Design and analysis of microarray experiments

Patty Solomon

Department of Applied Mathematics and
Centre for the Molecular Genetics of Development
University of Adelaide

26 July 2002

http://www.maths.adelaide.edu.au/people/psolomon

patty.solomon@adelaide.edu.au

Microarray Analysis Group

- University of Adelaide
  - Patty Solomon
  - Gary Glonek
  - Jonas Lloyd (Ph.D.)
  - Michael Calvert (Grad. Dip.)
  - Adam Kister (Grad. Dip./Ph.D.)
- Hanson Institute, IMVS
  - Anna Tsykin

Microarray Analysis Group

- University of Adelaide
  - Patty Solomon
  - Gary Glonek
  - Jonas Lloyd (Ph.D.)
  - Michael Calvert (Grad. Dip.)
  - Adam Kister (Grad. Dip./Ph.D.)
- Hanson Institute, IMVS
  - Anna Tsykin
- http://www.maths.adelaide.edu.au/MAG
  - ★ Preprints, talks, workshop materials, links, ...

Major collaborative projects

- Aim to understand the genetic basis of disease, and it is hoped, to find treatments or cures:
  - ◇ Osteoarthritis
    - ⋆ Nick Fazzalari (IMVS), David Findlay (RAH)
  - ◇ Leukaemia, cancer, asthma . . .
    - ⋆ Richard D'Andrea, Brenton Reynolds (WCH), Tom Gonda (Bionomics)
    - ⋆ Mark Guthridge (IMVS)
  - ◇ Angiogenesis
    - ⋆ Chris Hahn, Jenifer Gamble (Hanson Institute, IMVS)

Major collaborative projects

- Aim to understand the genetic basis of disease, and it is hoped, to find treatments or cures:
  - Osteoarthritis
    - Nick Fazzalari (IMVS), David Findlay (RAH)
  - Leukaemia, cancer, asthma . . .
    - Richard D'Andrea, Brenton Reynolds (WCH), Tom Gonda (Bionomics)
    - Mark Guthridge (IMVS)
  - Angiogenesis
    - Chris Hahn, Jenifer Gamble (Hanson Institute, IMVS)
  - Animal and plant breeding
    - Waite, Roseworthy
  - . . .

So, what do we do?

So, what do we do?

- *Design microarray experiments.*
  - Glonek & Solomon 2002.

So, what do we do?

- *Design microarray experiments.*
  - ⋆ Glonek & Solomon 2002.
- *Image processing:*
  - ⋆ quantifying expression.
- *Normalisation:*
  - ⋆ Removing dye and other systematic biases.
  - ⋆ Robust normalisation, Calvert et al.
- *Statistical analysis:*
  - ⋆ Detecting differential or co-expression in complex experiments
  - ⋆ Reynolds et al., Lloyd et al., Cox & Solomon 2002.
- *No single method of analysis can be appropriate for all experiments.*

What do we mean by 'design' for microarray experiments?

- Which mRNA samples should be competitively hybridised on the same slide?*
  - Should samples from individual animals or people be compared *directly* or via a common *reference* mRNA sample?*

What do we mean by 'design' for microarray experiments?

- Which mRNA samples should be competitively hybridised on the same slide?*
  - Should samples from individual animals or people be compared *directly* or via a common *reference* mRNA sample?*
- Should tissue samples from animals be *pooled* then compared, or should different animals be hybridised to different slides?*
- Which sample should be labelled with one dye and which with the other?
  - Should dye-swapped replicates be made on different extractions?

- How many replicates should there be of each gene within an array?
- How many times should each array be replicated?*
- …

- How many replicates should there be of each gene within an array?
- How many times should each array be replicated?*
- …
- *There are few published studies addressing these issues:*
  - *Glonek & Solomon 2002: *'The first careful treatment of optimal design for factorials'* (Yang & Speed 2002). Invited paper ASC 16.

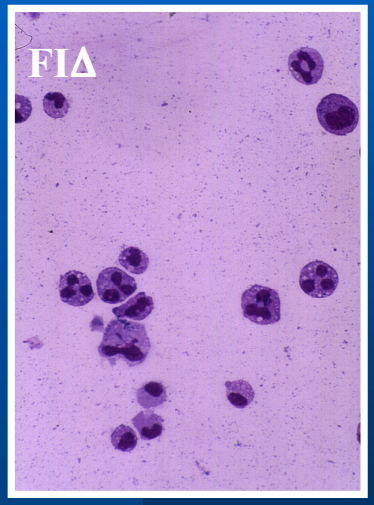
- How many replicates should there be of each gene within an array?
- How many times should each array be replicated?*
- ...

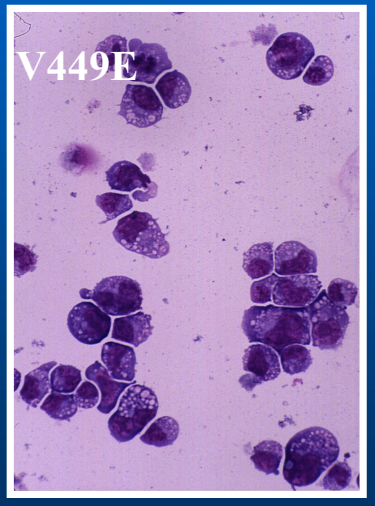
- *There are few published studies addressing these issues:*
  - *Glonek & Solomon 2002: *'The first careful treatment of optimal design for factorials'* (Yang & Speed 2002). Invited paper ASC 16.

- *Our premise is that the most appropriate way to find differentially expressed genes is to prescribe a design subject to*
  - the key contrasts and parameters of interest
  - and the practical constraints of the problem.

Design case study

Biological function of the activated mutants in FDB cells





⇒ Different signals generated by the two mutants

- *Aim: to identify genes that play an important role in receptor signalling and leukaemogenesis in mutant mice.*

A 2x2 factorial design

- *Hypothesis:* that there is a set of genes induced specifically in response to expression of V449E that results in its leukaemic effects.
  - ◇ Compare two cell populations: FIΔ and V449E at two times 0 hours and 24 hours

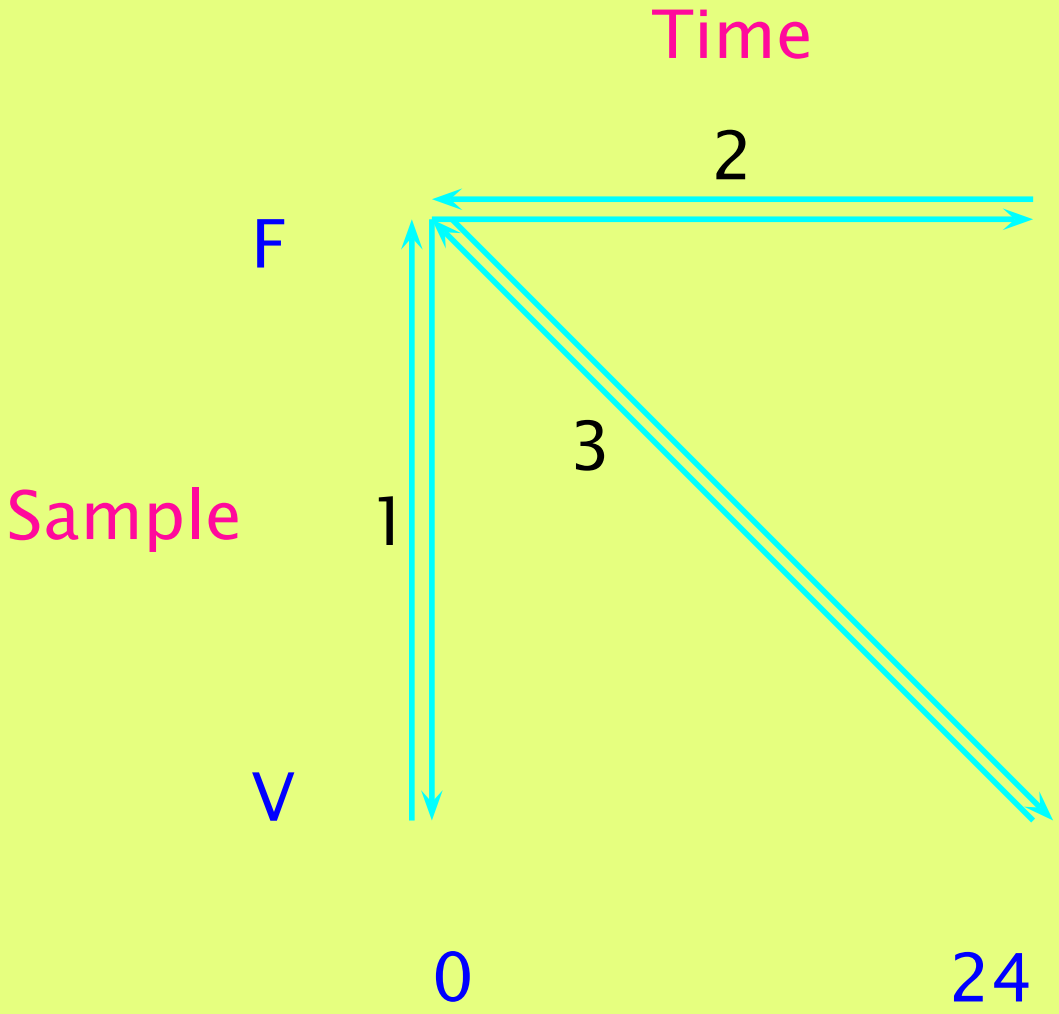
A 2x2 factorial design

- *Hypothesis:* that there is a set of genes induced specifically in response to expression of V449E that results in its leukaemic effects.
  - ◇ Compare two cell populations: FIΔ and V449E at two times 0 hours and 24 hours
    - ⋆ $\longrightarrow$ 4 combinations $F0, V0, F24, V24$.
    - ⋆ $\longrightarrow$ 2×2 factorial design of block size 2.

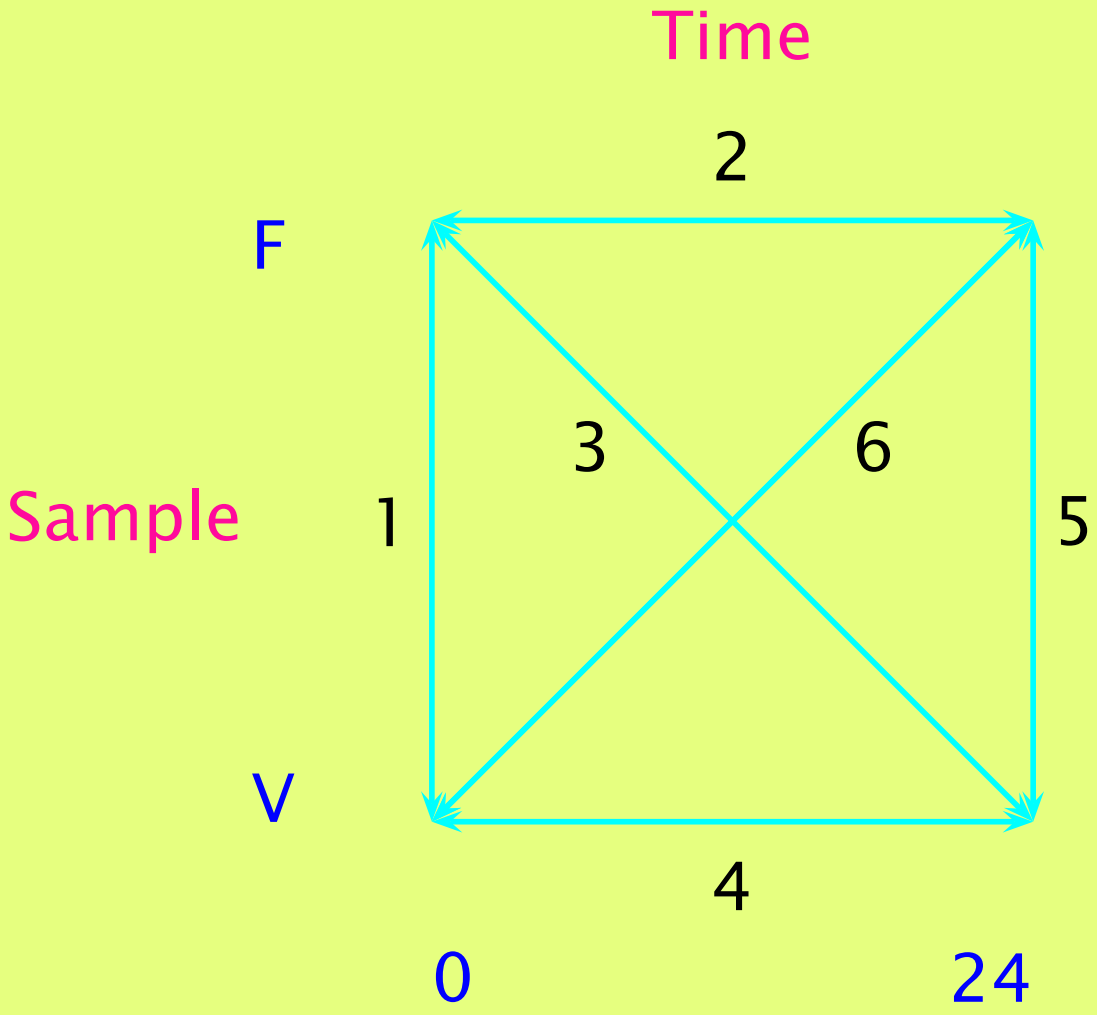
A 2x2 factorial design

- *Hypothesis:* that there is a set of genes induced specifically in response to expression of V449E that results in its leukaemic effects.
  - ◇ Compare two cell populations: FIΔ and V449E at two times 0 hours and 24 hours
    - ⋆ $\longrightarrow$ 4 combinations $F0, V0, F24, V24$.
    - ⋆ $\longrightarrow$ 2×2 factorial design of block size 2.
- *Interaction of primary interest:*
  e.g. genes that are differentially expressed in the two samples at time 24 hours, but not at time 0 hours.

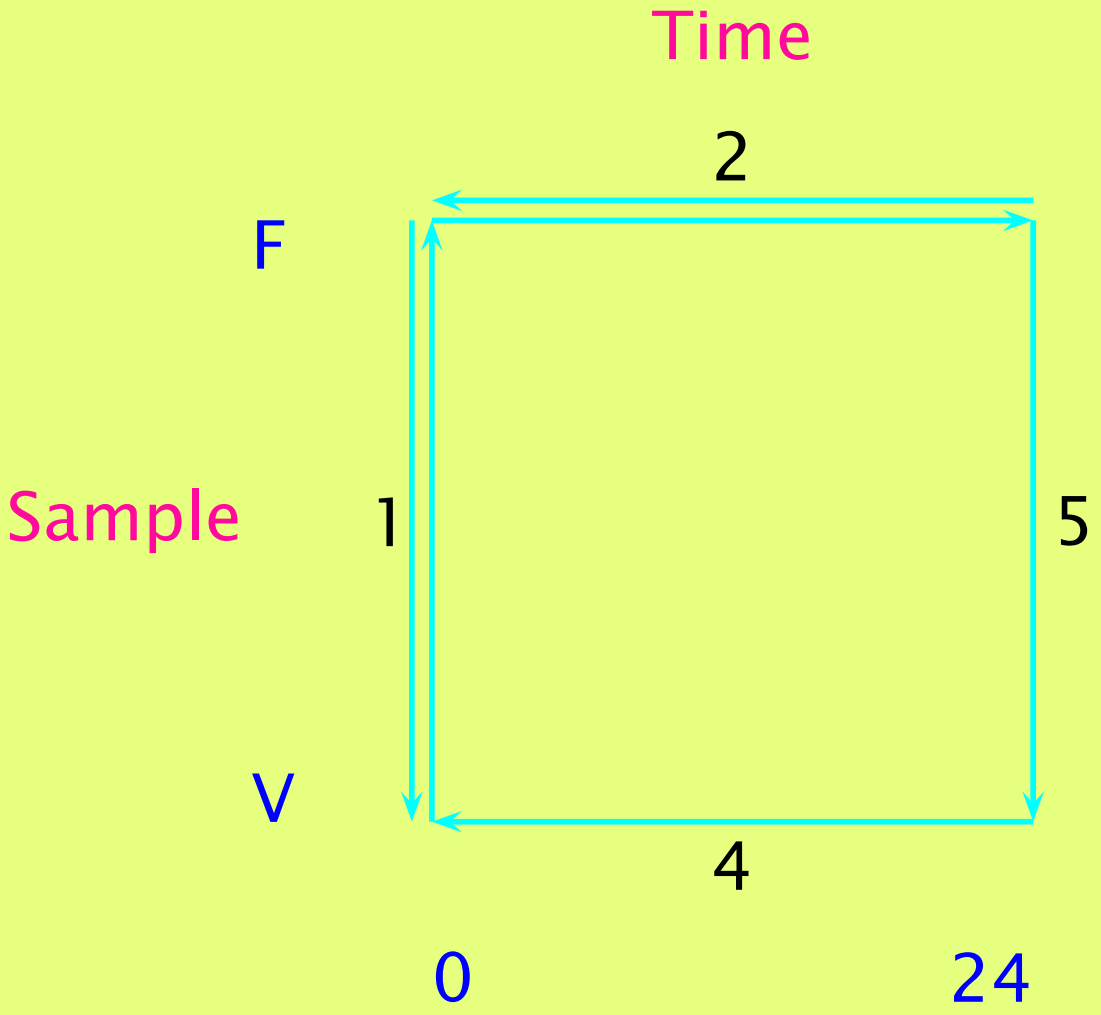
The reference design



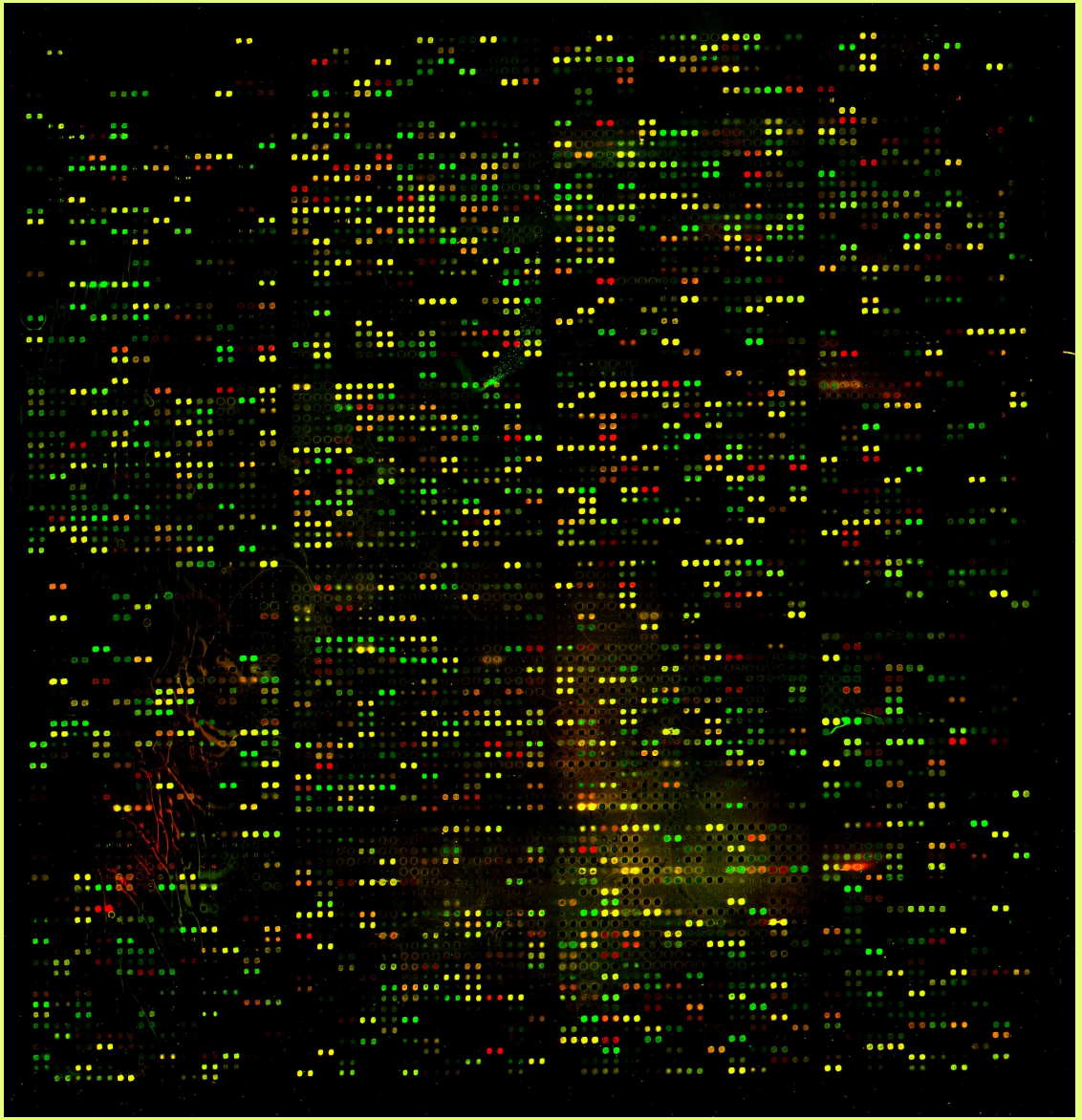
All comparisons design: 6 slides



Optimal design: 6 slides



Human clones with MCF-7 and Jurkatt probes



Quantifying expression

- For each spot on the array, calculate the background-adjusted intensities:

red intensity $R \rightarrow R - Rb$

green intensity $G \rightarrow G - Gb$

- and combine them in the log base 2 ratio:

$$M = \log_2(R/G) = \log_2 R - \log_2 G$$

Quantifying expression

- For each spot on the array, calculate the background-adjusted intensities:

red intensity $R \rightarrow R - Rb$

green intensity $G \rightarrow G - Gb$

- and combine them in the log base 2 ratio:

$$M = \log_2(R/G) = \log_2 R - \log_2 G$$

- *We use seeded region growing and morphological opening in Spot which runs within R.*

Quantifying expression

- For each spot on the array, calculate the background-adjusted intensities:

red intensity $R \rightarrow R - Rb$

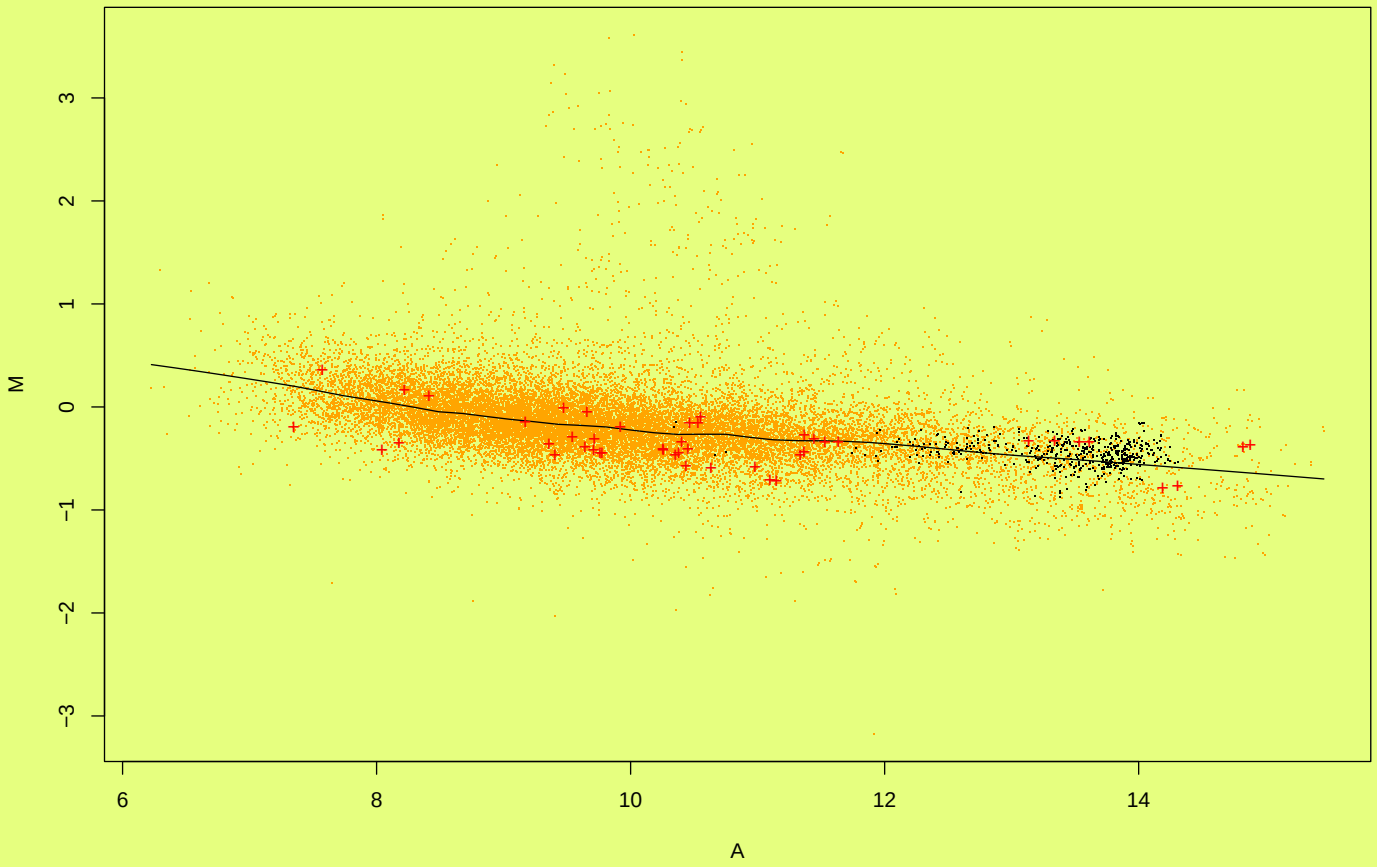
green intensity $G \rightarrow G - Gb$

- and combine them in the log base 2 ratio:

$$M = \log_2(R/G) = \log_2 R - \log_2 G$$

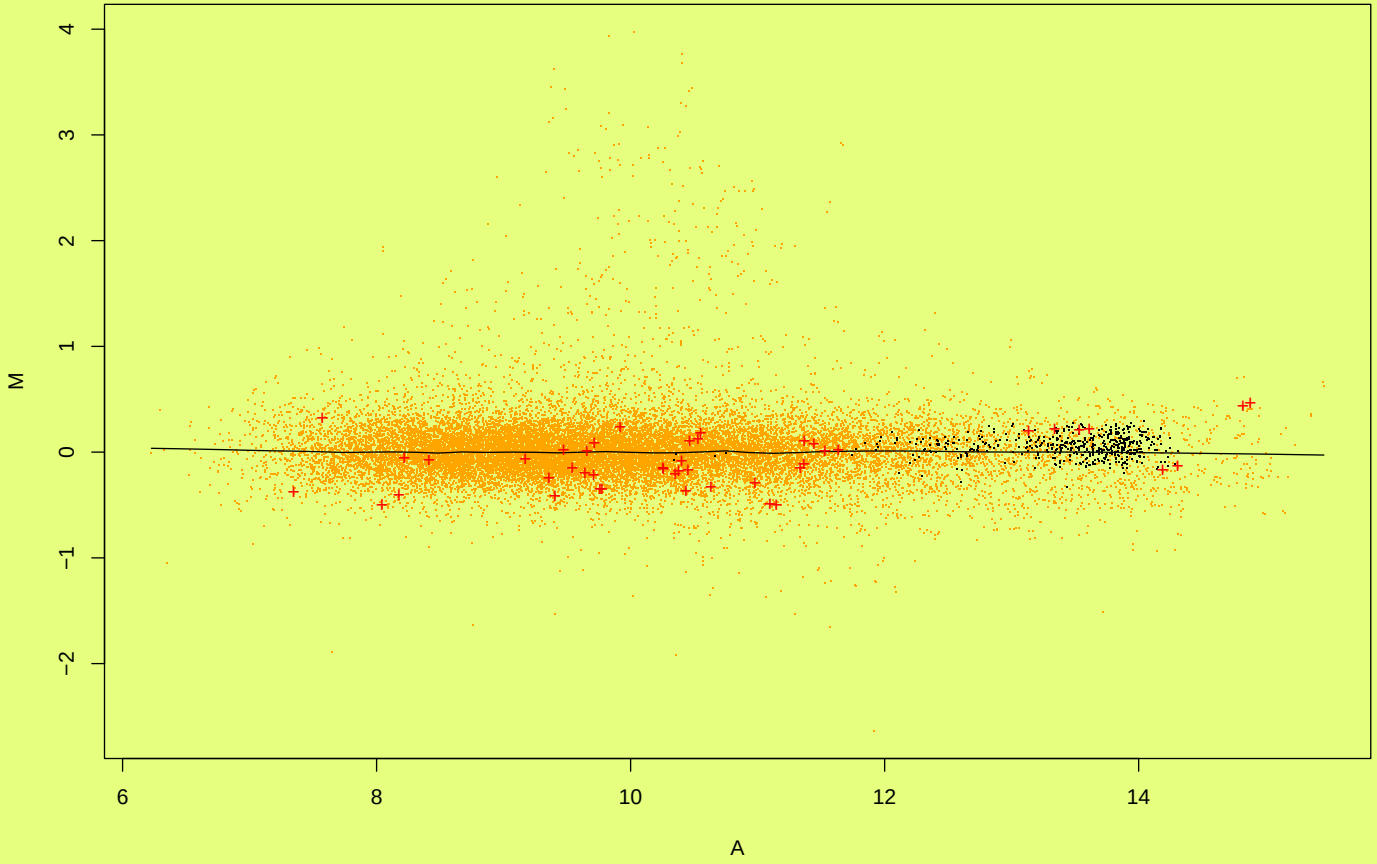
- *We use seeded region growing and morphological opening in Spot which runs within R.*
- *Background is a big issue.*

Pre-processing angiogenesis: time 0 versus .5 hour



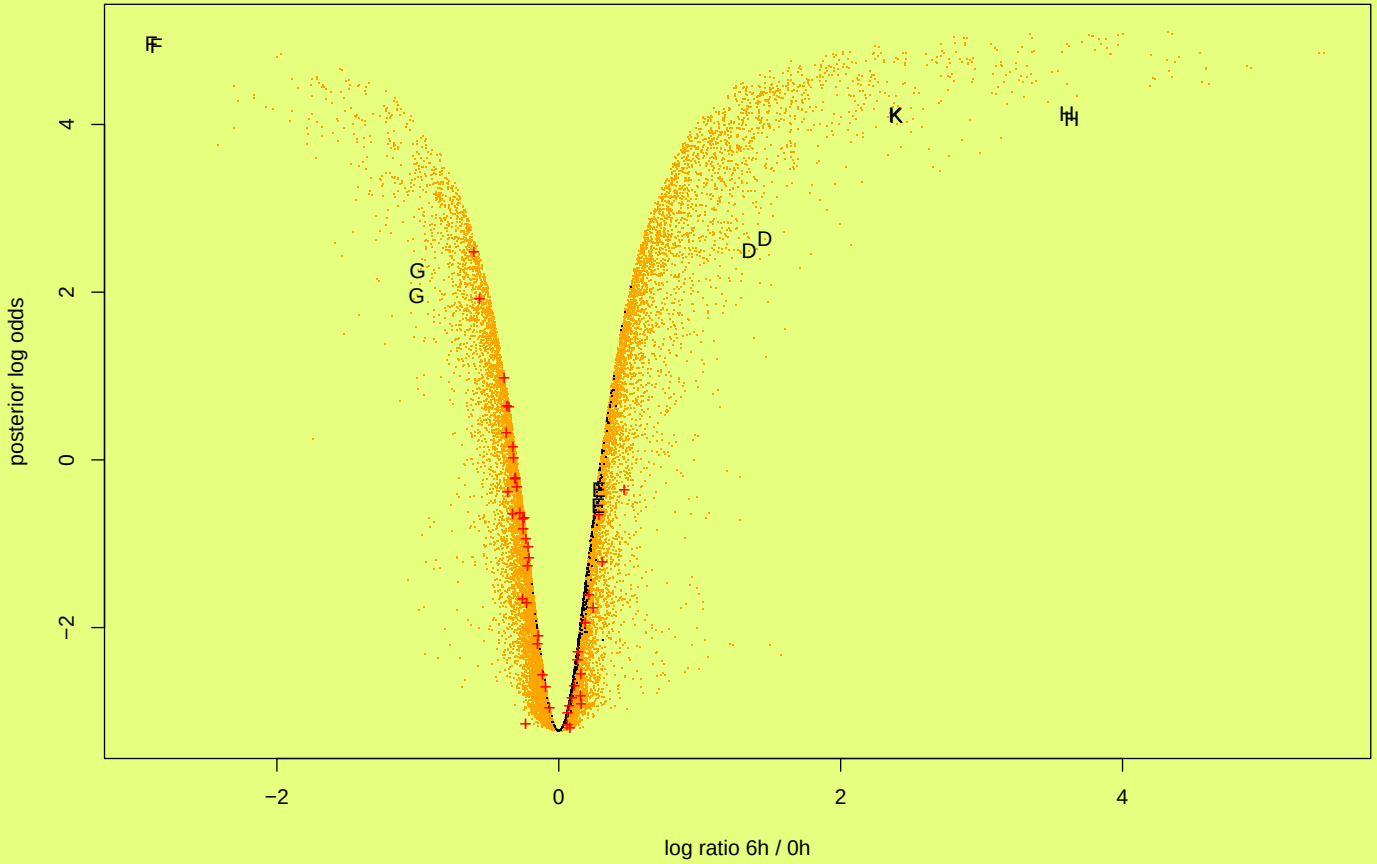
M is intensity and spatially dependent

Angiogenesis: after printtip normalisation



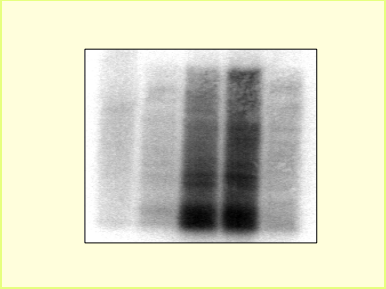
(Robust normalisation: Calvert et al.)

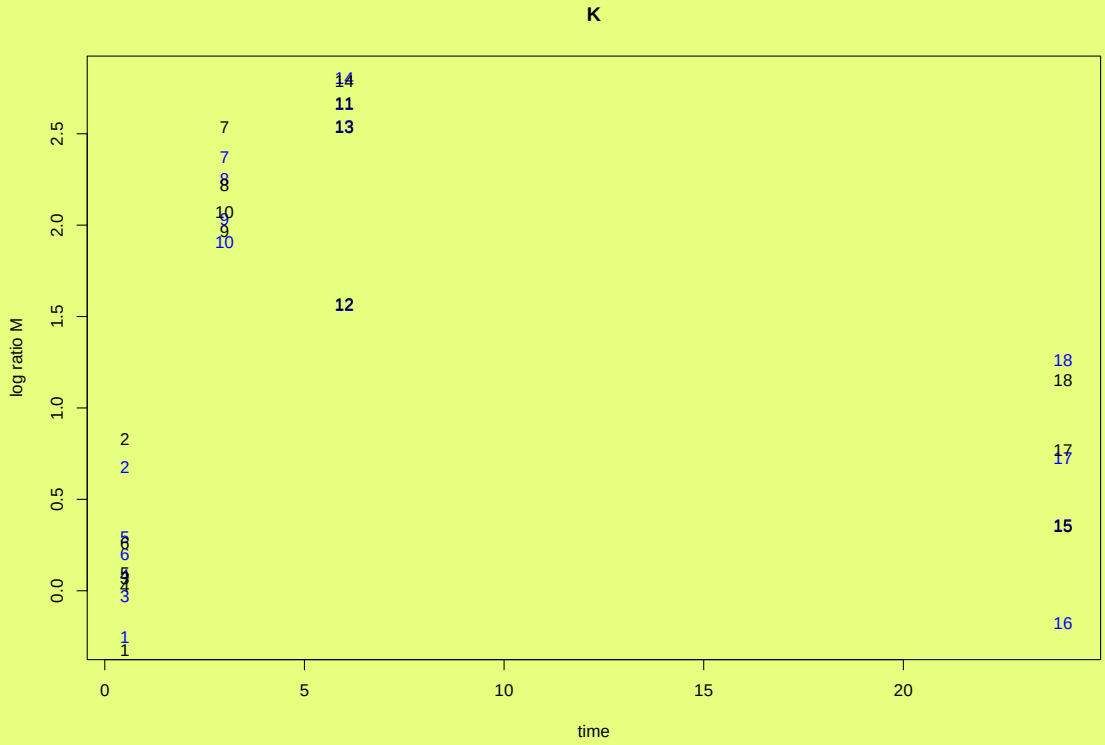
Bayesian analysis of differential expression



Angiogenesis: 0 vs 6 hours, 4 dye-swapped replicates

Angiogenesis: control K time course





Some ‘truths’

- ‘*When we entered the era of high technology, we entered the era of mathematical technology*’ [1].

Some 'truths'

- '*When we entered the era of high technology, we entered the era of mathematical technology*' [1].
- *The interface between statistics, biology, computer science and medicine has gone from information-poor to information-mega-rich*.
  - ◇ Statistics has a central role to play in processing that information and making it intelligible.

Some 'truths'

- '*When we entered the era of high technology, we entered the era of mathematical technology*' [1].
- *The interface between statistics, biology, computer science and medicine has gone from information-poor to information-mega-rich*.
  - ◇ Statistics has a central role to play in processing that information and making it intelligible.
- *Biology looks set to dominate statistics at the beginning of this century, just as it did at the beginning of the last one.*

---

[1] Ad hoc Committee on Resources for the Mathematical Sciences, US National Research Council, 1981.

Acknowledgements

- Adelaide Women's and Children's Hospital
  - Brenton Reynolds
  - Richard D'Andrea
- Hanson Institute, IMVS
  - Greg Goodall
  - Chris Hahn
  - Jenifer Gamble
  - Nick Fazzalari
  - Mark Guthridge
- Bionomics
  - Tom Gonda

Further reading and web sites

- Bioconductor Project: Open source bioinformatics using R. http://www.bioconductor.org
- MJ Buckley. The Spot user's guide. CSIRO Mathematical and Information Sciences, August 2000. http://www.cmis.csiro.au/IAP/Spot.htm
- G Glonek & PJ Solomon (2002). Factorial designs for microarray experiments. Submitted. http://maths.adelaide.edu.au/MAG
- R. Ihaka & R. Gentleman. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314, 1996.
- M Kerr & G Churchill (2001). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.

- Microarray Analysis Group. The University of Adelaide http://maths.adelaide.edu.au/MAG.
- sma v. 0.5.6 (November 2001), B Bolstad, S Dudoit, YH Yang. *http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html*.
- G Smyth, et al. (2002) Statistical issues in cDNA microarray data analysis. Research Report, WEHI.
- Statistical Science Web http://www.statsci.org/micrarra/index/html.
- The Chipping Forecast. *Supplement to Nature Genetics*, **21**, 1999.
- TP Speed. Preprints, information and software (sma) www.stat.berkeley.edu/users/terry/zarray
- TP Speed & YH Yang (2002). Direct versus indirect designs for cDNA microarray experiments. Preprint.

THE END