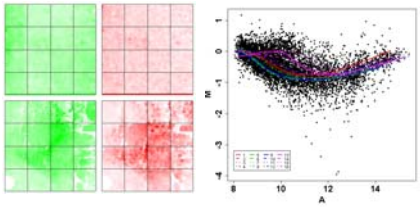
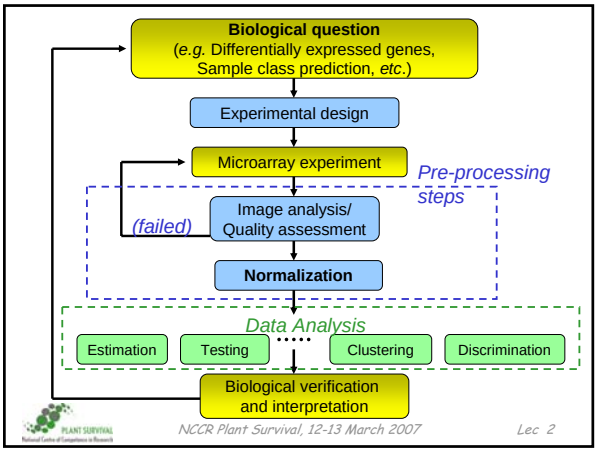
Statistics for cDNA microarrays

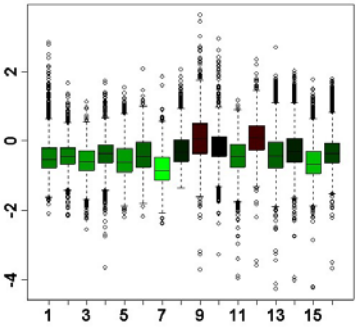
Exploratory data analysis; Normalization



`http://www.isrec.isb-sib.ch/~darlene/NCCR-PS/`

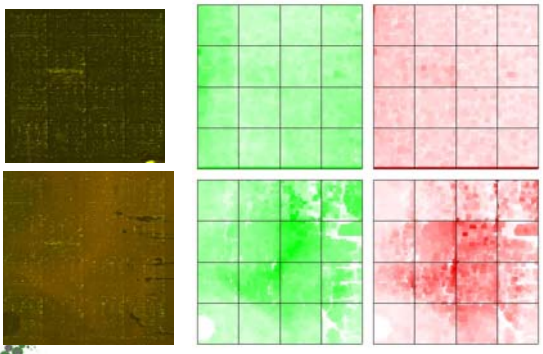


Boxplots of $\log_2 R/G$

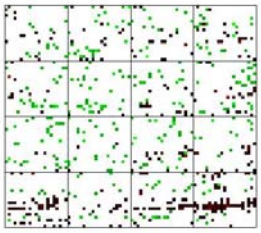


Liver samples from 16 mice: 8 WT, 8 ApoAI KO

Spatial plots: background from two slides

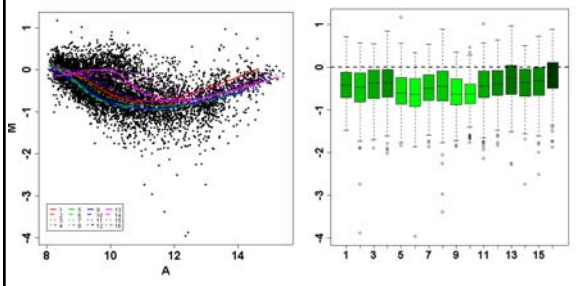


Highlighting extreme log ratios



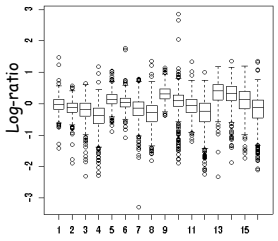
Top (black) and bottom (green) 5% of log ratios

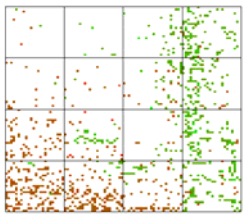
Pin group (sub-array) effects



Lowess lines through points from pin groups Boxplots of log ratios by pin group

Boxplots, highlighting pin group effects





Clear example of spatial bias

Preprocessing: Normalization

- *Why?*

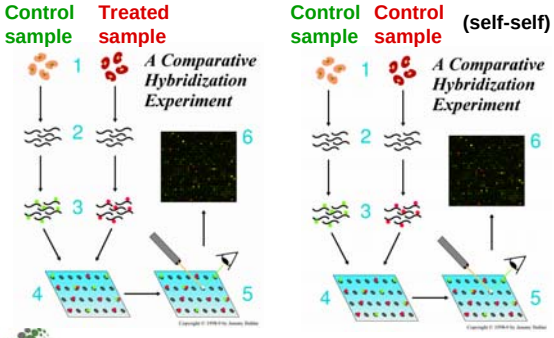
  To correct for *systematic differences* between samples on the same slide, or between slides, *which do not represent true biological variation* between samples
- *How do we know it is necessary?*

  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.*

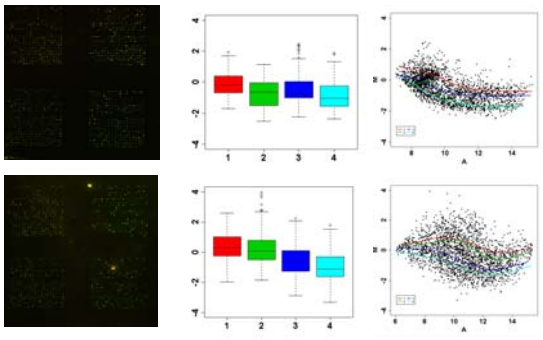
What is self-self hybridization?

- In dual channel (2-color) microarrays, such as cDNA arrays, two samples are each labeled with a different fluorescent dye
- In most studies, the samples are from *different sources* (*e.g.* cancer vs. normal)
- However, it is also possible to co-hybridize two samples from the *same source* (but differently labeled)

Dual channel co-hybridizations

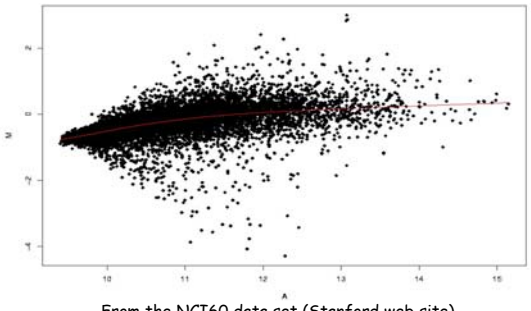


Self-self hybridizations



False color overlay | Boxplots within pin-groups | Scatter (MA-)plots

Similar patterns apparent in non self-self hybridizations



From the NCI60 data set (Stanford web site)

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 $k = \Sigma R_i / \Sigma G_i$

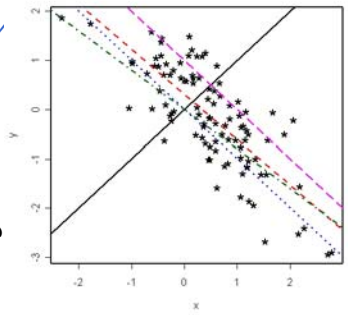
Normalization: intensity-dependent

- 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*

Simple linear modeling: which line?

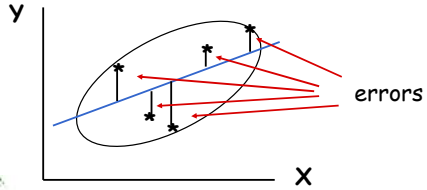
- There are *many possible lines* that could be drawn through the cloud of points in the scatterplot …
- How to choose?



Least Squares

- *Q*: Where does the regression equation come from?

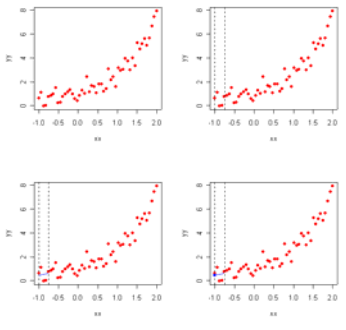
  A: It is the line that is 'best' in the sense that it *minimizes* the sum of the *squared* errors (residuals) in the vertical (*Y*) direction



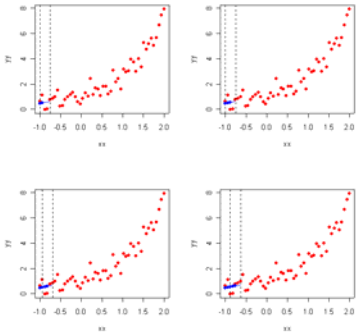
Local regression

- 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

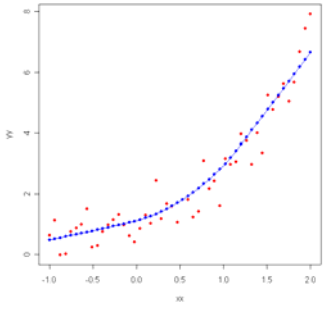
Getting local regression started



The neighborhood moves



Lowess line



Normalization: print-tip

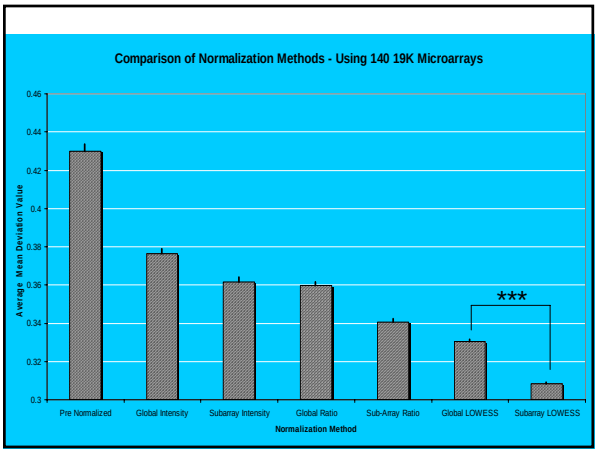
- *Intensity-dependent variation* and *spatial bias* can be significant sources of systematic error
- Global methods do *not* correct for spatial effects produced by hybridization artifacts or print-tip or plate effects during microarray construction
- Can correct for *both* print-tip and intensity-dependent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

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

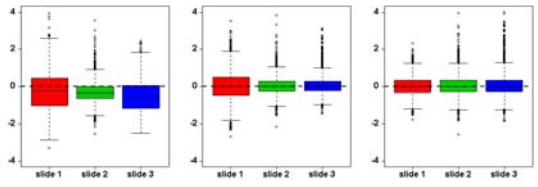
  where $c_i(A)$ is fit to the MA plot for grid *i* only

Comparison of Normalization Schemes (courtesy of Jason Goncalves)

- Experiment done to *assess* the common normalization methods
- Based on reciprocal labeling experimental data for a series of 140 replicate experiments on two different arrays each with 19,200 spots

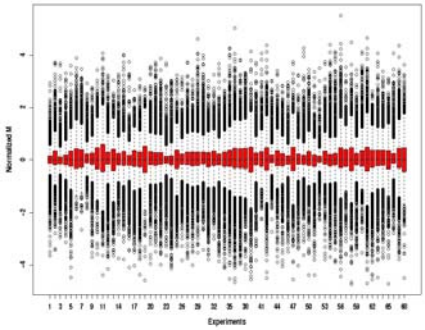


Scale normalization: between slides



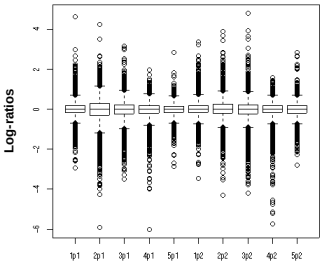
Boxplots of log ratios from 3 replicate self-self hybs
Left panel: before normalization
Middle panel: after within print-tip group normalization
Right panel: after a further between-slide scale normalization

NCI 60 experiments



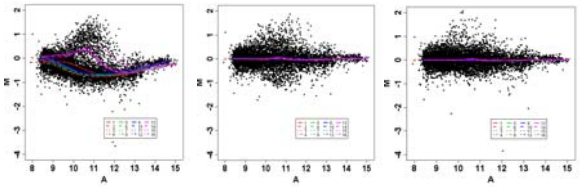
Should we use scale normalization here ??

Scale normalization: another data set



Should we use scale normalization here ??

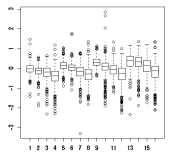
A comparison of three MA plots

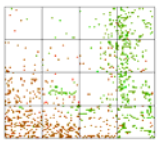


Unnormalized | Print-tip normalization | Print tip & scale normalization

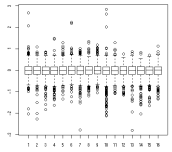
Same normalization on another data set

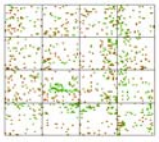
Before



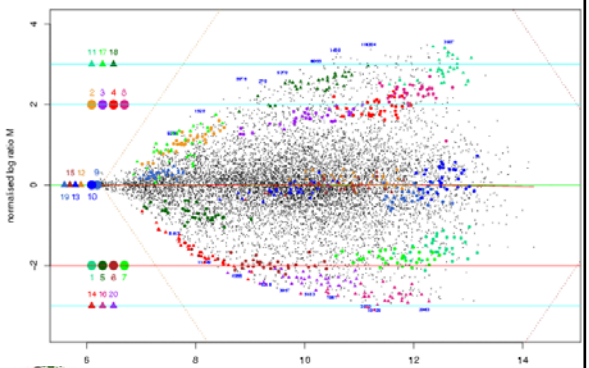


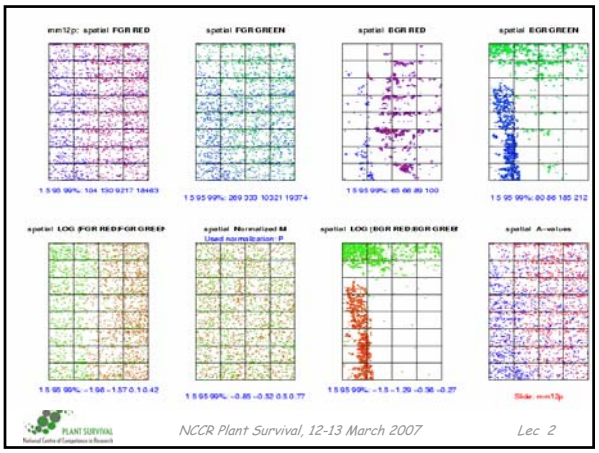
After

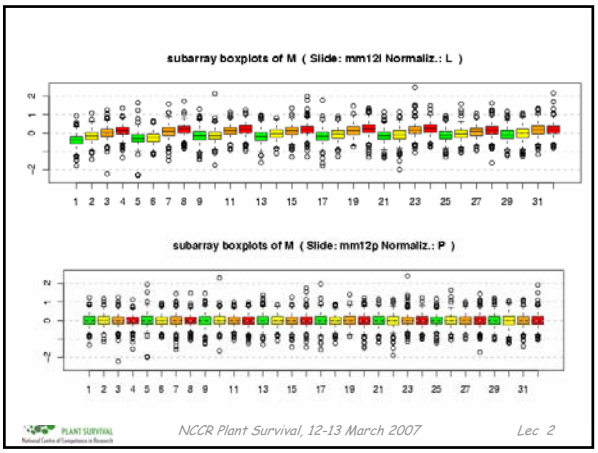


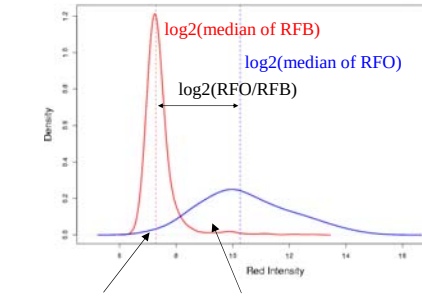


MvA plot from Agilent scanner





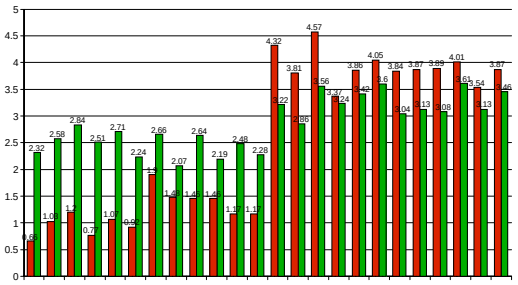


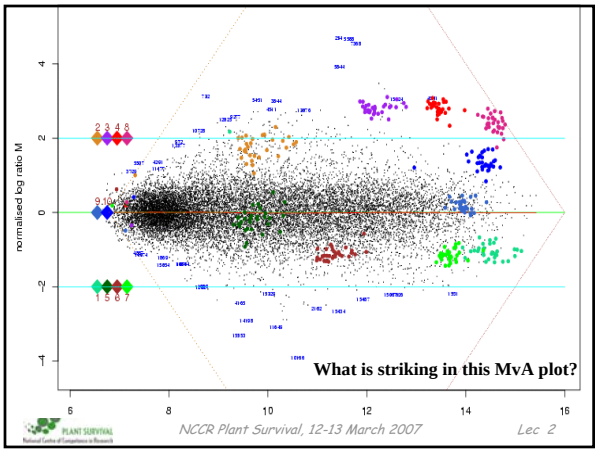


Number of good spots below the median of blanks

Number of good spots above the median of blanks

ratio of raw red fgr medians good / blanks (log scale) in red
ratio of raw green fgr medians good / blanks (log scale) in green





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
- 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 *p* 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)

Software for Microarray Analysis

- Very large number of commercial and free softwares
- There are several R packages for microarray analysis available as part of the open source BioConductor project http://www.BioConductor.org/
- BioC software often created by the author of the methodology
- We will be using some of the BioConductor packages in this course, especially the GUI versions of the package `limma`