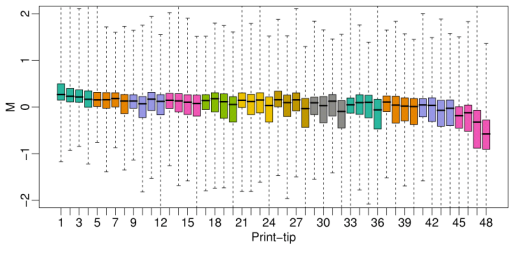
Quality Assessment

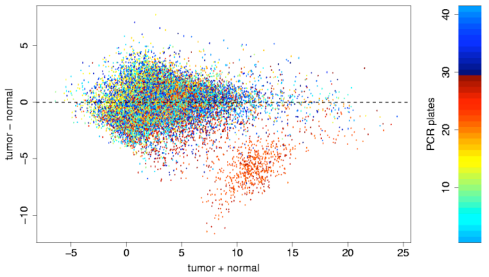
Credit for some of today's materials:
Ben Bolstad

Exploratory Data Analysis

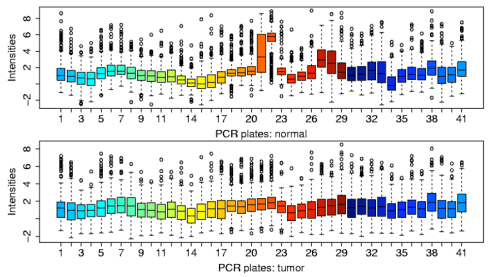
Print-tip Effect



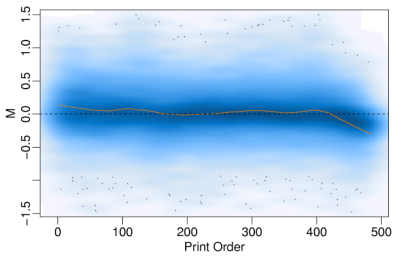
Bad Plate Effect



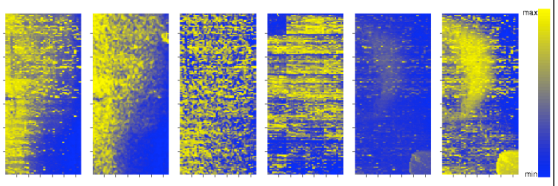
Bad Plate Effect



Print Order Effect



Spatial Effects



R Rb R-Rb
color scale by rank

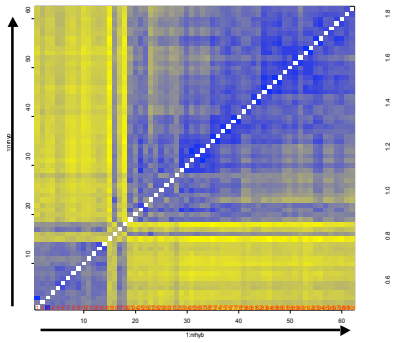
another array: print-tip

color scale ~ log(G)

color scale ~ rank(G)

spotted cDNA arrays, Stanford-type

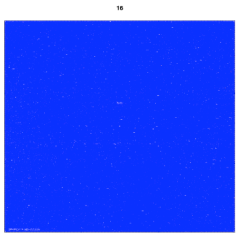
Batches: array to array differences $d_{ij} = mad_k(h_{ik} - h_{jk})$



arrays i=1…63; roughly sorted by time

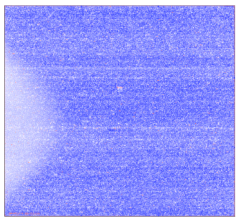
Affymetrix Chips

Images of probe level data



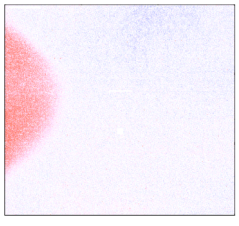
This is the raw data

Images of probe level data



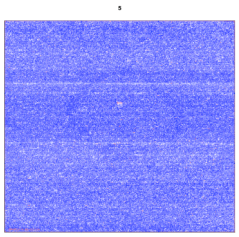
Log scale version much more informative

Images of probe level data



Residuals (or weights) from probe level model fits show problem clearly

Images of probe level data



Here is a more subtle artifact. Can you see it? The strong probe effect does not let you.

QC from probe level models

- **RMA fits a probe level model**
- **From these fits we can obtain residuals**
- **We can also get weights if we use formal robust regression procedures instead of median polish**
- **These probe-level residuals and summaries of their size can be used for quality control**
- **Software available: `affyPLM` Bioconductor package (Ben Bolstad)**

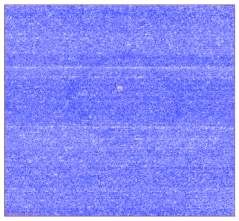
Using the robust fit

- **Assume additive model**

$$\log_2 y_{ij} = \mu_i + \alpha_j + \varepsilon_{ij}$$

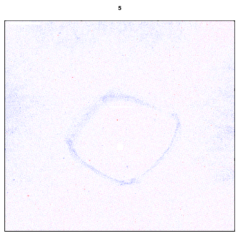
- **Estimate mu gives RMA**
- **Use M-estimators**
- **To avoid showing the variability introduced by expression and probe effect we plot the residuals**
- **We can also plot the weights used by the regression**

Images of probe level data



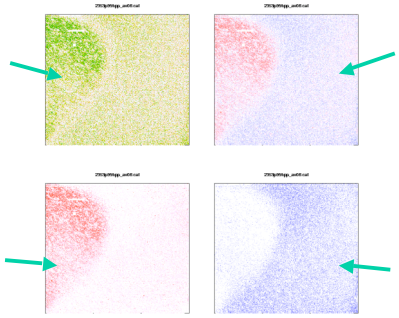
Here is a more subtle artifact. Can you see it?
The strong probe effect does not let you.

Images of probe level data



Probe level fit residuals really show it

Other pseudo-chip images



Can we report summaries?

RLE

- **Relative Log expression simply makes boxplots of the log expression - median across chip for all genes**
- **Notice this is not specific to RMA**

NUSE

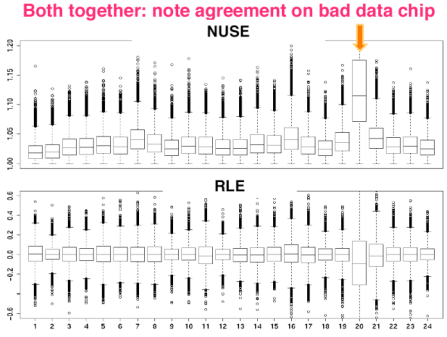
- **Normalized Unscales Standard Errors**

$$\hat{\mu}_i = \sum_j y_{ij} \cdot \frac{w_{ij}}{W_i} \quad \text{and} \quad SE(\hat{\mu}_i) = \frac{\hat{\sigma}}{\sqrt{W_i}}.$$

$$NUSE(\hat{\mu}_i) = \frac{\hat{USE(\mu_i)}}{\text{Median}_i\{USE(\hat{\mu}_n)\}} = \frac{1}{\sqrt{W_i}} \Big/ \text{Median}_n\left\{\frac{1}{\sqrt{W_n}}\right\}.$$

$$NUSE(\hat{\mu}_i) \approx \frac{1}{\sqrt{W_i}} \Big/ \frac{1}{\text{Median}_n\{\sqrt{W_n}\}} = \frac{\text{Median}_n\{\sqrt{W_n}\}}{\sqrt{W_i}} = \left(\frac{\sqrt{W_i}}{\text{Median}_n\{\sqrt{W_n}\}}\right)^{-1}$$

Both together: note agreement on bad data chip



NUSE