

Experimental design and linear models

Perry Moerland

Bioinformatics Lab, KEBB, AMC

`p.d.moerland@amc.uva.nl`

Microarray experiment

Control1
(normal)

Sample1
(cancer)

Control2
(normal)

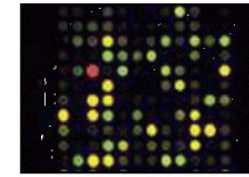
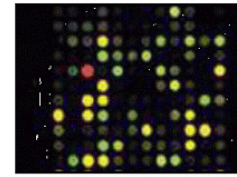
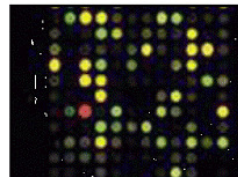
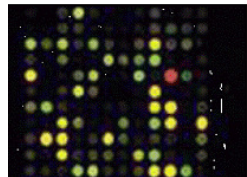
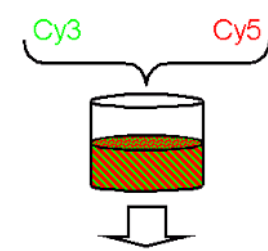
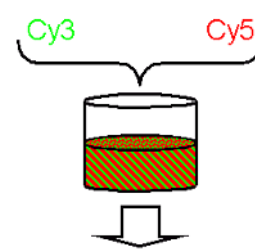
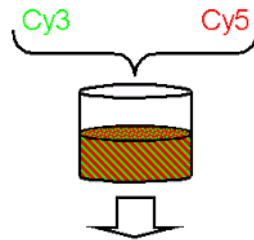
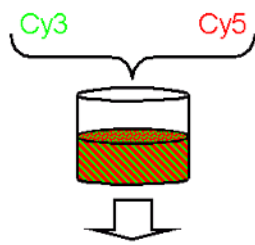
Sample 2
(cancer)

Control3
(normal)

Sample3
(cancer)

Control4
(normal)

Sample4
(cancer)



Which genes are different in **normal** tissue versus **cancer**?

Why is this a bad design?

- **Confounding**: the effects of two or more explanatory variables - on a response variable of interest - cannot be distinguished from one another
- Explanatory variables:
 - dye: {green, red}
 - patient status: {normal, cancer}
- Response variable: mRNA expression

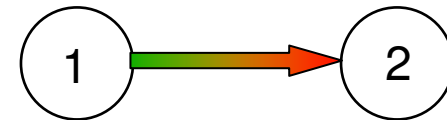
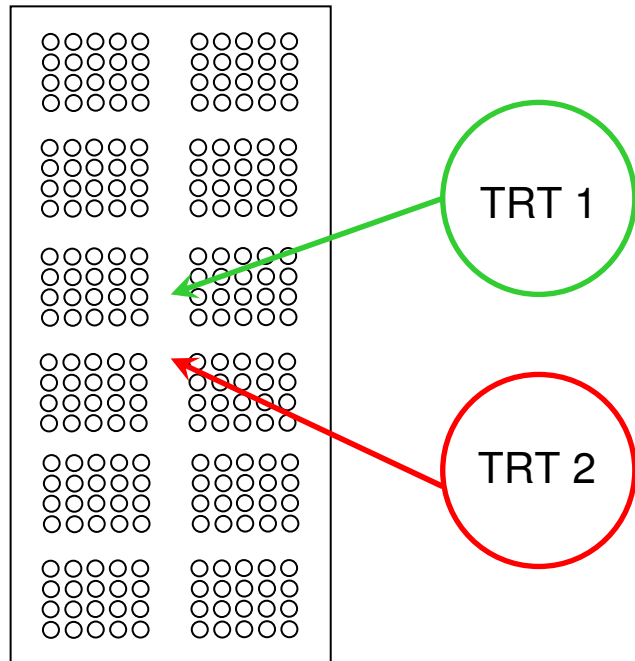
The trinity of experimental design (R.A. Fisher, 1935)

Randomization – random assignment of treatments (dye) to experimental units (patients)

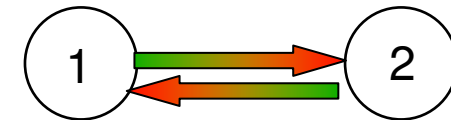
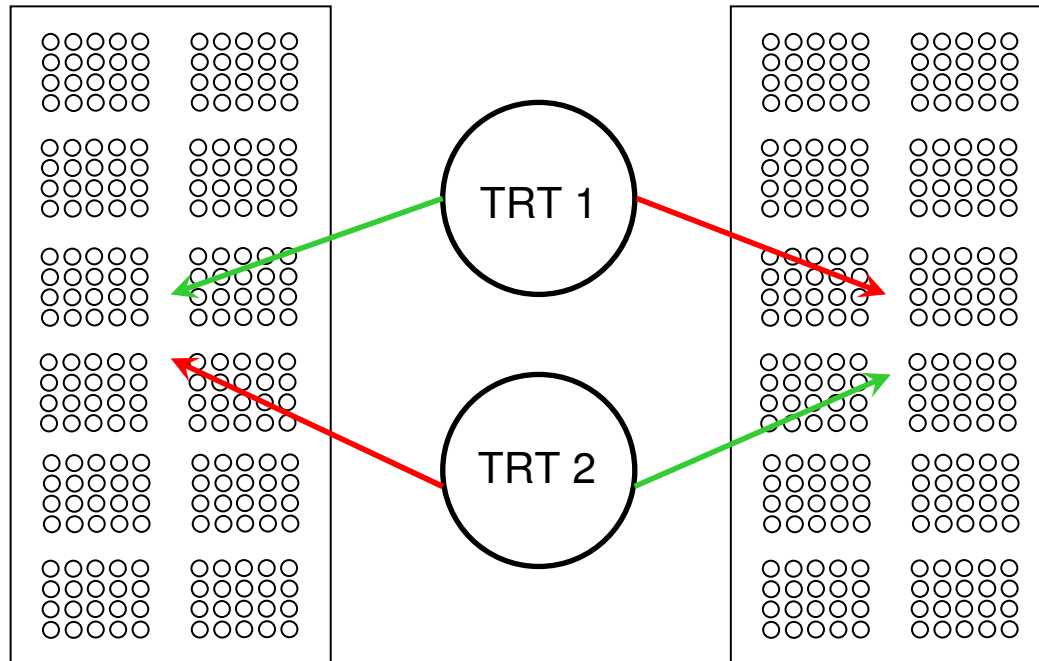
Blocking – grouping similar experimental units together and assigning different treatments within such groups of experimental units

Replication – applying a treatment independently to two or more experimental units

Microarrays: experimental design notation

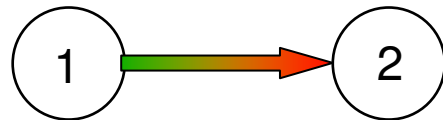
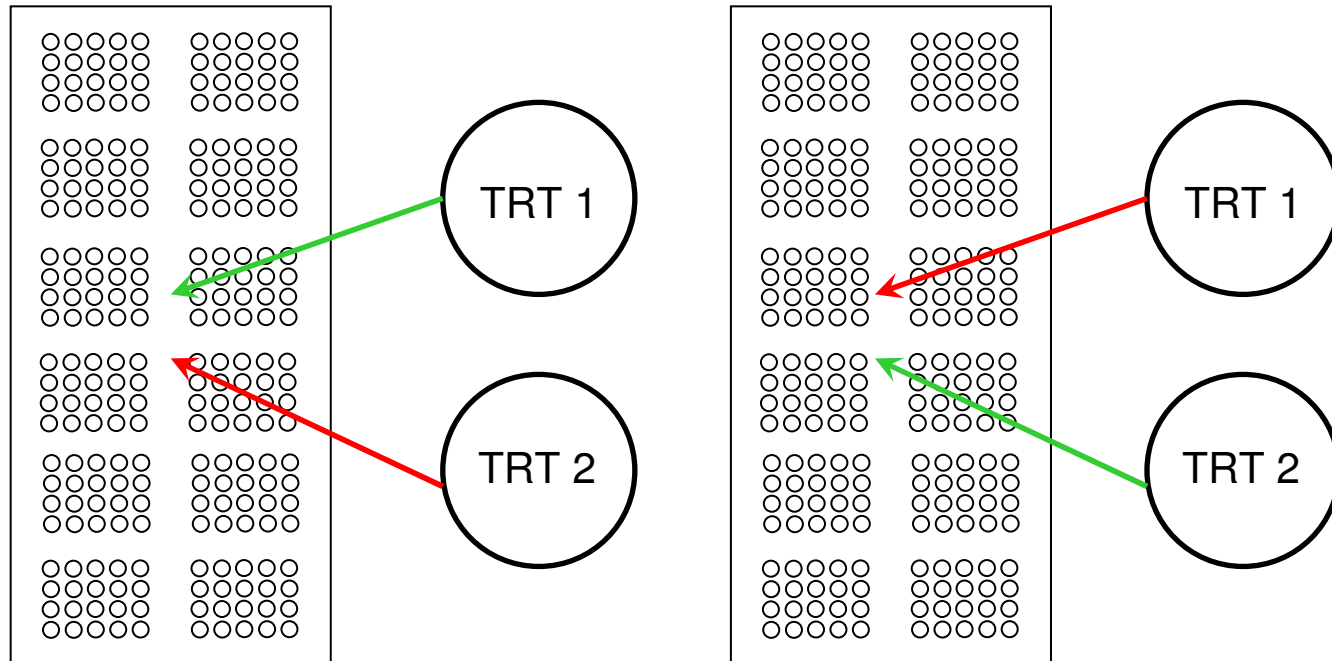


Microarrays: experimental design notation



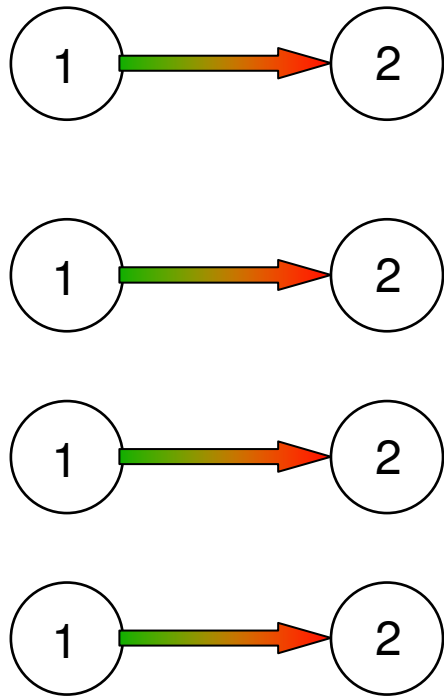
- dye swap
- same mRNA: technical replicates

Microarrays: experimental design notation

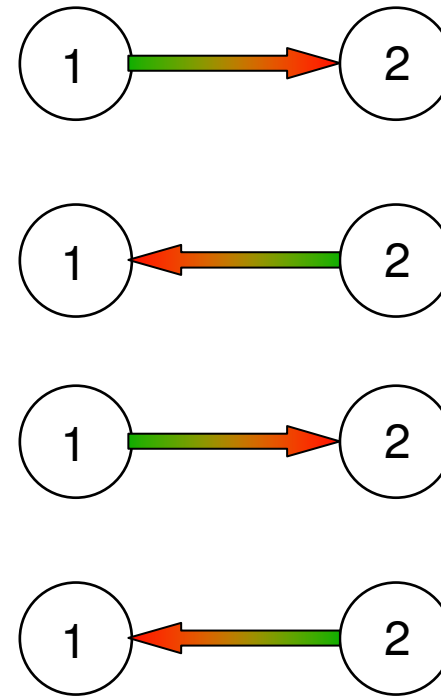


biological replicates

Back to the experiment

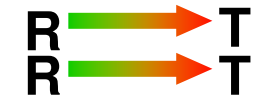


confounding



dye balanced wrt patient status

Reference



Common designs: reference

- Most widely used
- Extendable: new samples, new classes
- All comparisons have the same 'two-step' distance
- Robust to low-quality arrays
- Indirect comparison: dye balanced

Common designs: balanced block

N → T

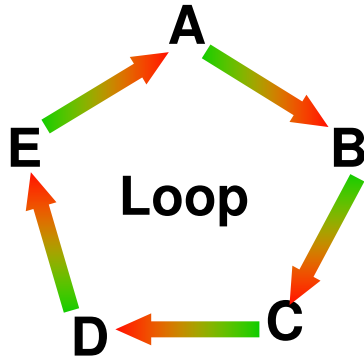
N ← T

T → N

T ← N

- No confounding of dye and treatment effect
- Efficient: 4 arrays to compare 4 samples of each group

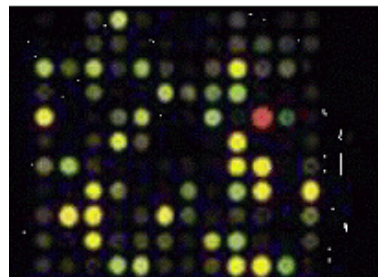
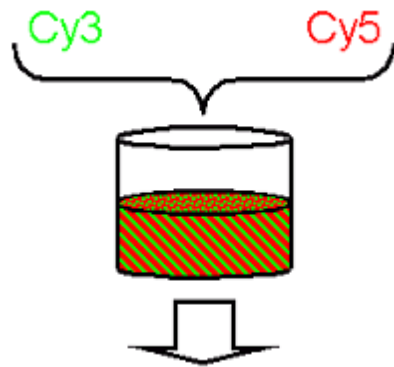
Common designs: loop



- Each variety sample is measured twice for the same number of arrays as with the common reference design
- Not robust to low-quality arrays
- Not easily expanded
- Some comparisons less precise than others

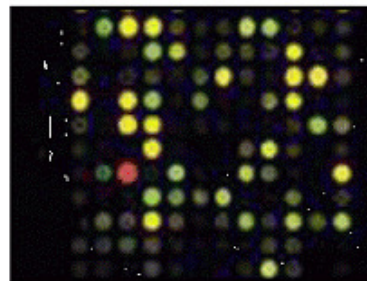
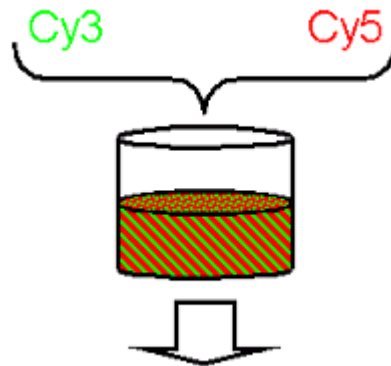
Differential expression

Reference sample 1
(group A)



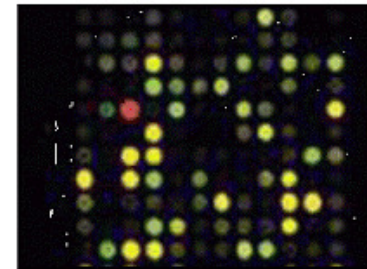
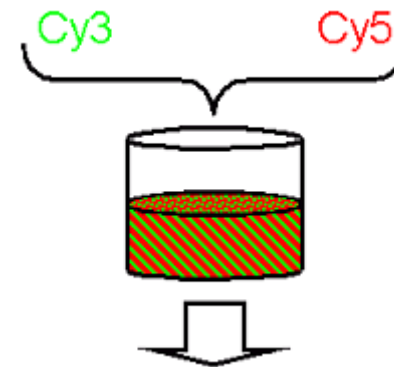
Array 1

sample 1 Reference
(group A)



Array 2

Reference Sample 3
(group B)



Array 3

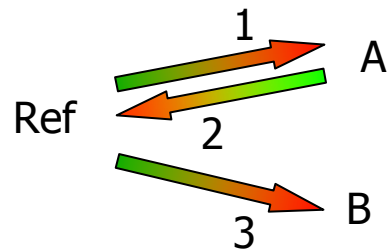
Which genes are different in Group A versus Group B?

Design matrix

Red	A	Ref	B
Green	Ref	A	Ref

Parameters

A - Ref	B - Ref
---------	---------



Observations

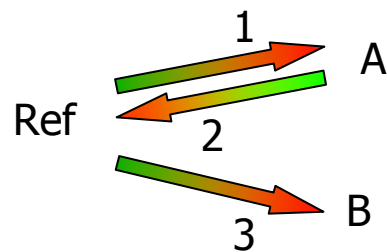
slide 1

slide 2

slide 3

Represent log-ratio from each slide by a parameter
→ specify the model for your data

Design matrix



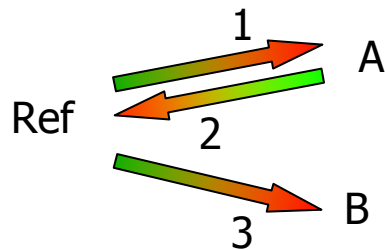
Observations

Parameters

	A - Ref	B - Ref
slide 1	1	0
slide 2		
slide 3		

Represent log-ratio from each slide by a parameter
→ specify the model for your data

Design matrix



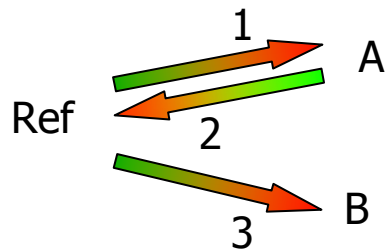
Observations

Parameters

	A - Ref	B - Ref
slide 1	1	0
slide 2	-1	0
slide 3		

Represent log-ratio from each slide by a parameter
→ specify the model for your data

Design matrix



Observations

Parameters

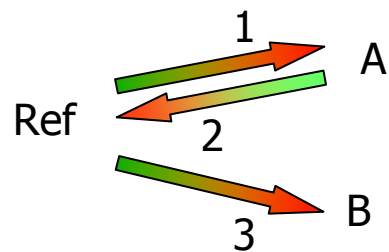
	A - Ref	B - Ref
slide 1	1	0
slide 2	-1	0
slide 3	0	1

Represent log-ratio from each slide by a parameter
→ specify the model for your data

Design matrix

Observed data modelled by these parameters

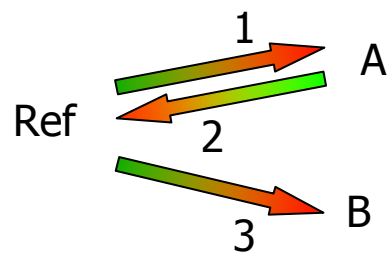
matrix notation



$$\text{Observed Data Matrix} = \text{Design Matrix} \times \text{Parameter Matrix}$$

Design matrix

matrix multiplication



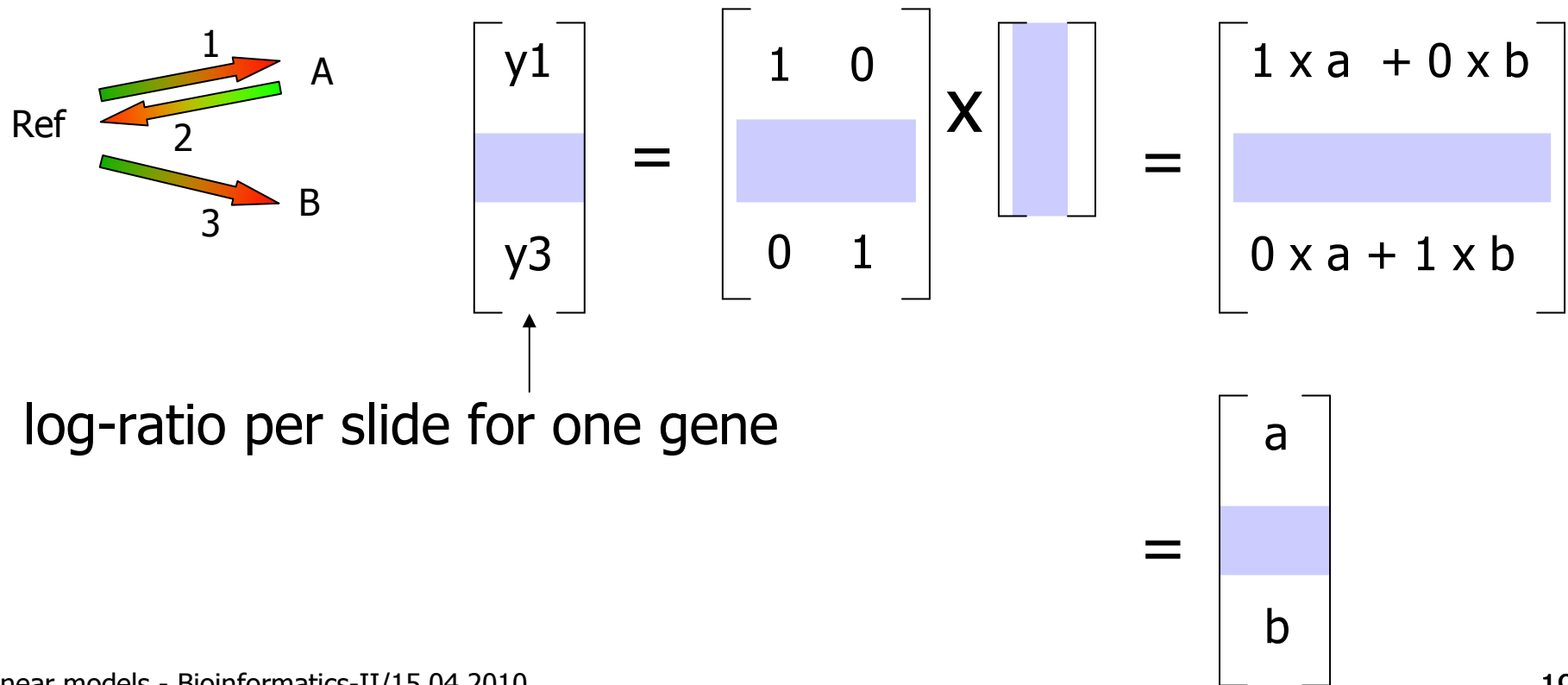
$$\begin{bmatrix} \text{shaded} \\ y2 \\ y3 \end{bmatrix} = \begin{bmatrix} \text{shaded} \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} \text{shaded} \\ \text{shaded} \end{bmatrix} = \begin{bmatrix} \text{shaded} \\ -1 \times a + 0 \times b \\ 0 \times a + 1 \times b \end{bmatrix}$$

log-ratio per slide for one gene

$$= \begin{bmatrix} \text{shaded} \\ -a \\ b \end{bmatrix}$$

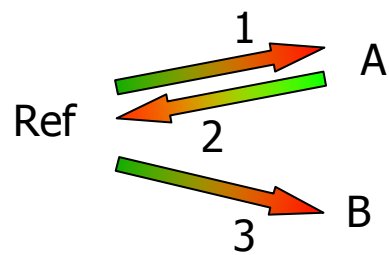
Design matrix

matrix multiplication



Design matrix

matrix multiplication



$$\begin{bmatrix} y1 \\ y2 \\ \text{[shaded]} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ \text{[shaded]} \end{bmatrix} \times \begin{bmatrix} \text{[shaded]} \\ \text{[shaded]} \\ \text{[shaded]} \end{bmatrix} = \begin{bmatrix} 1 \times a + 0 \times b \\ -1 \times a + 0 \times b \\ \text{[shaded]} \end{bmatrix}$$

log-ratio per slide for one gene

$$= \begin{bmatrix} a \\ -a \\ \text{[shaded]} \end{bmatrix}$$

Design matrix

FileNames	Cy3	Cy5
6Hs.195.1.gpr	b7-	b7+
6Hs.168.gpr	b7+	b7-
6Hs.166.gpr	b7+	b7-
6Hs.187.1.gpr	b7-	b7+
6Hs.194.gpr	b7-	b7+
6Hs.243.1.gpr	b7+	b7-

Samples

$$b7+ = a$$

$$b7- = b$$

Parameter

$$(b7+) - (b7-) = a - b$$

$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \times \begin{bmatrix} a - b \end{bmatrix}$$

Y
X
β

Design matrix

FileNames	Cy3	Cy5
6Hs.195.1.gpr	b7-	b7+
6Hs.168.gpr	b7+	b7-
6Hs.166.gpr	b7+	b7-
6Hs.187.1.gpr	b7-	b7+
6Hs.194.gpr	b7-	b7+
6Hs.243.1.gpr	b7+	b7-

Samples = 2

Parameter = 1

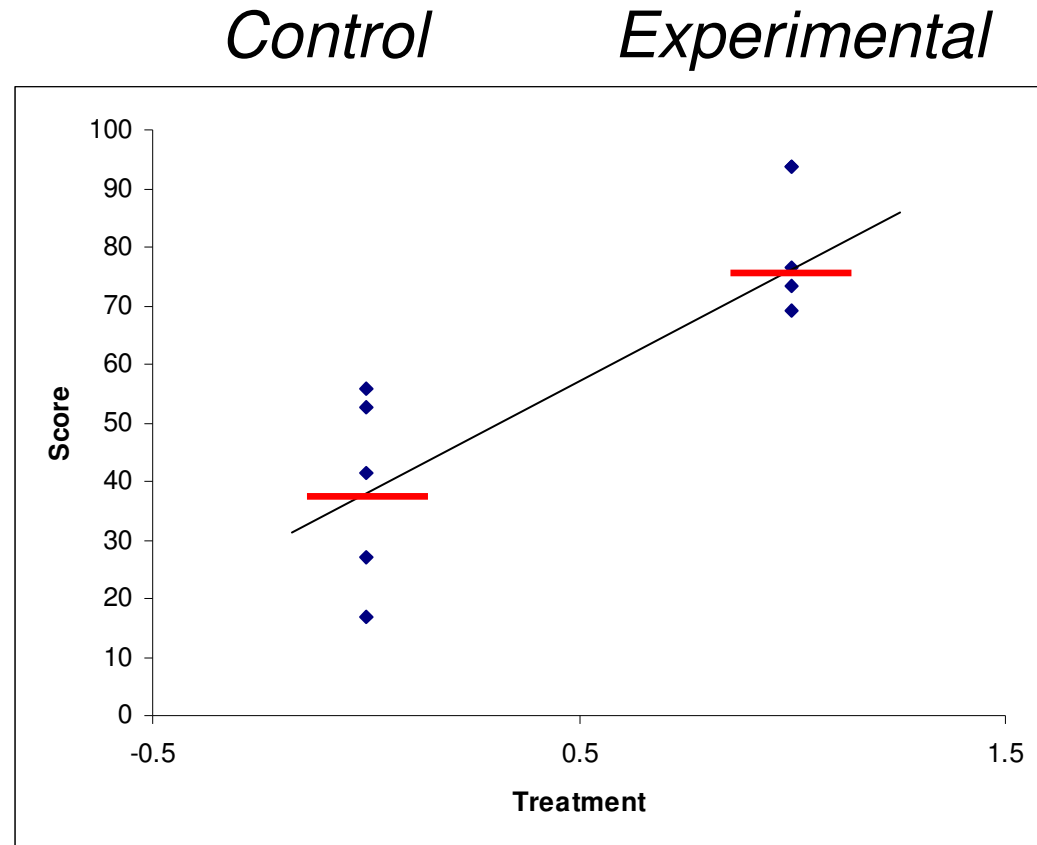
$$b7+ = a$$

$$(b7+) - (b7-) = a - b$$

$$b7- = b$$

$$\begin{matrix} E \\ Y \end{matrix}
 \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix}
 =
 \begin{matrix} X \\ X \end{matrix}
 \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 1 \\ -1 \end{bmatrix}
 \times
 \begin{matrix} \beta \\ \beta \end{matrix}
 \begin{bmatrix} a - b \end{bmatrix}
 =
 \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix}
 \begin{bmatrix} a - b \\ b - a \\ b - a \\ a - b \\ a - b \\ b - a \end{bmatrix}$$

Linear regression: t -test



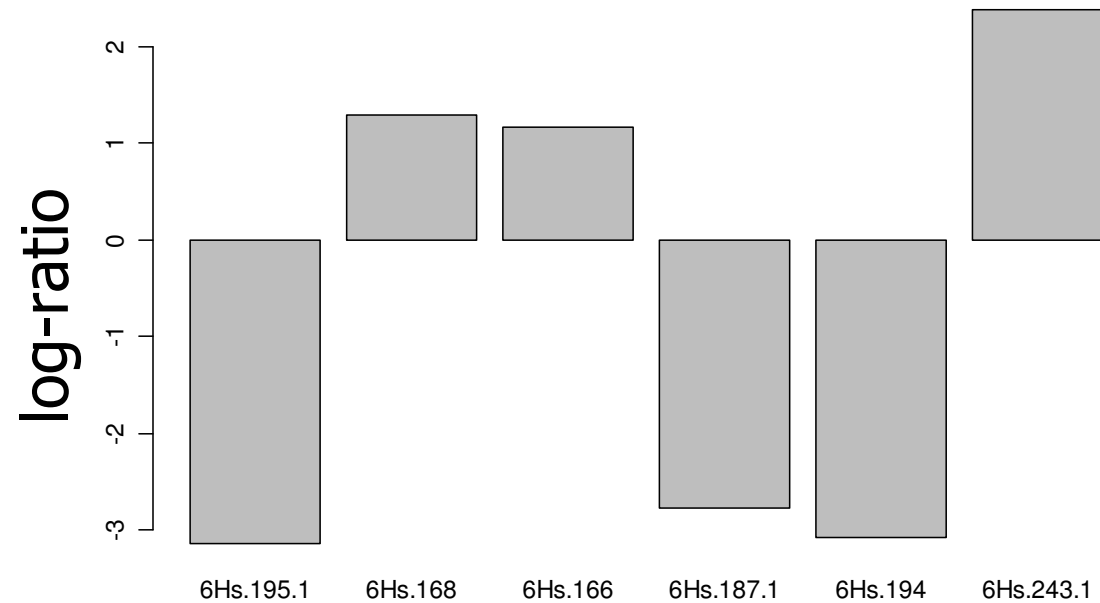
$$E(w_0, w_1) = \sum_{n=1}^N ((w_0 + w_1 x_n) - t_n)^2$$

R: limma

Fit linear model for each gene

```
library(limma)
design <- cbind(Beta7=c(1,-1,-1,1,1,-1))
fit <- lmFit(normdata, design)
```

```
barplot(normdata$M[6647,])
```



R: limma

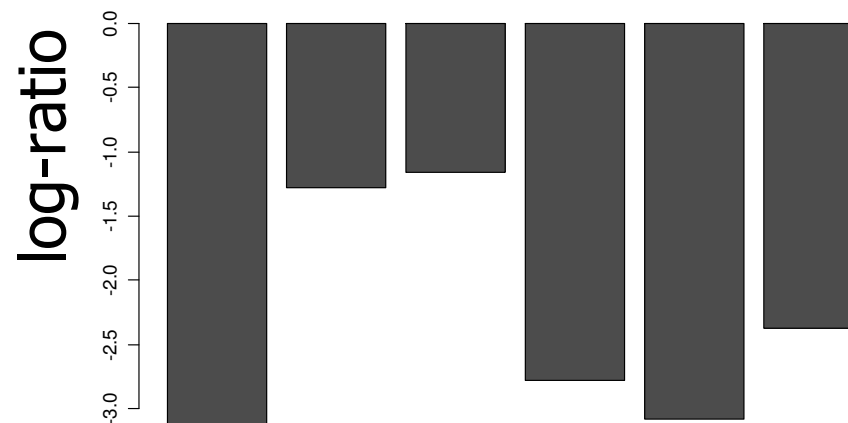
Fit linear model for each gene

```
design <- cbind(Beta7=c(1,-1,-1,1,1,-1))  
fit <- lmFit(normdata, design)
```

```
barplot(normdata$M[6647,]*t(design))
```

Coefficient estimates mean

```
> mean(normdata$M[6647,]*t(design))  
[1] -2.30042  
  
> fit$coef[6647]  
[1] -2.30042
```



R: limma

Calculate adjusted p-values (FDR):

```
tab <- topTable(fit, adjust="fdr", number=1)
tab$Name <- substring(tab$Name, 1, 10)
```

```
> tab[-c(1:4)]
```

	Name	logFC	AveExpr	t	P.Value	adj.P.Val	B
6647	GPR2 - G p	-2.30042	7.958596	-7.883697	2.511364e-05	0.3056057	0.8265352

Two-sample t -test for each gene

mean of treatment 1 observations for gene j

mean of treatment 2 observations for gene j

$$t_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{s_j^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

pooled variance estimate of σ_j^2

variance of trt 1 observations for gene j

$$s_j^2 = \frac{(n_1 - 1)s_{1j}^2 + (n_2 - 1)s_{2j}^2}{(n_1 - 1) + (n_2 - 1)}$$

variance of trt 2 observations for gene j

Problems with few degrees of freedom

- Variance estimates based on few degrees of freedom can be unreliable
- Problematic if our model for the data is not quite right
- Variances that are severely underestimated: false positives
- Variances that are severely overestimated: loss of power for detecting differentially expressed genes.


Solution: shrinkage

Standard t -statistics:

$$t_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{s_j^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Shrink the individual estimate s_j^2 towards s_0^2

degrees of freedom


$$\tilde{s}_j^2 = \frac{d s_j^2 + d_0 s_0^2}{d + d_0}$$

d_0 and s_0 estimated across all genes