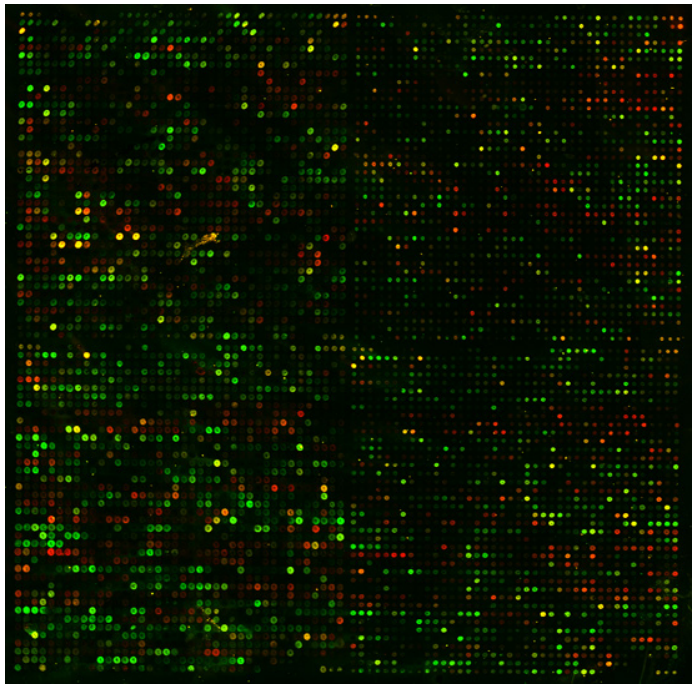
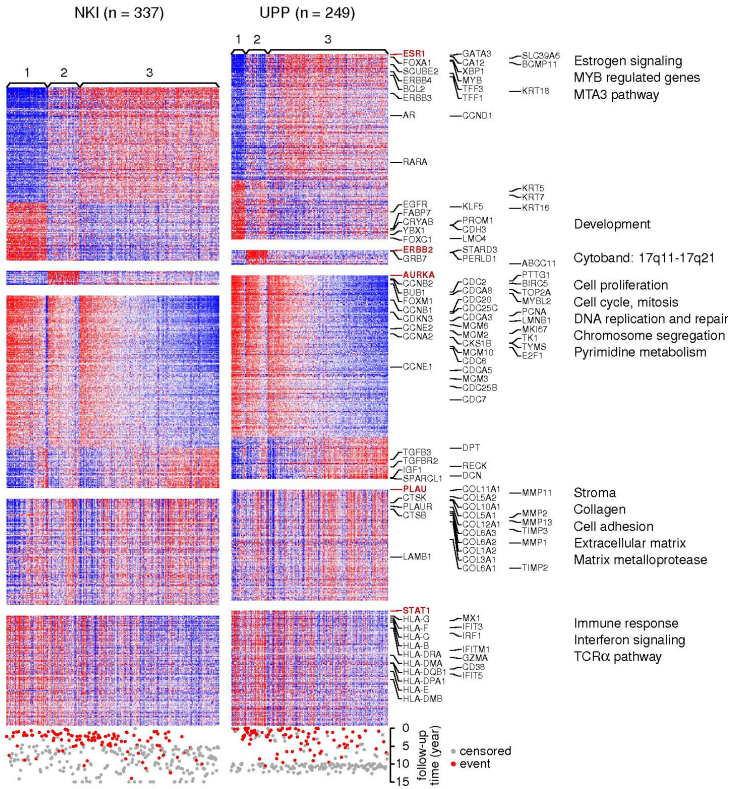
Statistics for Genomic Data Analysis

Miscellaneous topics; Review





http://moodle.epfl.ch/course/view.php?id=15271

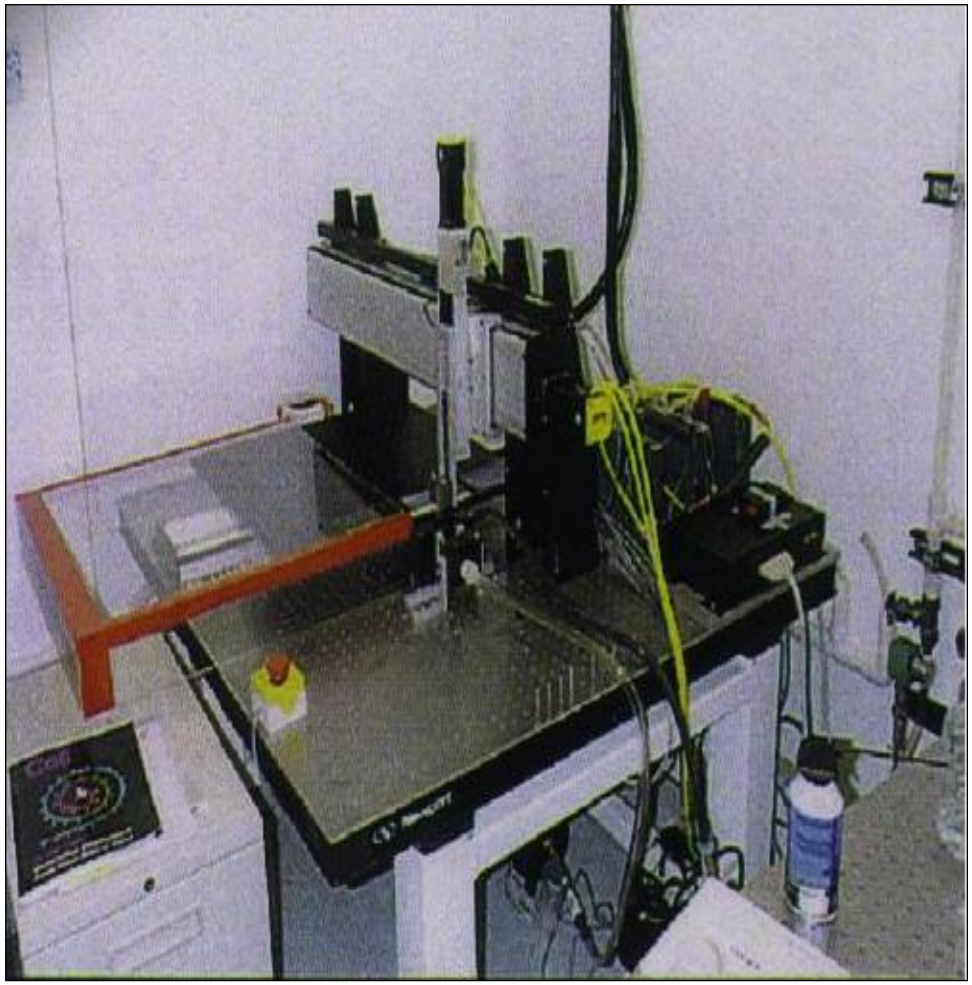
A few additional topics

- 2-channel microarrays
- Combining data
- Reproducible research/forensic bioinformatics
- Review for exam

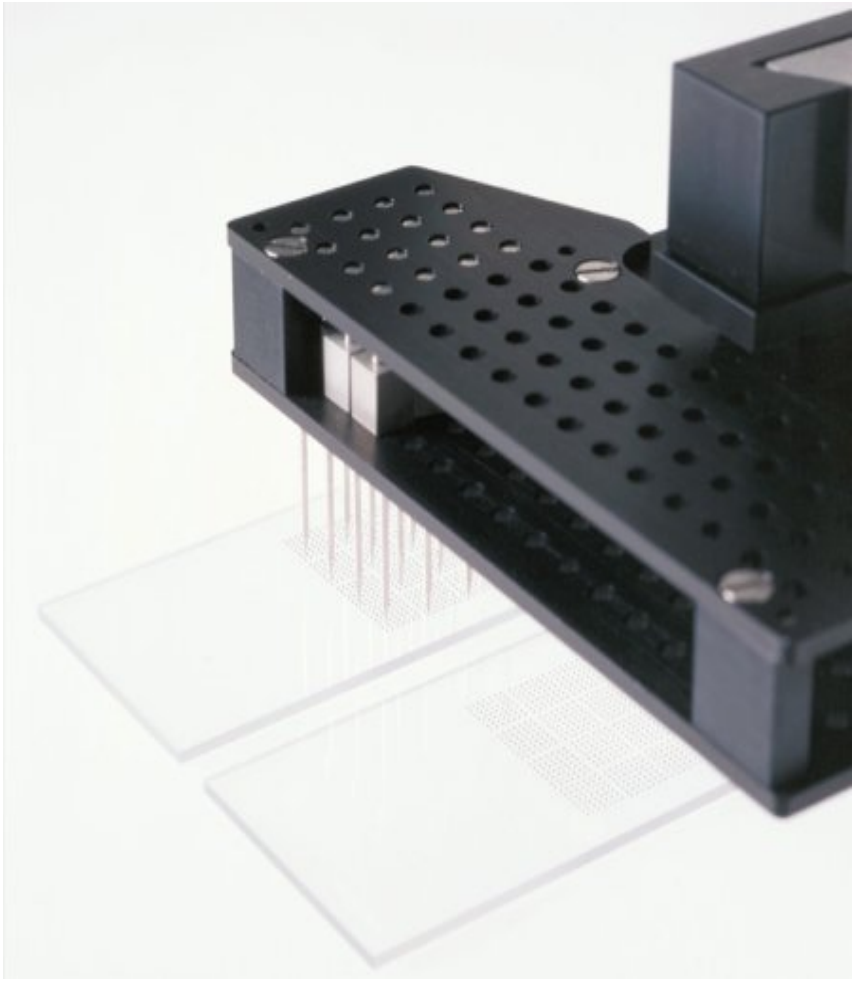
Producing a cDNA Microarray

- Probes are cDNA fragments, usually amplified by PCR
- Probes are deposited on a *solid support*, either positively charged nylon or coated glass slide
- Samples (normally poly(A)+ RNA) are labeled using *fluorescent dyes*
- (At least) *two samples* are hybridized to chip
- Fluorescence at *different wavelengths* measured by a scanner

Building the array

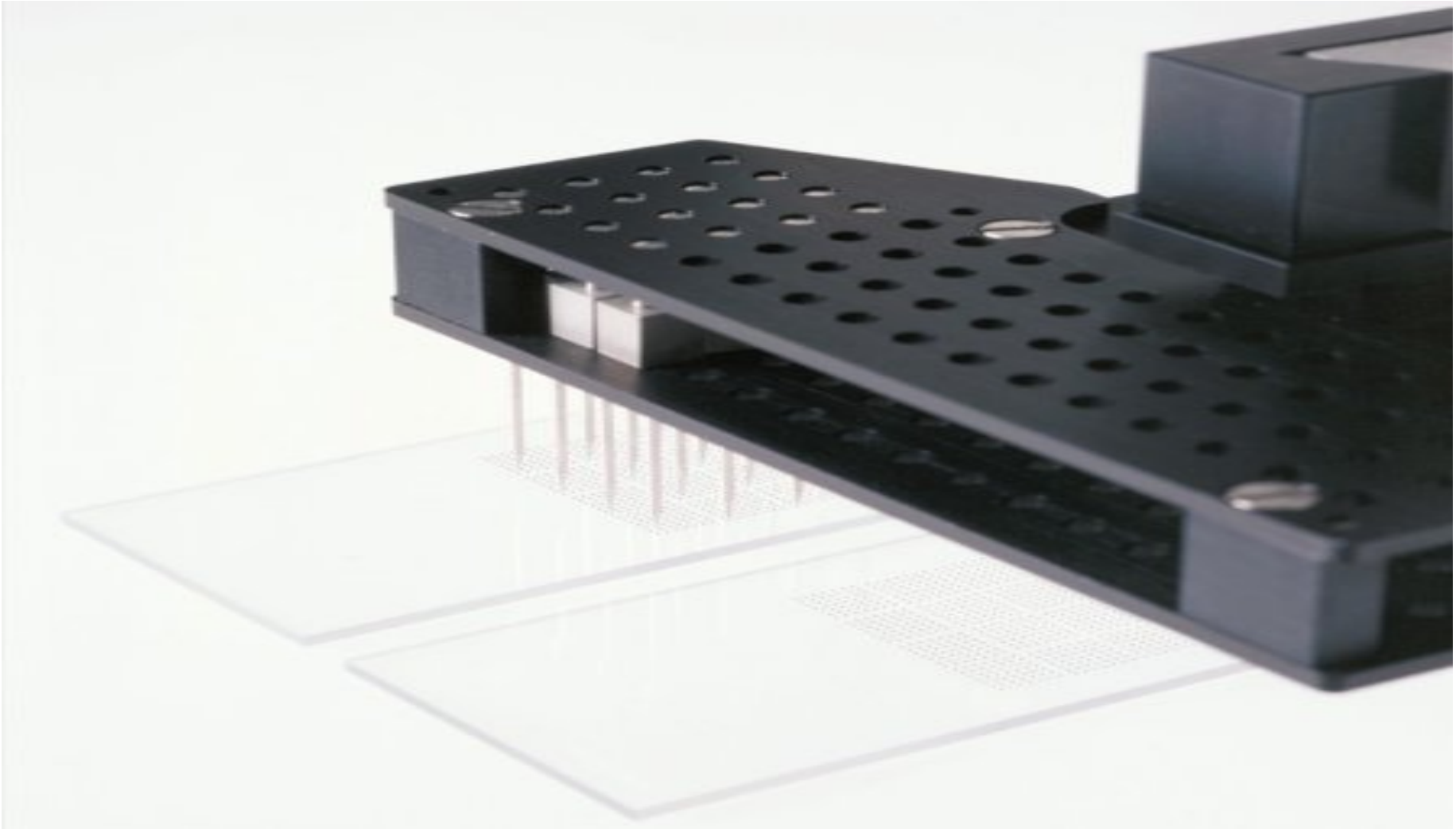


Ngai Lab arrayer , UC Berkeley

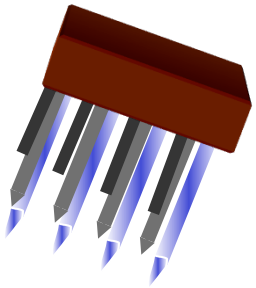


Print-tip head

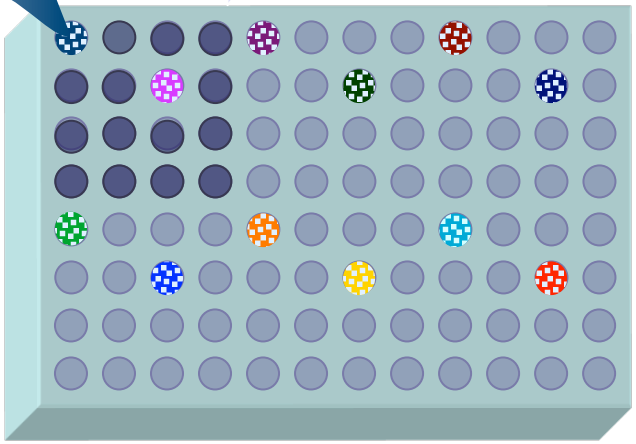
Print-tip head



How to make a cDNA array



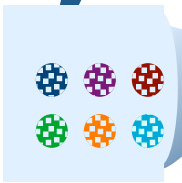
Pins collect cDNA from wells

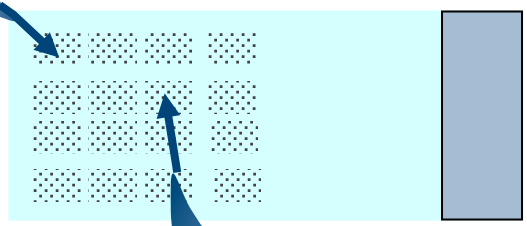


384 well plate --

Contains cDNA probes

Print-tip group 1





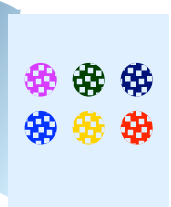
Glass Slide

Array of bound cDNA probes

cDNA clones

Spotted in duplicate

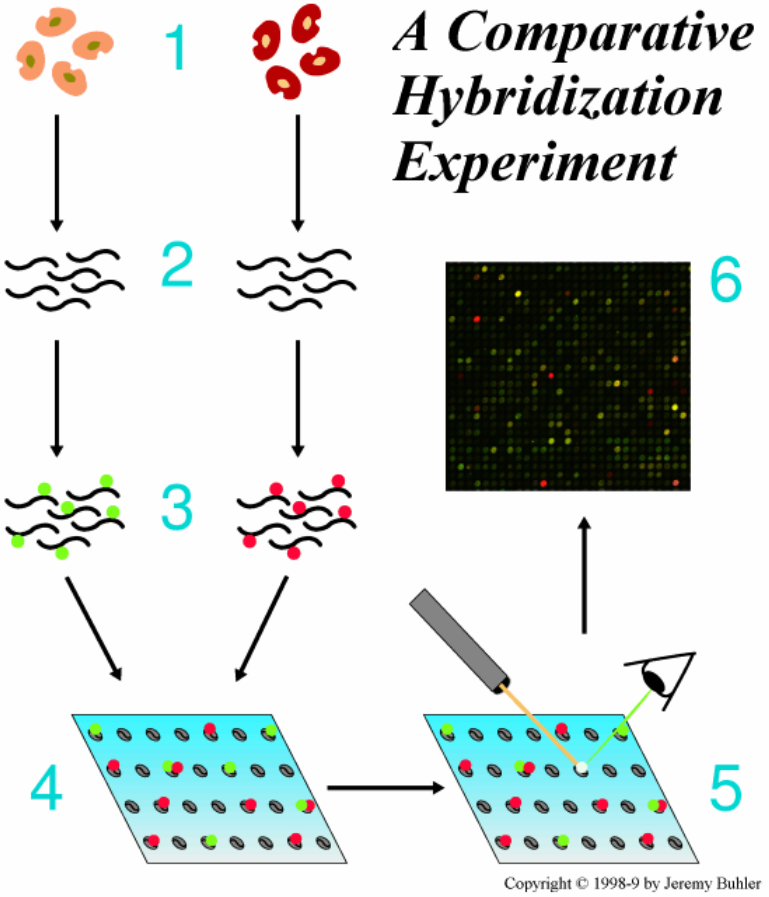
4x4 blocks = 16 print-tip groups



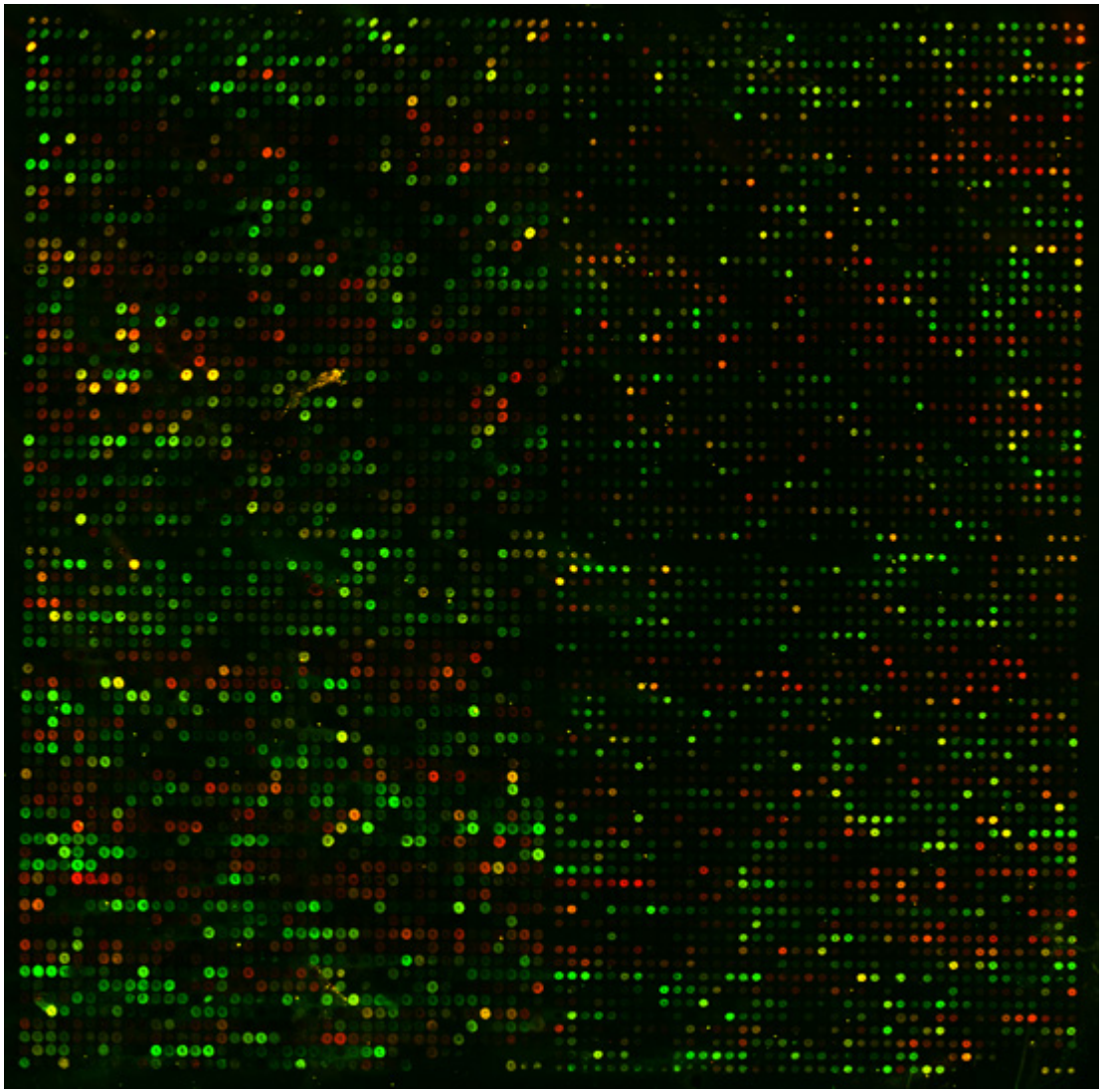
Print-tip group 7

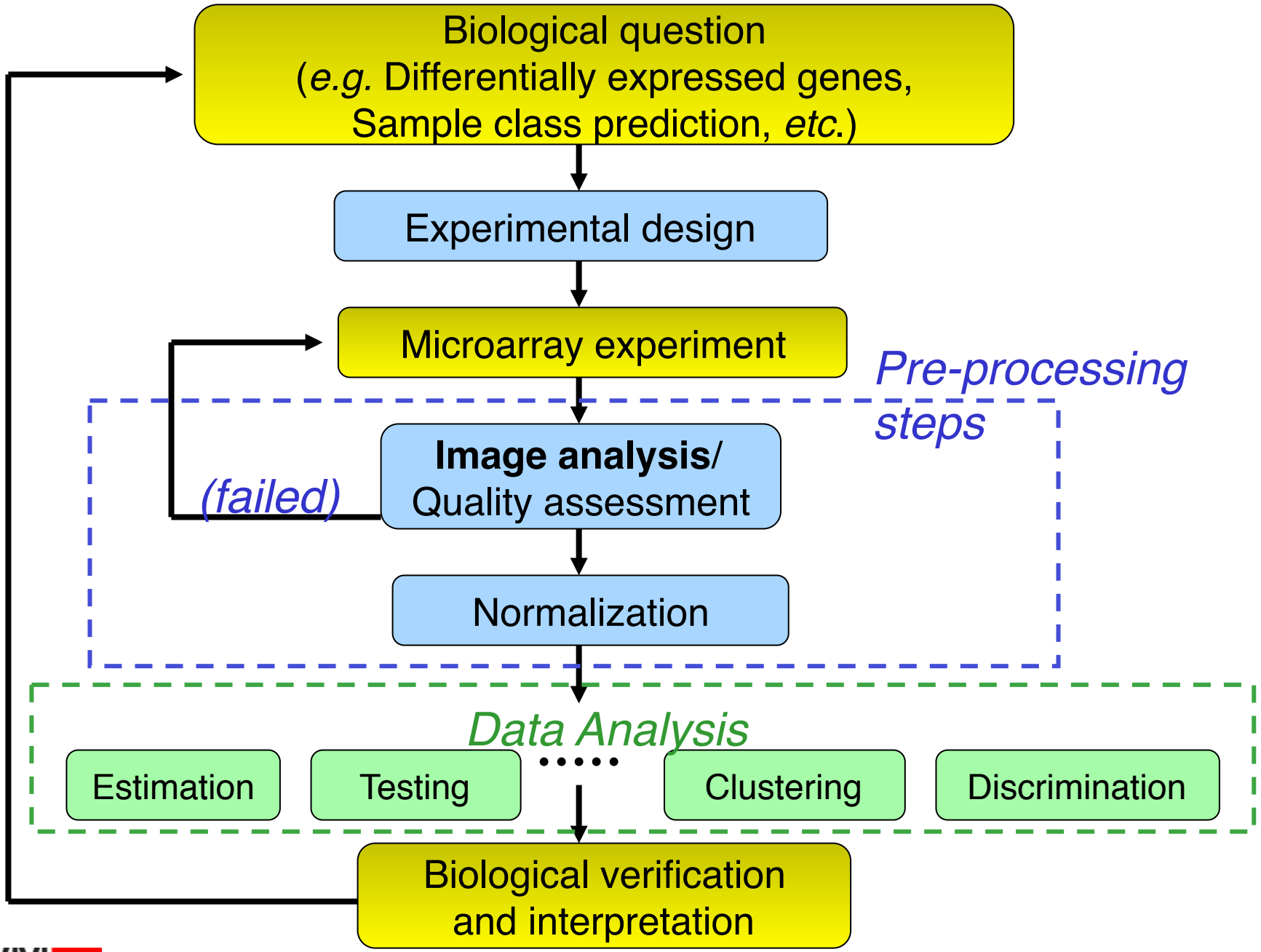
Standard Protocol for Comparative Hybridization

http://www.cs.wustl.edu/~jbuhler/research/array/

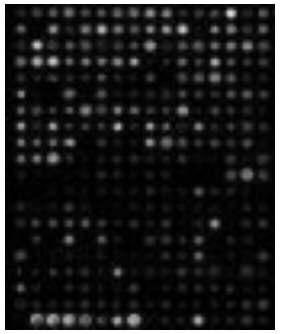


Yeast Genome on a Chip




Biological question
(e.g. Differentially expressed genes,
Sample class prediction, etc.)
Experimental design
Microarray experiment
Pre-processing steps
(failed)
Image analysis/
Quality assessment
Normalization
Data Analysis
Estimation
Testing
Clustering
Discrimination
Biological verification
and interpretation

Images from Scanner

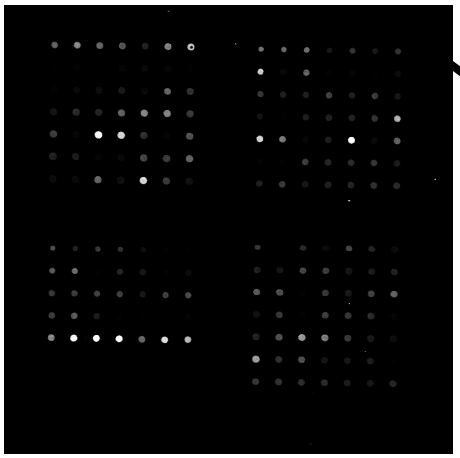


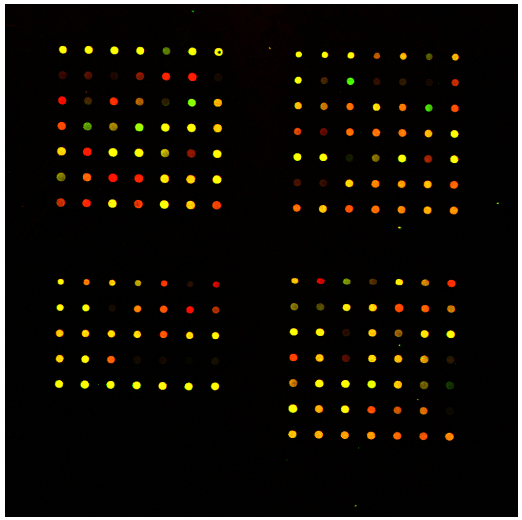
- Resolution
  - standard 10μm [currently, best ~ 5μm]
  - 100μm spot on chip = 10 pixels in diameter
- Image format
  - TIFF (tagged image file format) 16 bit (65,536 levels of gray) – also other formats
  - 1cm x 1cm image at 16 bit = 2Mb
- The two 16-bit images (Cy3, Cy5 for a typical dual channel microarray) are compressed into 8-bit images

Images : examples

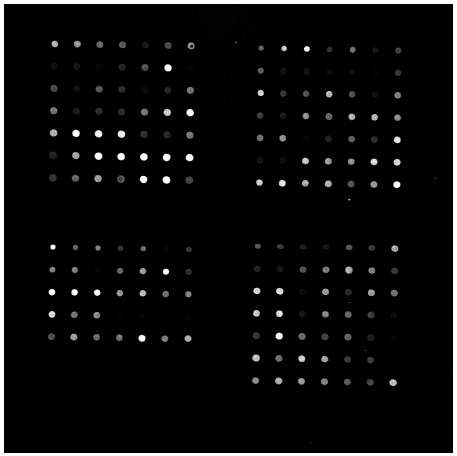
Pseudo-color overlay

Cy3





Cy5

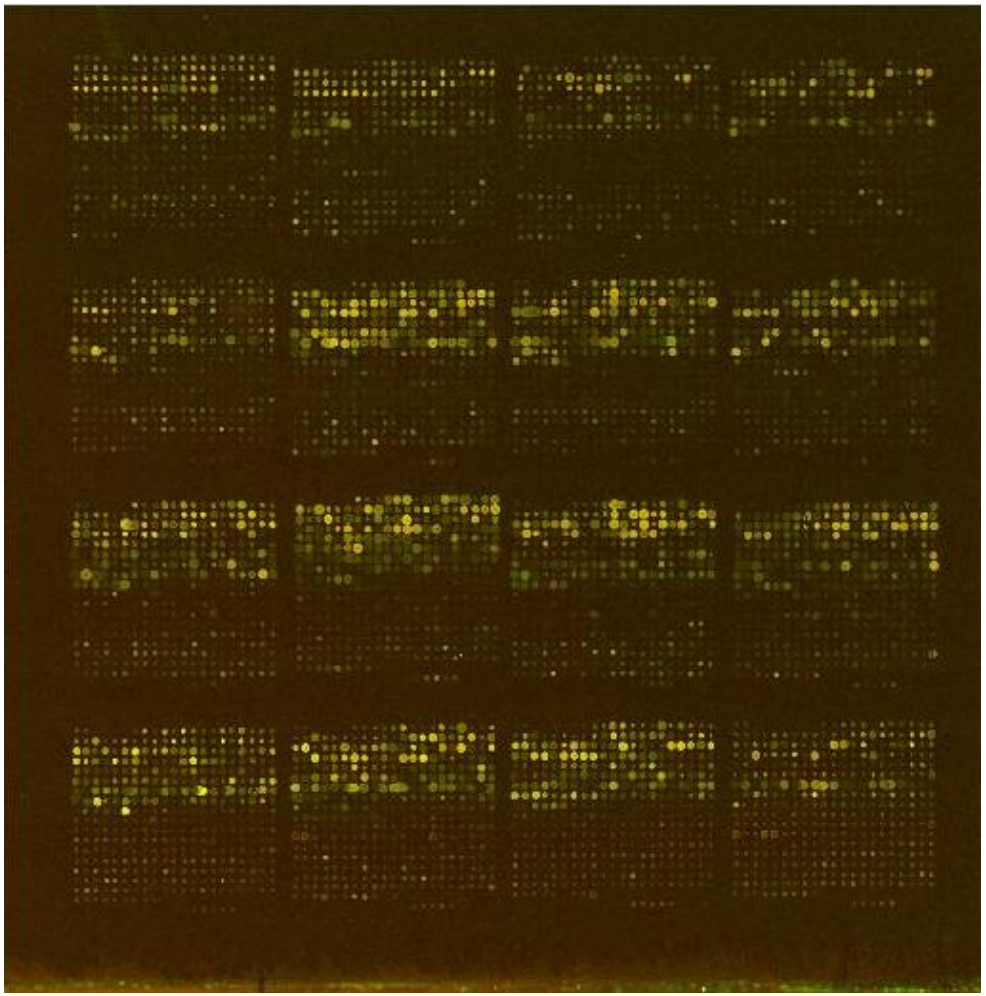


Spot color	*Signal strength*	*Gene expression*
yellow	Control = Treated	unchanged
red	Control < Treated	induced
green	Control > Treated	repressed

Steps in Images Processing

- *Addressing* (or *Gridding*)
  - Assigning coordinates to each spot
- *Segmentation*
  - *Classification of pixels* as either foreground (signal) or background
- *Information Extraction*
  - Foreground fluorescence intensity pairs (**R**,**G**)
  - Background intensities
  - Quality measures

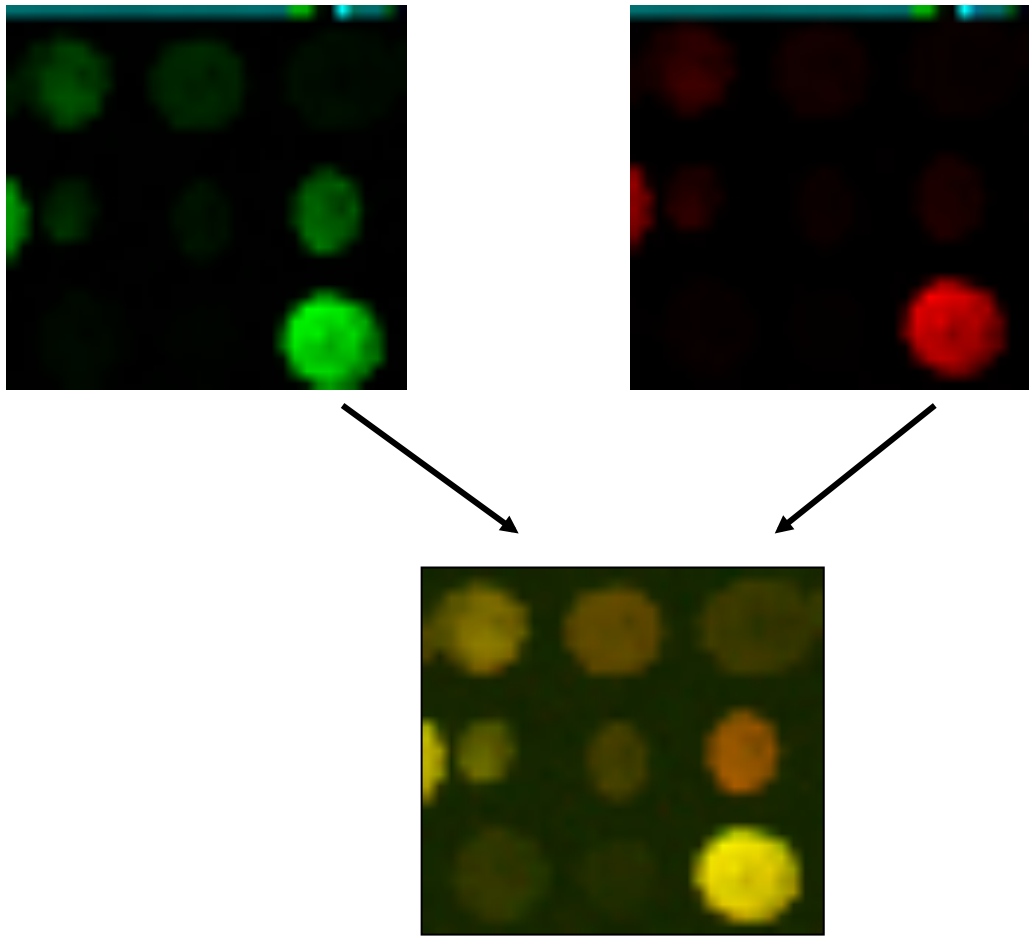
Addressing



- This is the process of *assigning coordinates* to each spot
- *Automating* this process permits high-throughput analysis

4 by 4 grids
19 by 21 spots per grid

Addressing — Registration



Steps in Images Processing

- *Addressing* (or *Gridding*)
  - Assigning coordinates to each spot
- *Segmentation*
  - *Classification of pixels* as either foreground (signal) or background
- *Information Extraction*
  - Foreground fluorescence intensity pairs (*R*,*G*)
  - Background intensities
  - Quality measures

Segmentation methods in some programs

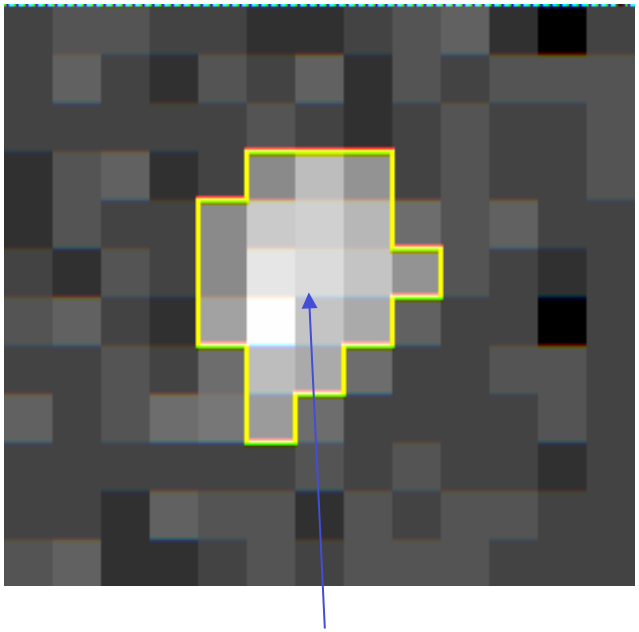
Fixed circle	ScanAlyze, GenePix, QuantArray
Adaptive circle	GenePix, Dapple, SignalViewer (uses ellipse)
Adaptive shape	Spot, region growing and watershed
Histogram	ImaGene, QuantArray, DeArray and adaptive thresholding

Steps in Images Processing

- *Addressing* (or *Gridding*)
  - Assigning coordinates to each spot
- *Segmentation*
  - *Classification of pixels* as either foreground (signal) or background
- *Information Extraction*
  - Foreground fluorescence intensity pairs (**R**,**G**)
  - Background intensities
  - Quality measures

Information Extraction

- *Spot Intensities*
  - mean of pixel intensities
  - median of pixel intensities
  - Pixel variation (e.g. IQR)
- *Background values*
  - None
  - Local
  - Constant (global)
  - Morphological opening
- *Quality Information*



Spot 'foreground' intensity

- The total amount of hybridization for a spot is proportional to the *total fluorescence* generated by the spot
- Spot intensity = sum of pixel intensities within the spot mask
- Since later calculations are based on *ratios* between Cy5 and Cy3, we compute the average* pixel value over the spot mask

  **alternative* : ratios of medians may be better than means if bright specks present

Background intensity

- The measured fluorescence intensity includes a contribution of *non-specific hybridization* and *other chemicals* on the glass
- Fluorescence from regions not occupied by DNA should be *different* from regions occupied by DNA

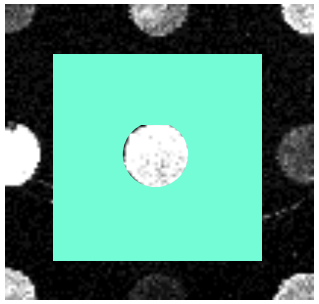
  → one solution is to use *local negative controls* (spotted DNA that should not hybridize)

BG: None

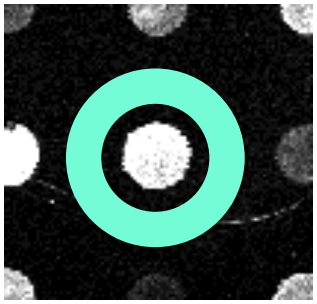
- Do not consider the background
  - Probably not accurate in many cases, but may be better than some forms of local background determination

BG: Local

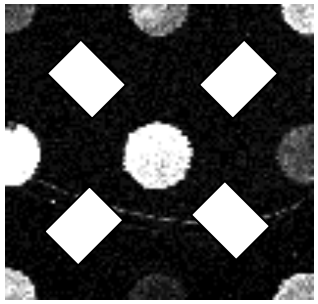
- Focus on small regions surrounding the spot mask
- Median of pixel values in this region
- Most software implements such an approach



ScanAlyze



ImaGene



Spot, GenePix

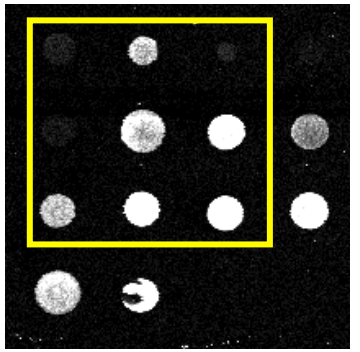
- By ignoring pixels immediately surrounding the spots, bg estimate is *less sensitive* to the performance of the segmentation procedure

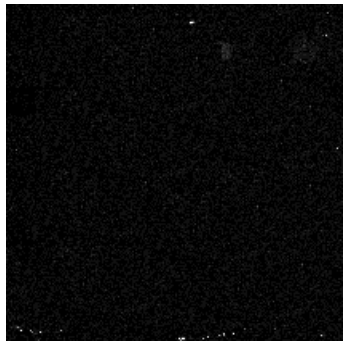
BG: Constant

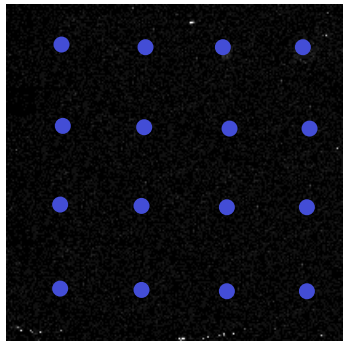
- *Global method* which subtracts a constant background for all spots
- Some evidence that the binding of fluorescent dyes to 'negative control spots' is lower than the binding to the glass slide
- → More meaningful to estimate background based on a *set of negative control spots*
  - If no negative control spots : approximation of the average background = third percentile of all the spot foreground values

BG: Morphological opening

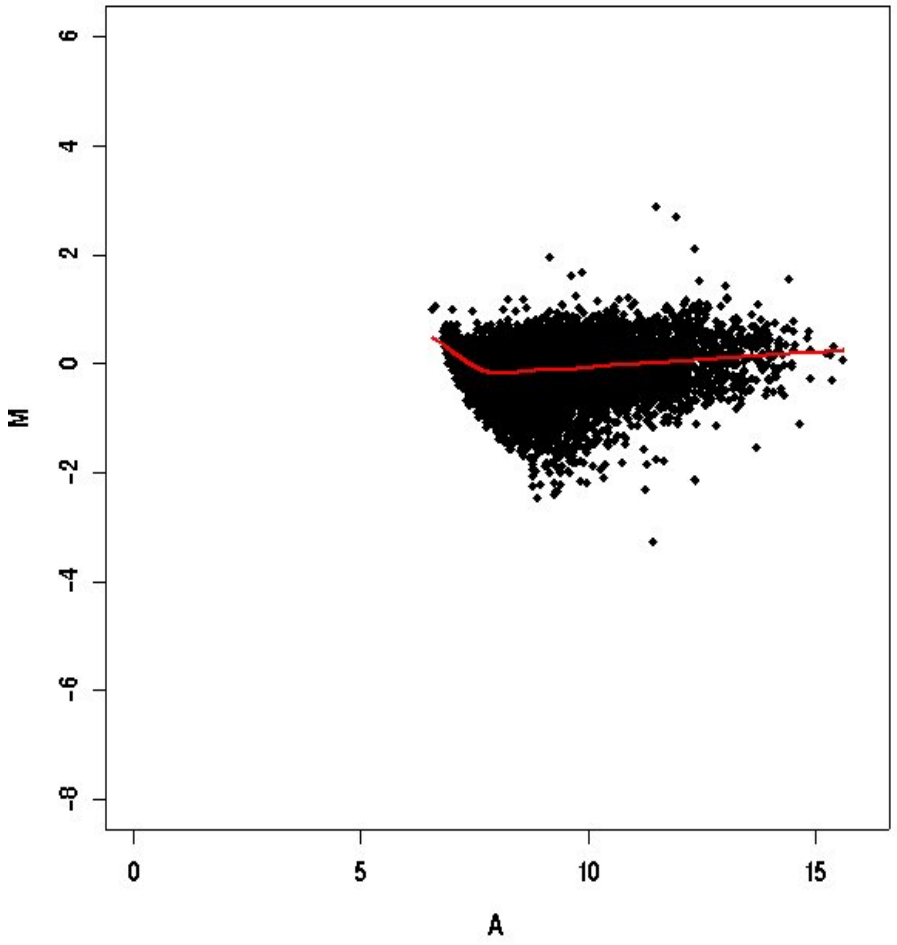
- Non-linear filtering, used in *Spot*
- Use a square structuring element with side length at least twice as large as the spot separation distance
- Compute local minimum filter, then local maximum filter
- *This removes all spots and generates an image that is an estimate of the background for the entire slide*
- For individual spots, the background is estimated by sampling this background image at the nominal center of the spot
- Lower, less variable bg estimate



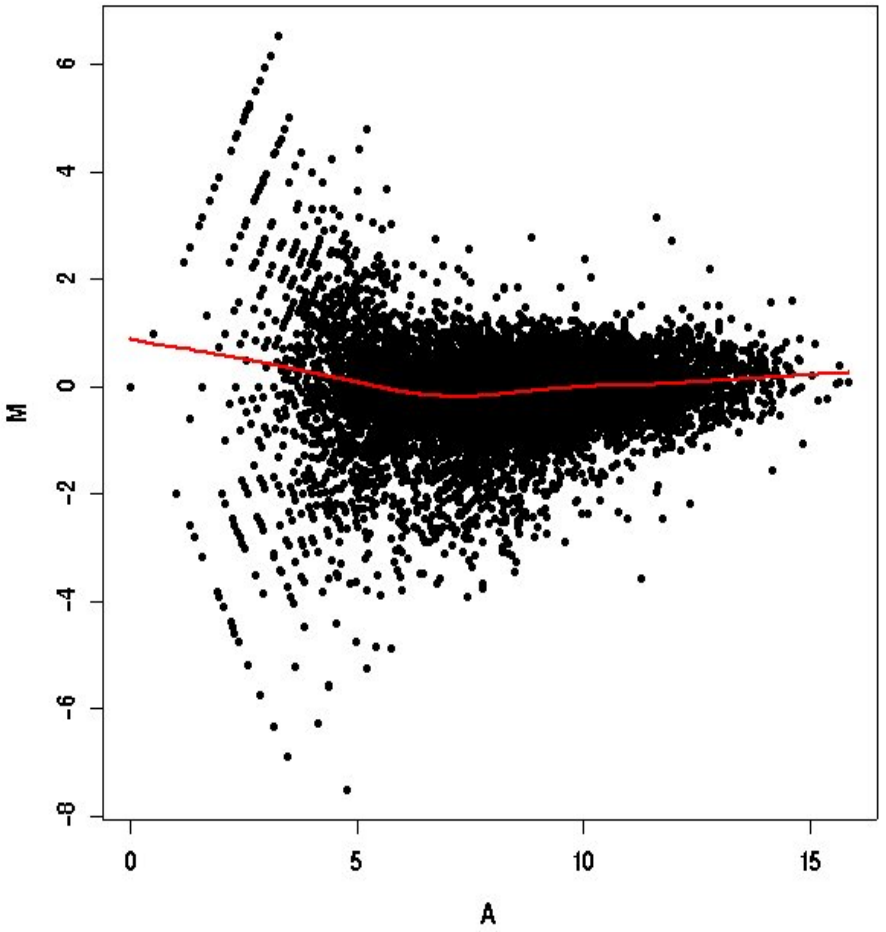




Background matters



From Spot



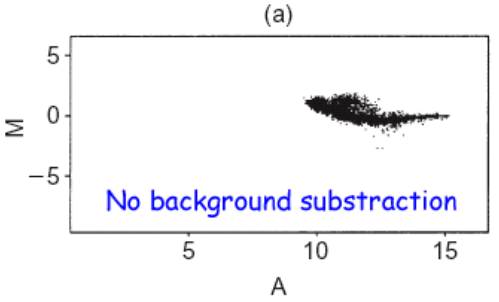
From GenePix

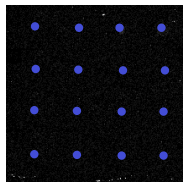
Summary

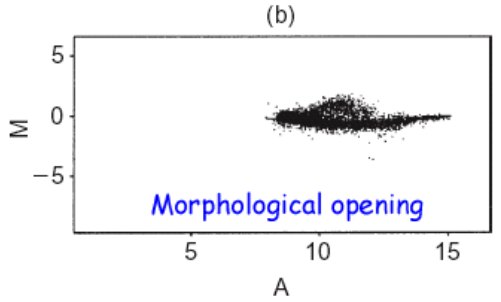
- Choice of bg correction method can have a larger impact on the log-intensity ratios than segmentation method
- Bg adjustment has *larger impact on low intensity spots*

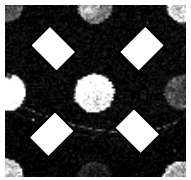
$M = \log_2 R/G$

$A = \log_2 \sqrt{(R \cdot G)}$

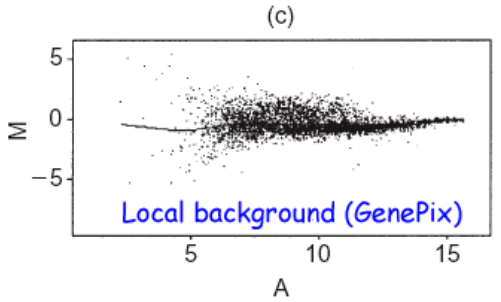


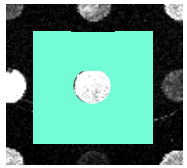




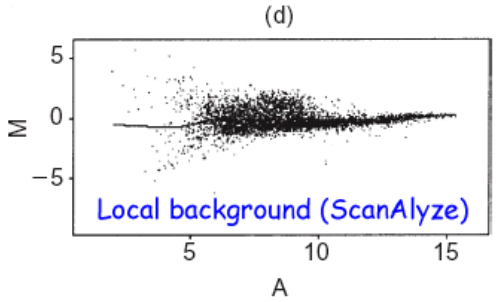


Spot, GenePix





ScanAlyze



Quantification of Expression

- For each spot on a dual channel microarray:

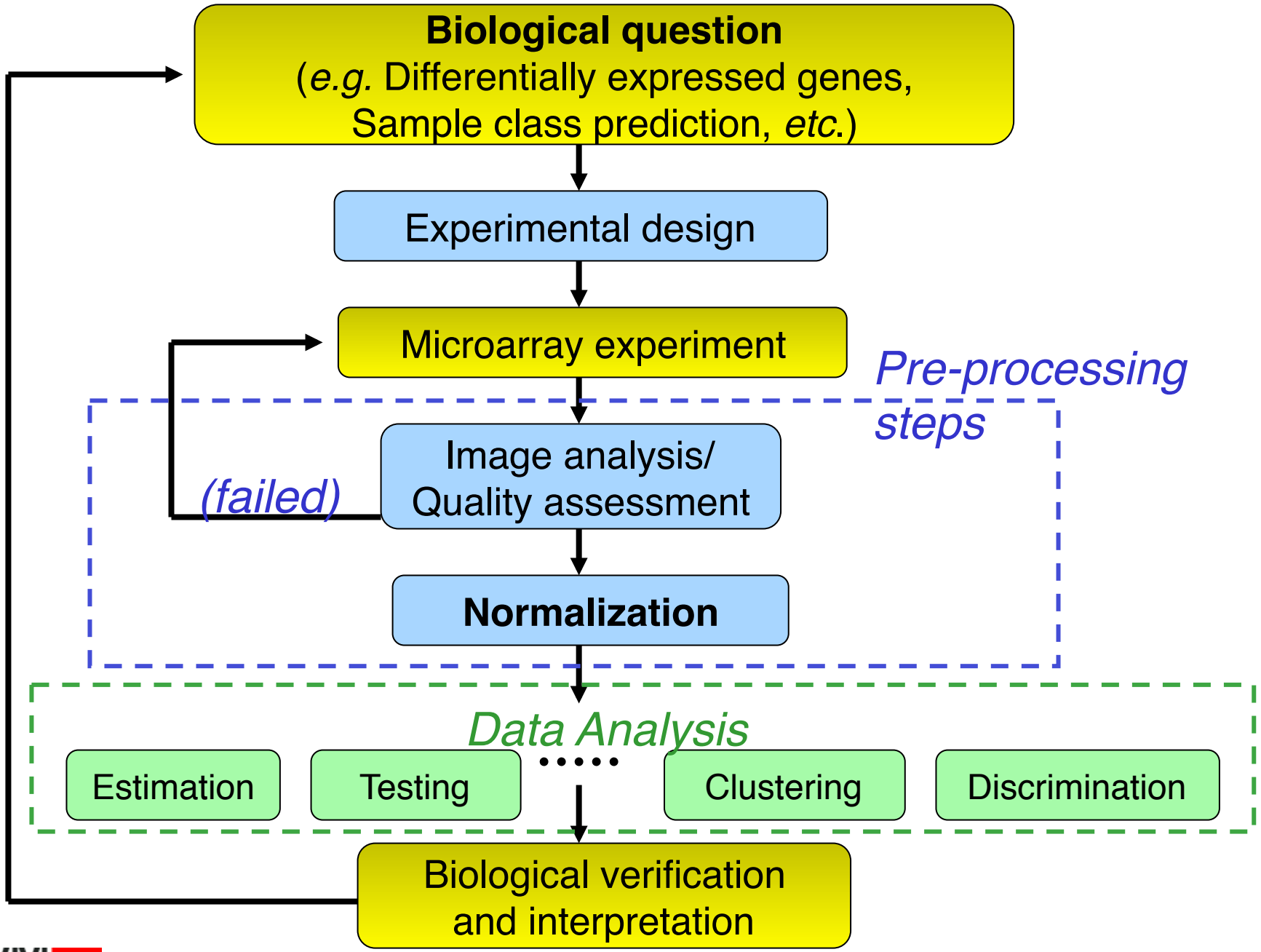
  Red intensity = Rfg - Rbg

  Green intensity = Gfg - Gbg

- and combine them in the log (base 2) ratio

  $\mathbf{Log_2(Red/Green)}$

- (Only need to do this once for single channel)
- Often, fg = mean and bg = median of relevant pixel intensities


Biological question
(e.g. Differentially expressed genes,
Sample class prediction, etc.)
Experimental design
Microarray experiment
Pre-processing steps
(failed)
Image analysis/
Quality assessment
Normalization
Data Analysis
Estimation
Testing
Clustering
Discrimination
Biological verification
and interpretation

Preprocessing: Data Visualization, Exploratory Data Analysis (EDA)

- Was the experiment a success?
- Are there any specific problems?
- What analysis tools should be used?

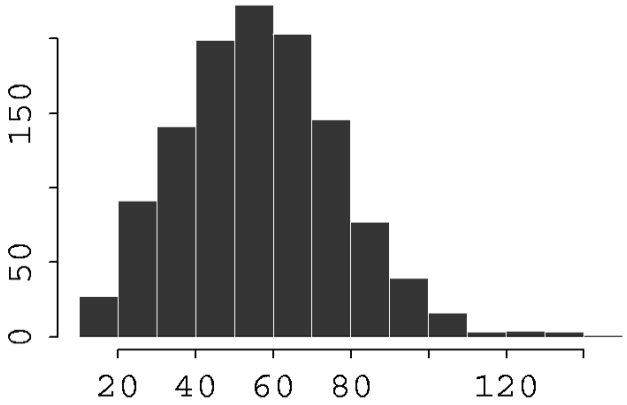
Quality Measurements

- Array
  - Correlation between spot intensities
  - Percentage of spots with no signals
  - Distribution of spot signal area
- Spot
  - Signal / Noise ratio
  - Variation in pixel intensities
  - Identification of "bad spot" (spots with no signal)
- Ratio (2 spots combined)
  - Circularity
- *Flag* or *weight* spots based on these (or other) criteria

Quality of Array

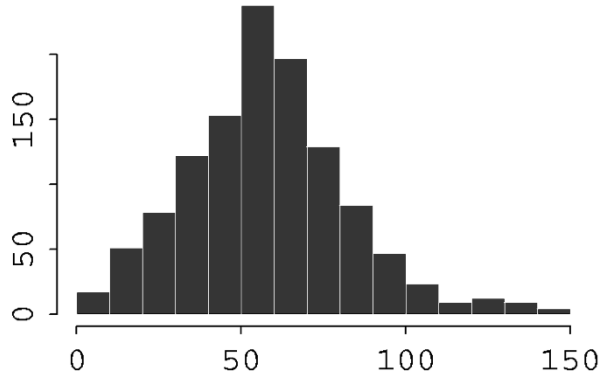
Distribution of areas

- Judge by eye
- Look at variation. (e.g. SD)



Cy3 area

- **mean 57**
- **median 56**
- **SD 20.67**



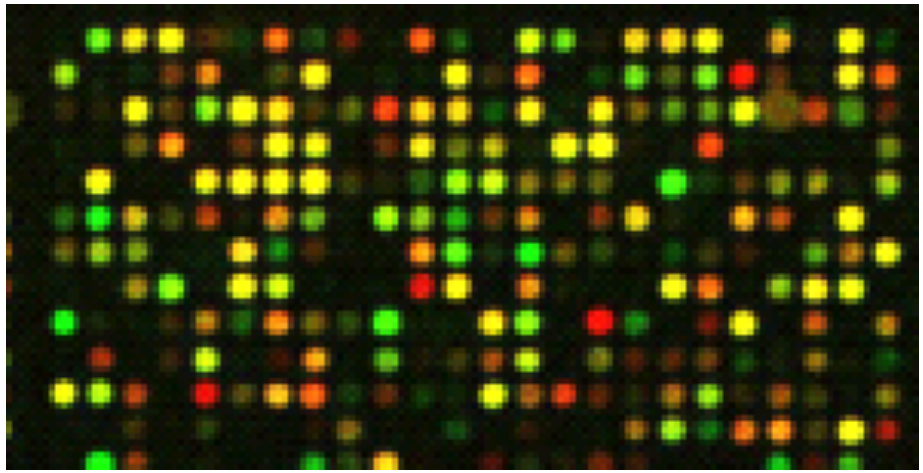
Cy5 area

- **mean 59**
- **median 57**
- **SD 24.34**

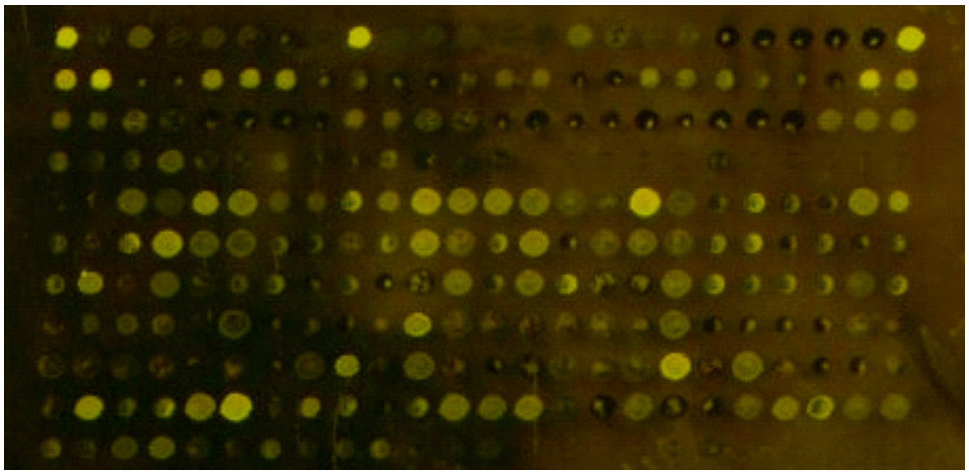
Red/Green overlay images

Co-registration and *overlay* offers a quick visualization, revealing information on color balance, uniformity of hybridization, spot uniformity, background, and artifacts such as dust or scratches

Good: low bg, detectable d.e.



Bad: high bg, ghost spots, little d.e.



Microarray data preprocessing

- Primary (dual channel) microarray data are pixel-level values from the *two tiff files*
- Prior to analysis, usually microarray data are *preprocessed* so that there is a *single value* for each spot
- The main preprocessing steps are
  - Image analysis
  - Normalization
- We assume now that image analysis has been carried out, so that pixel values are summarized as Rf, Rb, Gf, Gb

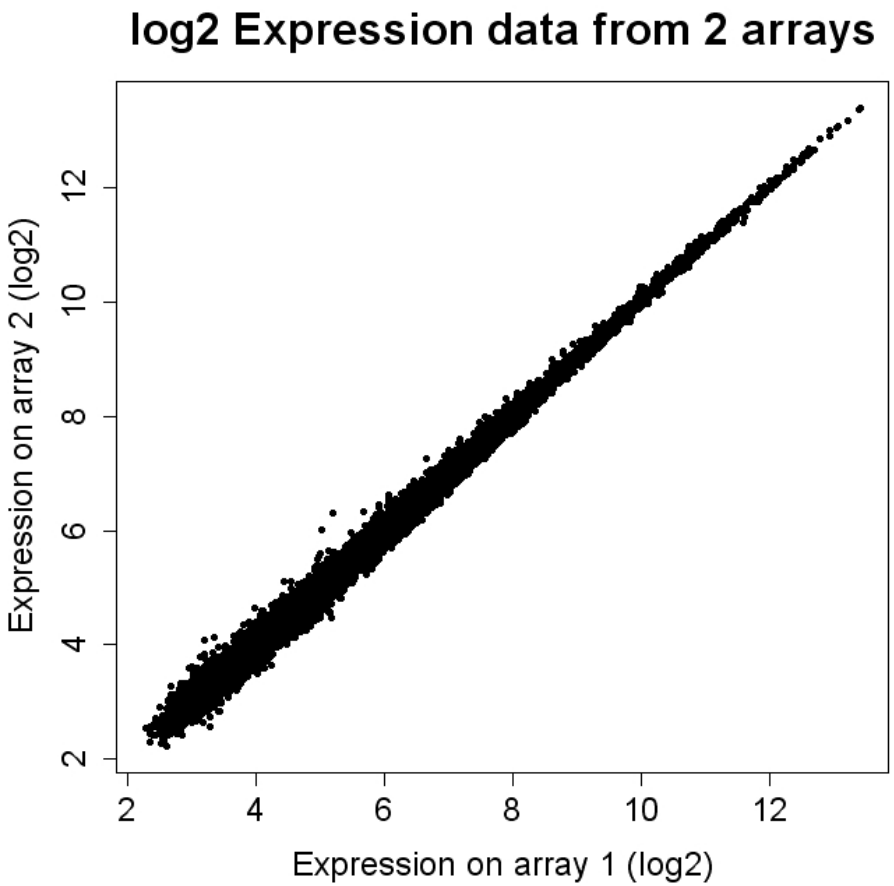
Artifacts in microarrays

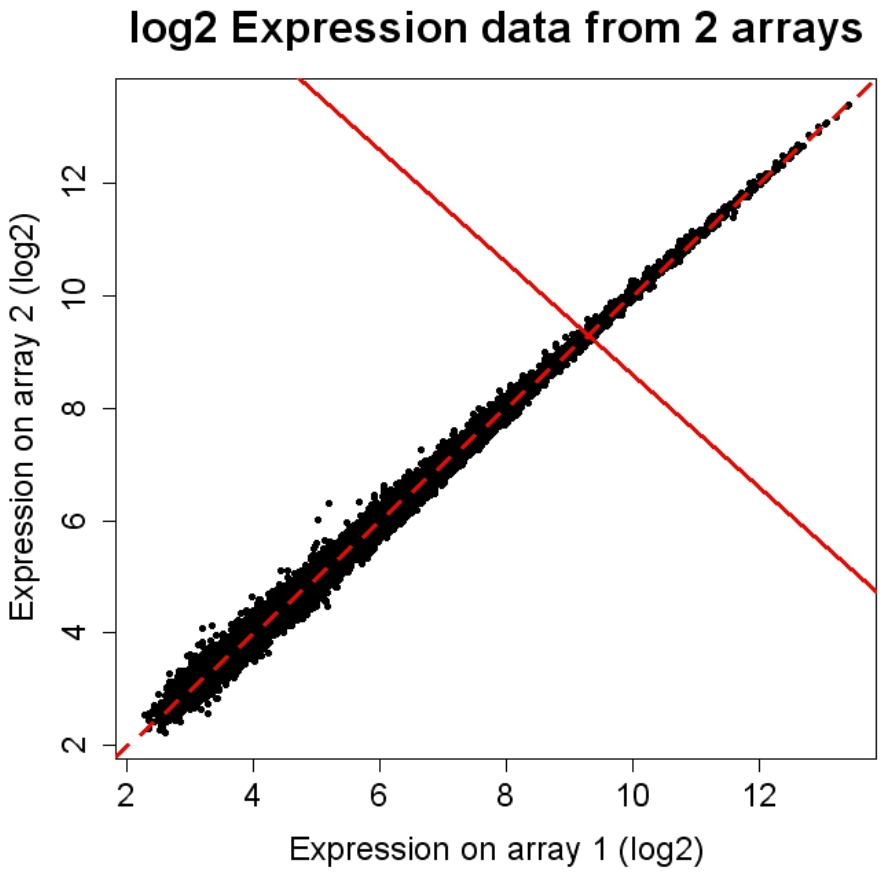
- We are interested in finding true *biologically meaningful differences* between sample types
- Due to other sources of systematic variation, there are also usually *artifactual differences*
- Sources of artifacts include:
  - print tips - differences in subarrays
  - plate effects – differences in rows within subarray
  - batch effects
  - hybridization artifacts

Looking for artifacts

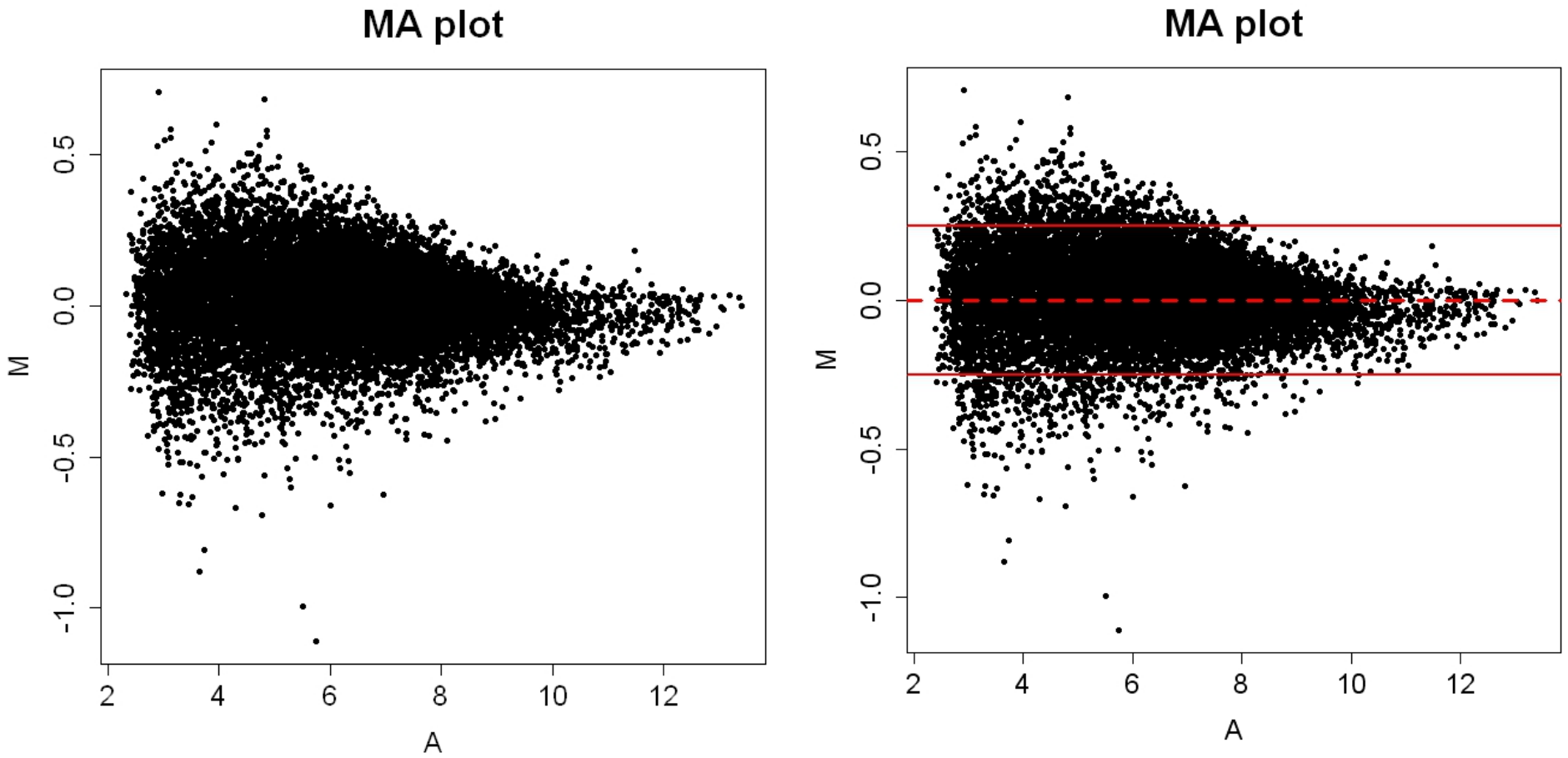
- Exploratory data analysis (EDA) is an important component of microarray data preprocessing
- EDA involves identifying data artifacts
- We will use several types of plots for data visualization, primarily
  - *scatterplots*
  - *boxplots*
  - *spatial plots*

Take logs…


log2 Expression data from 2 arrays
Expression on array 2 (log2)
Expression on array 1 (log2)
2
4
6
8
10
12

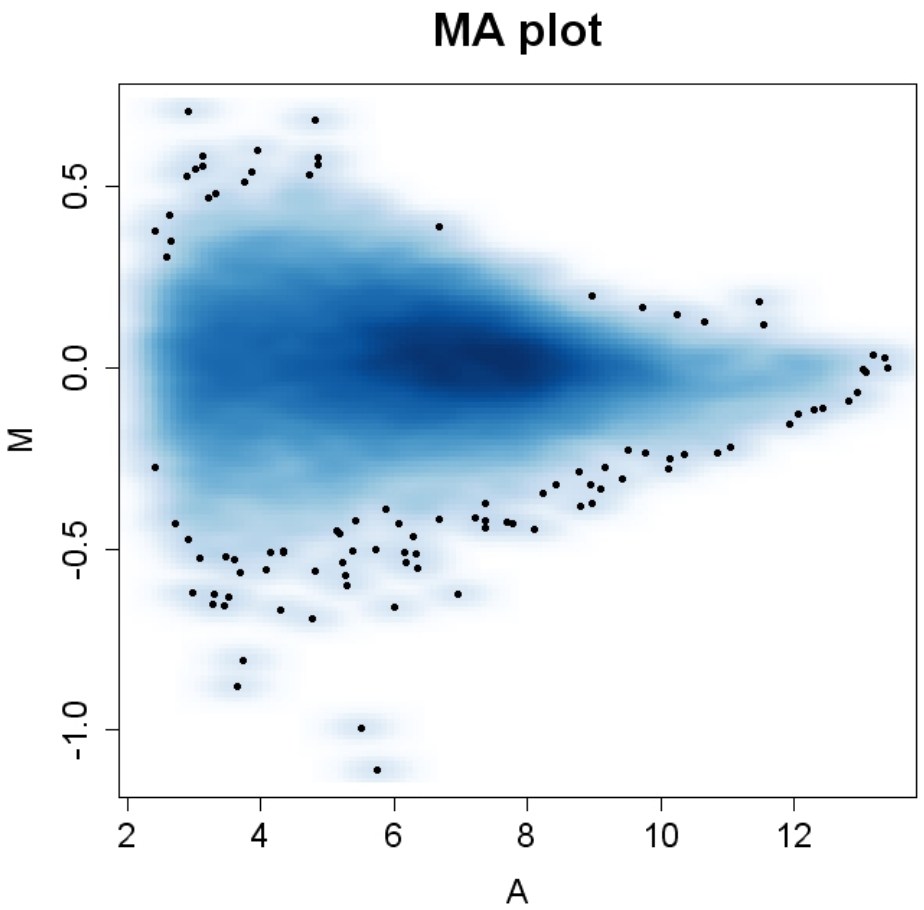

log2 Expression data from 2 arrays
Expression on array 2 (log2)
Expression on array 1 (log2)
2
4
6
8
10
12

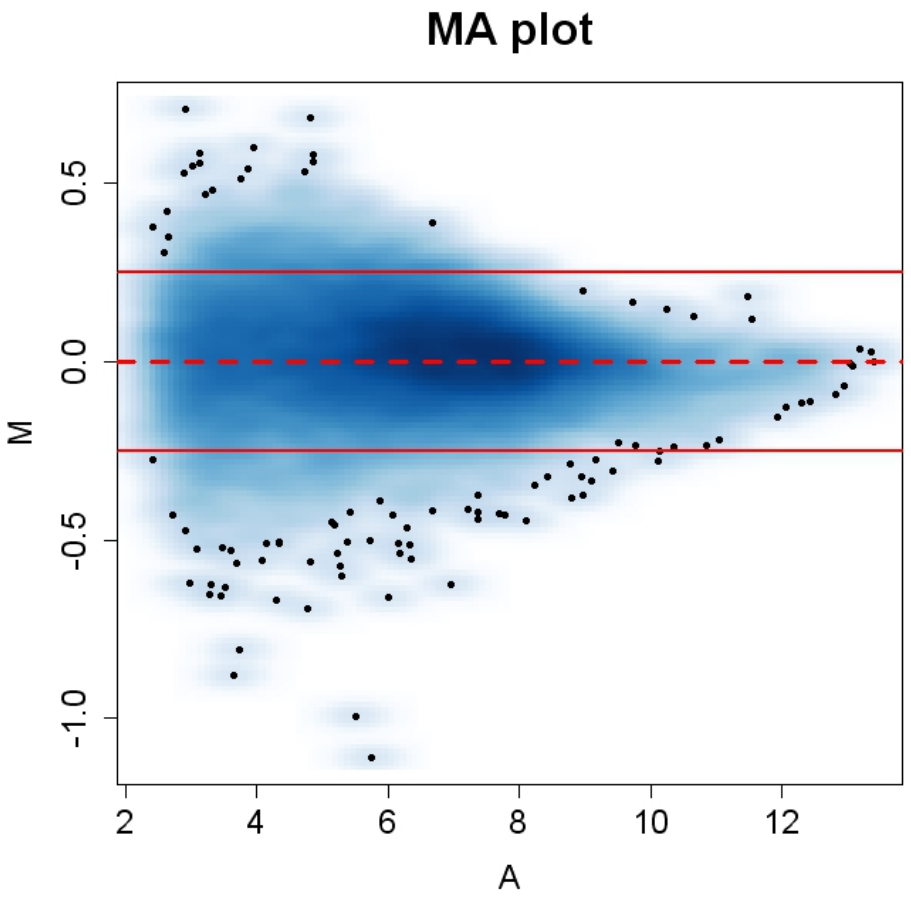
… and rotate



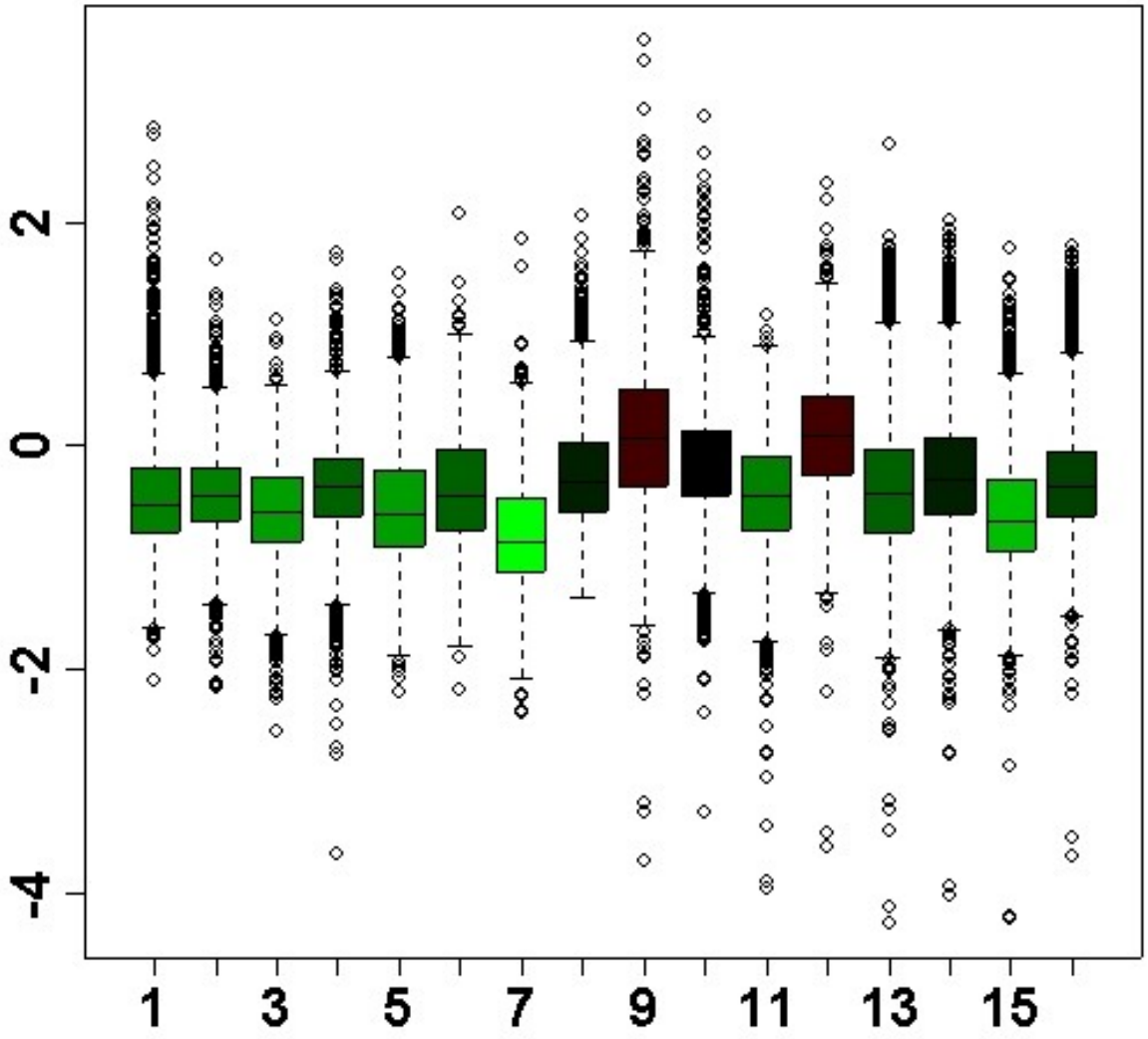
- M = 'minus' = log2 (expression 2) – log2 (expression 1)
- A = 'average' = [log2 (expression 1) – log2 (expression2)]/ 2

smoothScatter



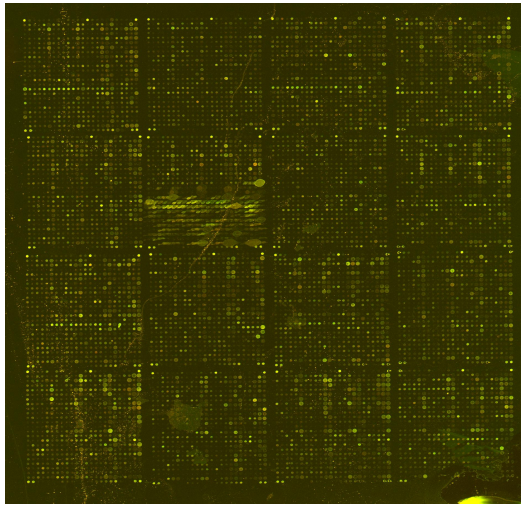


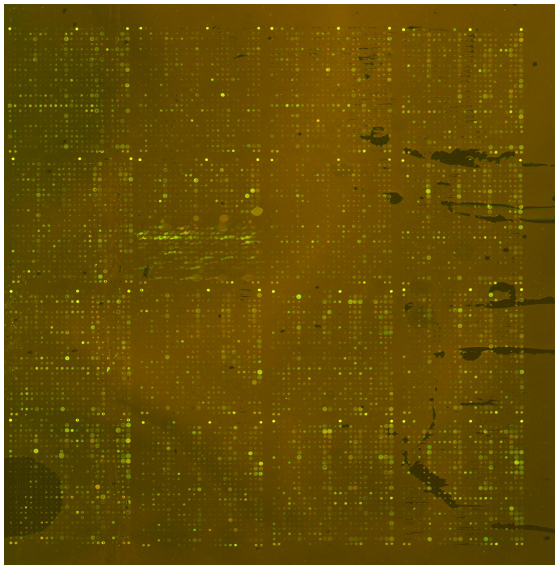
Boxplots of $\log_2 R/G$

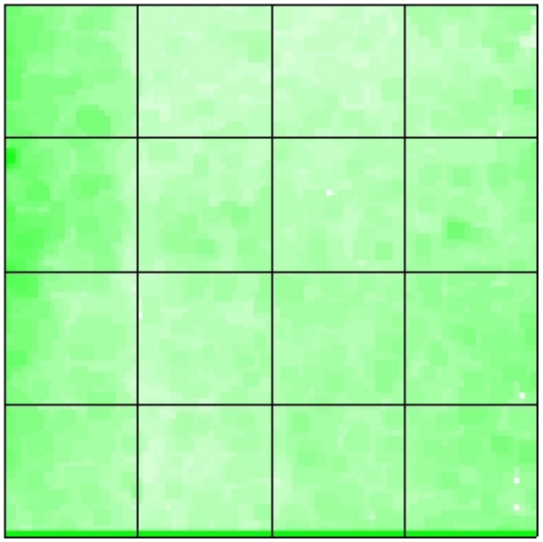


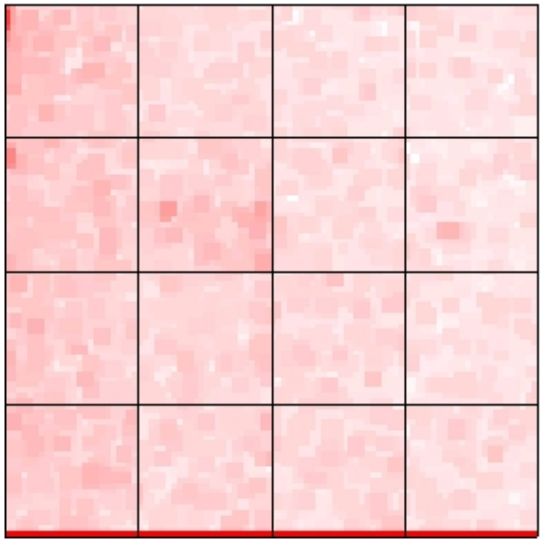
Liver samples from 16 mice: 8 WT, 8 ApoAI KO

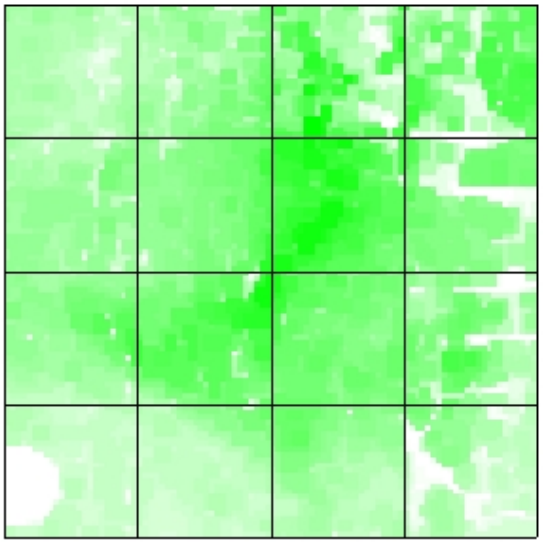
Spatial plots: background from two slides

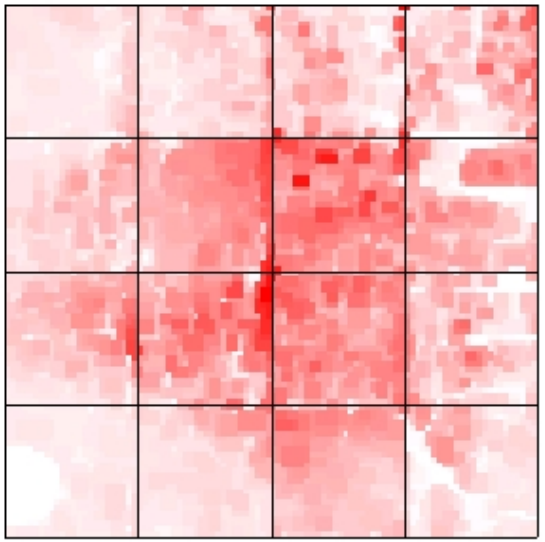




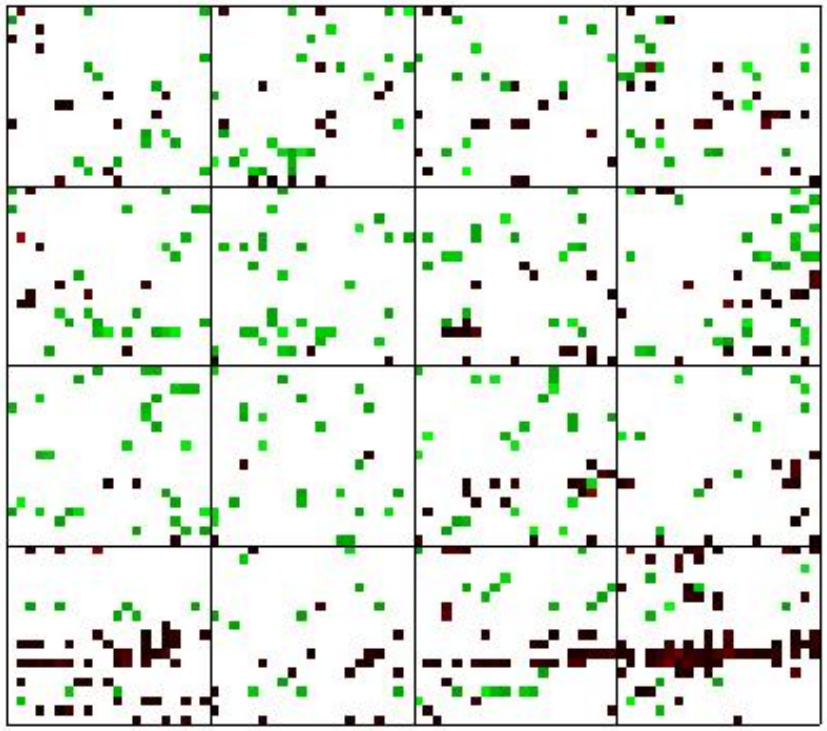






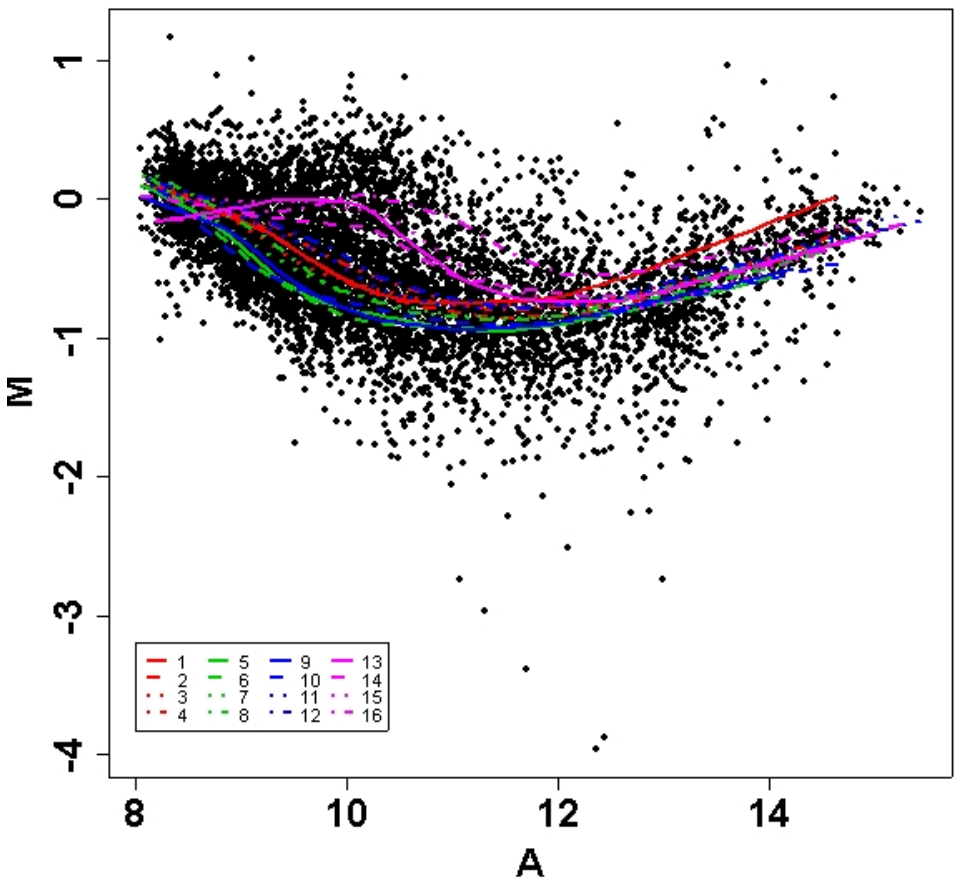


Highlighting extreme log ratios

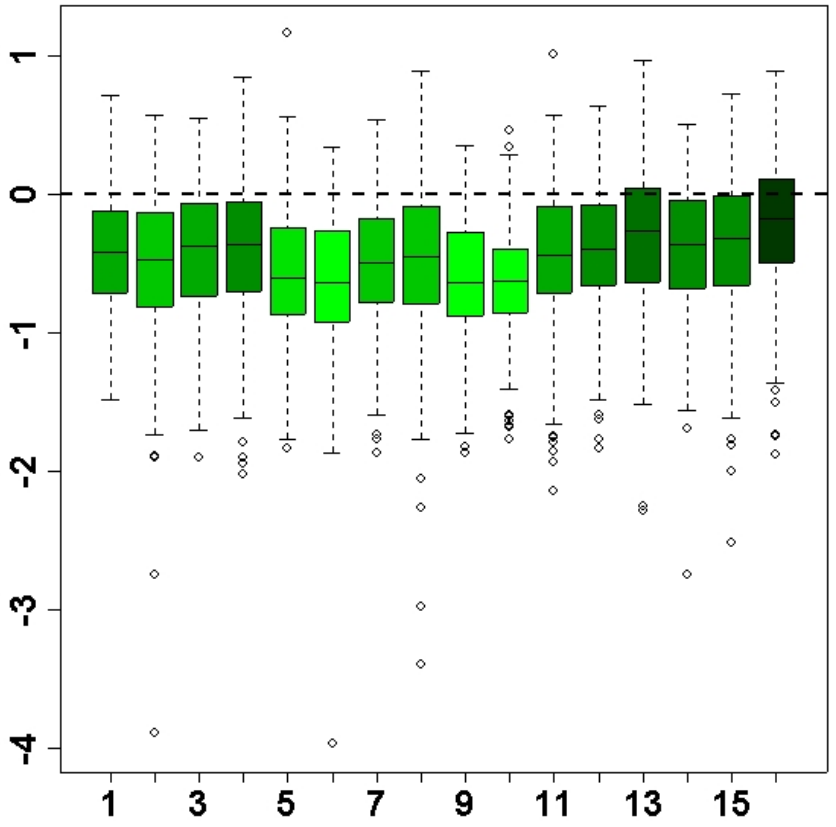


Top (black) and bottom (green) 5% of log ratios

Pin group (sub-array) effects

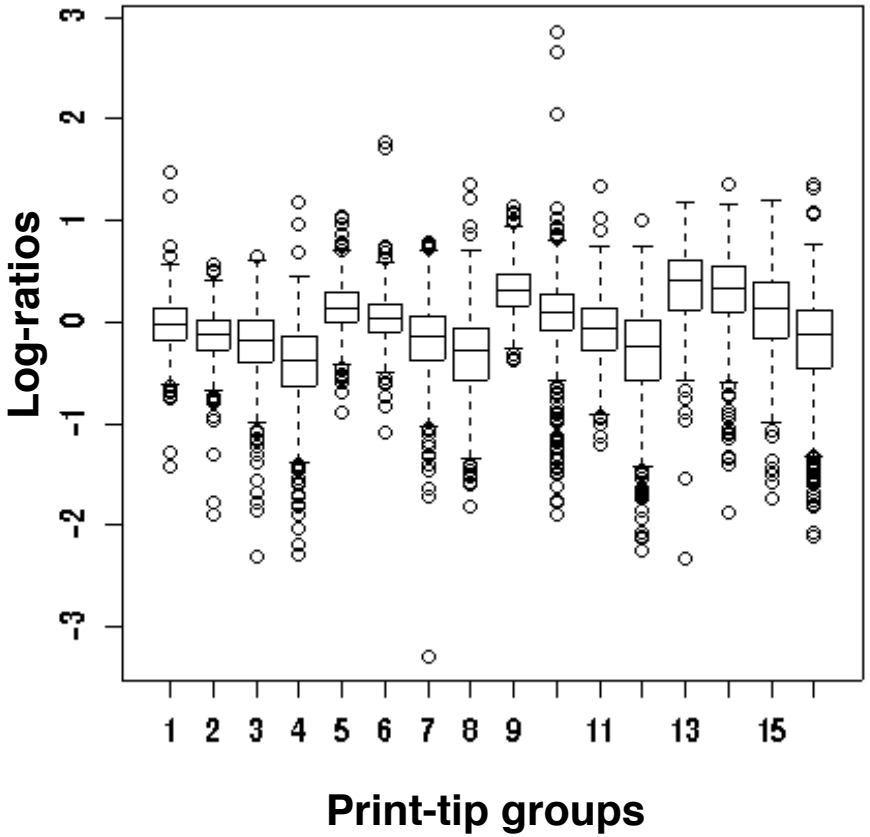


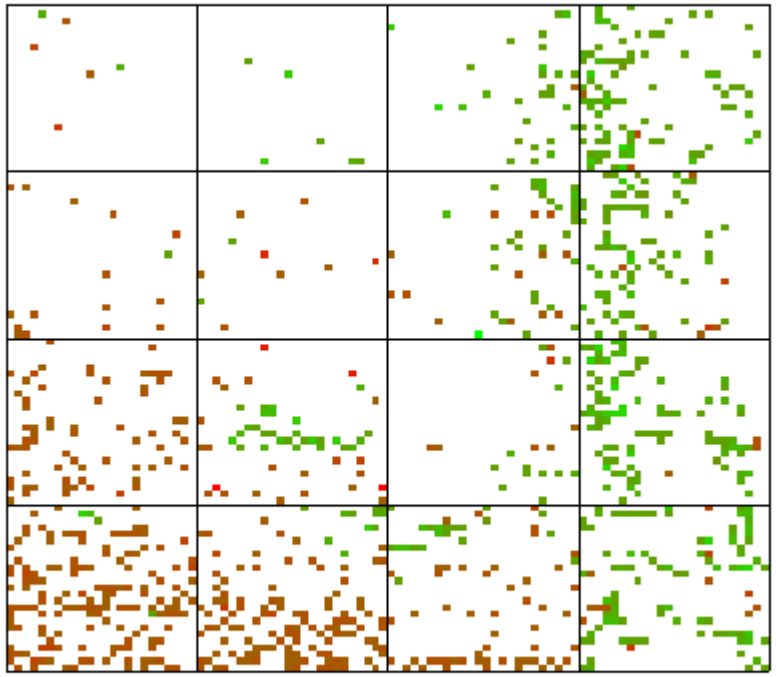
Lowess lines through points from pin groups



Boxplots of log ratios by pin group

Highlighting pin group effects





Clear example of spatial bias

(BREAK)

Preprocessing: Normalization

- *Why?*

  To correct for *systematic differences* between samples on the same slide, or between slides, *which do not represent true biological variation* between samples

- *How do we know it is necessary?*

  By examining *self-self hybridizations*, where no true differential expression is occurring. There are *dye biases* which vary with spot intensity, location on the array, plate origin, pins, scanning parameters, *etc.*

Normalization: global

- Normalization based on a *global adjustment*

$$\log_2 R/G \rightarrow \log_2 R/G - c = \log_2 R/(kG)$$

- Common choices for k or $c = \log_2 k$ are c = *median* or *mean* of log ratios for a particular gene set (*e.g.* all genes, or control, or 'housekeeping' genes)
- Another possibility is *total intensity* normalization, where $k = \sum R_i / \sum G_i$

Normalization: intensity-dependent

- Here, run a line through the middle of the MA plot, shifting the M value of the pair (A,M) by c=c(A), *i.e.*

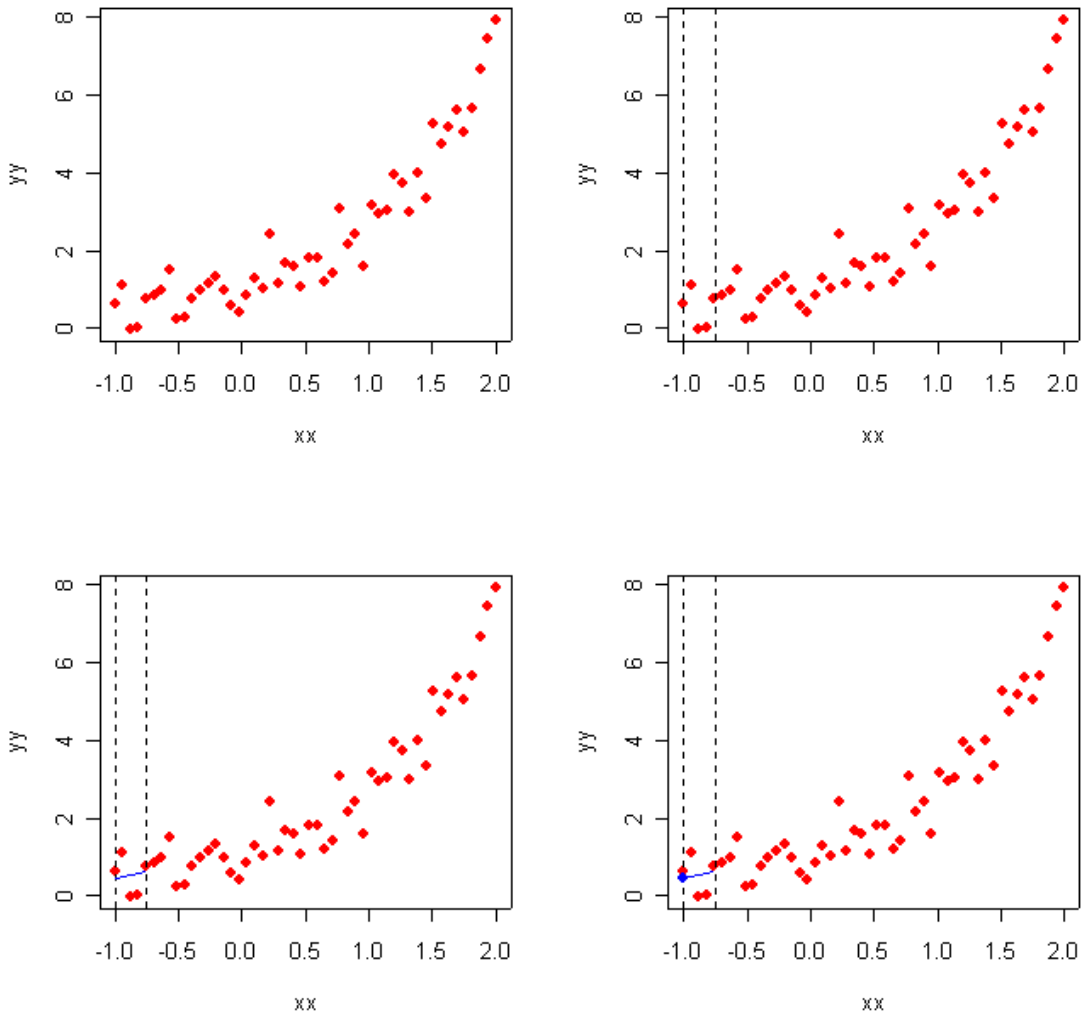
  $\log_2 R/G \rightarrow \log_2 R/G - c(A) = \log_2 R/(k(A)G)$

- One estimate of c(A) is made using the LOWESS (or loess) function of Cleveland (1979): *LOcally WEighted Scatterplot Smoothing*

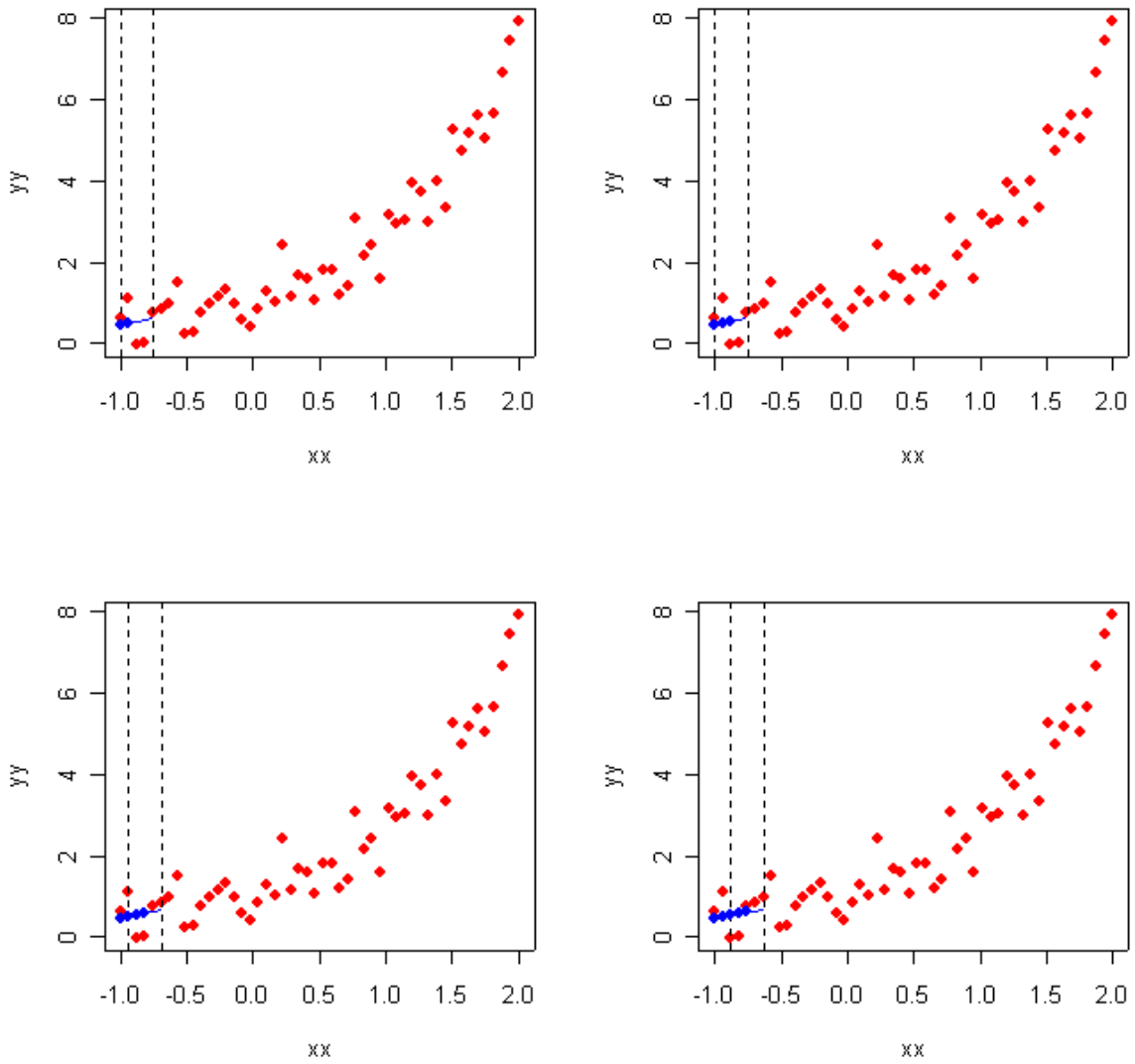
Local regression

- Classical (global) regression: draws a *single line* to the entire set of points
- *Local regression* : draws a *curve* through noisy data by *smoothing*
- Linear (or polynomial) function of the predictor(s) is created in a *local neighborhood*, points are *weighted*
- As you move through values of the predictor, the neighborhood moves as well

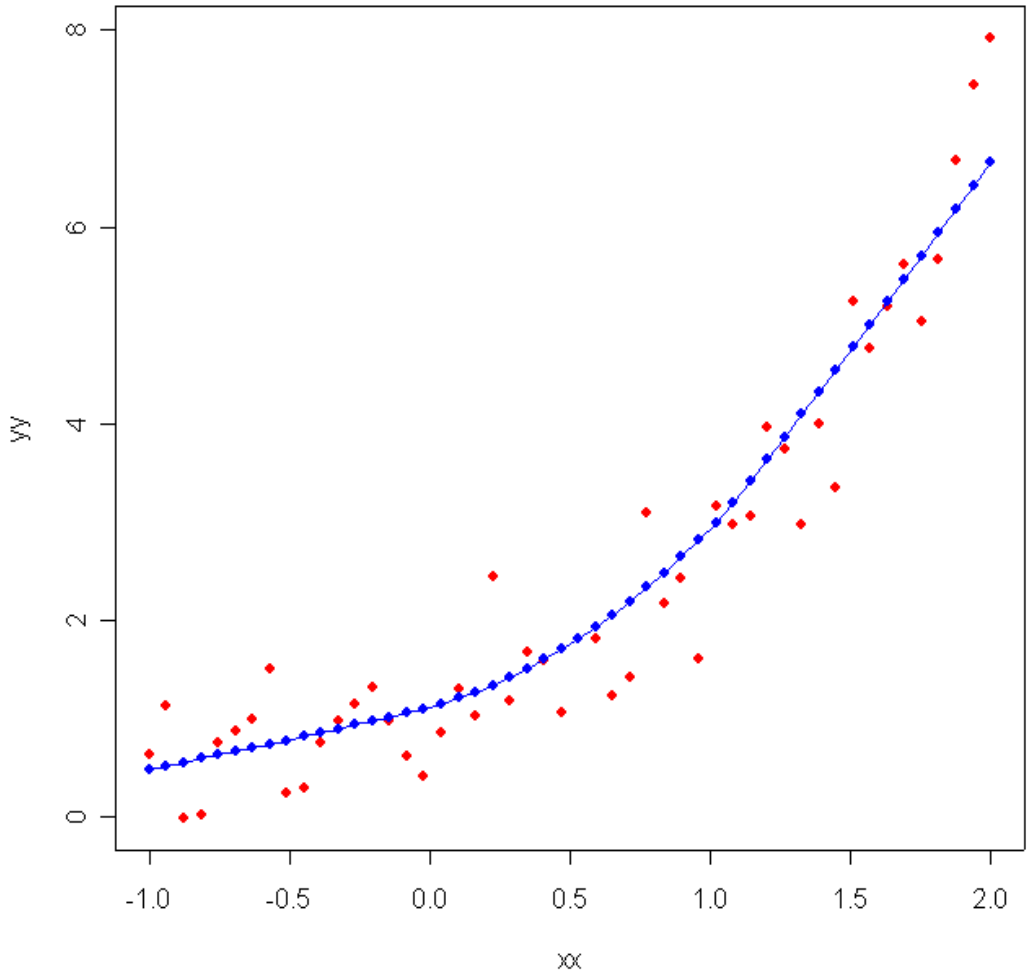
Getting local regression started



The neighborhood moves



Lowess line



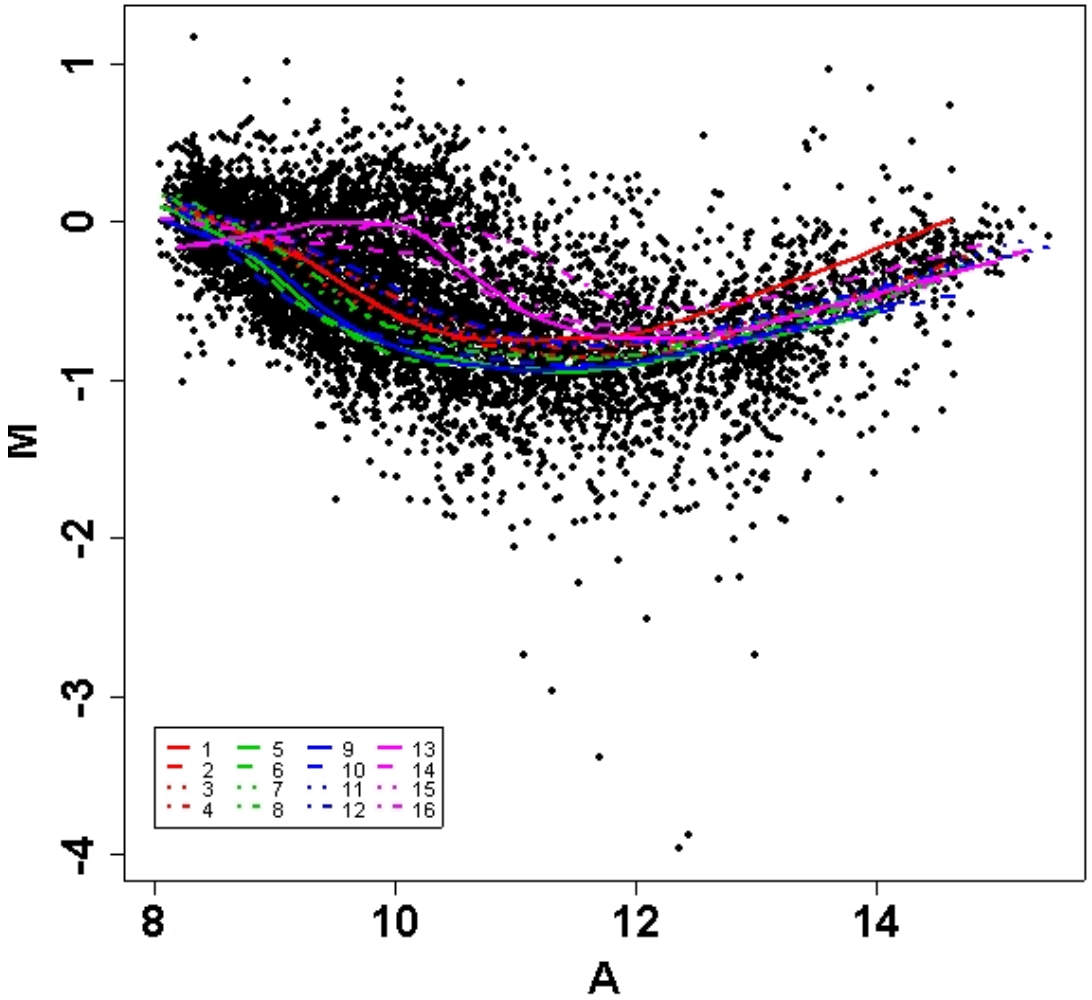
Normalization: print-tip

- *Intensity-dependent variation* and *spatial bias* can be significant sources of systematic error
- Global methods do *not* correct for spatial effects produced by hybridization artifacts or print-tip or plate effects during microarray construction
- Can correct for *both* print-tip and intensity-dependent bias by performing LOWESS fits to the data *within print-tip (pin) groups*, *i.e.*

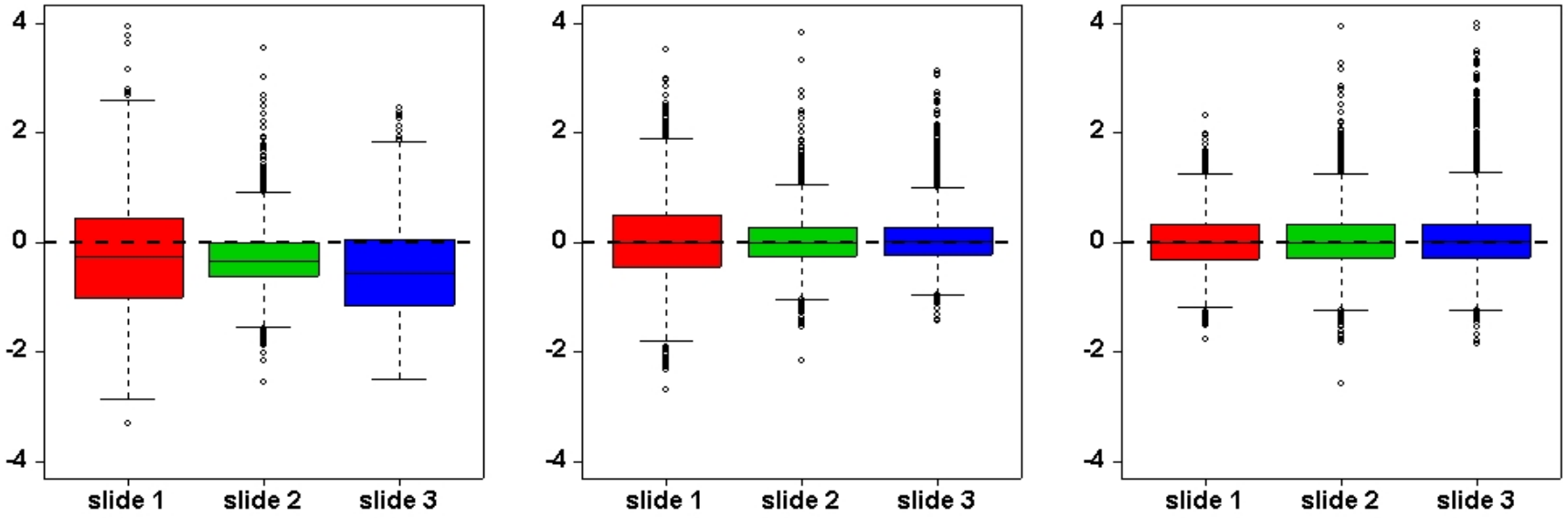
$$\log_2 R/G \rightarrow \log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G),$$

where $c_i(A)$ is fit to the MA plot for grid *i* only

MA plot with print-tip (pin) loess



Scale normalization: between slides



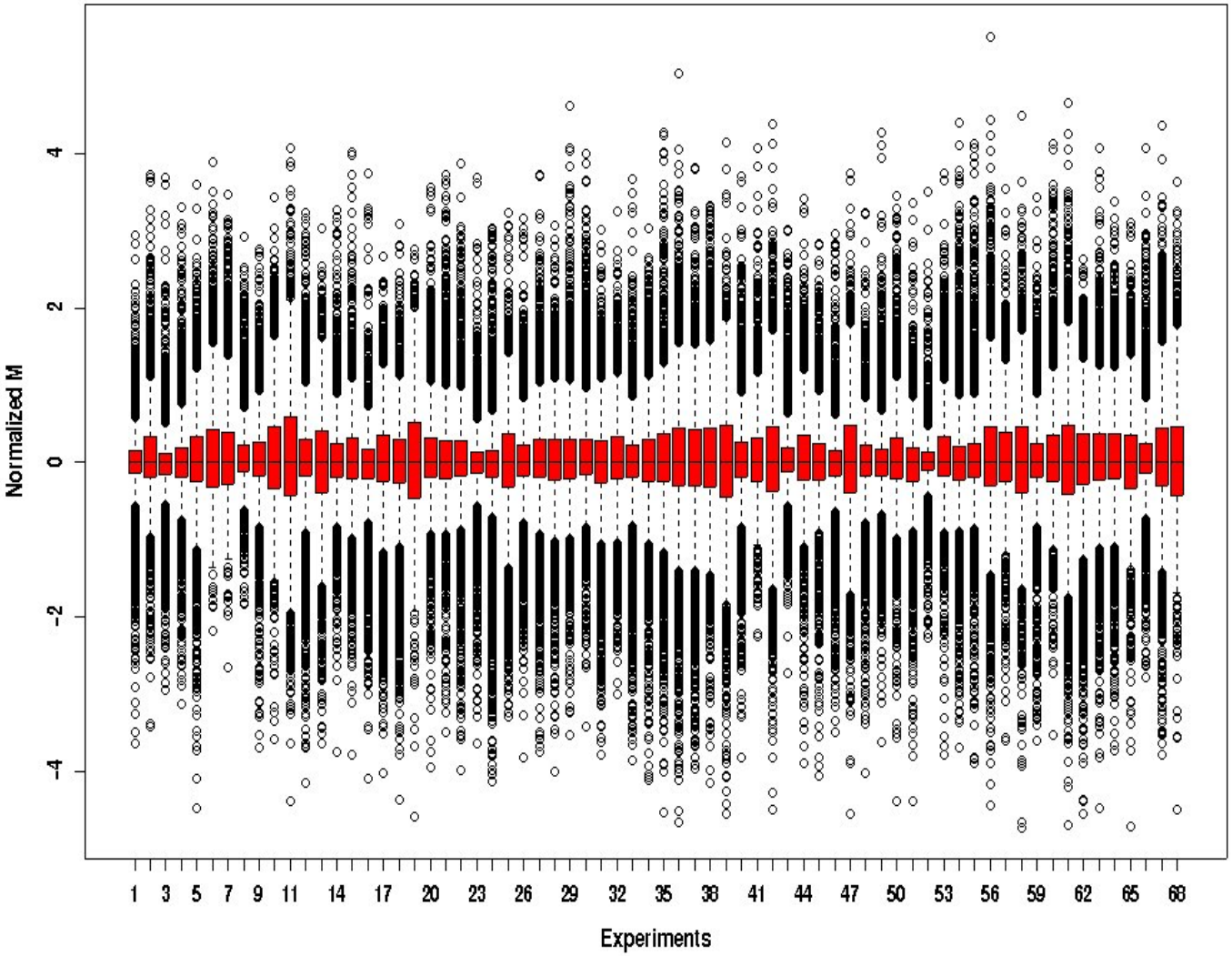
Boxplots of log ratios from 3 replicate self-self hybs

Left panel: before normalization

Middle panel: after within print-tip group normalization

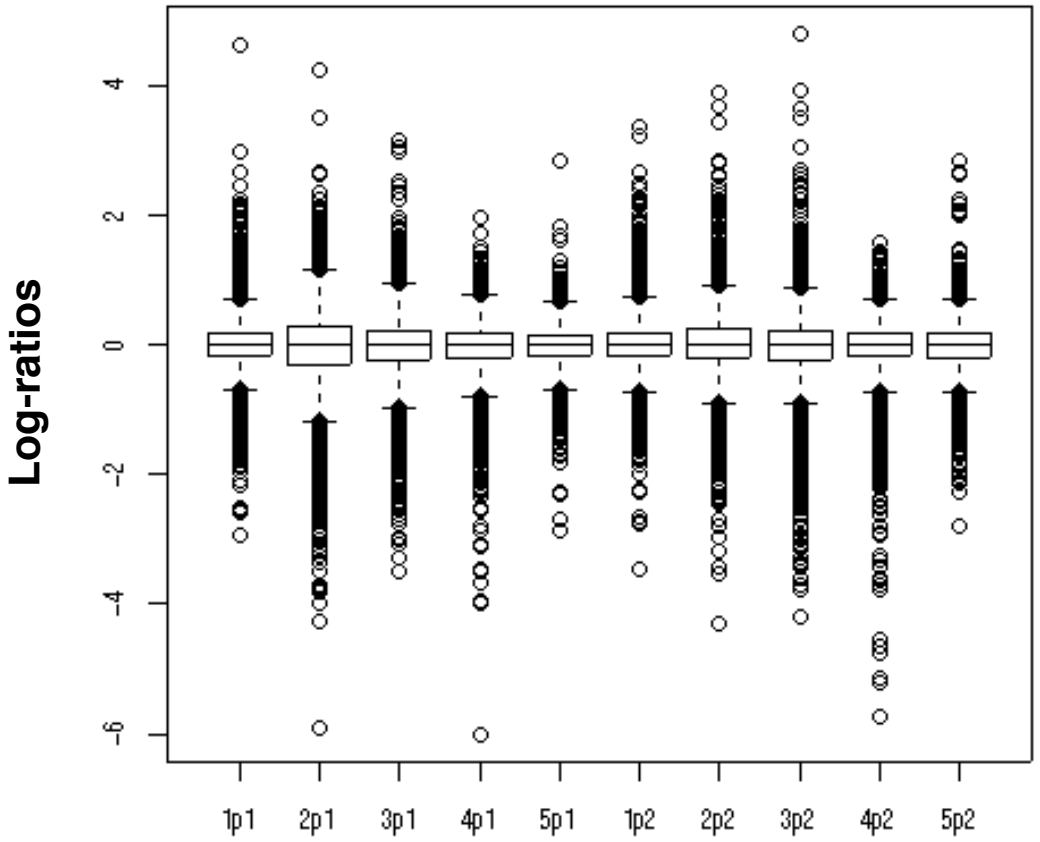
Right panel: after a further between-slide scale normalization

NCI 60 experiments



Should we use scale normalization here ??

Scale normalization: another data set



Should we use scale normalization here ??

Taking scale into account

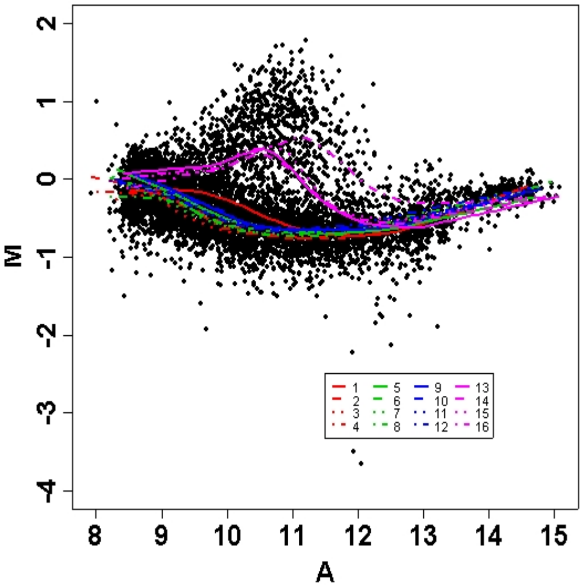
Assume: All slides have the same spread in M

- True log ratio is m_{ij} where *i* represents different *slides* and *j* represents different spots
- Observed is M_{ij}, where $M_{ij} = a_i m_{ij}$
- Robust estimate of a_i is

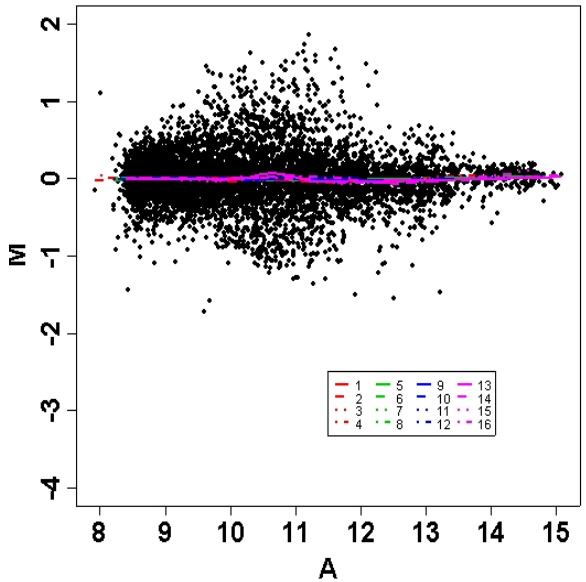
$$MAD_i = median_j \{ |y_{ij} - median(y_{ij}) | \}$$

- Could instead make same assumption for *print tip groups* (rather than *slides*)

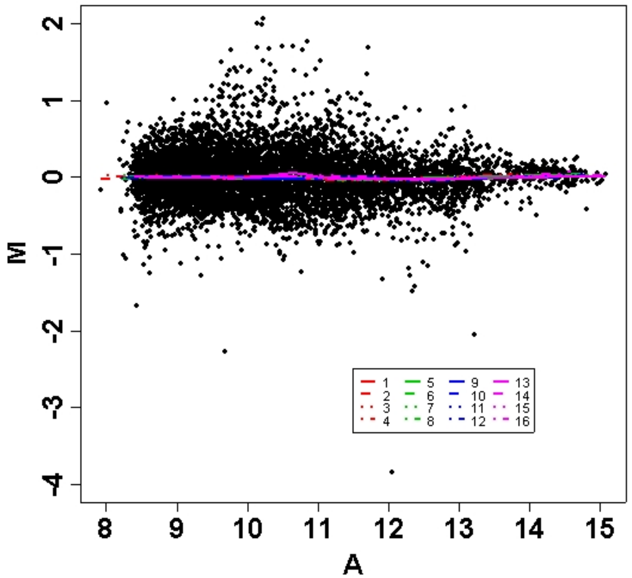
A comparison of three MA plots



Unnormalized



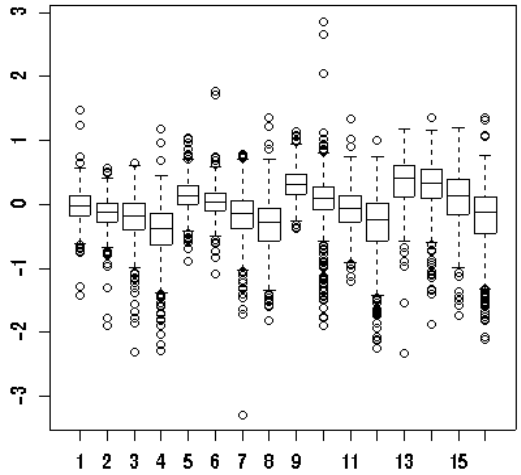
Print-tip normalization

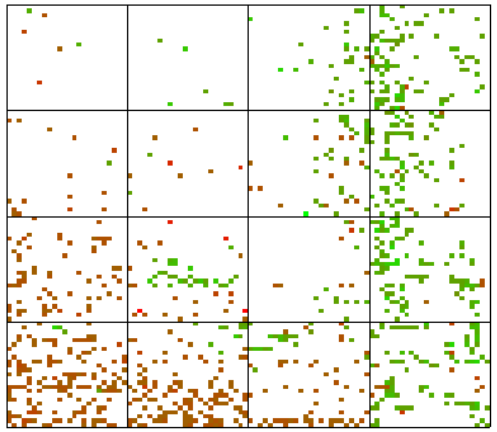


Print tip & scale normalization

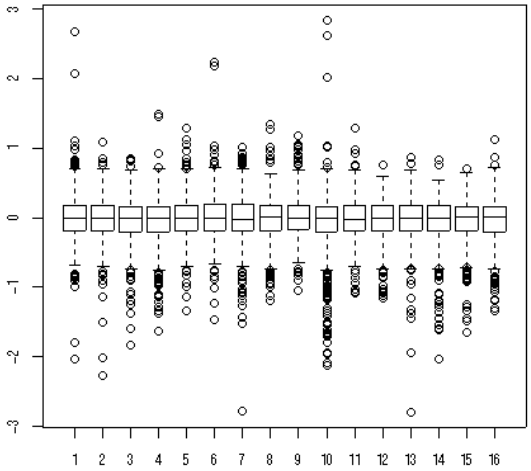
Same normalization on another data set

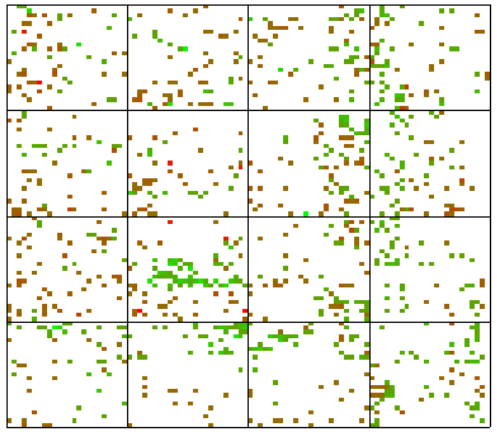
Before





After





Normalization: which spots to use?

- The lowess/loess lines can be run through many different sets of points
- Each strategy has its own *assumptions*
- Global lowess/loess can be justified by supposing that, when stratified by mRNA abundance,
  - most genes are not differentially expressed,

OR

  - over- and under-expression are *equally likely*
- For pin-specific (print-tip) lowess, assumptions should hold *within each pin group*

When lowess is not justified

- Assumptions are most likely to hold on arrays with *full genome representation* and samples from the *same tissue*
- For *specialized arrays* containing mainly genes of a certain class (*e.g.* lipid metabolism genes), lowess is ***NOT warranted*** :
  - the assumption that most genes are not differentially expressed is likely to be violated
  - here, would want to have many control genes spotted for normalization

Normalization: Summary

- Reduces *systematic* (not random) effects
- Makes it possible to compare several arrays
- For 2-color arrays:
  - Use logratios (MA plots)
  - Lowess normalization (dye bias)
  - Pin-group location normalization
  - Pin-group scale normalization
  - Between slide scale normalization

cDNA gene expression data

Data on *G* genes for *n* samples:

mRNA samples

Genes	sample1	sample2	sample3	sample4	sample5	…
1	0.46	0.30	0.80	1.51	0.90	…
2	-0.10	0.49	0.24	0.06	0.46	…
3	0.15	0.74	0.04	0.10	0.20	…
4	-0.45	-1.03	-0.79	-0.56	-0.32	…
5	-0.06	1.06	1.35	1.09	-1.09	…

Gene expression level of gene *i* in mRNA sample *j*

= (normalized) Log_2(Red intensity / Green intensity)

Combining data

- We have looked at analyzing one type of data (microarray data) from a single planned study
- Often, however, interested in
  - combining many types of data together, or
  - combining similar types of data from several studies
- Combined analysis of *different types of data* is an area of active research

Meta-analysis

- Meta-analysis can be thought of as an 'analysis of analyses'
- Not a new idea
  - R. A. Fisher (1944) had the idea of combining p-values
  - W. G. Cochran (1953) discusses a method of averaging means across independent studies, inverse variance weighting, homogeneity testing
- Meta-analysis focuses on the *direction* and *magnitude* of the effects across studies

Strengths of meta-analysis

- Formal methods for synthesizing research
- Capable of finding relationships across studies that may be obscured in other approaches
- Provides some protection against over-interpreting differences across studies
- Can handle a large numbers of studies

Weaknesses of meta-analysis

- May require substantial effort
- Difficult to capture qualitative distinctions between studies
- Risk of including poor studies
- *Selection bias*
  - negative and null finding studies that you were unable to find
  - outcomes for which there were negative or null findings that were not reported

Overview of meta-analytic data analysis

- Transformations, Adjustments and Outliers
- The Inverse Variance Weight
- The Mean Effect Size and Associated Statistics
- Homogeneity Analysis
- Fixed Effects Analysis of Heterogeneous Distributions
- Random Effects Analysis of Heterogeneous Distributions

The inverse variance weight

- Studies generally vary in size
- An effect size (ES) based on 100 subjects gives a more precise estimate of the population ES than an ES based on 10 subjects
- Therefore, larger studies should carry more weight in the analysis than smaller studies
- Simple approach: weight each ES by its sample size
- Better approach: weight by the inverse variance

Homogeneity analysis

- Homogeneity analysis tests whether all of the effect sizes are estimating the same population effect size
- If homogeneity is rejected, the data suggest that there are real differences between studies and they should therefore not be combined as just mentioned
- Two possibilities:
  - model between study differences
  - fit a random effects model

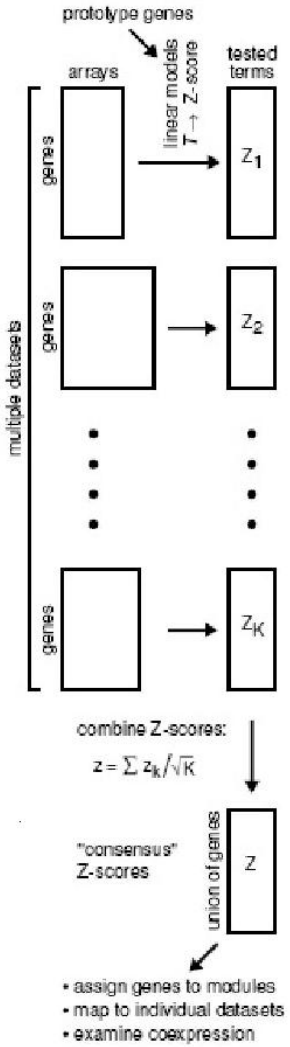
Example: Tumor profiling in breast cancer

- A *gene expression signature* is a pattern of expression for a particular set of genes
- Look for genes with reproducible patterns of expression
- Several prognostic signatures for breast cancer have been proposed in the literature
- All use *different sets of genes*
- Want to *combine information* across studies to form a more robust composite signature
- Collected large sized publicly available breast cancer survival datasets (12 studies, 2865 tumor expression profiles, 17198 genes in total)

Gene modules

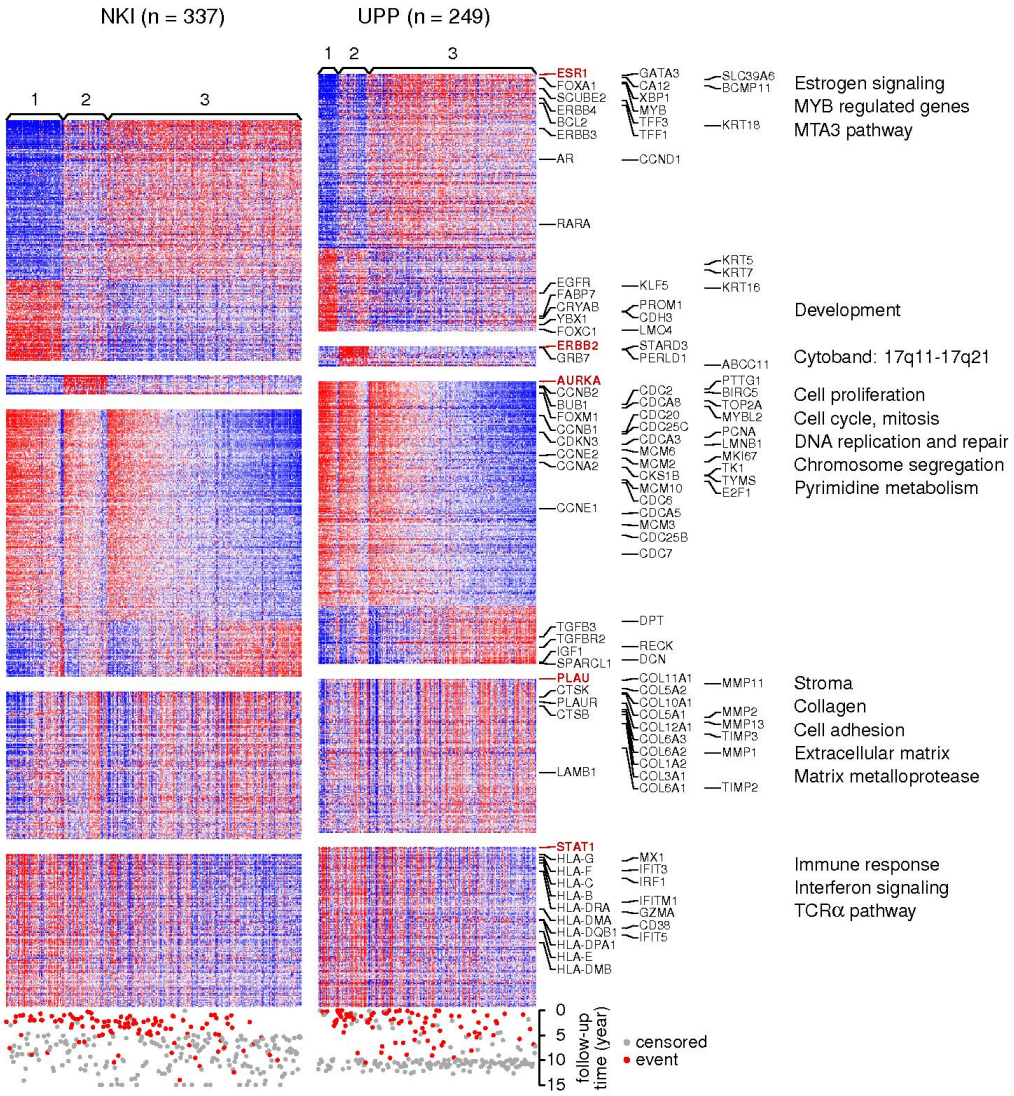
- Select *prototype genes* based on biological knowledge about breast cancer
- Prototypes
  - Estrogen receptor signalling (ESR1)
  - ERBB2 amplification (ERBB2)
  - Proliferation (AURKA)
  - Invasion (PLAU)
  - Immune response (STAT1)
- Each prototype forms the core of a *module*
- Additional genes added to modules based on association of expression with prototype
- Module score based on linear combination

Creation of the modules



- For each gene *i* separately, fit multiple regression modeling expression Y_i as a function of prototype expression
- *t*-test for each coefficient -> *Z*-score
- Combine Z-scores across studies with (equal weights) inverse normal method
- Select for each prototype the genes most strongly associated

Visualization of module coexpression



Data analysis review - exam

- Background/Introduction
- Quality assessment
- Affymetrix gene expression
- Statistical analysis
  - Linear model
  - Ranking genes for DE
  - Multiple hypothesis testing
  - Cluster analysis

Graphs

- Describe what each figure depicts
- Need to be *precise:* exactly what quantity are you plotting?
  - *`Boxplots of chips …'* **NO!!!!!**
  - *`Heatmap of genes …'* **NO!!!!!**
- Make sure to explain the color scheme, what the reference lines indicate, *etc.*
- Figure layout should be nice-looking and easy to read

Final report - General

- I should *receive* your report by *Friday 3 February 2017*, 12.00 (midi) at the latest
- If you have *any questions*, you should ask *ONLY ME* – do *NOT* talk to *anyone else* about your exam
- Make sure that you follow *all* the instructions (page limit, etc.)
- *GOOD LUCK!!*

BONNES VACANCES!!!