Normalizing cDNA microarray data

- There are many sources of systematic variation in microarray experiments which affect the measured gene expression level.
- *Normalization is the term used to describe the process of removing bias due to*
  - differential incorporation of dyes
  - different amounts of mRNA
  - different scanning properties or parameters
  - spatial effects
    e.g., bent pin heads → print-tip effects
  - ...
- Aim is to balance the red and green intensities.

How should we normalize?

- It can be done in a number of ways, depending on the experimental setup.
- We distinguish[1]
  - *location* and *scale* normalization within a single slide;
  - *location* and *scale* normalization across multiple slides;
  - *self-normalization* for dye-swapped experiments;
  - *microarray sample pool normalization* based on a control sample ensemble; and
  - *composite normalization*.

[1] Yang et al 2001, 2002

Within-slide normalization

- **Location:**
  - ◇ standard practice is *global normalization* which forces the M's to have 0 mean or median;
    - ★ it is assumed that intensities are related by a constant factor ($R = kG$), so that

$$\log_2 \frac{R}{G} \rightarrow \log_2 \frac{R}{G} - \log_2 k$$

    - ★ But this is inadequate in situations where dye biases depend on *overall spot intensity* and *location* on the array.

- ⋆ *Why?*
  Because interest is in differential expression, but the differential is intensity and location dependent.

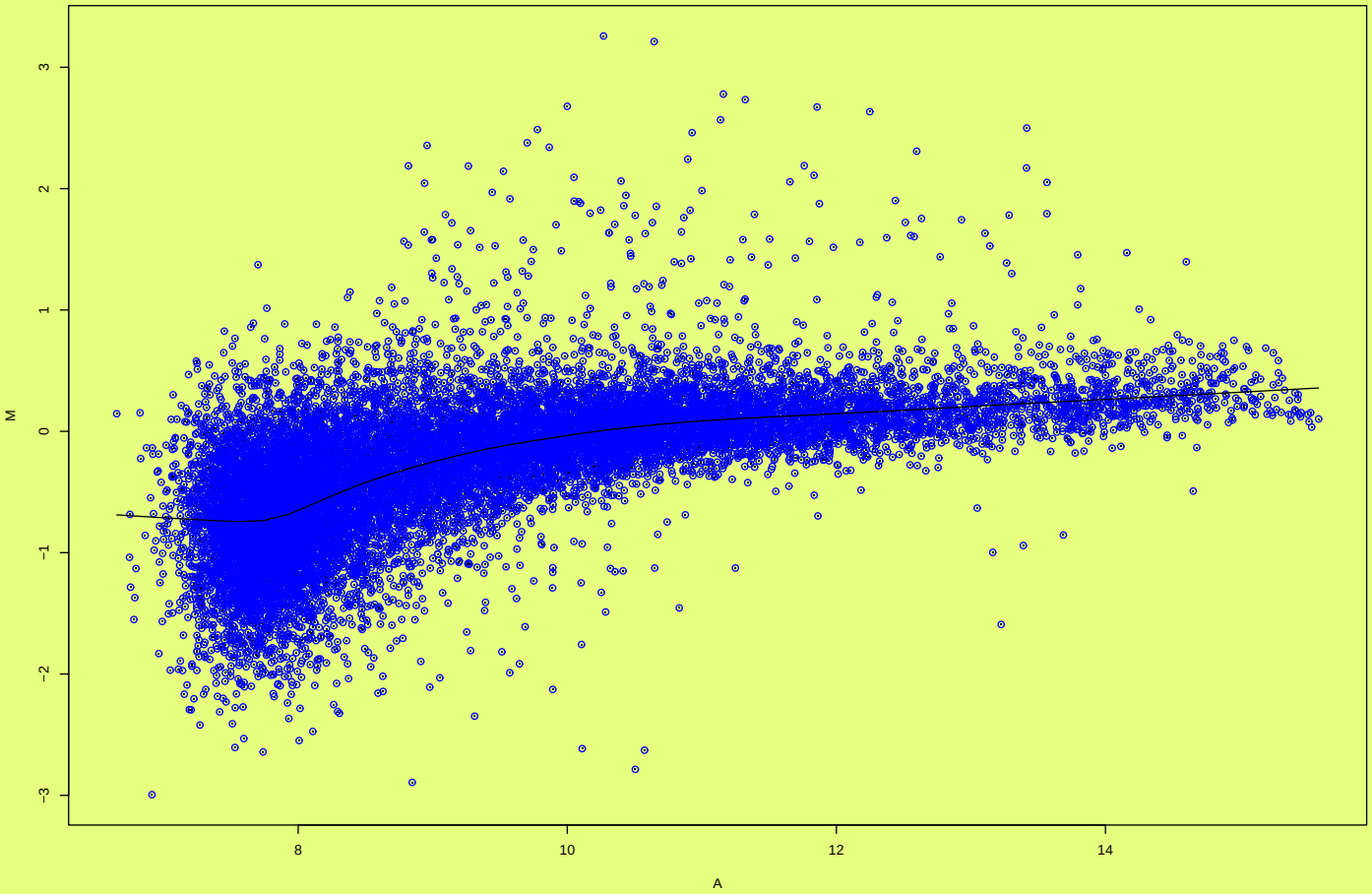
◇ We normalize in an *intensity-dependent* way:

- ⋆ in R, fit a robust scatterplot smoother called `lowess` to the M versus A plot:

$$\log_2 \frac{R}{G} \rightarrow \log_2 \frac{R}{G} - c(A)$$

  where $c(A)$ is the `lowess` fit to the M versus A plot.
- ⋆ The `lowess` curve becomes the new zero line.

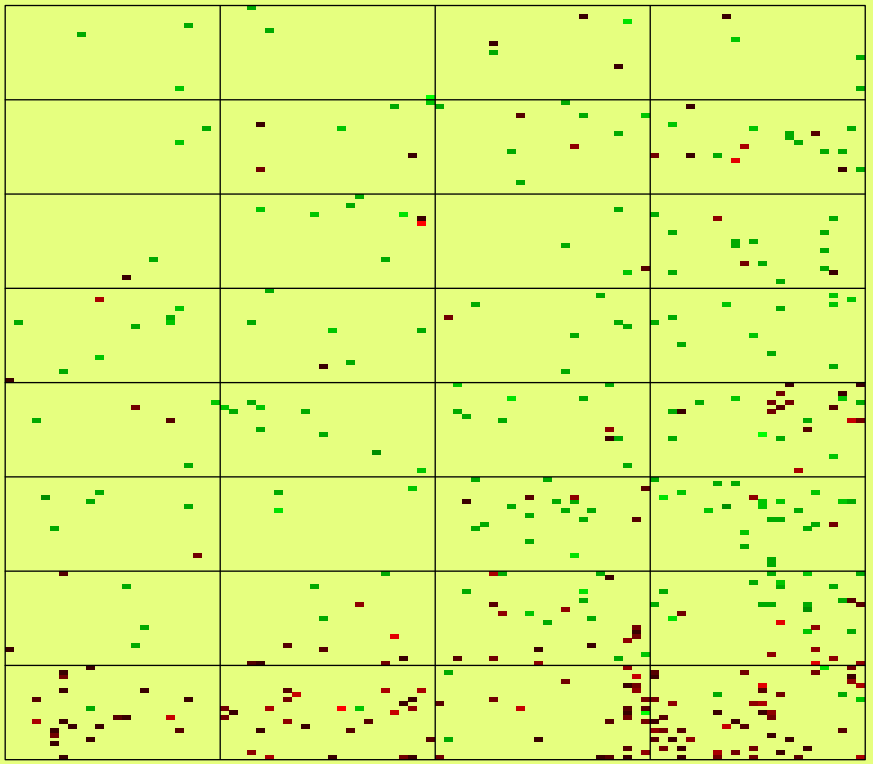
M is intensity dependent



We normalise in an A-dependent way.

Dr Patty Solomon, University of Adelaide ©2002

M can be spatially dependent



Print-tip effects

Dr Patty Solomon, University of Adelaide ©2002

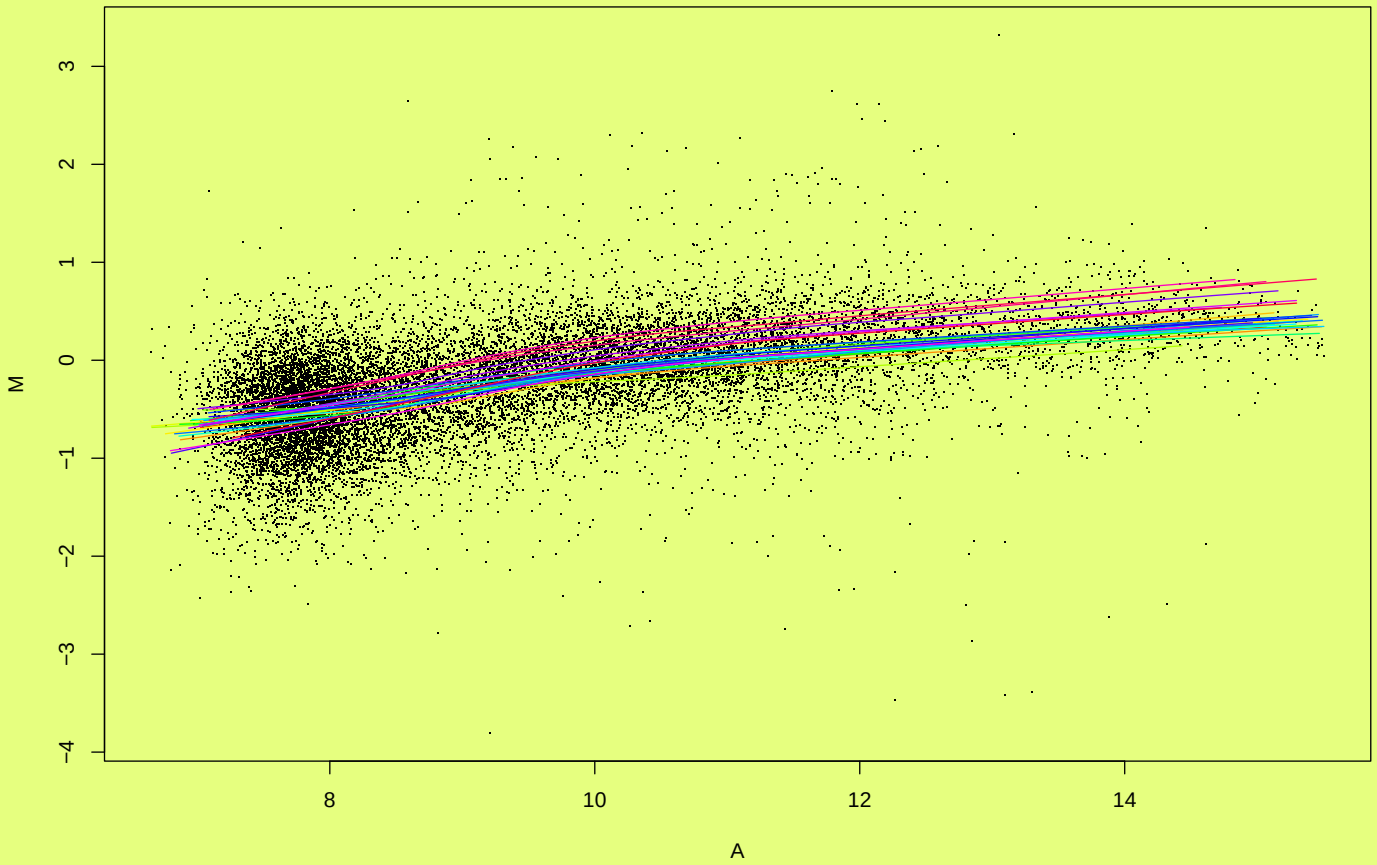
Print-tip normalization

- **Location:**
  - ◇ fit lowess curve to each print-tip group.
- **Scale:**
  - ◇ Assume all the log ratios from ith print tip group $\sim N(0, a_i^2\sigma^2)$
    - ⋆ Estimate the scale factors a_i by maximum likelihood:

$$\hat{a}_i = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{\sqrt[I]{\prod_{k=1}^{J}\sum_{j=1}^{n_i} M_{kj}^2}}$$

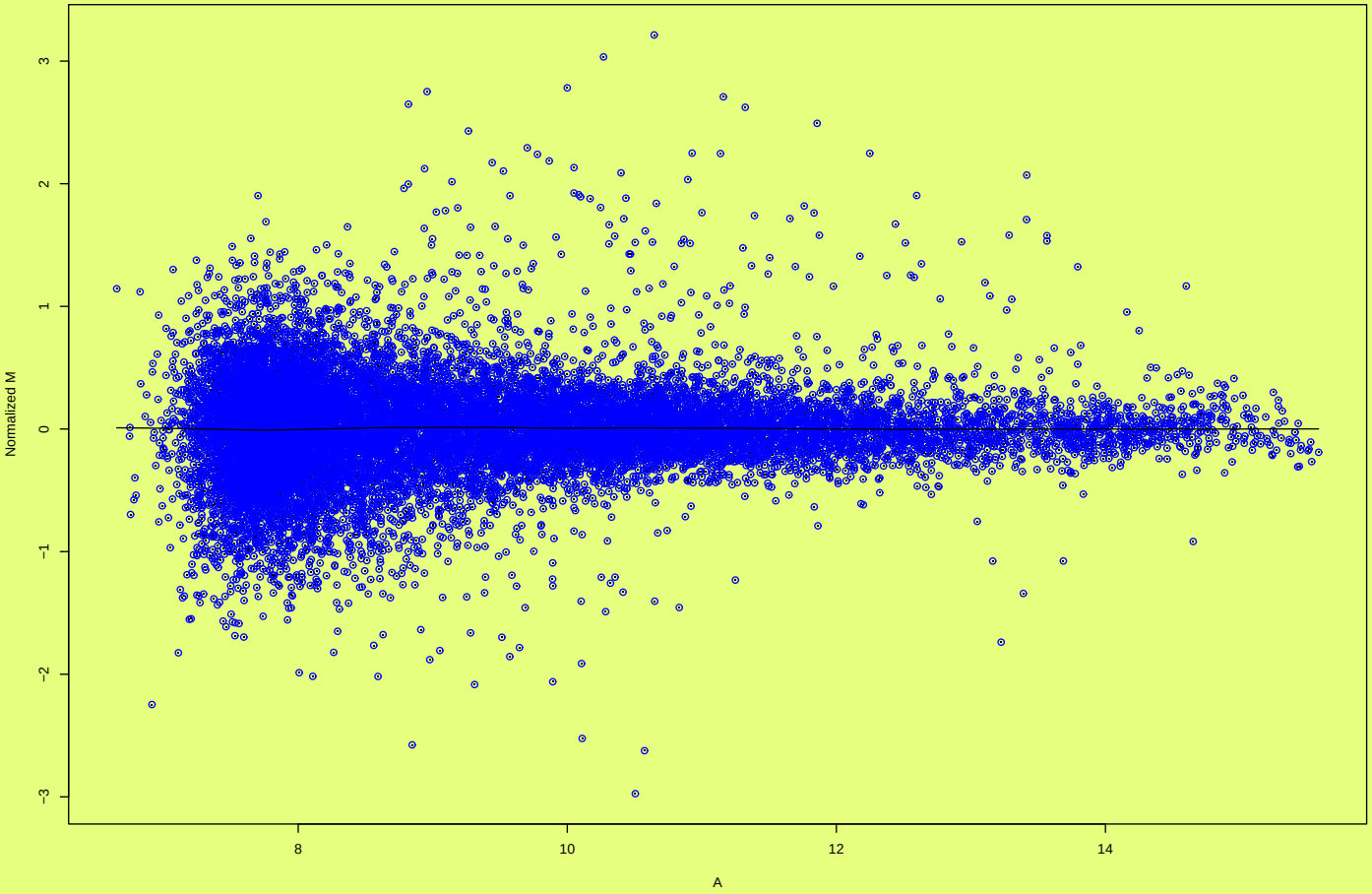
    - ⋆ In practice, we use a robust estimate, then eliminate the $\hat{a}_i$'s.

Lowess curves fitted to each print-tip group



Dr Patty Solomon, University of Adelaide ©2002

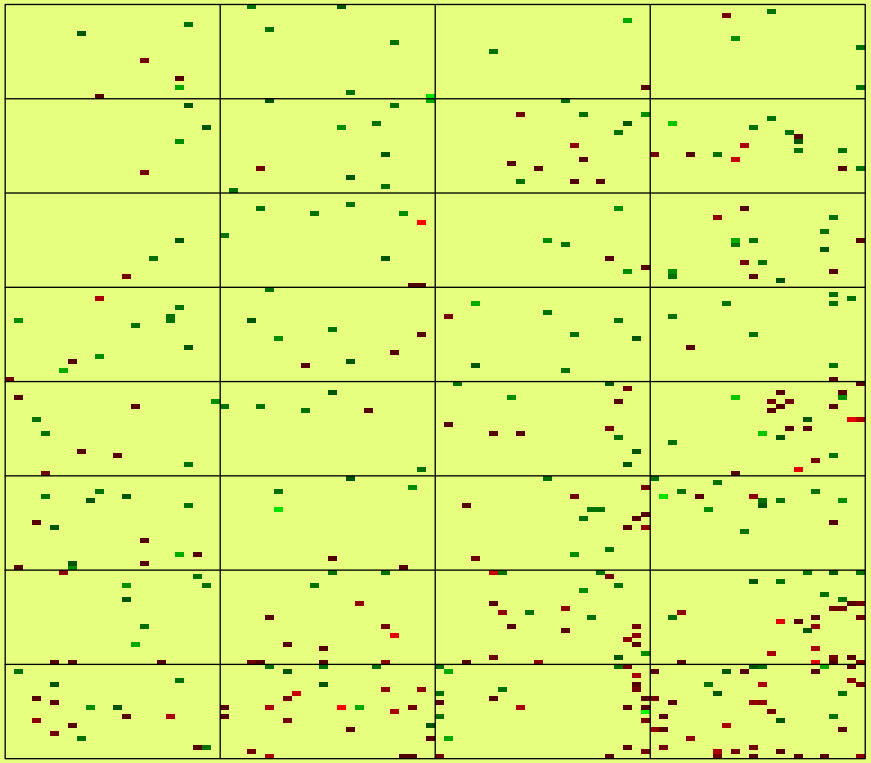
After print-tip normalization



Changes are roughly symmetric about zero.

Dr Patty Solomon, University of Adelaide ©2002

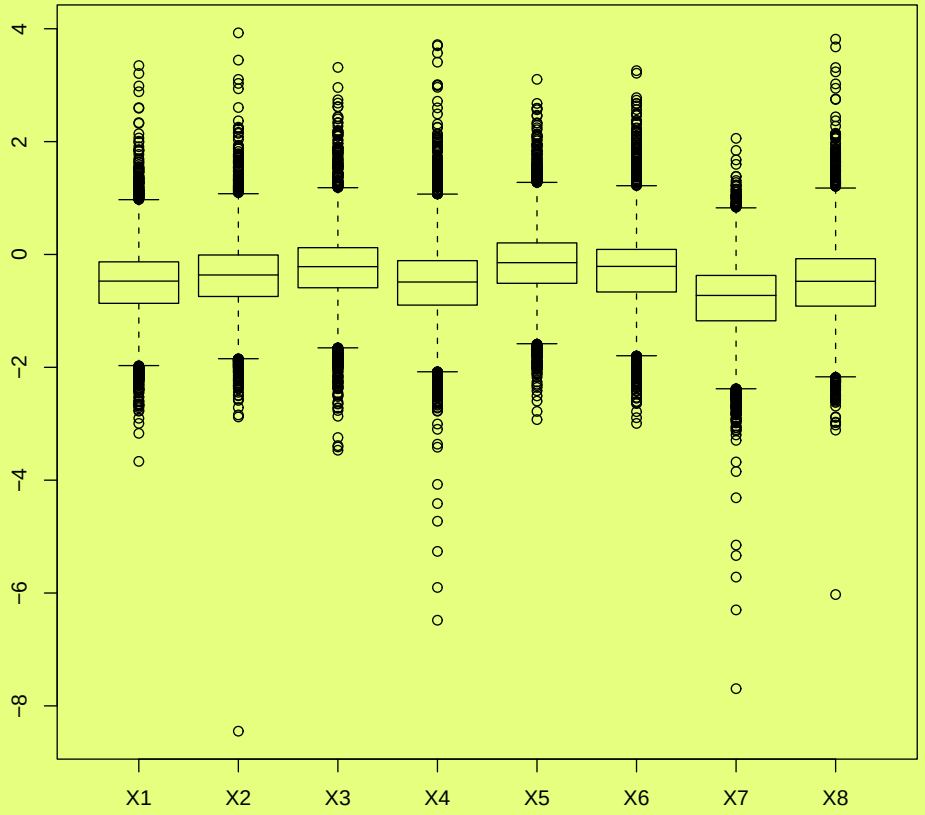
After print-tip normalization



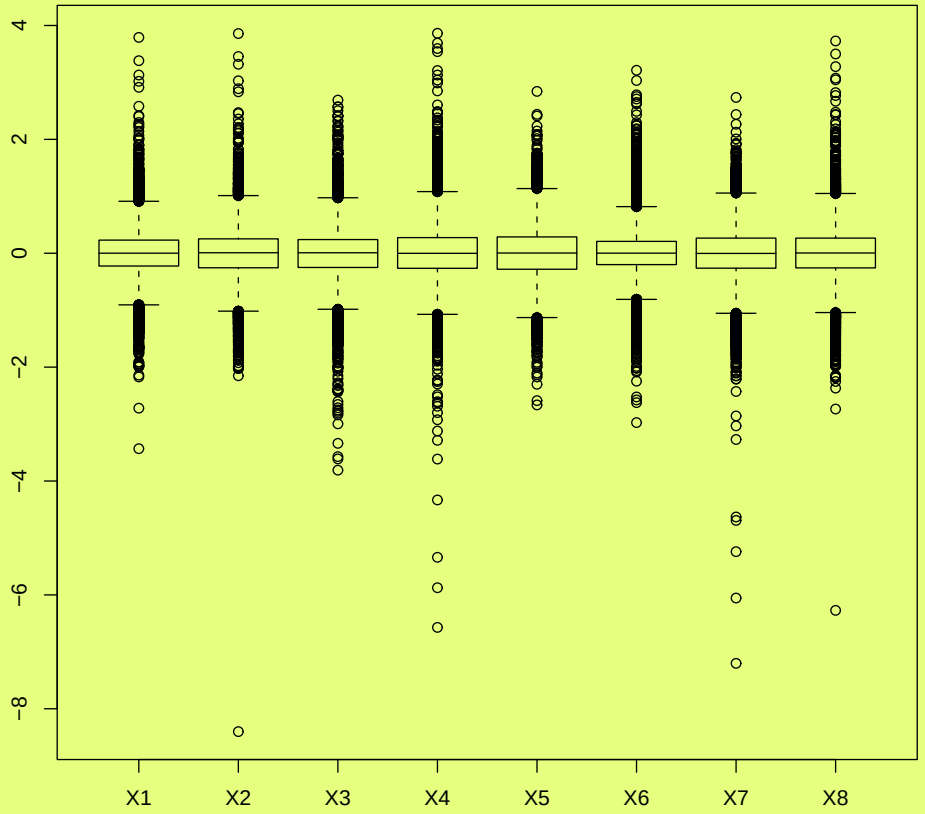
Normalizing across arrays

- After within-slide normalization, all log ratios will be centred around zero.
- If arrays have different *spreads*, may need to perform *scale* normalization as well.
  - Can apply same principles used for *within-slide print-tip scale normalization*.
  - In practice, the need for scale adjustment across slides is determined empirically.
  - Research is underway to develop improved procedures for scale adjustment.
- → *Bias–variance trade-off.*

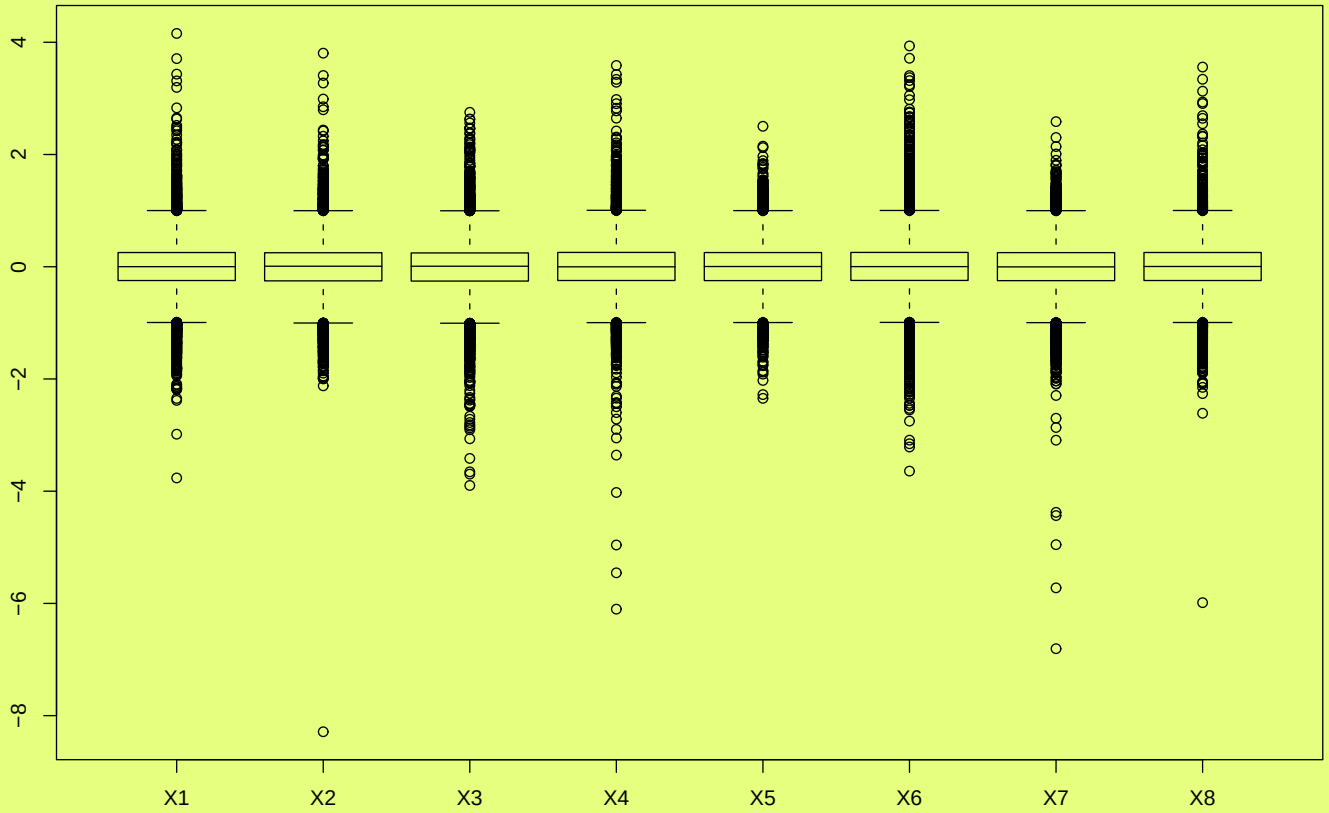
Multiple arrays before location and scale normalization



Multiple arrays after print-tip location and scale normalization



Multiple arrays after scale normalization



Self-normalization

- If the experiment is replicated (and it should be) use *dye-swapped replicates*:
  - ◇ *2 hybridizations for 2 mRNA samples with dye assignment reversed in the second hybridization (Latin square)*
    - ⋆ For each gene, get M and M'.
    - ⋆ Dye-swapped replicates are like ordinary replicates, but in addition, allow direct measurement of the dye bias.
  - ◇ *Self-normalization*: assuming the normalization function is the same in the two slides, we can estimate the combined relative expression level by

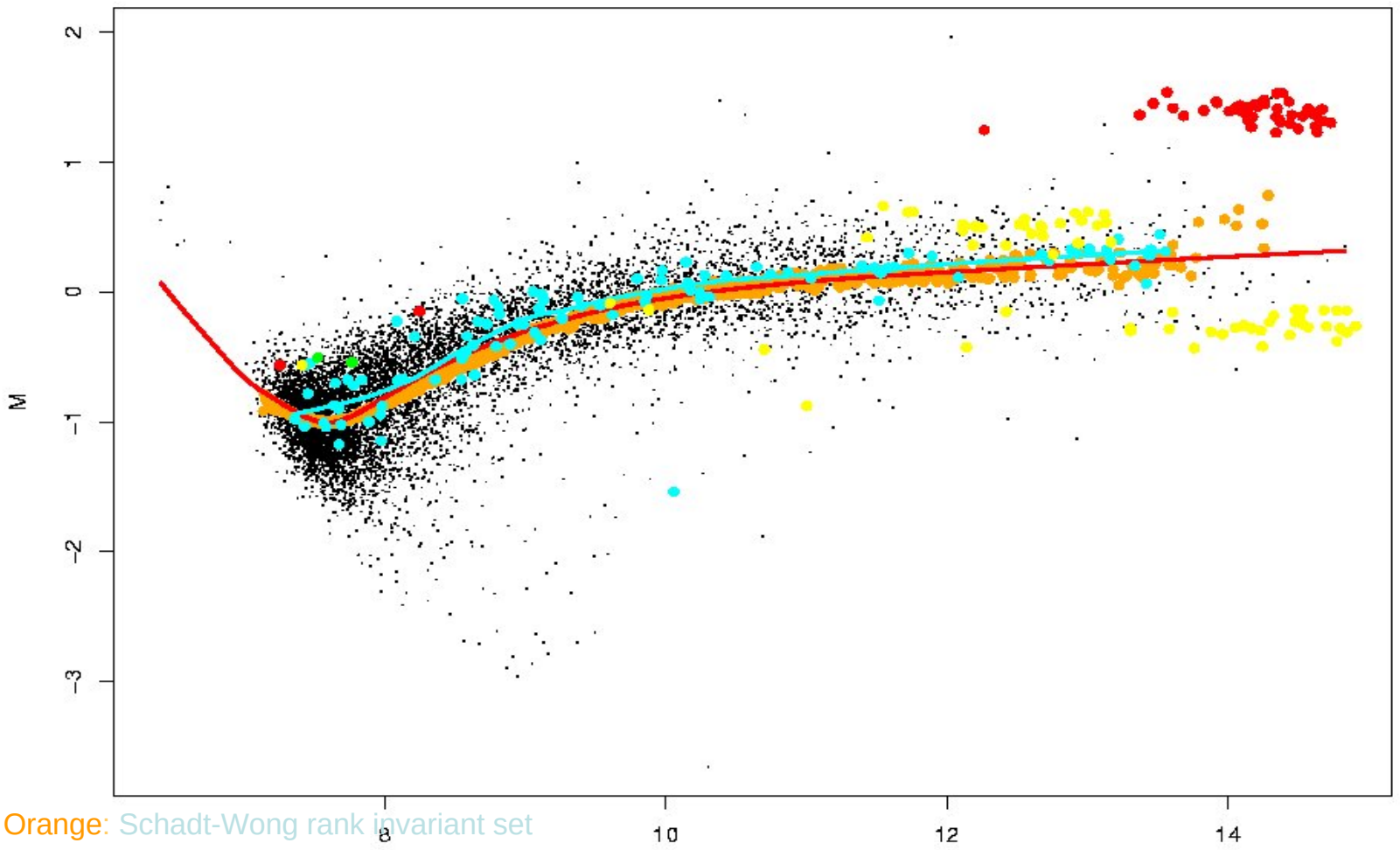
$$(M - M')/2$$

Which genes to use for normalization

- *When*
  - only a small proportion expected to be differentially expressed in the two samples, *or*
  - there is symmetry in the expression levels of the up/down regulated genes,
  - use all genes on array *or* self-normalization.
- *When many genes expected to change*, can use
  - self-normalization based on dye-swapped replicates, *or*
  - Microarray Sample Pool (MSP) controls which span the intensity range and are 'constant' across biological samples.

Dr Patty Solomon, University of Adelaide ©2002

Microarray Sample Pool



Orange: Schadt-Wong rank invariant set

Red dots : 18S rRNA, Red line: lowess smooth; **Yellow: GAPDH, tubulin**;, Light blue: DNA pool/titration.

Microarray sample pool

- Advantages:
  - Mimics yeast genomic DNA.
  - Titration series covers whole intensity range.
  - Relatively constant expression level.
  - Potentially, all labelled cDNA sequences can hybridize → minimal sample-specific bias.
- Disadvantages:
  - May produce less stable estimates in context of spatial normalization, since have only small number of MSP spots per print-tip group.
- This leads to *composite normalization*.

Composite normalization

- It is a *weighted average* of the MSP lowess curve $g(A)$ and the within-print-tip group lowess curve $f_i(A)$ for the ith print-tip group:

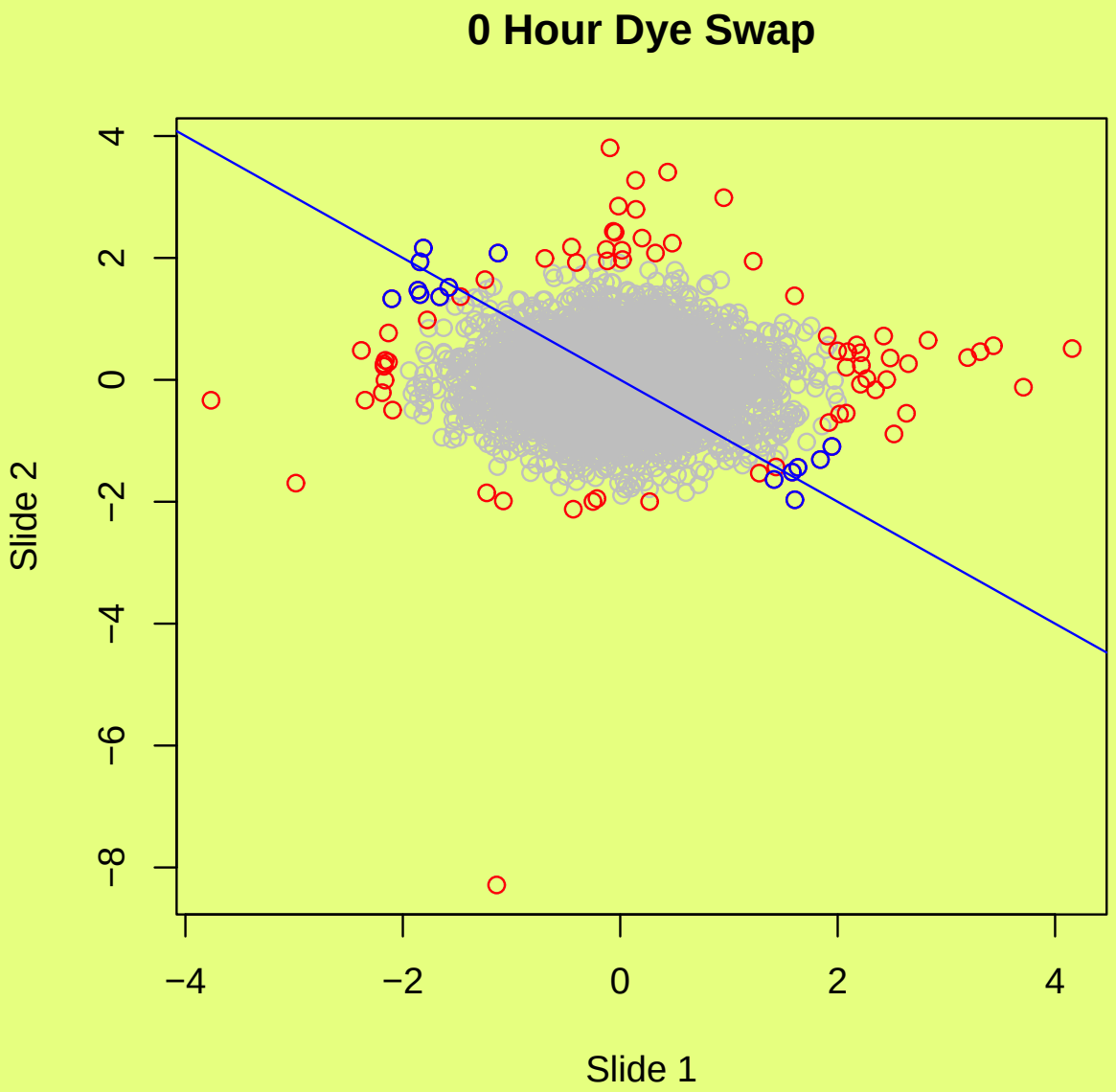
$$c_i(A) = \alpha_A \hat{g}(A) + (1 - \alpha_A)\hat{f}_i(A)$$

where α_A is proportion of genes less than a given intensity A.

  - ◇ *In practice, composite normalization recommended for genetically divergent mRNA samples*
    - ★ evident in *increased* spread of log ratios at high intensities.

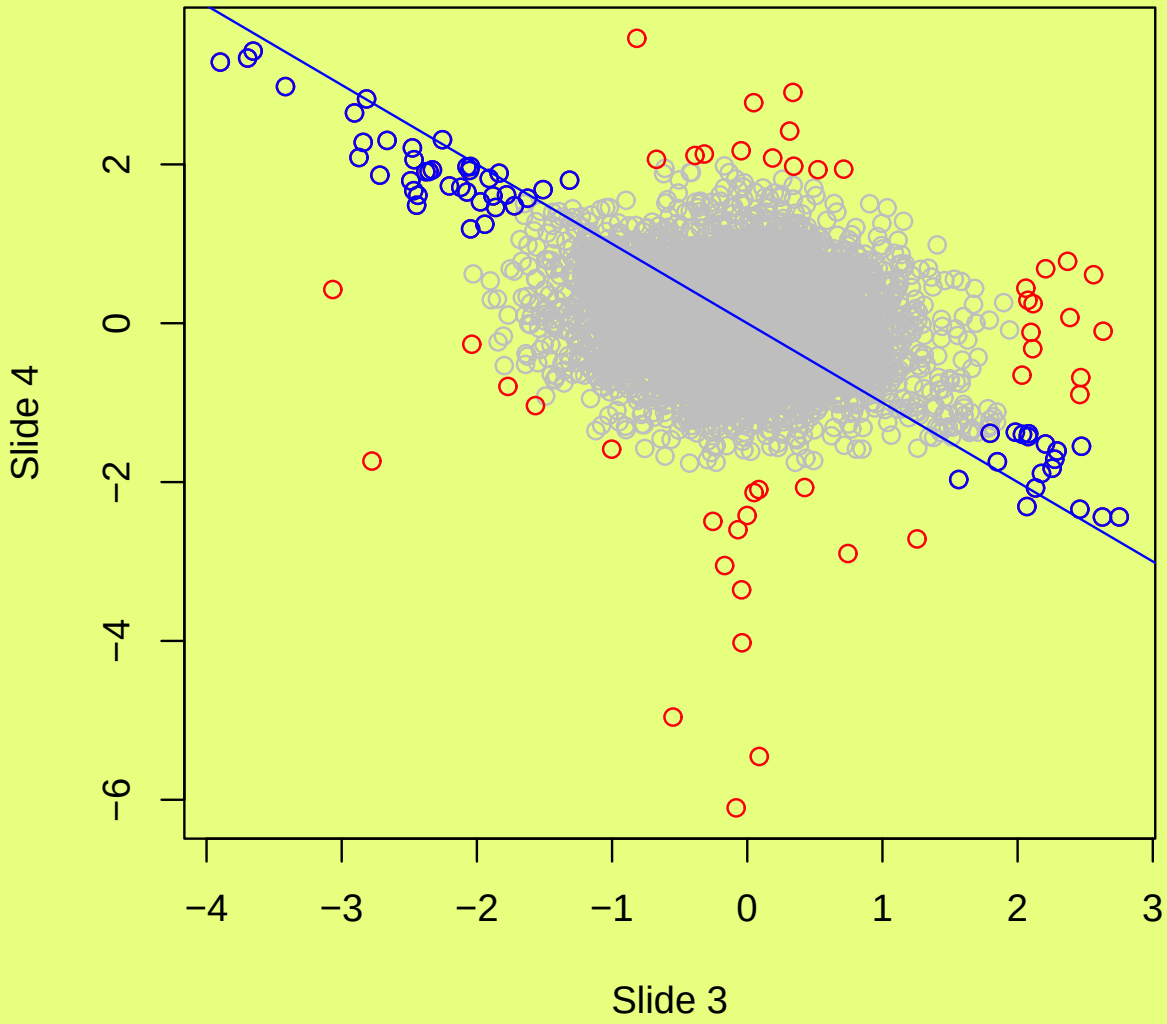
A simple discriminant analysis

- Comparison of 2 mutant cells lines N and L in leukaemic mice at time 0 hours and time 24 hours based on *Mahalanobis distance*:
  - $N0$ and $L0$ compared using dye-swapped replicates
    - Slide 1 N is labelled G and L is labelled R
    - Slide 2 N is labelled R and L is labelled G;
  - $N24$ and $L24$ compared using dye-swapped replicates
    - Slide 3 N is labelled G and L is labelled R
    - Slide 4 N is labelled R and L is labelled G.



Dr Patty Solomon, University of Adelaide ©2002

24 Hour Dye Swap



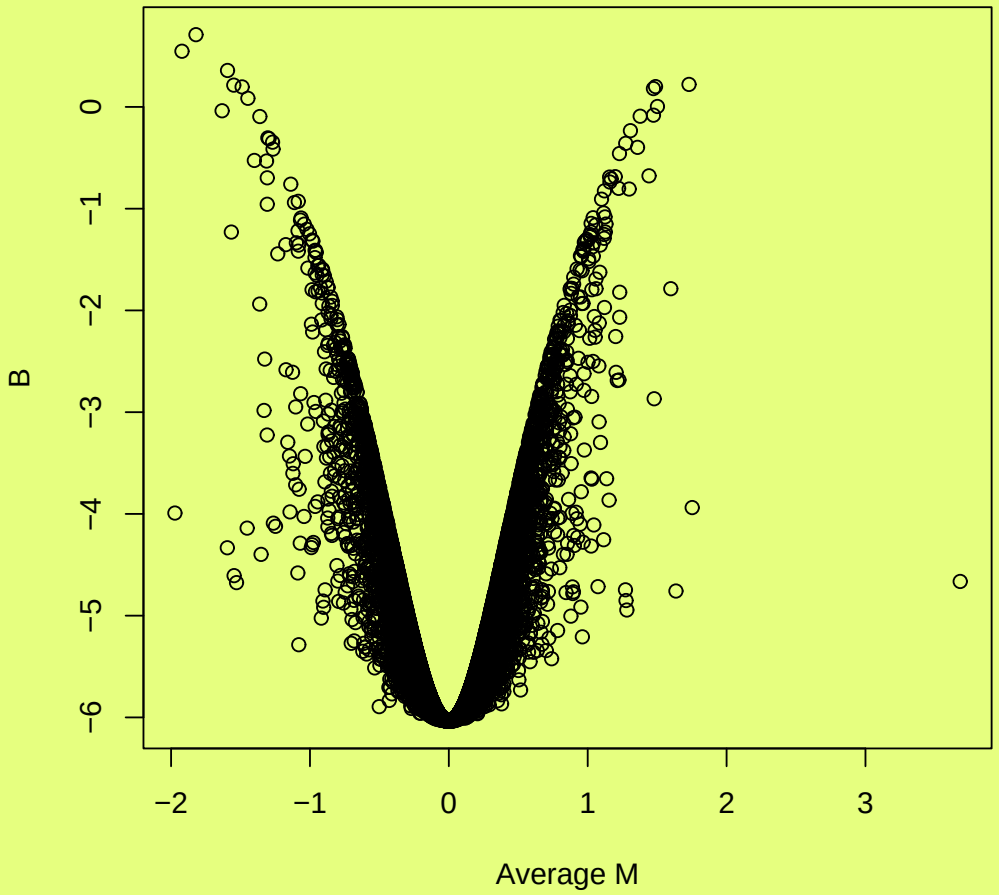
Dr Patty Solomon, University of Adelaide ©2002

An empirical Bayes analysis

- *Idea:* the information from all genes is combined to estimate a statistic B for *each* gene.
  - B is the *log posterior odds of differential expression*[2] and provides an alternative estimator to M, or to statistics based on M.
  - Useful when have small number of replicates per gene, and many genes.
  - *Consider previous example, reversing sign on one of the dye-swapped replicates.*

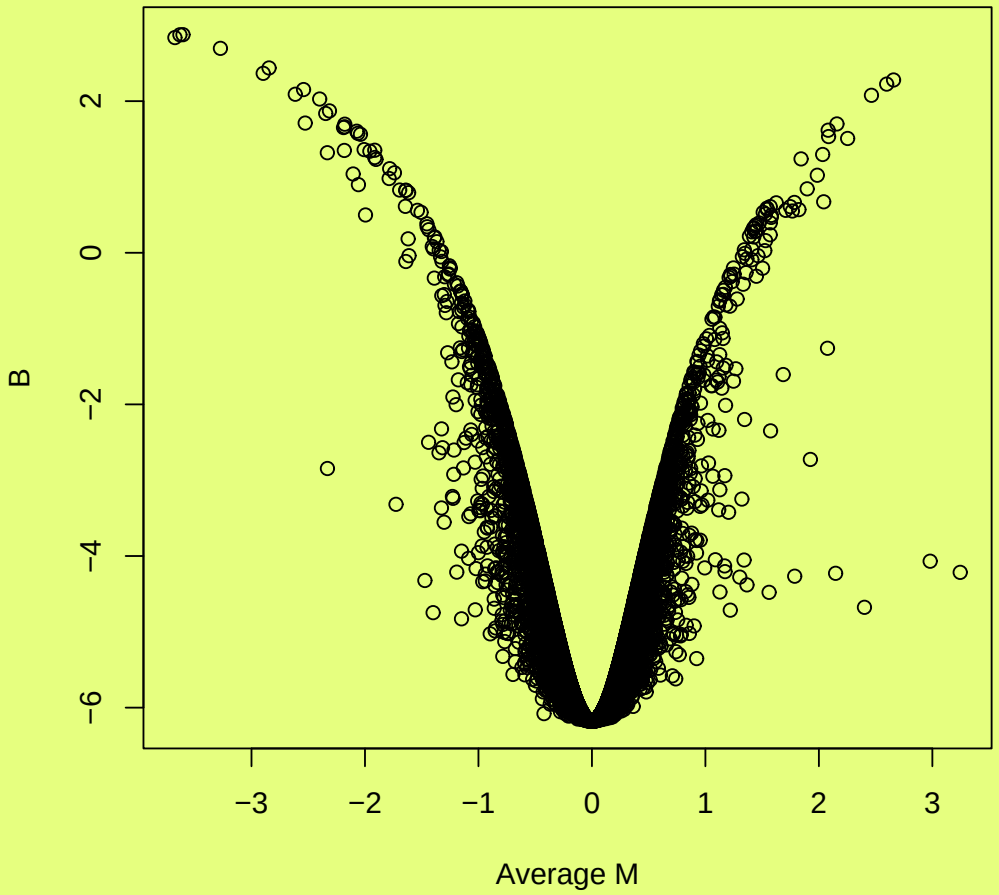
[2] Lönnstedt & Speed 2002

Comparing cell samples at time 0 hours



Dr Patty Solomon, University of Adelaide ©2002

Comparing cell samples at time 24 hours



Dr Patty Solomon, University of Adelaide ©2002

Further references

- Bioconductor Project. Open source bioinformatics using R: http://www.bioconductor.org
- sma v. 0.5.6 (November 2001), B Bolstad, S Dudoit, YH Yang. http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html
- S Dudoit, YH Yang, MJ Callow. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 2002.
- S Dudoit, YH Yang, B Bolstad. Using R for the analysis of DNA microarray data. *R News* **2**, 24–32, 2002.
- I Lönnstedt, TP Speed. Replicated microarray data. *Statistica Sinica*, in press.

- TP Speed. Wald Lectures, 2001: Slide 16. www.stat.berkeley.edu/users/terry/zarray
- YH Yang, S Dudoit, P Luu, TP Speed. Normalization for cDNA microarray data. In ML Bittner, Y Chen, AN Dorsel, ER Gougherty (editors) *Microarrays: optical technologies and informatics* **4266** *Proceedings of SPIE*, May 2001.
- YH Yang, S Dudoit, P Luu, DM Lin, V Peng, J Ngai, TP Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, 2002.

THE END

Dr Patty Solomon, University of Adelaide ©2002