Planning Microarray Experiments

Patty Solomon
School of Mathematical Sciences
University of Adelaide

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

What are microarrays ?

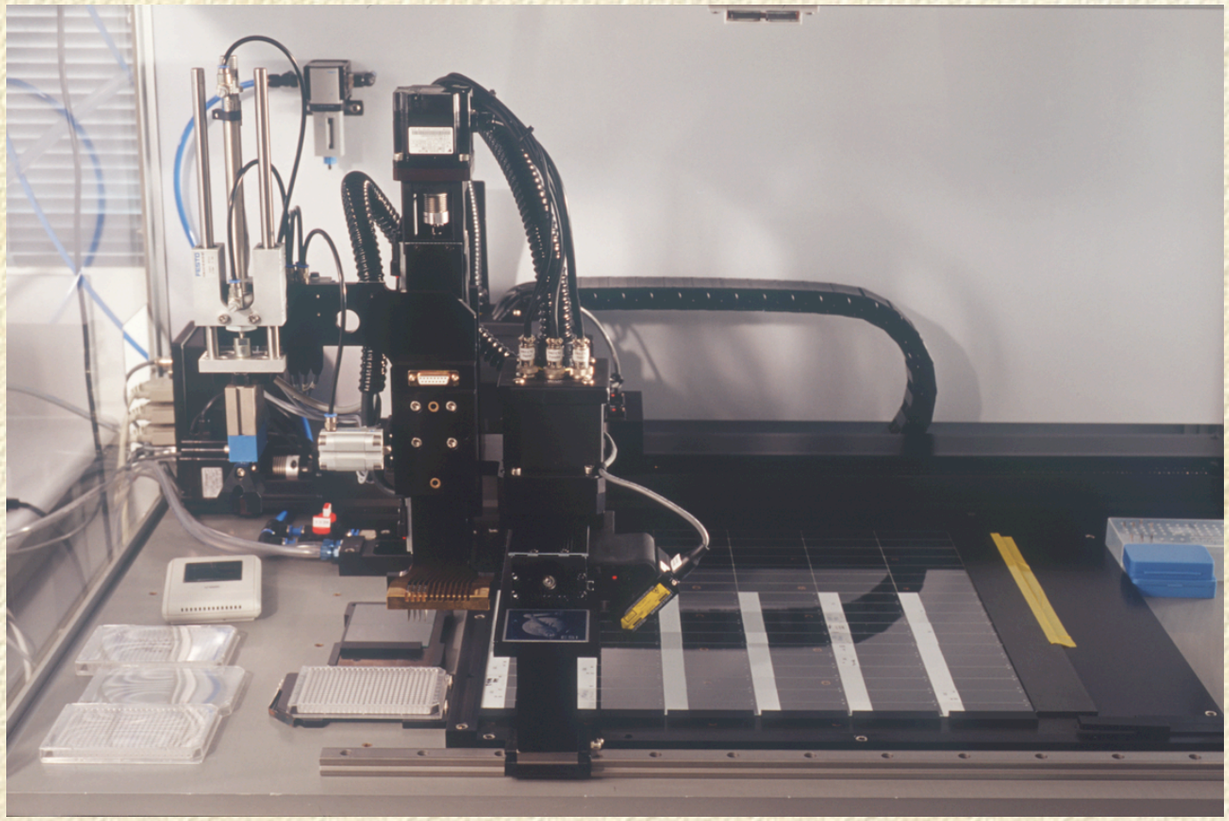
- *A new technology for surveying the expression levels of thousands of genes simultaneously*

- Detect and measure gene expression at the mRNA or protein level, mutational analysis, genetic mapping studies, to (re)sequence DNA, to locate chromosomal changes, and more ...

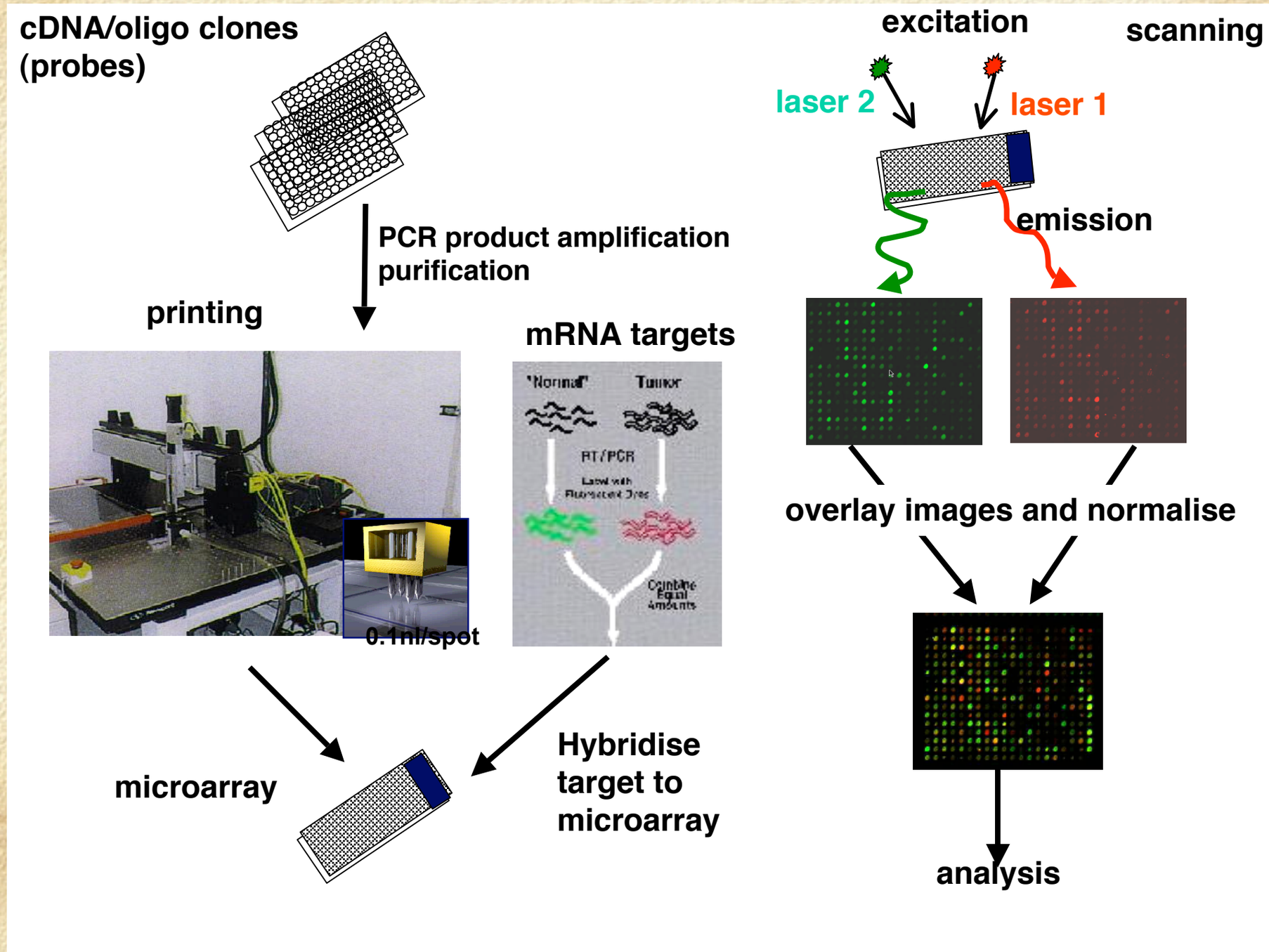
- *Experiments limited only by the imagination of the biologists!*

There are many different types of microarrays

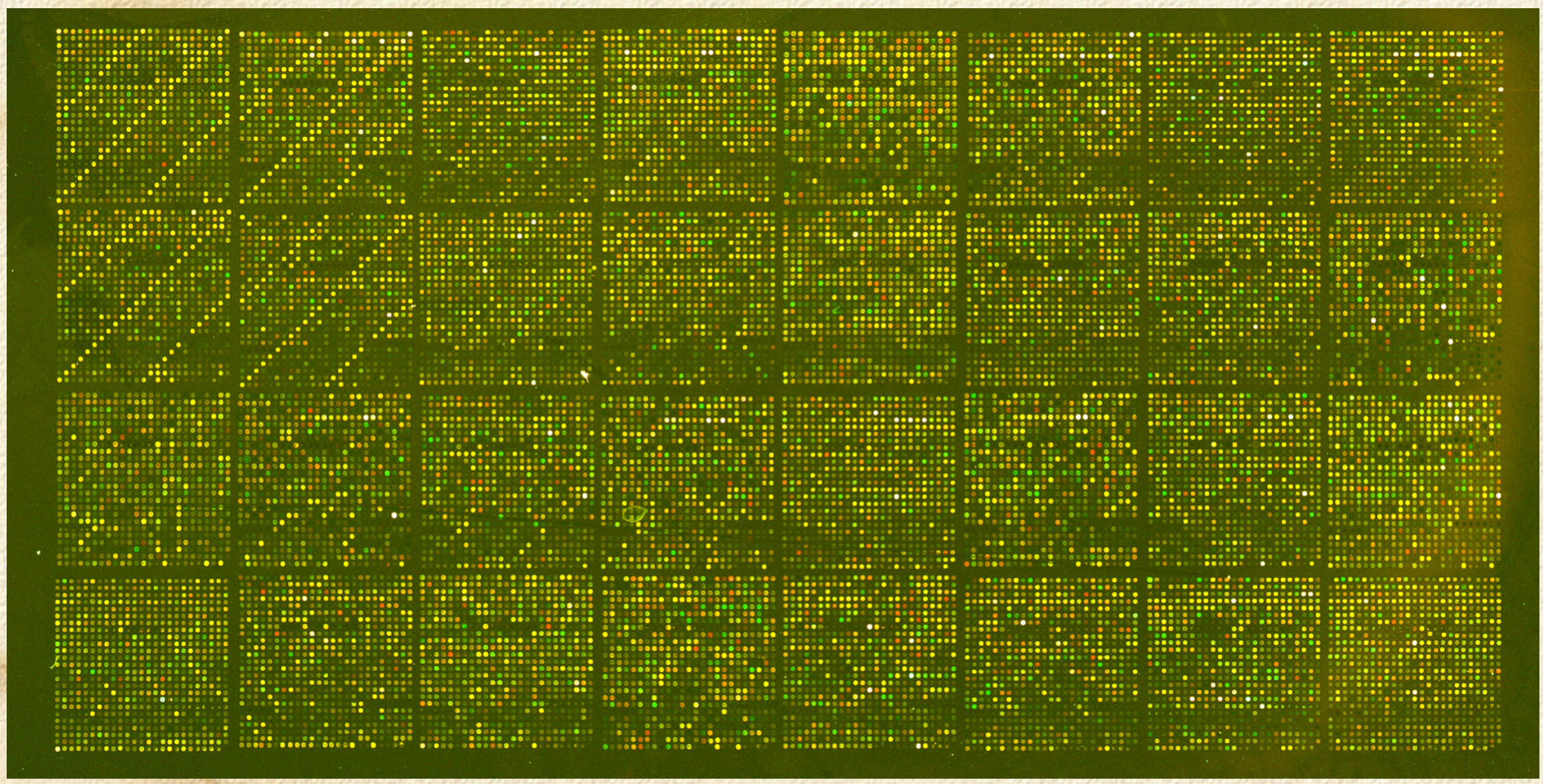
The Adelaide Microarray Facility



The microarray process



A human oligonucleotide array



This is the raw data

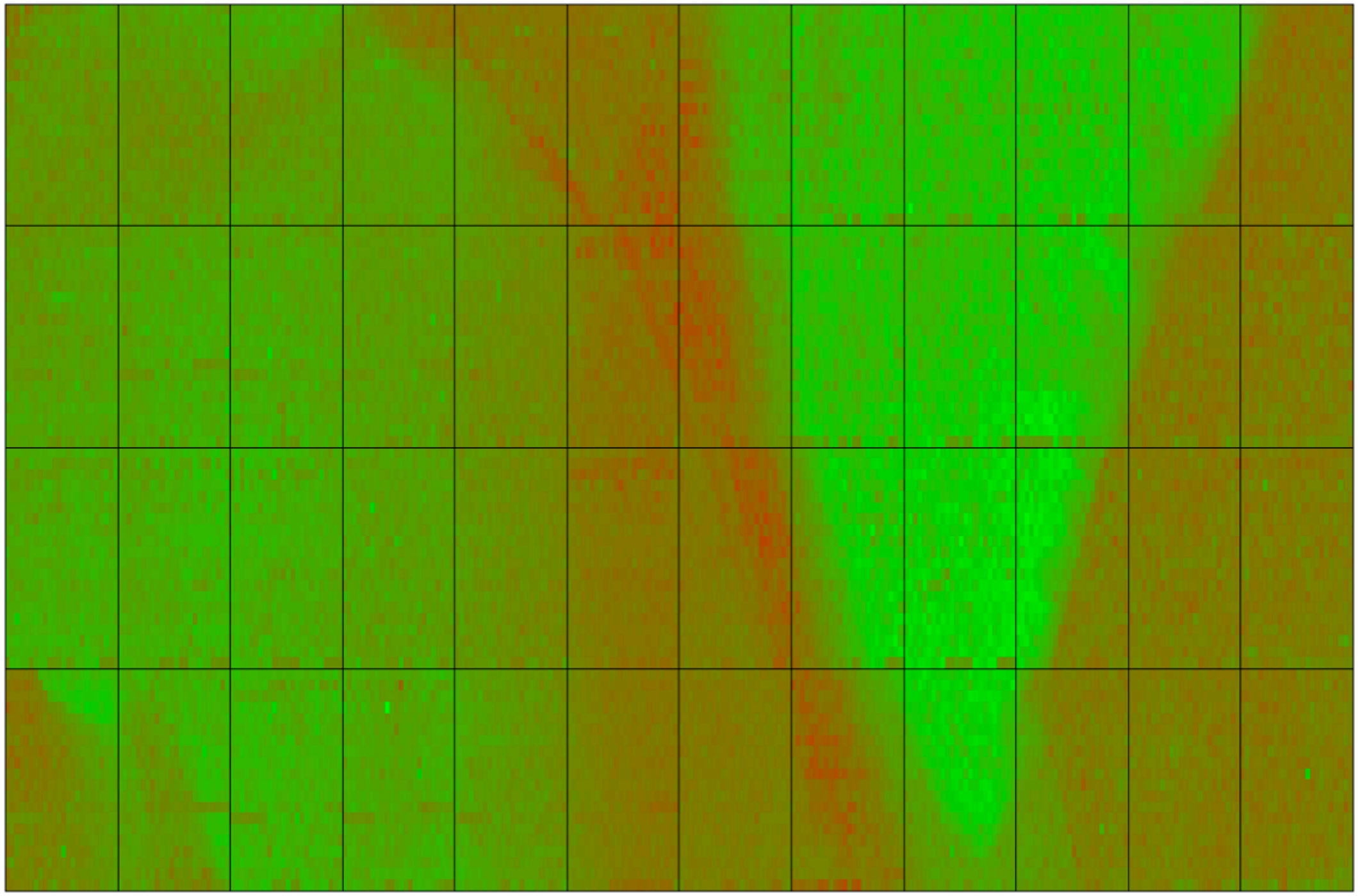
The experimental process leaves a 'global fingerprint' on the gene expression data

- Owing to amplification effects, dye effects, hybridization, scanning, ...

- These systematic biases need to be removed in a data pre-processing step known as *normalisation*.

- *Approaches which attempt to adjust for these biases in a 'global' model are not correct.*

Biases can be extreme ...

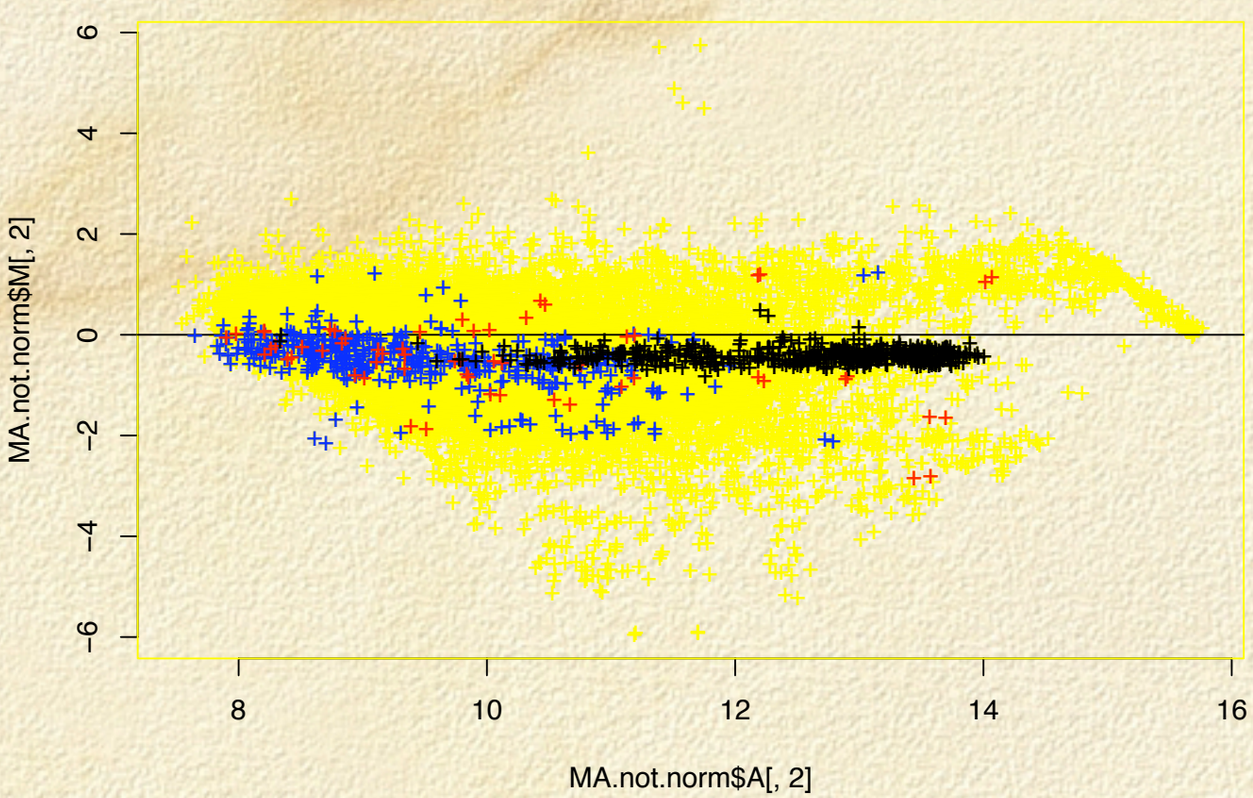


Two key aspects of design

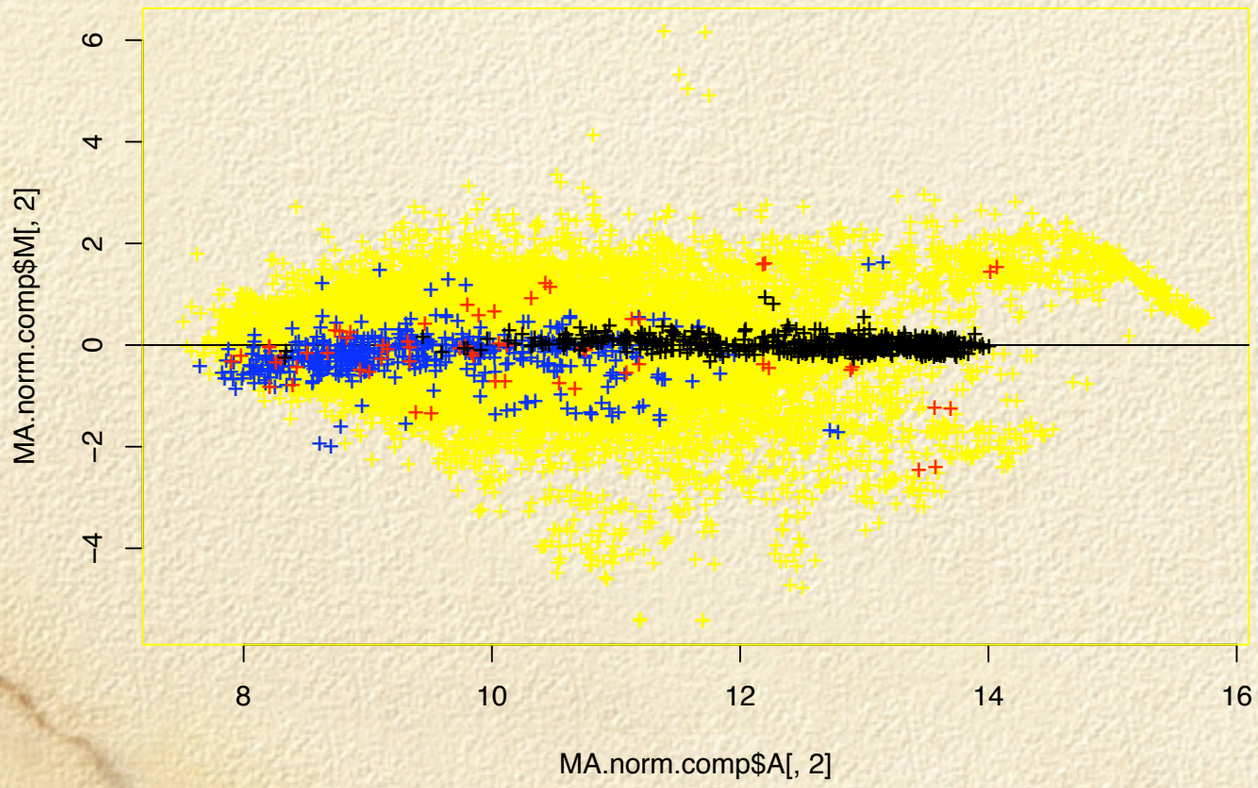
- What to spot on the array
- What to hybridize to the array

Angiogenesis

Huvec array:
heart, hrp



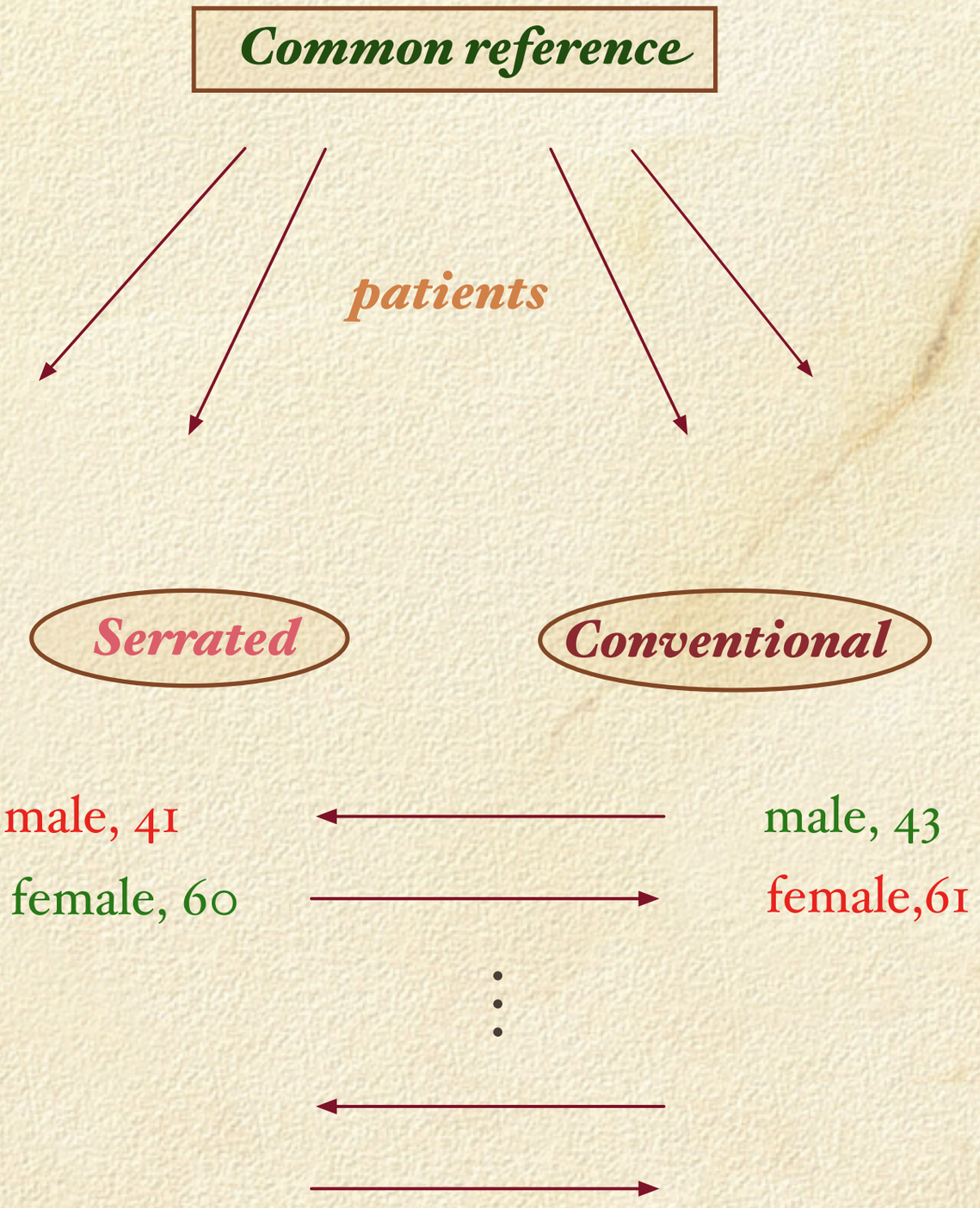
black points: msp



Sometimes, there isn't a choice

- *Extensibility,* important in studies of individual cancer risk or diagnosis

- Only enough mRNA for a single hybridization: *match conventional polyps versus serrated polyps* in colon cancer study



But often, there is a choice

- Time course experiments, factorial experiments
- *Our approach: identify the parameters of interest, and seek designs which minimize the variance of the corresponding estimates, subject to resource constraints*
- We have introduced classes of *admissible designs* (*Glonek & Solomon 2004*)

Admissible design

For a given number of hybridizations, a design is admissible if there is no other design which has a smaller variance for all contrasts of interest.

For each gene, fit the linear model

$$E(M) = X\gamma$$

We know

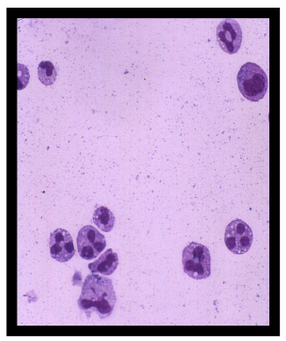
$$\mathrm{Var}(\gamma_i) = \sigma^2 c_i$$

It makes sense to choose a design with the smallest values of c

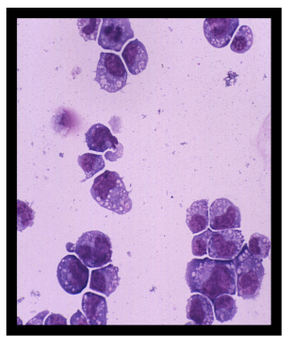
Acute myeloid leukaemia

Bi-Potential Properties of FDB-1 Cells

FI Δ
37aa duplication in extracellular domain

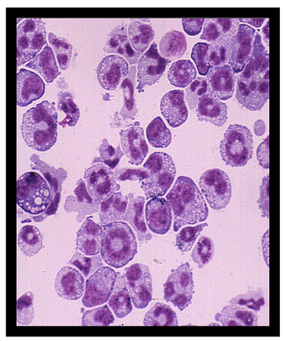


V449E
Point mutation in Transmembrane domain

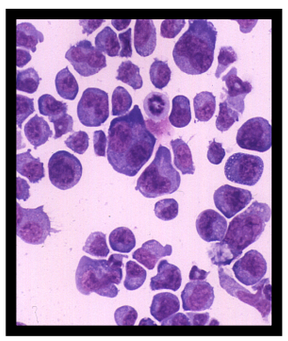


Factor Independent Survival

FDB-1 + GM-CSF



FDB-1 + IL-3



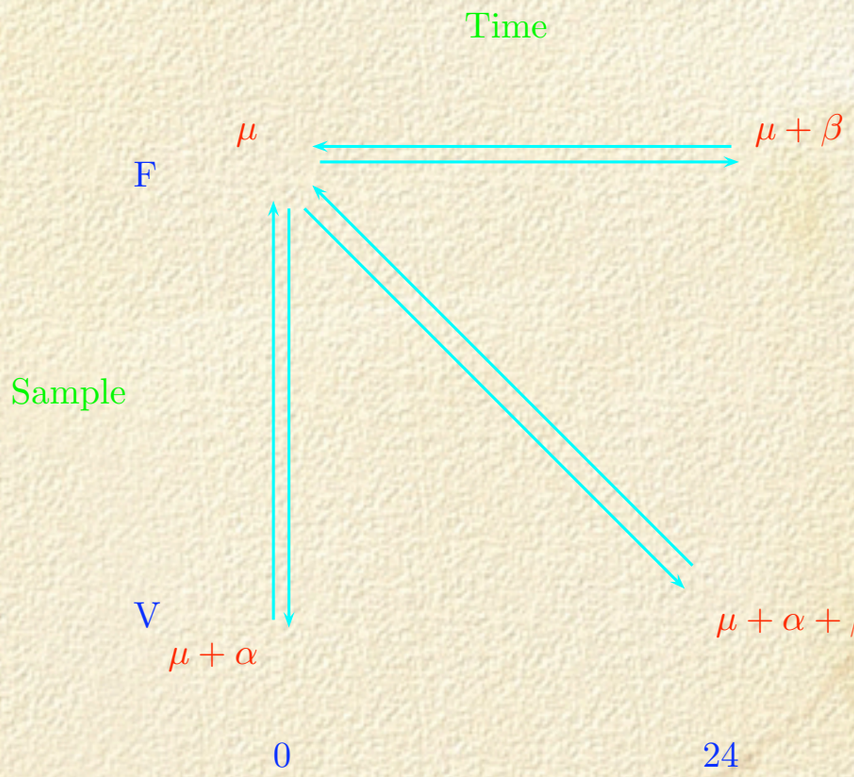
Factor Dependant Survival

Differentiation Pathways

Proliferation Pathways

Illustration: 2x2 experiment

- For each gene, interest centres on changes in expression between F and V over 0 to 24 hours



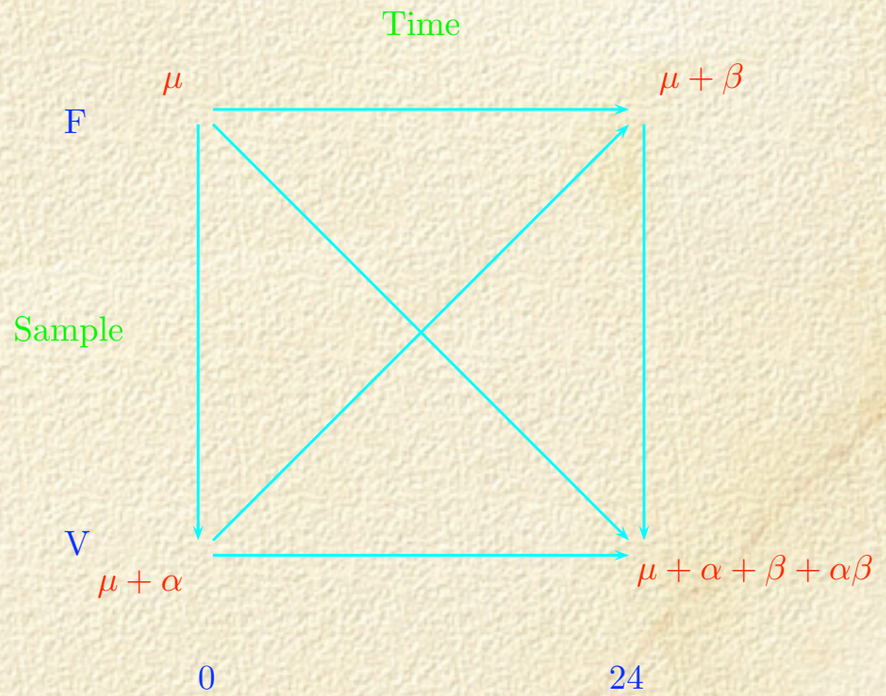
The interaction is the parameter of primary importance

Reference design
6 slides

$$c_\alpha = c_\beta = 0.5, \quad c_{(\alpha\beta)} = 1.5$$

Classical design

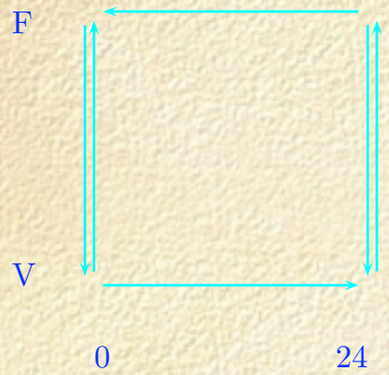
- This would be the choice of many classical experimental design folk
- But doesn't make sense to devote effort to estimating with high precision contrasts of no biological interest
- *Inadmissible in our formulation*



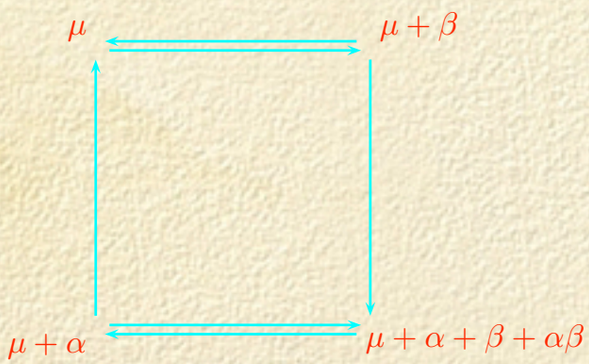
All pairwise comparisons
6 slides

$$c_\alpha = c_\beta = 0.5, \quad c_{(\alpha\beta)} = 1.0$$

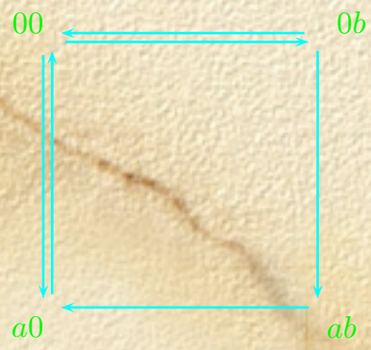
Optimal admissible designs



$$c_\alpha = 0.42, \quad c_\beta = 0.67, \quad c_{(\alpha\beta)} = 0.67$$



$$c_\alpha = 0.67, \quad c_\beta = 0.42, \quad c_{(\alpha\beta)} = 0.67$$



$$c_\alpha = c_\beta = 0.42, \quad c_{(\alpha\beta)} = 0.75$$

What's happening ...

Contrast			Estimate		Lost
A	*oo* *ob*	*ao* *ab*	β	$\beta + \alpha\beta$	α
B	*oo* *ao*	*ob* *ab*	α	$\alpha + \alpha\beta$	β
C	*oo* *ab*	*ao* *ob*	$\alpha + \beta + \alpha\beta$	$\beta - \alpha$	$\cdot\, \alpha\beta$

Balanced confounding: 1/3 1/3 1/3

Interaction confounding: 1/2 1/2 0

Thank you Sir David!

Extensions to ...

- 2^m factorial designs with block size 2
- $2^m 3^n$ designs with block size 2
- Admissible time course designs *(more anon)*

It's not the whole story ...

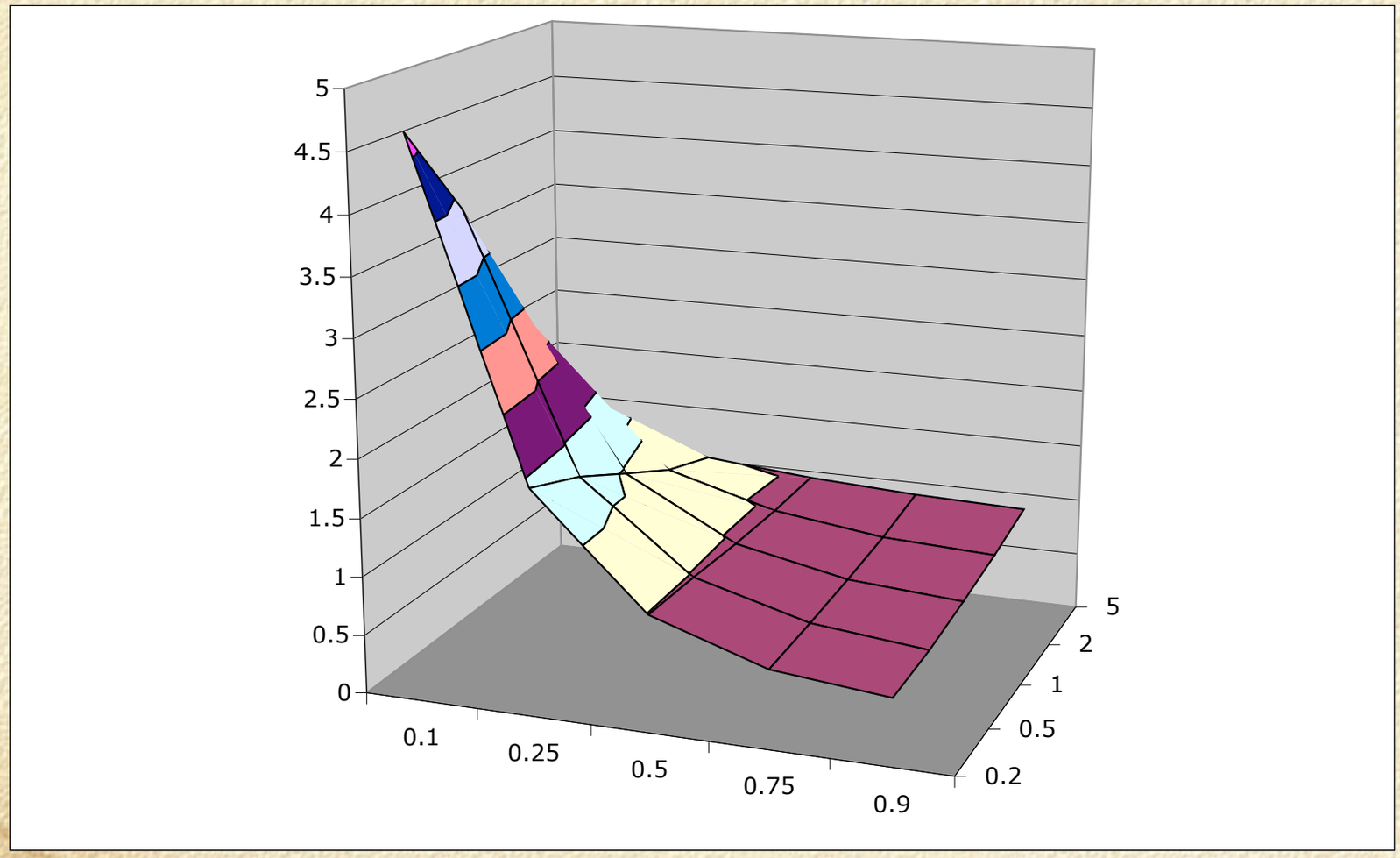
- *Definition of replication not straightforward in microarrays*
- 'Technical replicates' improve efficiency of common reference design (*Speed & Yang 2002,2003*)
- This relates to current work with Sir David Cox on design of studies to estimate components of variance in hierarchical arrangements *(Cox & Solomon 2003,2004)*

Illustration: 2 nested components

τ_ξ *mRNA samples*

τ_ϵ *sub-samples*

n_1 n_2



Of course, any recommendation will depend on the costs involved

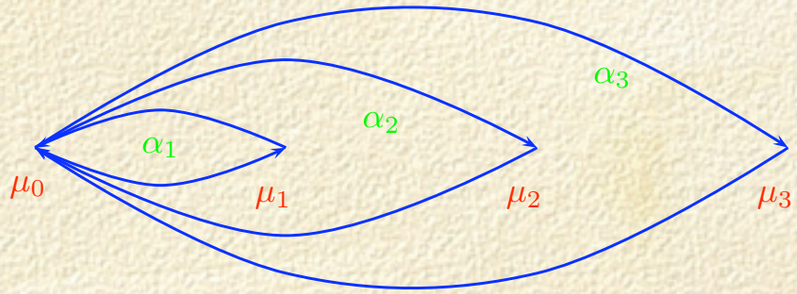
Time course experiments

We distinguish three situations:

- *Time zero is a meaningful baseline, and want to measure (smooth) profiles over time*
- Want to measure short-term or sudden changes
- Time profiles of interest specified *a priori*

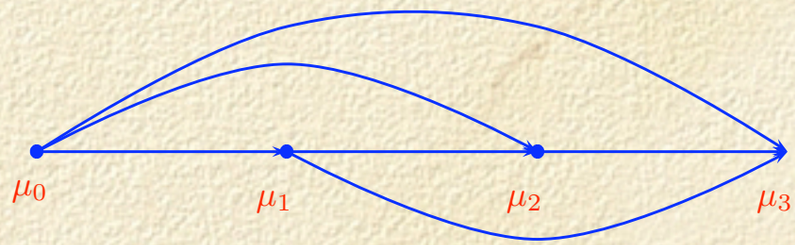
Smooth time profiles

- Time course experiment with 4 time points, 6 slides



$$c_{\alpha_1} = c_{\alpha_2} = c_{\alpha_3} = 0.5$$

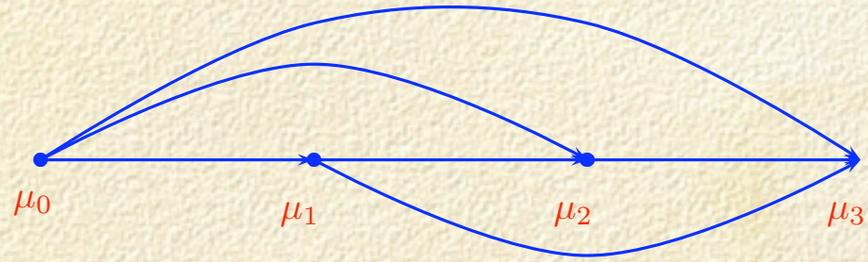
2 admissible designs with equal variances for all parameters



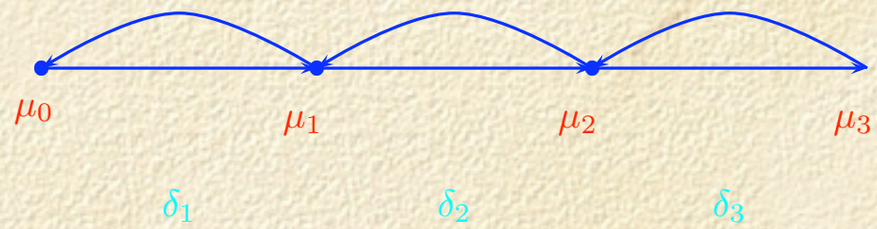
$$c_{\alpha_1} = c_{\alpha_2} = c_{\alpha_3} = 0.5$$

Adjacent time points

- Two admissible designs with equal variances for all parameters



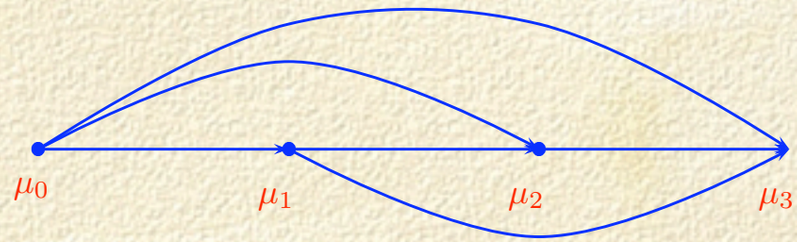
$$c_{\delta_1} = c_{\delta_2} = c_{\delta_3} = 0.5$$



$$c_{\delta_1} = c_{\delta_2} = c_{\delta_3} = 0.5$$

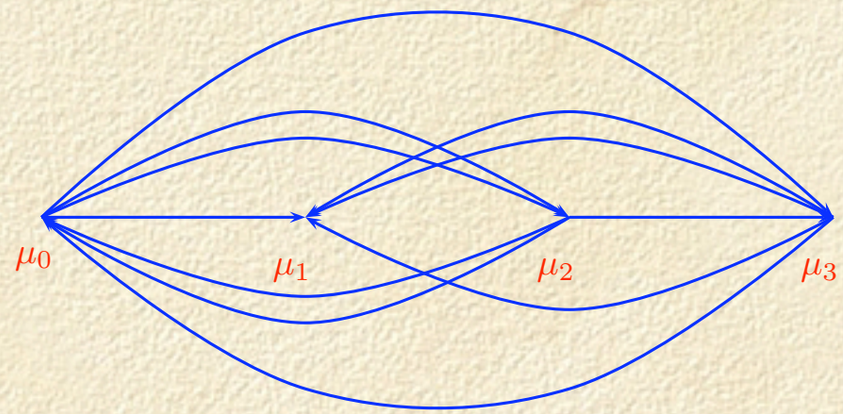
Quadratic time profiles

- Parametrize via *orthogonal polynomials*, linear and quadratic terms



$$c_{\beta_1} = c_{\beta_2} = c_{\beta_3} = 0.25$$

One admissible design for 6 slides with equal variances for $\hat{\beta}_1, \hat{\beta}_2$

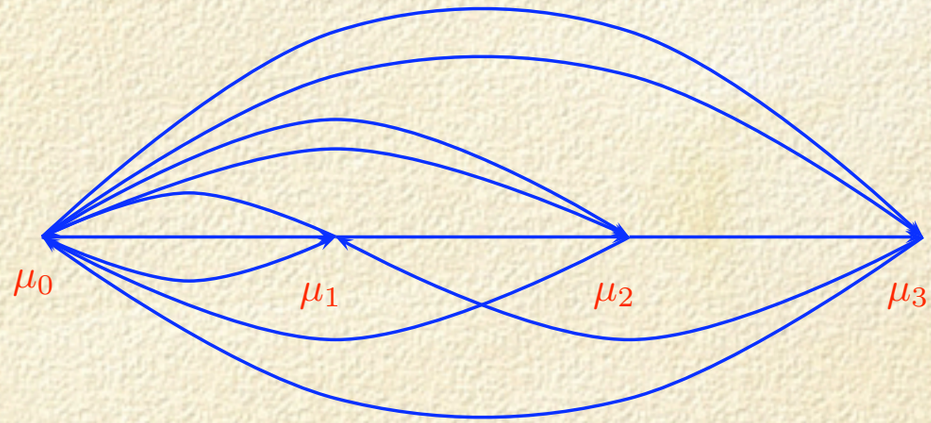


12 slides,
2 admissible designs

$$c_{\beta_1} = c_{\beta_2} = 0.1, \quad c_{\beta_3} = 0.29$$

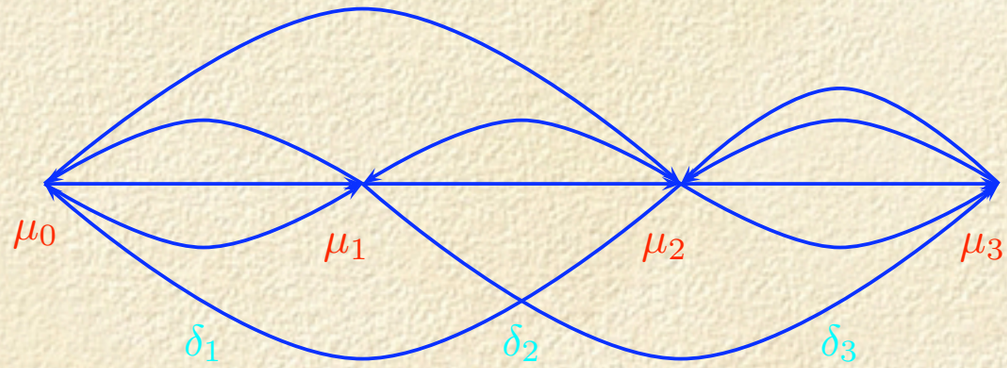
Admissible designs for 12 slides

Smooth time profiles:
alpha parameters,
1 admissible design



$$c_{\alpha_1} = c_{\alpha_2} = c_{\alpha_3} = 0.22$$

Adjacent time points:
delta parameters,
7 admissible designs



$$c_{\delta_1} = 0.24 \quad c_{\delta_2} = 0.25, \quad c_{\delta_3} = 0.21$$

Recommendation

- *Designs which allocate equal numbers of each of the 6 possible slide types perform well in all three situations*

- *Achieve (dye) balance, replication, and near optimal efficiency*

Further work

- *Evaluating performance of optimal admissible time course designs for measuring pluripotence and 'stemness' in embryonic stem cells*
- Types of 'replication' in complex microarray experiments
- Role of constraints in invariance to re-parametrisation. Search engines: Pareto simulated annealing
- *Statistical bioinformatics: 'gene equivalence'; design and analysis of protein arrays*

Next step

Evaluate performance of optimal admissible time course designs for measuring pluripotence and 'stemness' in embryonic stem cells

Web sites and reading

- Bioconductor in R http://www.bioconductor.org
- MAG http://www.maths.adelaide.edu.au/MAG
- Statistical Science Web http://www.statsci.org/micrarra/
- Terry Speed's web page http://www.stat.berkeley.edu/users/terry
- Glonek & Solomon *Biostatistics* **5**:89-111, 2004
- Cox & Solomon *Components of Variance*, 2003

Ack

Micr alth
Grou Institute

Gary
Chris D'Andrea
Anna Reynolds

Rath entre for
esearch/

Joy R
Rebe oodall