Design of microarray experiments

Ulrich Mansmann

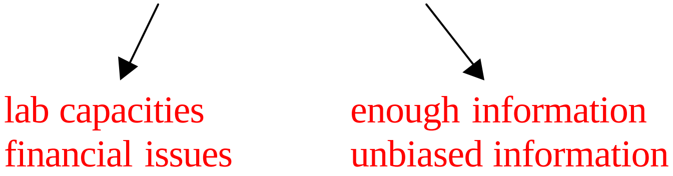
mansmann@imbi.uni-heidelberg.de

Practical microarray analysis
September 2002
Heidelberg

Motivation

The lab biologist and theoretician need to make a concerted effort to design experiments that can be *realised* and *analysed*.

Vingron M. (2001), Bioinformatics, 17:389-390



lab capacities
financial issues

enough information
unbiased information

Proper experimental design is needed to ensure that questions of interest *can be answered* and that this can be done *accurately*, given experimental constraints, such as costs of reagents and availability of mRNA

Dudoit S. (2002), Bioconductor short course

Besides economical aspects, the main task of experimental design is to remove bias - systematic error which may invalidate the result of the data analysis.

A randomisation design to elicit responses to sensitive questions

Adipose men with hypertension were asked the following question:

Are you able to get an erection during sexual activities?

In order to improve the rate of correct answers the following rule was introduced: Toss a coin, in case of *head* answer the IIEF question correctly with yes/no correctly, in case of *tail* answer an innocuous question correctly with yes/no:

Does your telephone number end with an even digit?

π = unknown proportion with erection during sexual activity, which is the parameter to be estimated

λ = known proportion with telephone number ending with an even digit

p = observed proportion of yes responses

$\frac{1}{2}\cdot\pi + \frac{1}{2}\cdot\lambda = p$, which provides an estimate $\pi^* = 2\cdot p - \lambda$

Do not believe in naive measurements.

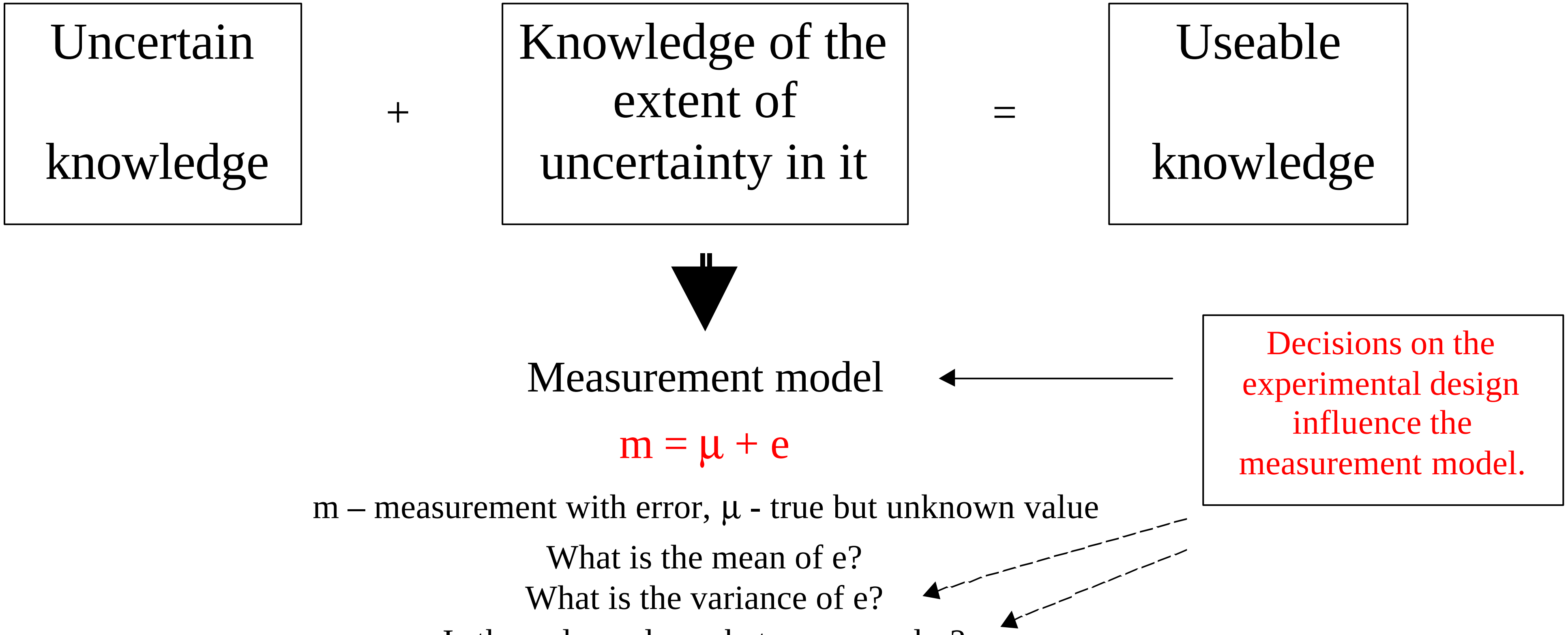
Experimental design issues for microarrys

- Design of the array itself:
  - which cDNA probe sequences to print
  - whether to use replicated probes
  - which control sequences
  - how many and where these should be printed

- Allocation of target samples to the slides
  - pairing of mRNA samples for hybridization
  - dye assignments
  - type and number of replicates

Taken from Dudoit S (2002) Bioconductor short course

Statistical thinking



Uncertain knowledge + Knowledge of the extent of uncertainty in it = Useable knowledge

Measurement model

$m = \mu + e$

Decisions on the experimental design influence the measurement model.

m – measurement with error, μ - true but unknown value

What is the mean of e?

What is the variance of e?

Is there dependence between e and μ?

What is the distribution of e (and μ)?

Typically but not always: $e \sim N(0,\sigma^2)$

Gaussian / Normal measurement model

Useable Knowledge

- Quantitative knowledge:

parameter estimates together with $(1-\alpha)\cdot 100\%$ confidence intervals (CI) for a normal measurement model: $\mu \in [m - z_{1-\alpha/2}\cdot SE, m + z_{1-\alpha/2}\cdot SE]$

SE: standard error

CI gives information on the precision of estimates, how close estimate and true but unknown value are.

- Qualitative knowledge:

Rejection of a null hypothesis, statistical test

No evidence for a difference is not evidence for no difference

Taming of Uncertainty

How to handle variances and variability?

1. $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ for two independent random variables X and Y
2. $\mathrm{Var}(c \cdot X) = c^2 \cdot \mathrm{Var}(X)$ for c, a real constant, and random variable X

Application: Variance of the arithmetic mean (SE^2)

$$m = \frac{1}{n} \sum_{i=1}^{n} X_i$$

with X_i independent and identical distributed (iid) with $\mathrm{Var}(X_i) = \sigma^2$

$$\mathrm{Var}(m) = n^{-2} \cdot \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = n^{-2} \cdot \sum_{i=1}^{n} \mathrm{Var}(X_i) = n^{-1}\,\sigma^2$$

The Statistical test

- Question of interest (*Alterative*): Is the gene G differentially expressed between two cell populations?
- Answer question via *reductio ad absurdum* (proof by contradiction): Show that there is no evidence to support the logical contrary of the *alternative.* The logical contrary of the *alternative* is called *null hypothesis*.
- *Null hypothesis*: The gene G is not differentially expressed between two cell populations of interest.
- A *test statistic* **T** is introduced which measures the fit of the observed data to the *null hypothesis.*
  A *test distribution* **P** is introduced which quantifies the variability of the *test statistic* T in case the *null hypothesis* is true.
  It will be checked if the *test statistic* evaluated at the observed data t_{obs} behaves typically (not extreme) with respect to the *test distribution*. The *p-value* is calculated: $\mathbf{P}(T \geq t_{obs}) = p$.
  A criteria is needed to asses not typical or extreme behaviour of the test statistic via the *p – value* which is called the *level of the test:* **a .**
- The observed data does not fit to the null hypothesis if $p < \mathbf{a}$ or $t_{obs} > t_{1-\mathbf{a}}$ where $t_{1-\mathbf{a}}$ is the 1-α quantile of the test distribution P. $t_{1-\mathbf{a}}$ is also called the *critical value.* The conditions $p < \mathbf{a}$ and $t_{obs} > t_{1-\mathbf{a}}$ are equivalent.
- If $p < \mathbf{a}$ or $t_{obs} > t_{1-\mathbf{a}}$ the null hypothesis will be rejected. If $p \geq \mathbf{a}$ or $t_{obs} \geq t_{1-\mathbf{a}}$ the null hypothesis can not be rejected – this does not mean that it is true.

The statistical test is a decision problem

True state of nature **Test result**	Gene is differentially expressed between two cell populations	Gene is not differentially expressed between two cell populations
p < a	**OK**	**false positive decision** happens with probability α
p ³ a	**false negative decision** happens with probability β	**OK**

Two sources of error: false positive rate a, false negative rate b

Power of a test: Ability to detect a difference if there is a true difference
Power – true positive rate.

Power = 1 - b

Controlling the power – sample size calculations

The test should produce a significant result (level α) with a power of $1-\beta$ if the true difference in expression is $\Delta = \mu_1 - \mu_2$.

A normal measurement model is assumed: $m_i \sim N(\mu_i, \sigma^2/n_i)$

m_i - observed arithmetic mean (of log-transformed expression) in cell population i versus a reference population

μ_i - true but unknown (log-transformed) expression level,
σ^2 - variability in individual observation,
n_i – number of probes from cell polulation i (i = 1, 2,)

Test statistics: $D = m_1 - m_2$

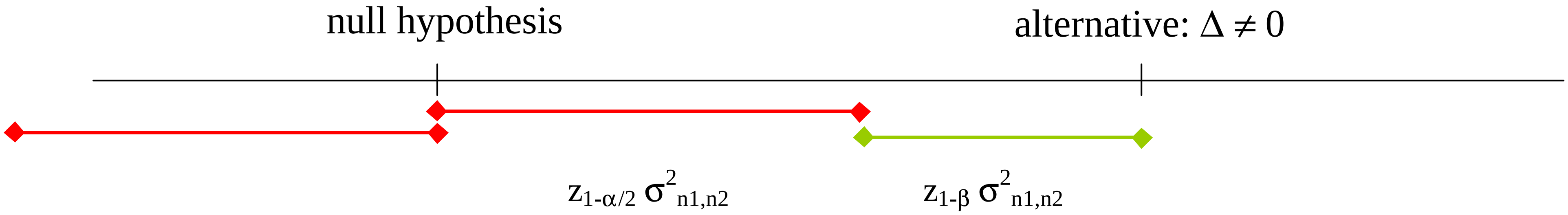
$$\mathrm{Var}(m_1 - m_2) = \mathrm{Var}(m_1) + \mathrm{Var}(m_2)$$
$$= \sigma^2 \cdot [1/n_1 + 1/n_2] = \sigma^2_{n1,n2}$$

Distribution of D under:

null hypothesis $(\Delta=0)$: $D \sim N(0, \sigma^2_{n1,n2})$
alternative $(\Delta\neq0)$: $D \sim N(\Delta, \sigma^2_{n1,n2})$

Controlling the power – sample size calculations

The test should produce a significant result (level α) with a power of $1-\beta$ if the true difference in expression is $\Delta = \mu_1 - \mu_2$.



null hypothesis

alternative: $\Delta \neq 0$

$z_{1-\alpha/2}\ \sigma^2_{n1,n2}$

$z_{1-\beta}\ \sigma^2_{n1,n2}$

The above requirement is fulfilled if: $\Delta = (z_{1-\alpha/2} + z_{1-\beta})\cdot\sigma^2_{n1,n2}$

or

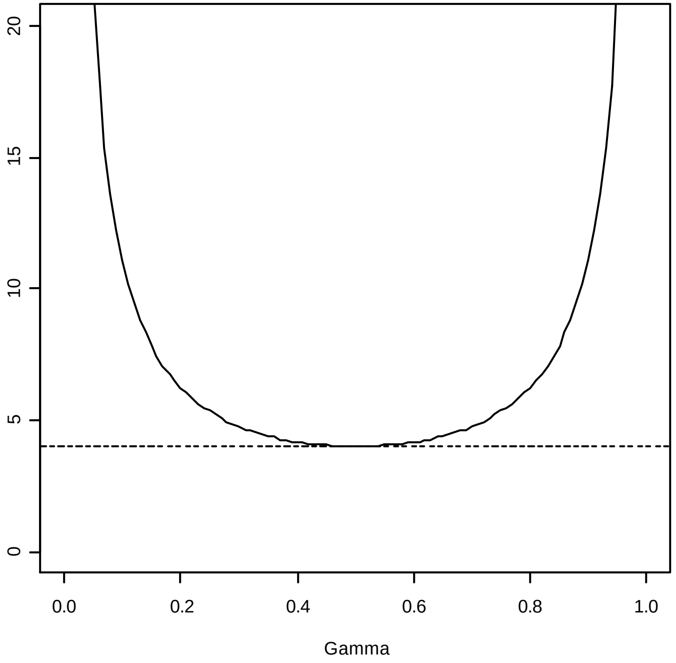
$$\frac{n_1 \cdot n_2}{n_1 + n_2} = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \cdot \sigma^2}{\Delta^2}$$

Controlling the power – sample size calculations

$$\frac{n_1 \cdot n_2}{n_1 + n_2} = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \cdot \sigma^2}{\Delta^2}$$

$n_1 = n \cdot \gamma$ and $n_2 = n \cdot (1-\gamma)$ with n – total size of experiment and $\gamma \in\ ]0,1[$

$$n = \frac{1}{\gamma \cdot (1-\gamma)} \cdot \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \cdot \sigma^2}{\Delta^2}$$



The size of the experiment is minimal if $\gamma = ½$.

Measurement model for cDNA arrays

Gene expression under condition 2 – intensity of red colour,
Gene expression under condition 1 – intensity of green colour

$$\text{Measurement: } m_{\text{Red2}} = \text{Log}_2\left(\frac{R_{\text{Gene–C2}}}{G_{\text{Gene–C1}}}\right) = \gamma_{12} + \delta + e$$

γ_{12} – log-transformed true differential expression of gene between condition 1 and 2
δ - dye effect, e – measurement error with E[e] = 0 and Var(e) = σ^2

$$\text{If colour is swapped C2} \rightarrow \text{green, C1} \rightarrow \text{red: } m_{\text{Red1}} = \text{Log}_2\left(\frac{R_{\text{Gene–C1}}}{G_{\text{Gene–C2}}}\right) = -\gamma_{12} + \delta + e$$

Consider $m = ½\,(m_{\text{Red2}} - m_{\text{Red1}})$ with $E[m] = \gamma_{12}$ and $\text{Var}(m) = ½\cdot\sigma^2$

Dye effect is removed and precision is increased
without increasing the actual sample size

Consequences of a design desicion

- n arrays used without dye swap:

γ_{12} is estimated by $\frac{1}{n}\sum_{i=1}^{n} m_i$ with precision σ^2/n and possible bias δ.

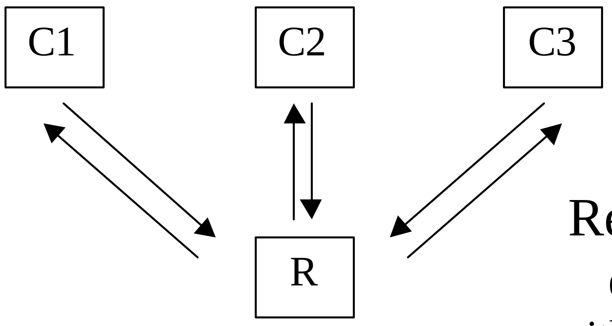
- n arrays used with dye swap:

This results in n/2 pairs of arrays

γ_{12} is estimated by $\frac{2}{n}\sum_{i=1}^{n/2} 0.5 \cdot (m^{i}_{Red2} - m^{i}_{Red1})$

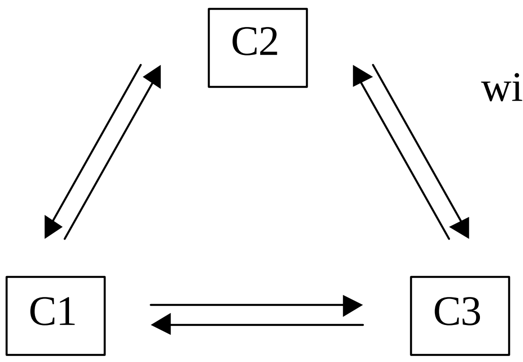
with precision $[4/n^2]\cdot[n/2]\cdot(1/4)\cdot 2\cdot\sigma^2 = \sigma^2/n$ and no bias δ.

Graphical representations of experiments: Multi – digraphs



Reference design
with dye swap

most information is collected on a condition which is of no interest



All pairs design
with dye swap

- *Vertices* mRNA samples
- *Edges* hybridization
- *Direction* Dye assignment

Green → Red

Which design gives the most precise estimate of the contrast $\gamma_{12}, \gamma_{13}, \gamma_{23}$?

Comparing the *reference* and *all pair* design

- Reference design:
  - Each pair of slides estimates γ_{RC} with precision $\frac{1}{2}\cdot\sigma^2$.
  - To get γ_{CaCb} it is necessary to subtract the estimate of γ_{RCa} from γ_{RCb}

    $\gamma_{CaCb} = \log_2[E_b/E_a] = \log_2[E_b\cdot E_R /E_a\cdot E_R] = \log_2[E_b/E_R] - \log_2[E_a/E_R] = \gamma_{RCb} - \gamma_{RCa}$
  - The estimate of γ_{CaCb} has precision σ^2.
  - The six slides used give the estimates looked for with precision σ^2.
  - If every pair of slides is replicated and estimates of two equal pairs are combined by taking the average, the resulting precision of the estimated γ_{CaCb} is $\frac{1}{2}\cdot\sigma^2$.

- All pair design:
  - Each pair of slides estimates γ_{CaCb} with precision $\frac{1}{2}\cdot\sigma^2$.

- Summary: For the same precision the *reference design* requires two times as many hybridizations or slides as the *all pair* design.

Graphical representation – summary

- The structure of the graph determines which effects can be estimated and the precision of the estimates:
  - Two mRNA samples can be compared only if there is a path joining the corresponding two vertices
  - The precision of the estimated contrast then depends on the number of paths joining the two vertices and is inversely related to the length of the paths.
- Direct comparisons within slides yield more precise estimates than indirect ones between slides.

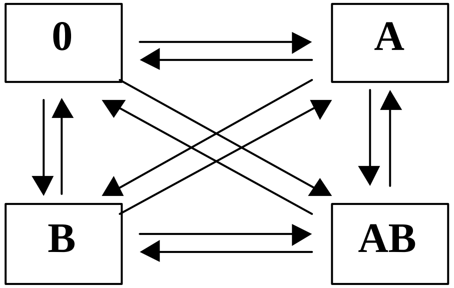
2x2 factorial experiments

two factors, two levels each

Study the **joint** effect of two conditions / treatments, A and B, on the gene expression of a cell population of interest.

There are four possible condition / treatment combinations:

AB:	both treatments/conditions are applied
A:	only treatment/condition A is applied
B:	only treatment/condition B is applied
0:	cells are not treated or exposed



Design with 12 slides

2x2 factorial experiments

two factors, two levels each

For each gene, consider a linear model for the joint effect of A and B on the expression:

ν: baseline effect

α: main effect if treatment/condition A is applied

β: main effect if treatment/condition B is applied

ψ: interaction between A and B

$$\mu_0 = \nu$$

$$\mu_A = \nu + \alpha$$

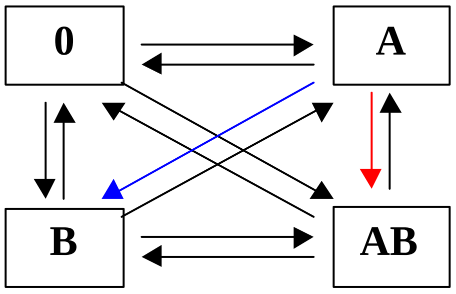
$$\mu_B = \nu + \beta$$

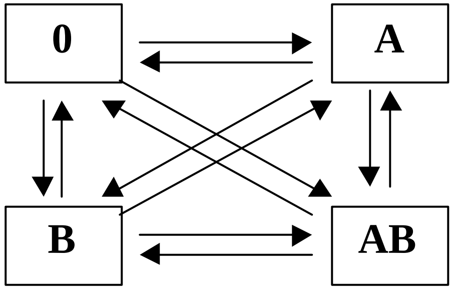
$$\mu_{AB} = \nu + \alpha + \beta + \psi$$

Log-ratio M for hybridisation A → AB estimates $\mu_{AB} - \mu_A = \beta + \psi$

Log-ratio M for hybridisation A → B estimates $\mu_B - \mu_A = \beta - \alpha$

+ 10 others





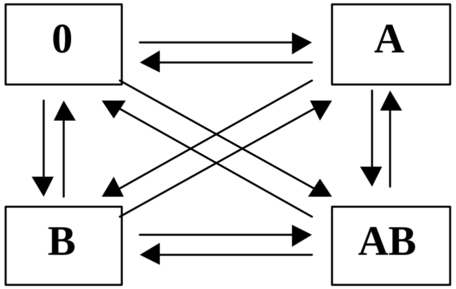
Regression analysis

$$E\begin{pmatrix} M_{0\to A} \\ M_{0\leftarrow A} \\ M_{0\to B} \\ M_{0\leftarrow B} \\ M_{0\to AB} \\ M_{0\leftarrow AB} \\ M_{A\to AB} \\ M_{A\leftarrow AB} \\ M_{B\to AB} \\ M_{B\leftarrow AB} \\ M_{A\to B} \\ M_{A\leftarrow B} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \\ 1 & 1 & 1 \\ -1 & -1 & -1 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \\ 1 & 0 & 1 \\ -1 & 0 & -1 \\ -1 & 1 & 0 \\ 1 & -1 & 0 \end{pmatrix} \bullet \begin{pmatrix} \alpha \\ \beta \\ \psi \end{pmatrix}$$

- For parameter $\theta = (\alpha, \beta, \psi)$ define the design matrix X such that $E(M) = X\theta$.
- For each gene, compute least square estimate $\theta^* = (X'X)^{-1}X'M$ (BLUE)
- Obtain measures of precision of estimated effects.
- Use all possibilities of the theory of linear models.

Design problem:

- Assume each measurement M is made with variability σ^2. How precise can we estimate the components or contrasts of θ?
  Answer: Look at $(X'X)^{-1}$



2 x 2 factorial designs

```
> x.mat
     alpha beta   psi
0>A      1    0     0
0<A     -1    0     0
0>B      0    1     0
0<B      0   -1     0
0>AB     1    1     1
0<AB    -1   -1    -1
A>AB     0    1     1
A<AB     0   -1    -1
B>AB     1    0     1
B<AB    -1    0    -1
A>B     -1    1     0
A<B      1   -1     0
```

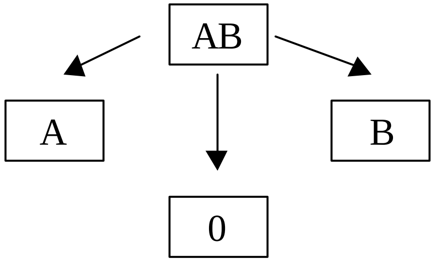
```
> precision.rfc(x.mat)
$inv.mat

         alpha   beta   psi
  alpha  0.250  0.125 -0.25
  beta   0.125  0.250 -0.25
  psi   -0.250 -0.250  0.50

$effects
alpha  beta   psi   A-B
 0.25  0.25  0.50  0.25
```

Var(A-B) = Var(A) + Var(B) - 2·Cov(A,B)

2 x 2 factorial designs

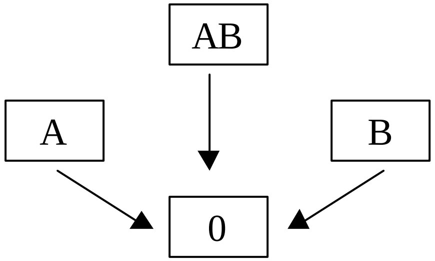
Design I

Common ref.



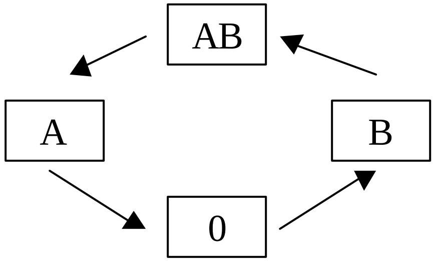
Design II

Common ref.



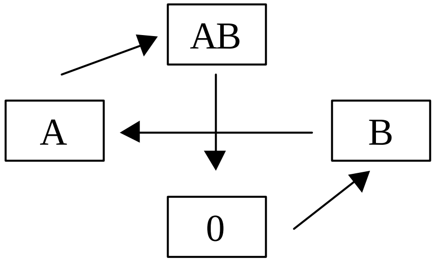
Design III

Connected



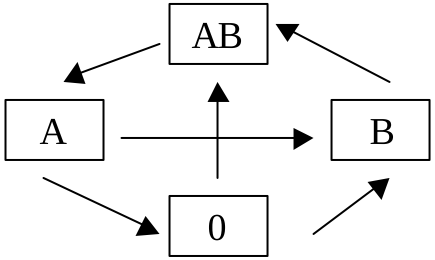
Design IV

Connected



Design V

All-pairs



Scaled variances of estimated effects

```
compare.2.by.2.designs.rfc()
      D.I D.II  D.III  D.IV  D.V D.tot
alpha   2    1   0.75  1.00  0.5  0.25
beta    2    1   0.75  0.75  0.5  0.25
psi     3    3   1.00  2.00  1.0  0.50
A-B     2    2   1.00  0.75  0.5  0.25
```

Experimental Design - Conclusions

- Designs for *time course* experiments: Yee Hwa Yang (2002)
- In addition to experimental constraints, design decisions should be guided by knowledge of which effects are of greater interest to the investigator.
- The experimenter should decide on the comparisons for which s/he wants the most precision and these should be made within slides to the extent possible.
- Efficiency of an experimental design can be measured in terms of different quantities (number of slides, units of biological material)
- Issues:

  - Replication, type of replication

    within or between, biological or technical,

    generalizibility vs. reproducibility

  - Sample size and power calculations
  - Dye assignment

- Fundamental principles of good design: balance and replication

  Balance *insures that the effects of interest are not confounded with other sources of variation.* ***Replication*** *improves the precision of estimates and provides degrees of freedom for error estimation.*

- Further reading: Kerr MK, Churchill GA (2001) *Experimental design for gene expression microarrays,* Biostatistics, 2:183-201