hybridization

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



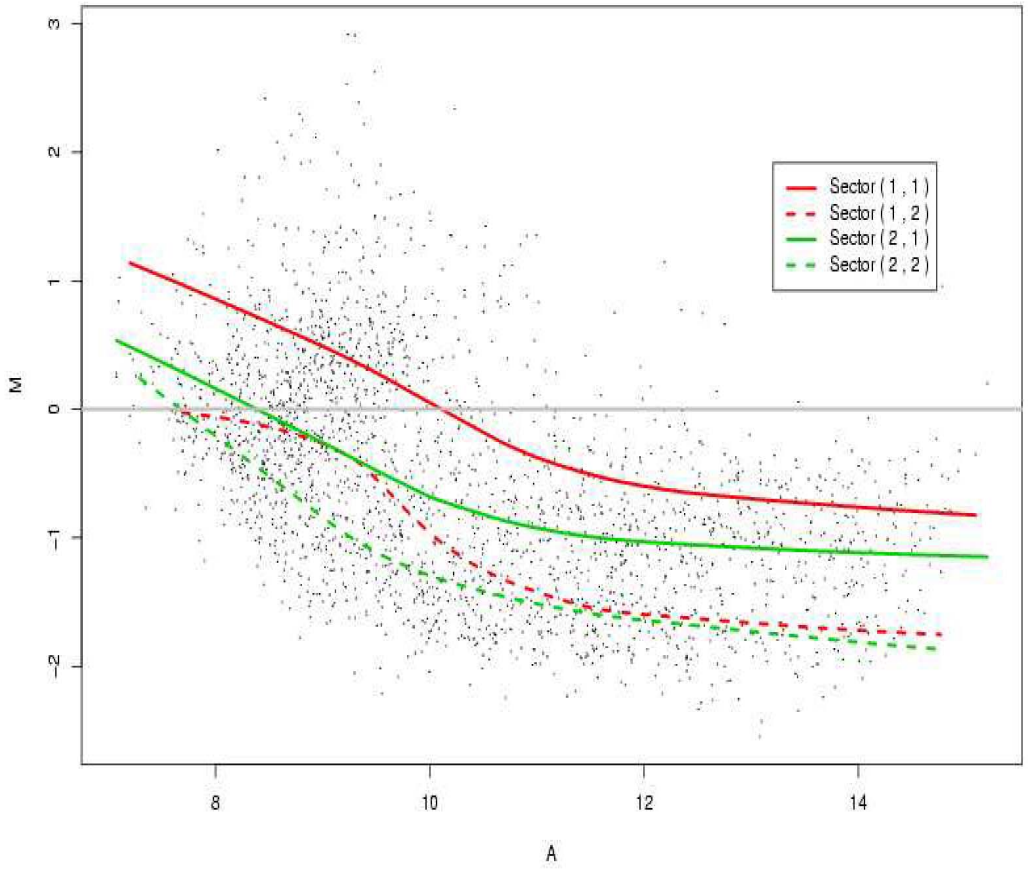
M vs. A



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

Self-self hybridization

M vs. A plot



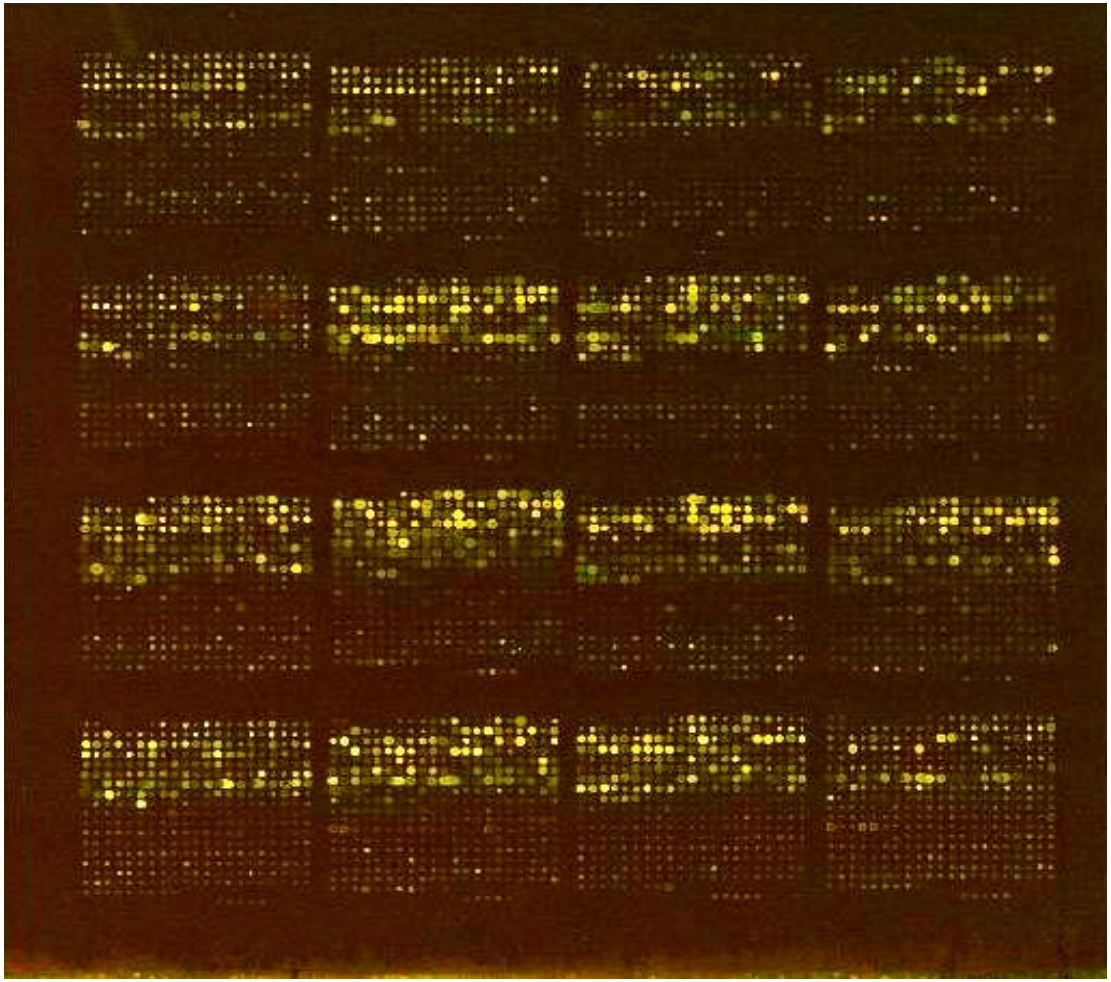
Robust local regression
within sectors
(print-tip-groups)
of intensity log-ratio M
on average log-intensity A.

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

Apo AI experiment

- Goal. Identify genes with altered expression in the livers of apo AI knock-out mice compared to inbred C57Bl/6 control mice.

- 8 treatment (trt) mice and 8 control (ctl) mice.
- 16 hybridizations: target mRNA from each of the 16 mice is labeled with Cy5, pooled mRNA from control mice is labeled with Cy3.
- Probes: 6,384 spots, including 257 genes related to lipid metabolism.



Target mRNA samples:

R = apo A1 ko mouse liver mRNA

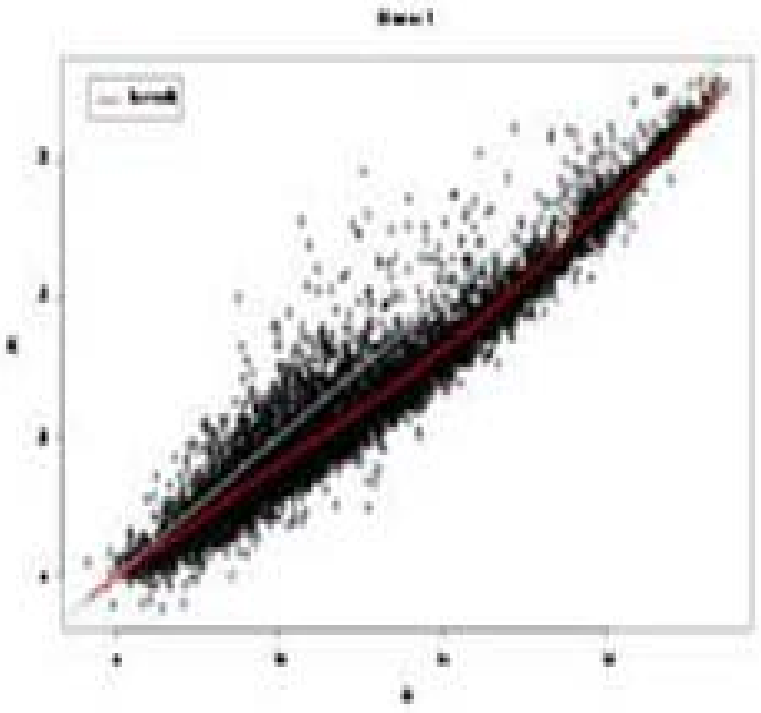
G = pooled control C57Bl/6 mouse liver mRNA

Probes: 6,384 spots, including 257 genes related to lipid metabolism.

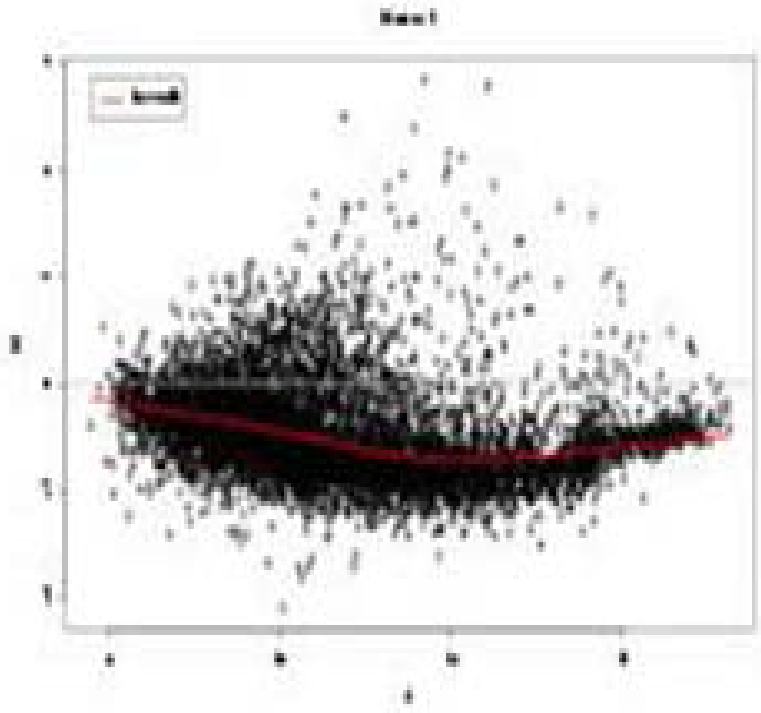
Diagnostic plots

- Diagnostics plots of various spot statistics, such as red and green log-intensities, intensity log-ratios M, average log-intensity A, spot area, etc.
  - Boxplots;
  - 2D images or spatial plots;
  - Scatter plots, e.g. MA-plots;
  - Density plots.
- Stratify plots according to layout parameters, e.g. sector.

MA-plot

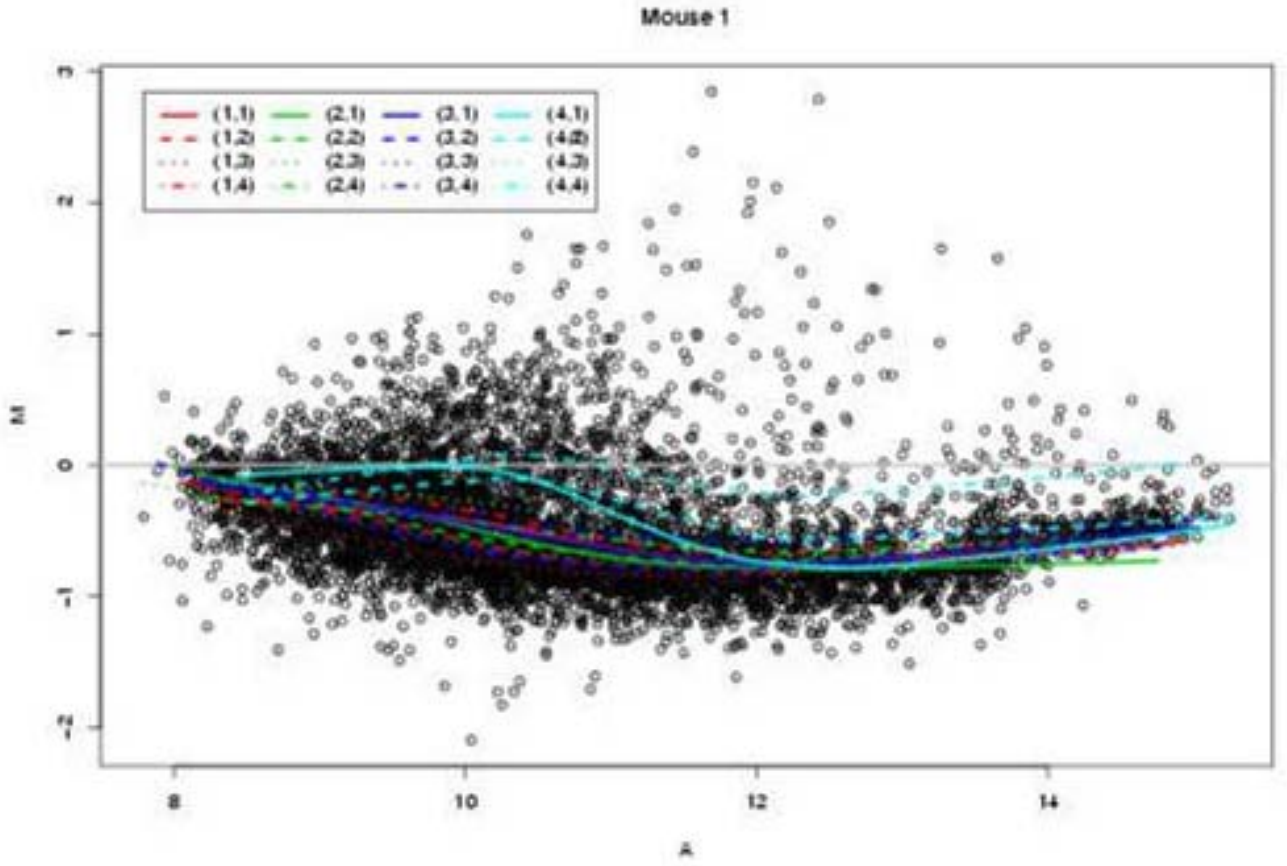


$\log_2 R$ vs. $\log_2 G$

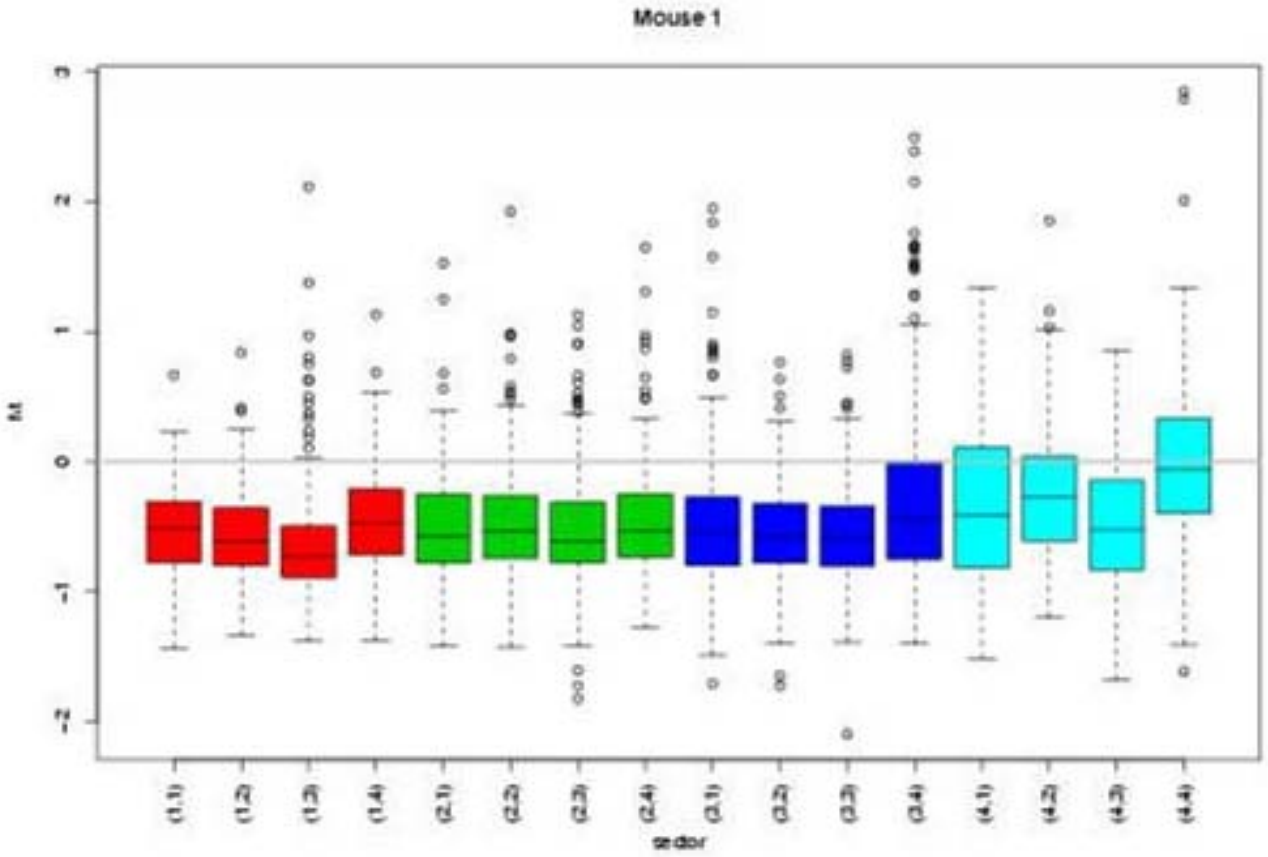


M vs. A

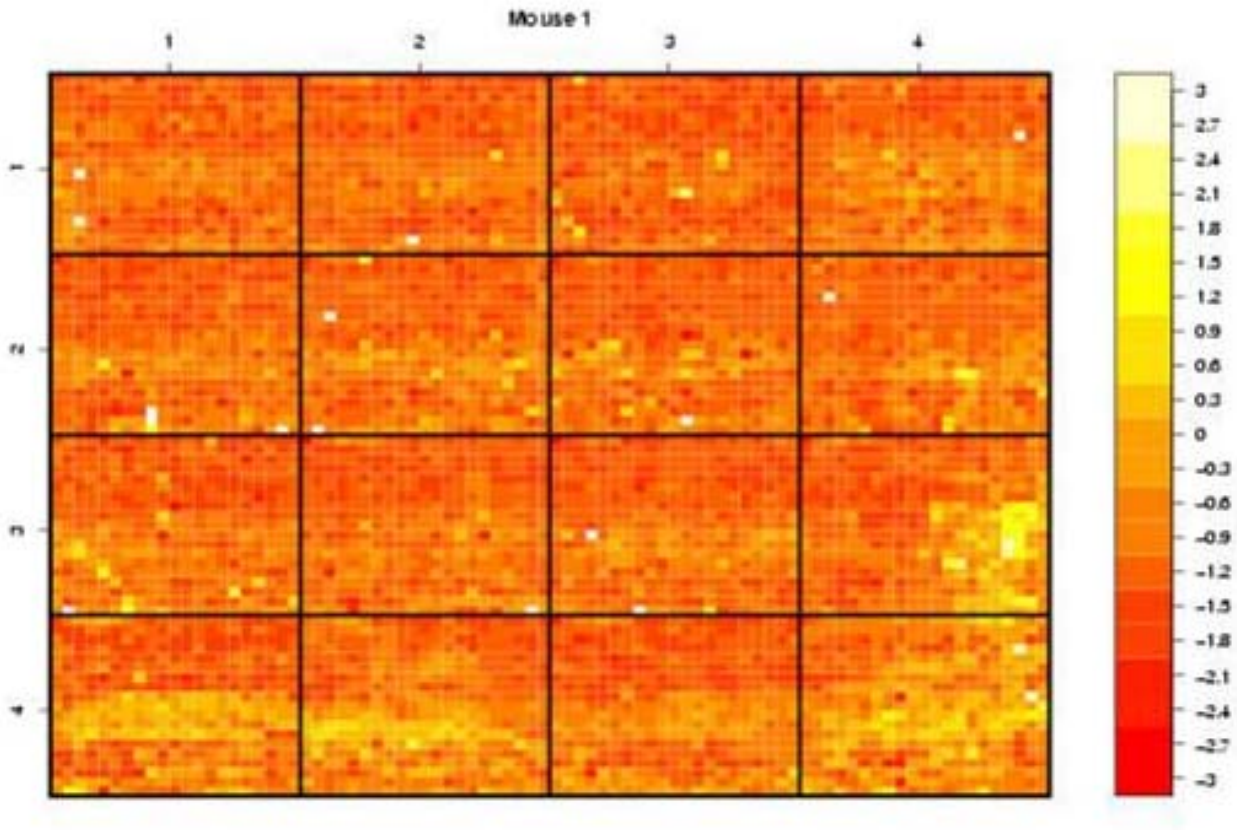
MA-plot by sector



Boxplots by sector



2D image



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.

Location normalization

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

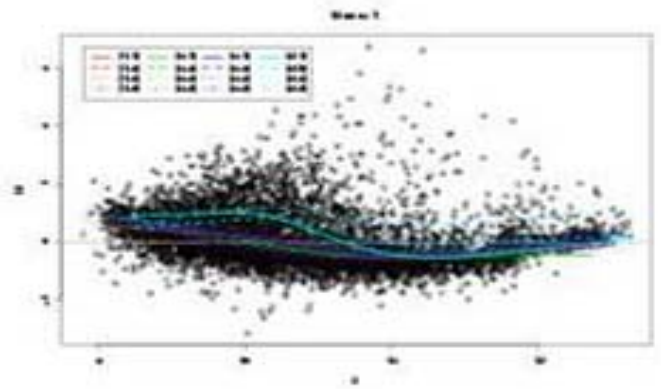
- **Global normalization.** Normalization function l is **constant** across the spots and equal to the mean or median of the log-ratios M.
- **Adaptive normalization.** Normalization function l depends on a number of **predictor variables**, such as spot intensity, location, plate origin.
- The normalization function can be obtained by **robust locally weighted regression** of the log-ratios M on the predictor variables.

  E.g. lowess or loess smoothers.

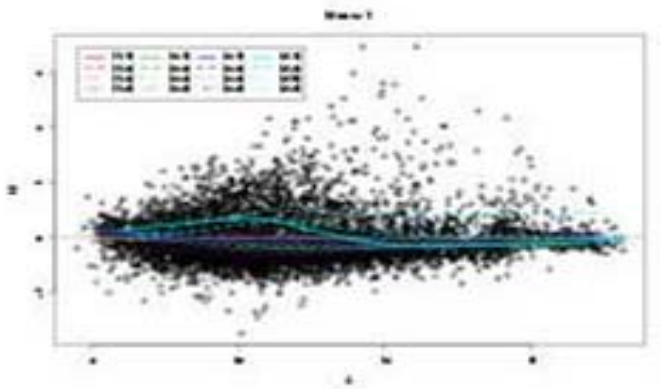
Location normalization

- **Intensity-dependent normalization.**
  Regression of M on A.
- **Intensity and sector-dependent normalization**.
  Same as above, for each sector separately.
- **Spatial normalization**.
  Regression of M on 2D-coordinates.
- Other variables: time of printing, plate, etc.
- **Composite normalization**.

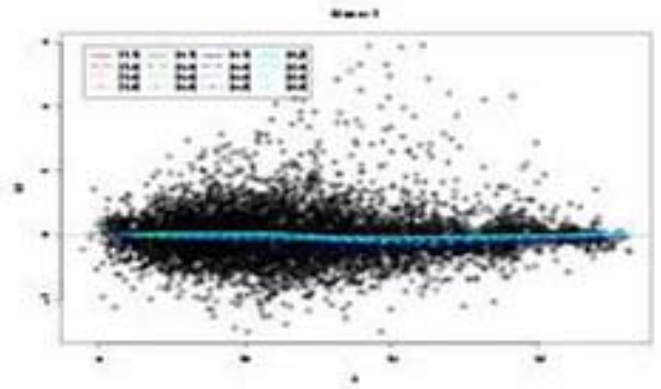
Post-normalization MA-plots



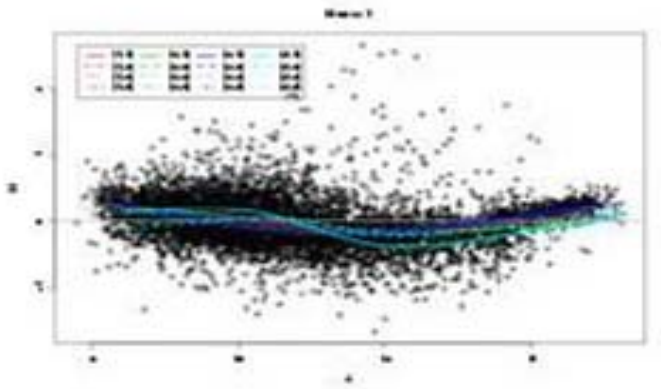
Median



A-dependent

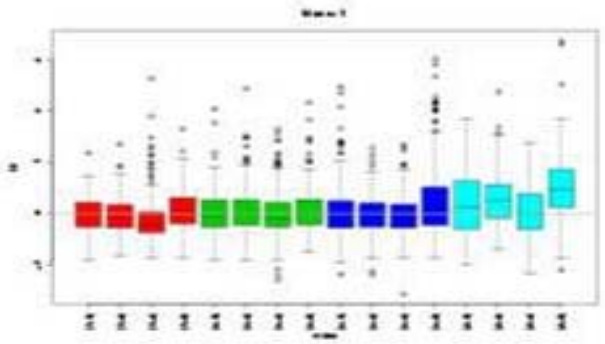


A+sector-dependent

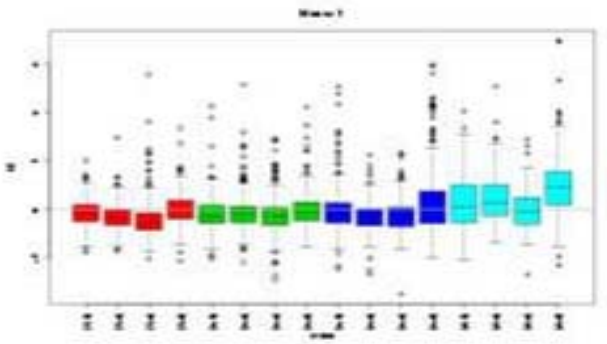


2D

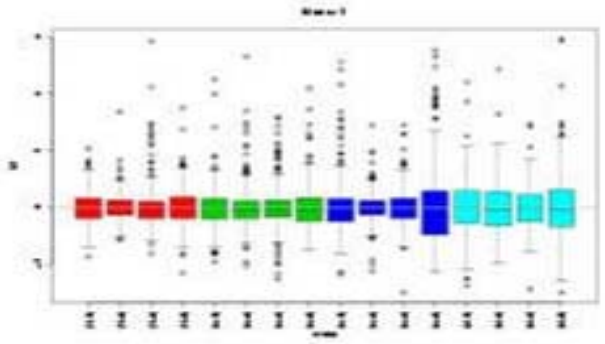
Post-normalization boxplots



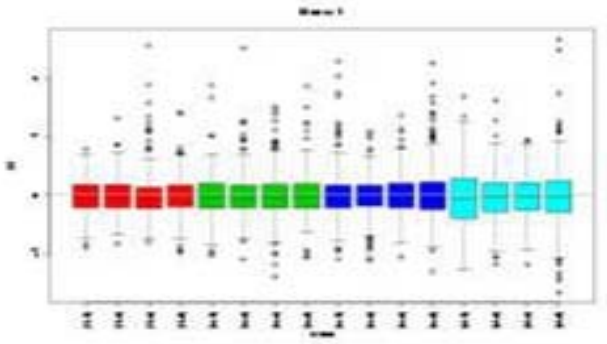
Median



A-dependent

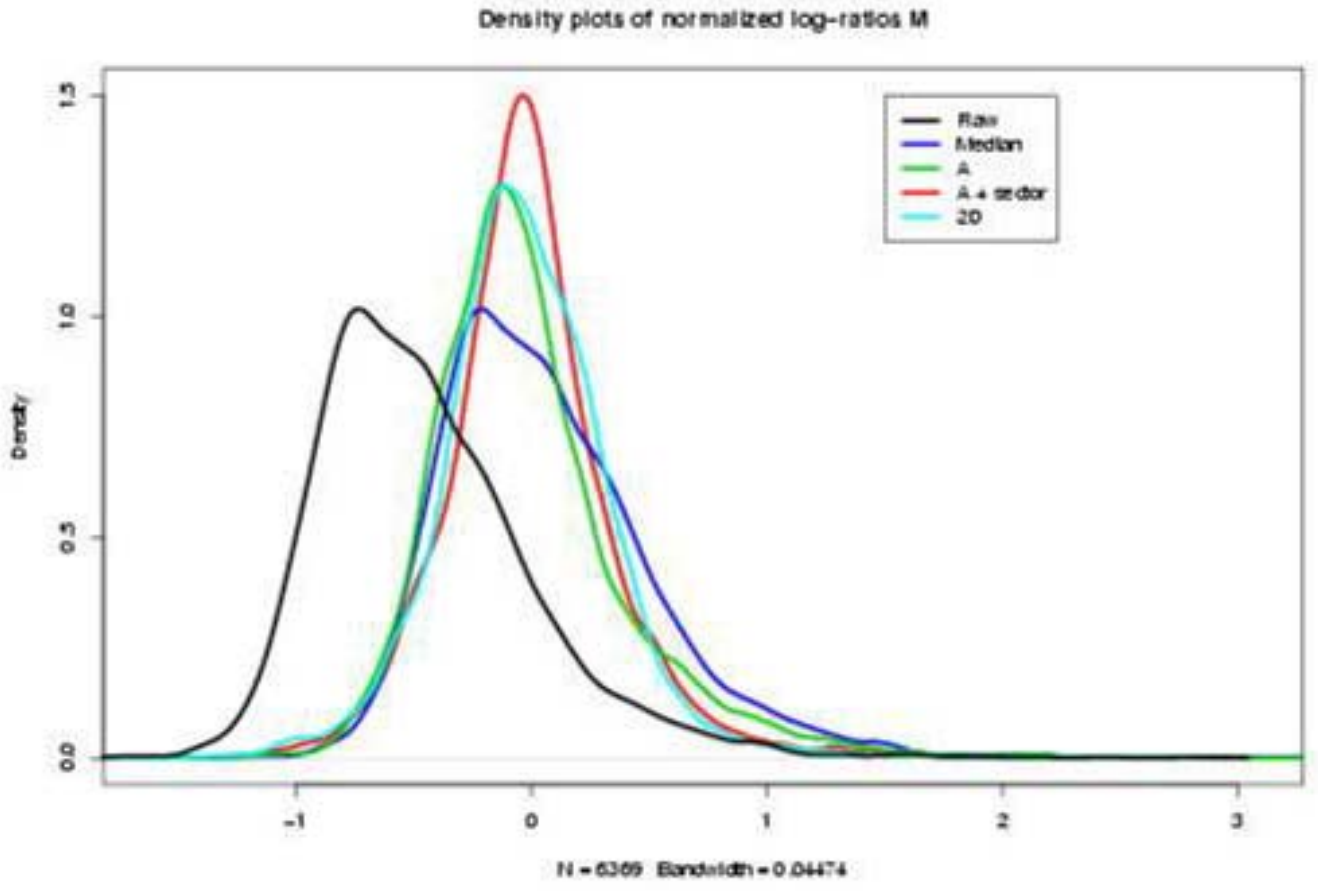


A+sector-dependent

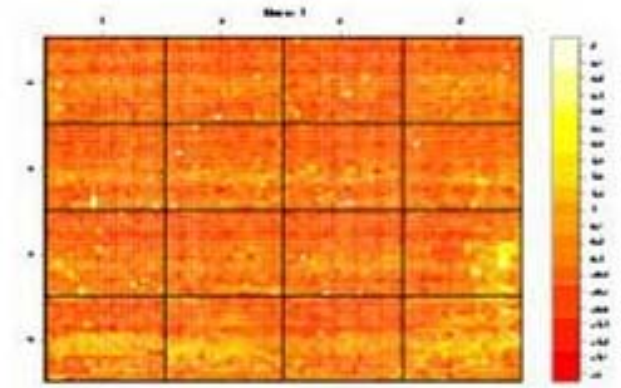


2D

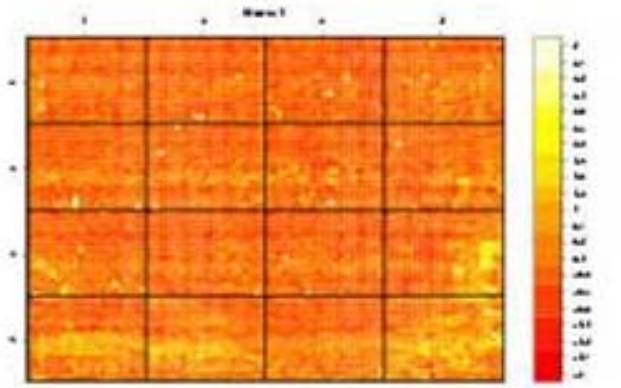
Post-normalization densities



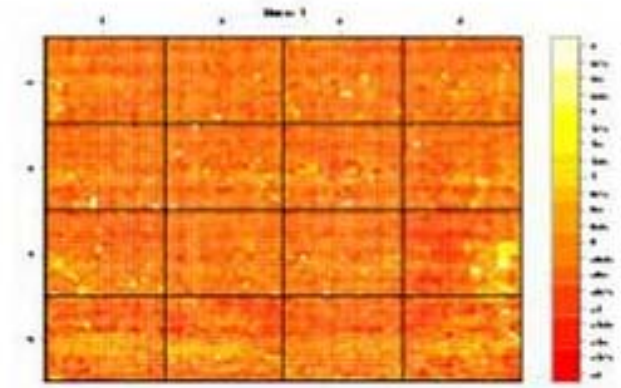
Post-normalization images



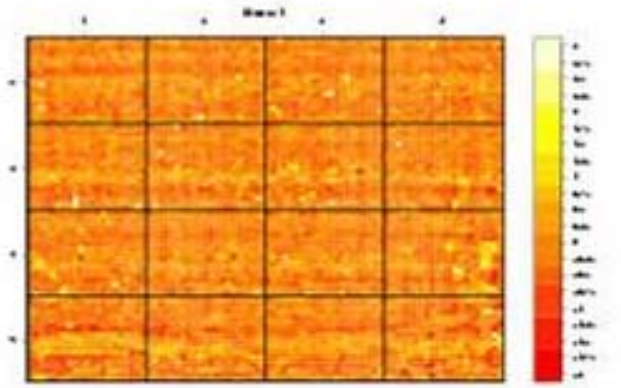
Median



A-dependent

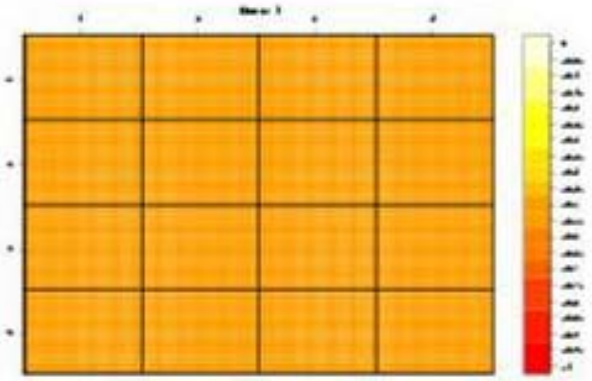


A+sector-dependent

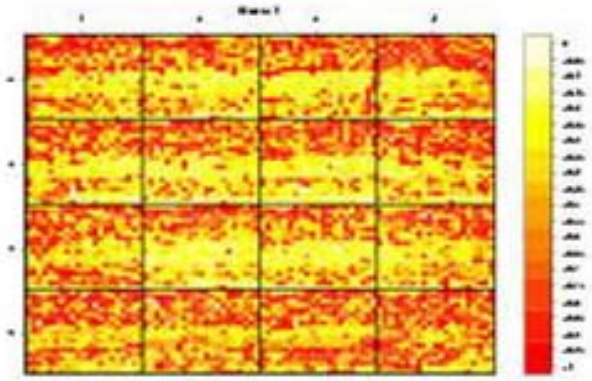


2D

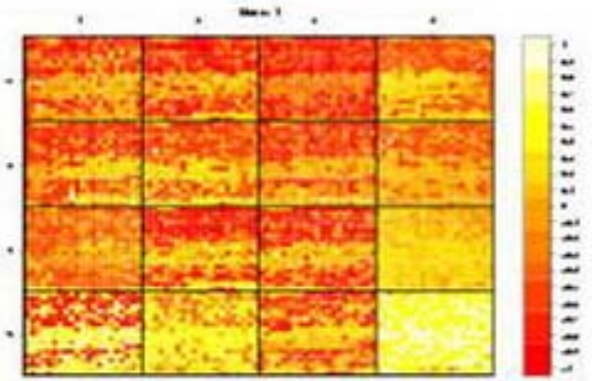
Images of normalization functions



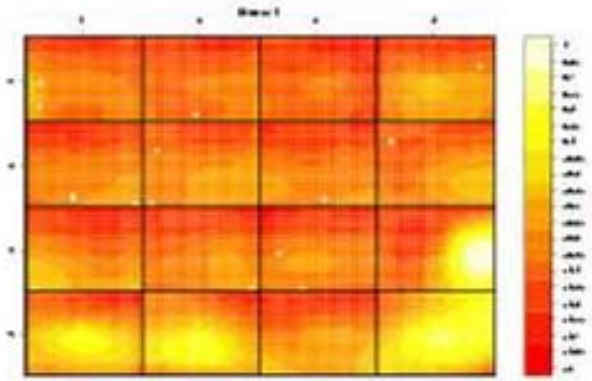
Median



A-dependent



A+sector-dependent



2D

Scale normalization

- The log-ratios M from different sectors or plates 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 like the **median absolute deviation (MAD)**

  MAD = median | M – median(M) |.

Which genes to use?

- All spots on the array.
- Constantly expressed genes (housekeeping).
- Controls
  - Spiked controls (e.g. plant genes);
  - Genomic DNA titration series;
  - Microarray sample pool (MSP).
- Rank invariant set.

Follow-up dye-swap experiment

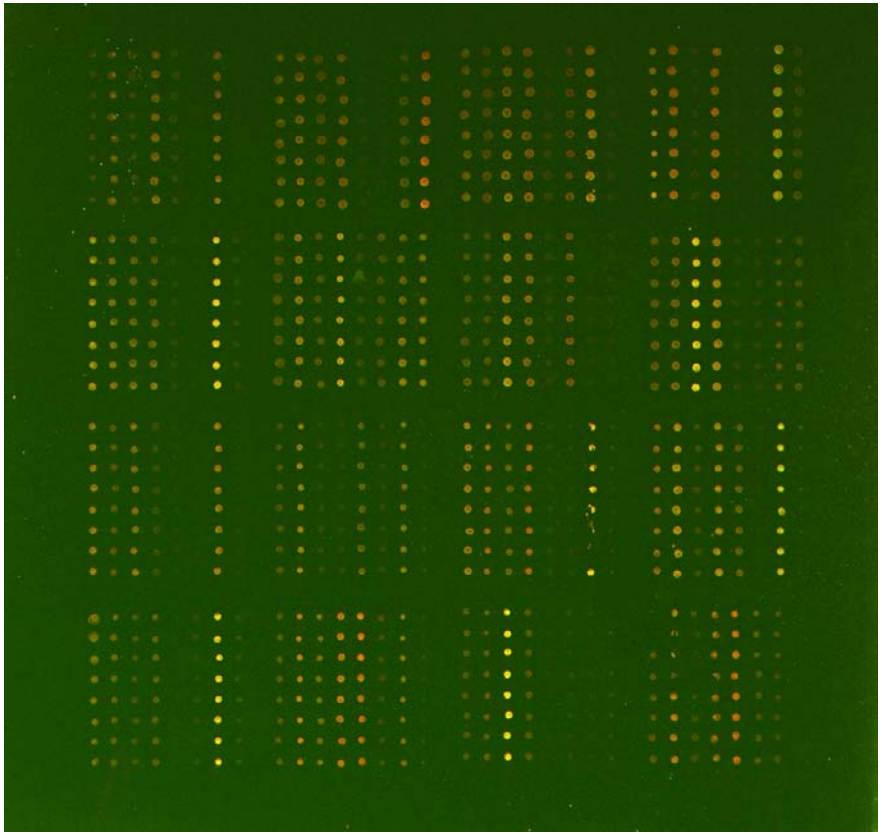
- Probes
  - 50 distinct clones with largest absolute t-statistics from apo AI experiment.
  - 72 other clones.
- Spot each clone 8 times .

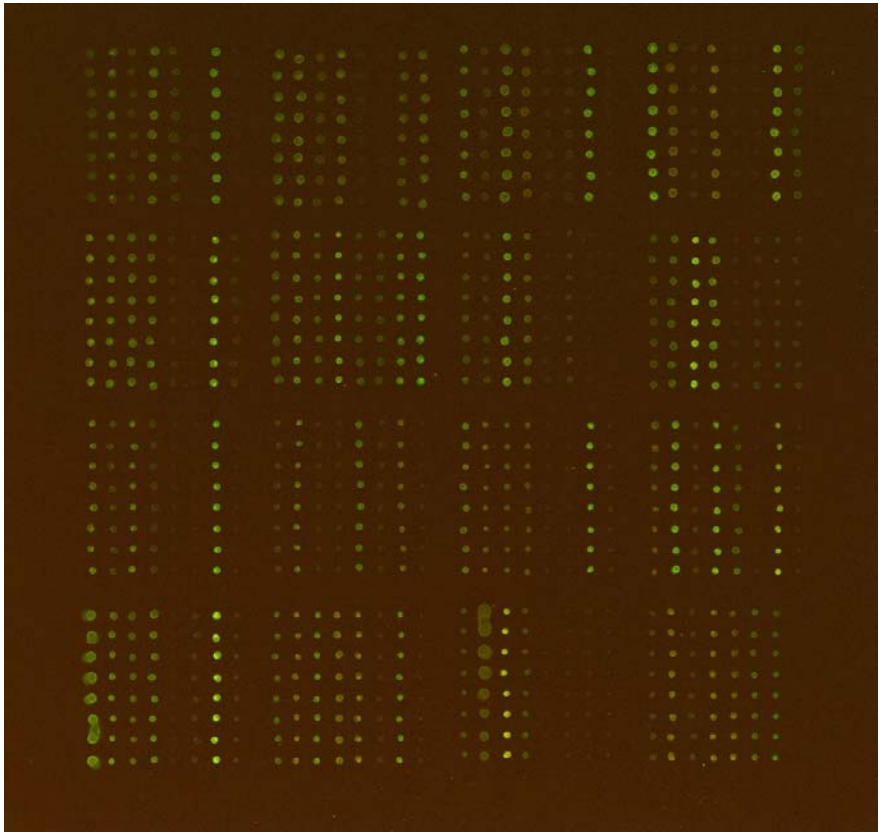
- Two hybridizations with dye-swap:

  Slide 1: trt $\rightarrow$ red, ctl $\rightarrow$ green.

  Slide 2: trt $\rightarrow$ green, ctl $\rightarrow$ 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:

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

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

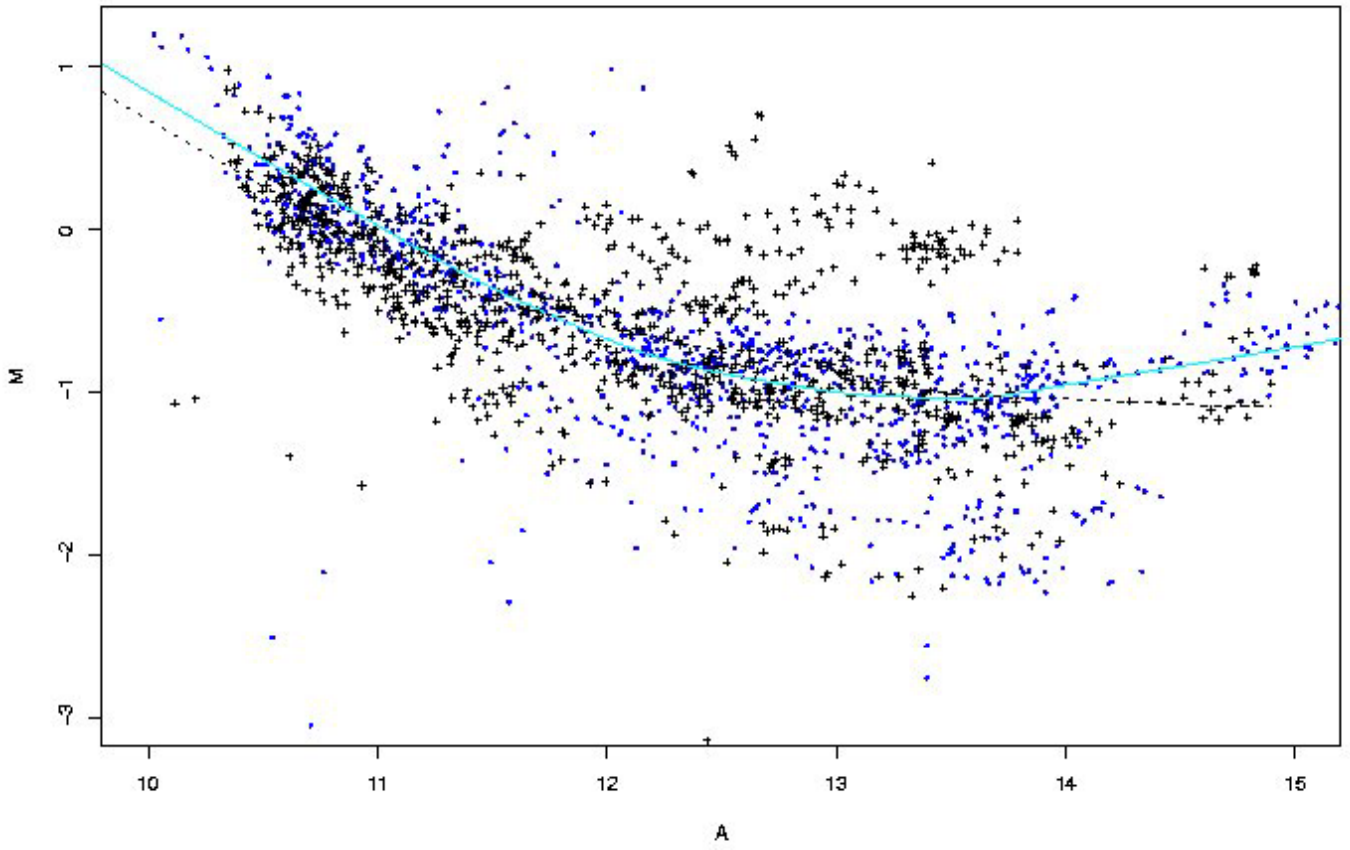
$$\approx [\log_2 (RG'/GR')] / 2$$

provided $l = l'$.

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

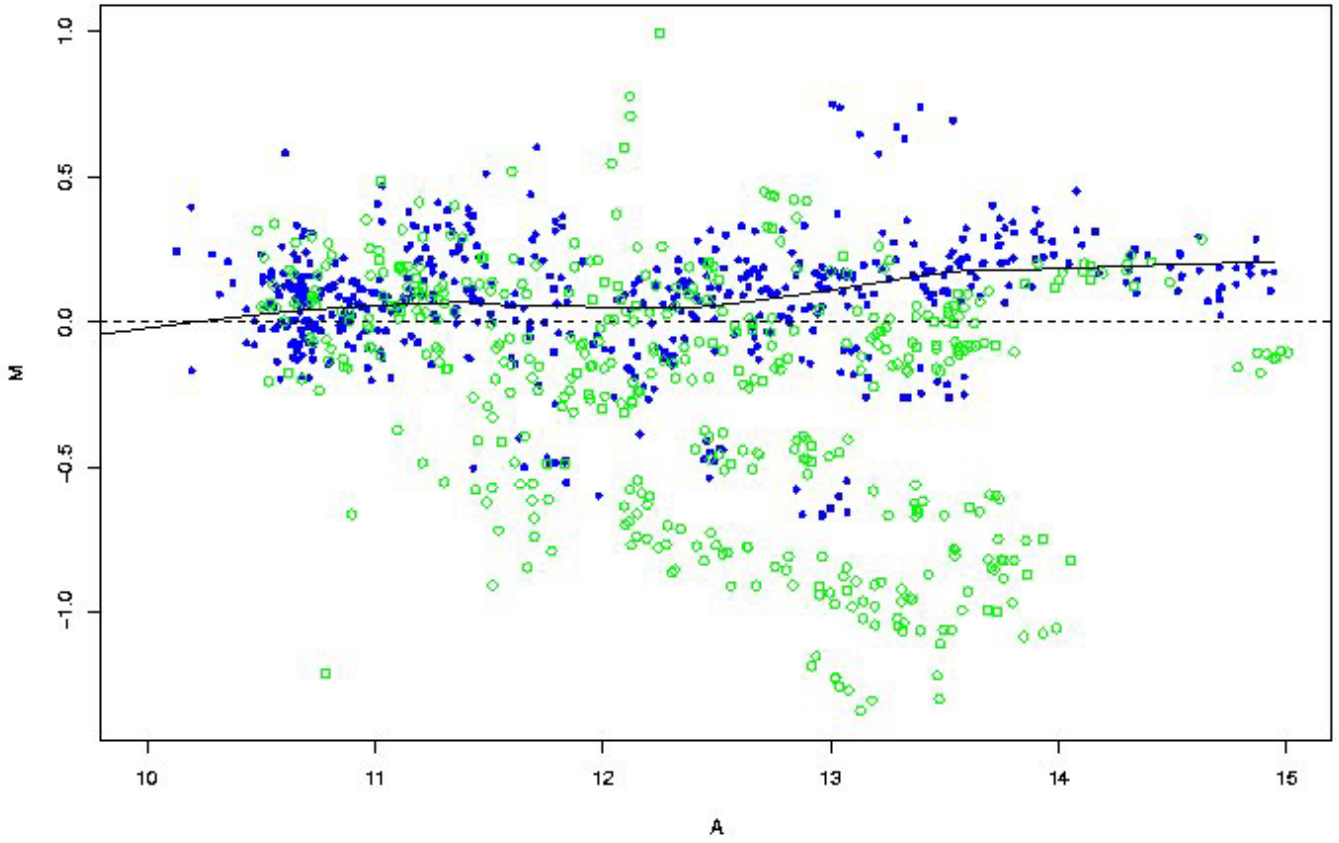
Checking the assumption

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.

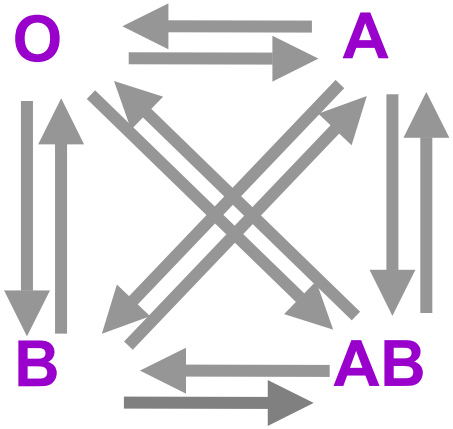
Case 2. Many genes expected to change.

- Paired-slides: Self-normalization.
- Use of controls or known information.

MarrayNorm Bioconductor package

- Class definitions for microarray data;
- Functions for diagnostic plots;
- Functions for normalization.

Experimental design



Combining data across slides

Data on *G* genes for *n* hybridizations

→ *G* x *n* genes-by-arrays data matrix

Arrays

Genes		Array1	Array2	Array3	Array4	Array5	...
	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	...
	...	...	...	...	...	...	

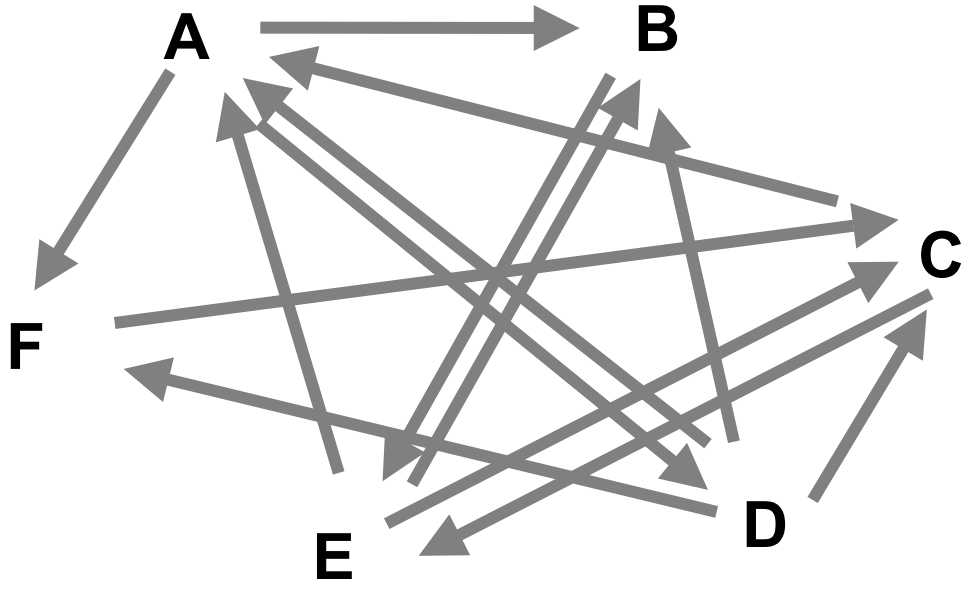
M = $\log_2$(Red intensity / Green intensity)

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?

Linear models



Experimental design

Proper experimental design is needed to ensure that questions of interest ***can*** be answered and that this can be done **accurately**, given experimental constraints, such as cost of reagents and availability of mRNA.

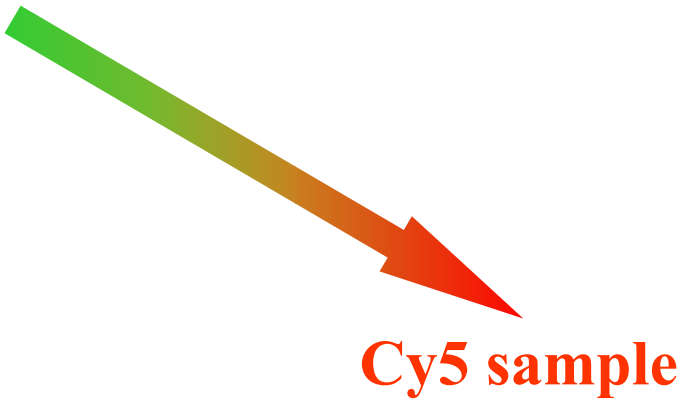
Experimental design

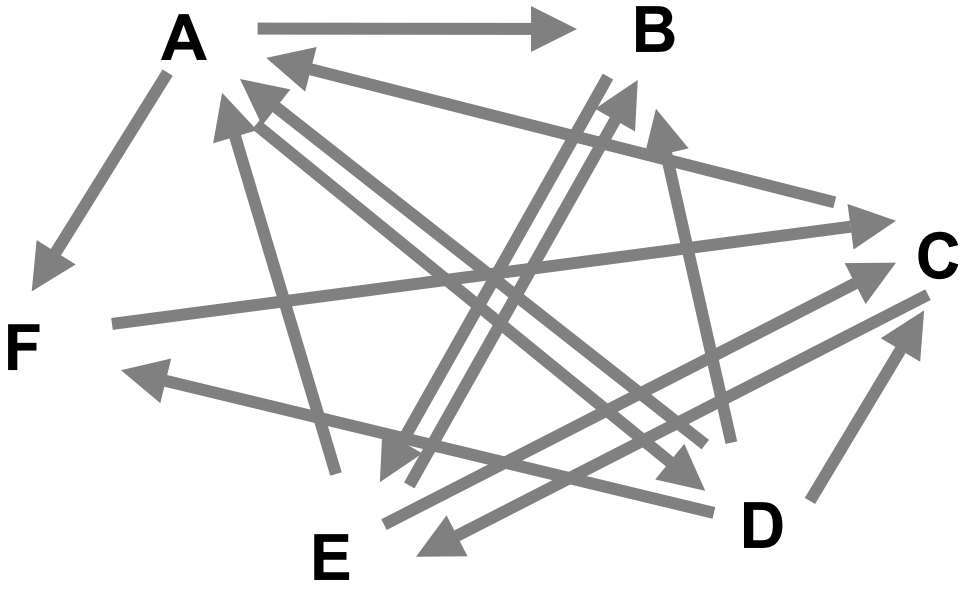
- Design of the array itself
  - which cDNA probe sequences to print;
  - whether to use replicated probes;
  - whether to use control sequences;
  - how many and where these should be printed.
- Allocation of mRNA samples to the slides
  - pairing of mRNA samples for hybridization;
  - dye assignments;
  - type and number of replicates.

Graphical representation

Multi-digraph

- ***Vertices*:** mRNA samples;
- ***Edges:*** hybridization;
- ***Direction:*** dye assignment.





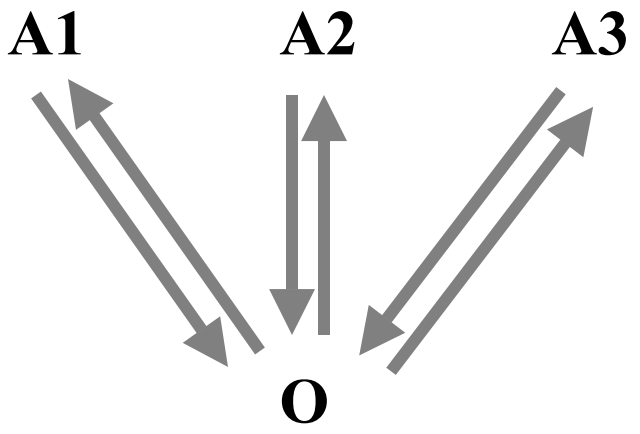
A design for 6 types of mRNA samples

Graphical representation

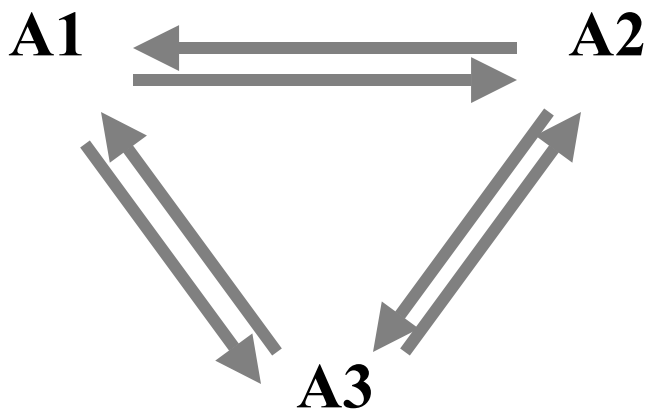
- The structure of the graph determines which effects can be estimated and the precision of the estimates.
  - Two mRNA samples can be compared only if there is a **path** joining the corresponding two vertices.
  - The precision of the estimated contrast then depends on the **number of paths** joining the two vertices and is inversely related to the **length of the paths**.
- Direct comparisons **within slides** yield more precise estimates than indirect ones between slides.

Comparing K treatments

(i) Common reference design



(ii) All-pairs design



Question: Which design gives the most precise estimates of the contrasts A1-A2, A1-A3, and A2-A3?

Comparing K treatments

- **Answer:** The all-pairs design is best, because comparisons are done **within slides**.

  For the same precision, the common reference design requires three times as many hybridizations as the all-pairs design.

- In general, for K treatments

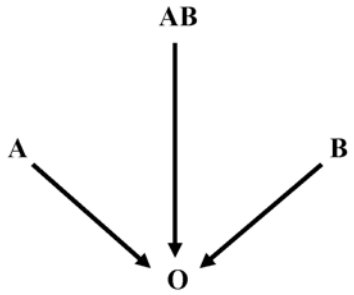
  Relative efficiency

  $= 2K/(K-1) = 4,\ 3,\ 8/3, \ldots \rightarrow 2.$

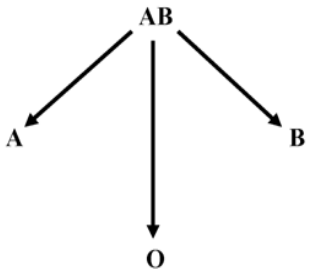
  For the same precision, the common reference design requires 2K/(K-1) times as many hybridizations as the all-pairs design.

2 x 2 factorial experiment

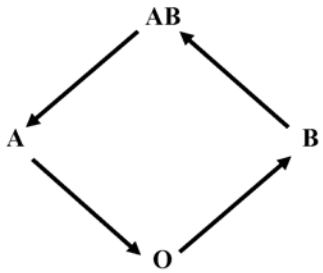
two factors, two levels each



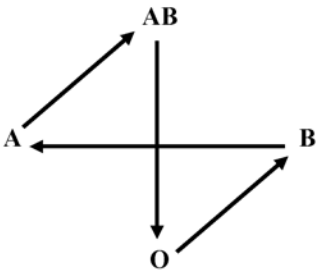
(1) Common ref.



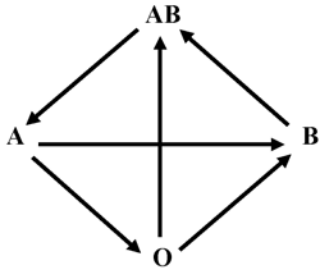
(2) Common ref.



(3) Connected



(4) Connected



(5) All-pairs

Scaled variances of estimated effects

	(1)	(2)	(3)	(4)	(5)
Main effect A	**1**	**2**	**1**	**4/3**	**1**
Main effect B	**1**	**2**	**1**	**1**	**1**
Interaction AB	**3**	**3**	**4/3**	**8/3**	**2**
Contrast A-B	**2**	**2**	**4/3**	**1**	**1**

Experimental design

- In addition to experimental constraints, design decisions should be guided by the knowledge of which effects are of greater interest to the investigator.

  E.g. which main effects, which interactions.
- The experimenter should thus decide on the comparisons for which he wants most precision and these should be made **within slides** to the extent possible.

Issues in experimental design

- Replication.
- Type of replication:
  - *within* or *between* slide replicates;
  - *biological* or *technical* replicates.
- Sample size and power calculations.
- Dye assignments.
- Combining data across slides and sets of experiments:

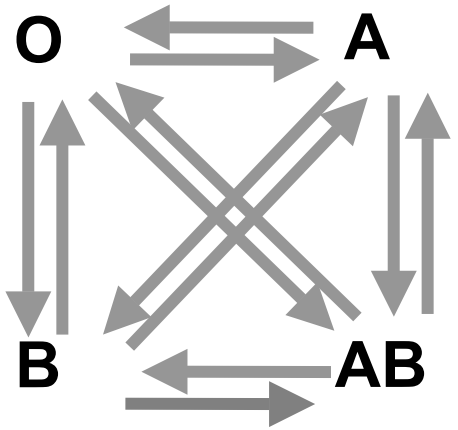
  regression analysis … more later.

2 x 2 factorial experiment

Study the **joint** effect of two treatments (e.g. drugs),
A and B, say, on the gene expression response of tumor cells.

There are four possible treatment combinations

AB: both treatments are administered;
A : only treatment A is administered;
B : only treatment B is administered;
O : cells are untreated.



2 x 2 factorial experiment

For each gene, consider a linear model for the joint effect of treatments A and B on the expression response:

$$\mu_{AB} = \mu + \alpha + \beta + \gamma$$

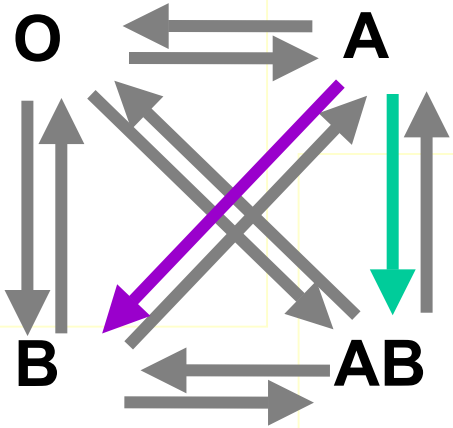
$$\mu_{A} = \mu + \alpha$$

$$\mu_{B} = \mu + \beta$$

$$\mu_{O} = \mu$$

μ: baseline effect;
α: treatment A main effect;
β: treatment B main effect;
γ: interaction between treatments A and B.

2 x 2 factorial experiment



Log-ratio M for hybridization

A → **AB**

estimates

$$\mu_{AB} - \mu_A = \beta + \gamma$$

Log-ratio M for hybridization

A → **B**

estimates

$$\mu_B - \mu_A = \beta - \alpha$$

etc.

Regression analysis

- For parameters $\theta = (\alpha, \beta, \gamma)$, define a design matrix X so that $E(M)=X\theta$.
- For each gene, compute least squares estimates of θ.

$$E\begin{pmatrix} M_{11} \\ M_{12} \\ M_{21} \\ M_{22} \\ M_{31} \\ M_{32} \\ M_{41} \\ M_{42} \\ M_{51} \\ M_{52} \\ M_{61} \\ M_{62} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 0 \\ 1 & 1 & 1 \\ -1 & -1 & -1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \\ 1 & 0 & 1 \\ -1 & 0 & -1 \\ -1 & 1 & 0 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$$

$$\hat{\theta} = (X'X)^{-1}X'M$$

Regression analysis

- Combine data across slides for complex designs (can “link” different sets of hybridizations).
- Obtain unbiased and efficient estimates of the effects of interest (BLUE).
- Hypothesis testing.
- Use estimated effects in pattern discovery and recognition.
- Extensions:
  - generalized linear models,
  - robust weighted regression, etc.

Acknowledgments

- **Brown Lab,** Biochemistry, Stanford;
- **Terry Speed**, **Yee Hwa (Jean) Yang**, Statistics, UC Berkeley;
- **Matt Callow**, LBNL;
- **Ngai Lab**, MCB, UC Berkeley.

References

Tech. reports may be downloaded from

http://www.stat.berkeley.edu/~sandrine