Design of microarray experiments

Ulrich Mansmann

mansmann@imbi.uni-heidelberg.de

Practical microarray analysis
October 2003
Heidelberg

Experiments

Scientists deal mostly with experiments of the following form:

- A number of alternative conditions / treatments
- one of which is applied to each experimental unit
- an observation (or several observations) then being made on each unit.

The objective is:

- Separate out differences between the conditions / treatments from the uncontrolled variation that is assumed to be present.
- Take steps towards understanding the phenomena under investigation.

Statistical thinking

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

Measurement model

$m = \mu + e$

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)?

Decisions on the experimental design influence the measurement model.

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

Gaussian / Normal measurement model

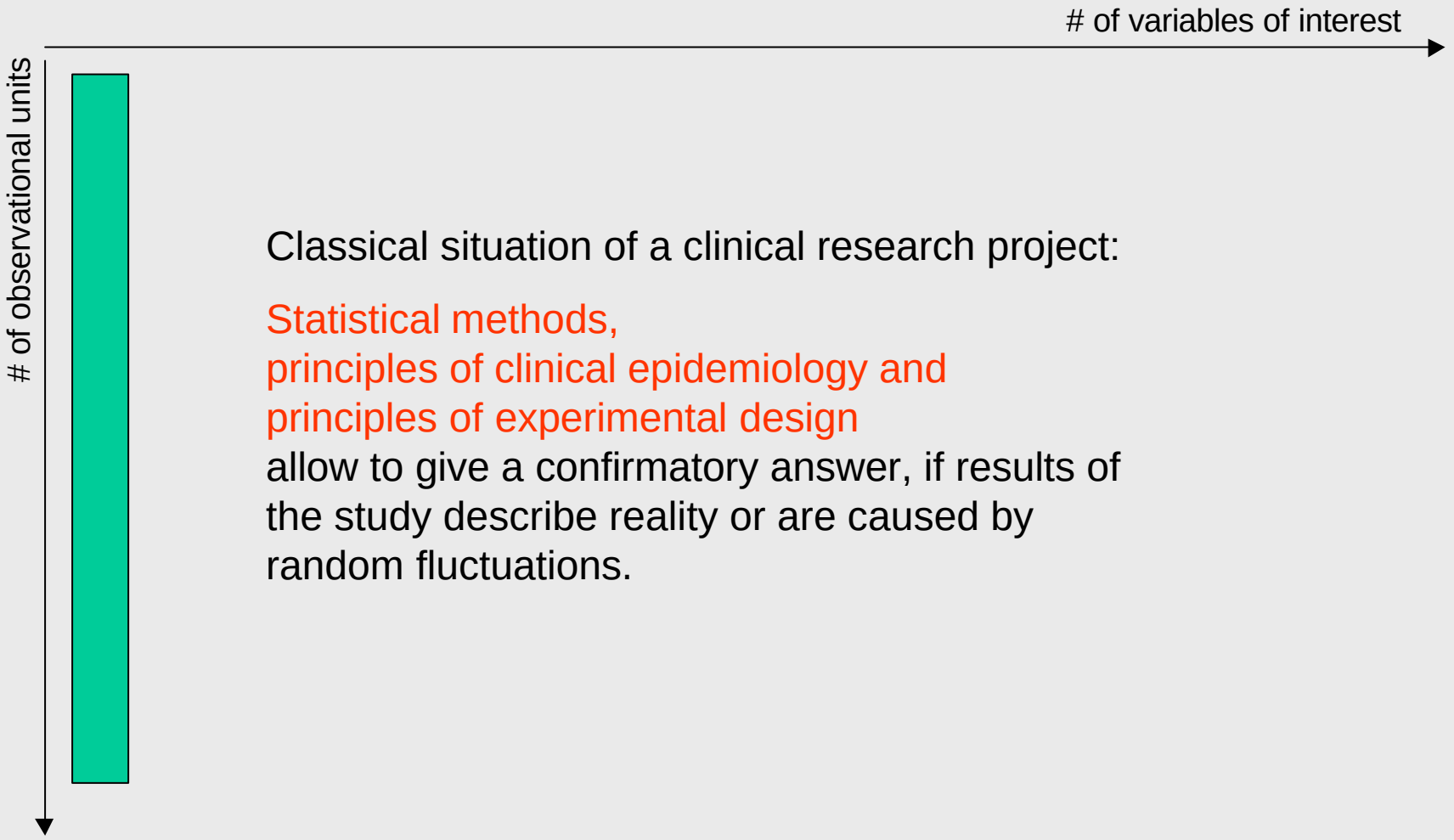
Main requirements for experiments

Once the *conditions / treatments*, *experimental units*, and the *nature of the observations* have been fixed, the main requirements are:

- Experimental units receiving different treatments should differ in no systematic way from one another – Assumptions that certain sources of variation are absent or negligible should, as far as practical, be avoided;
- Random errors of estimation should be suitably small, and this should be achieved with as few experimental units as possible;
- The conclusions of the experiment should have a wide range of validity;
- The experiment should be simple in design and analysis;
- A proper statistical analysis of the results should be possible without making artificial assumptions.

Taken from Cox DR (1958) *Planning of experiments*, Wiley & Sons, New York (page 13)

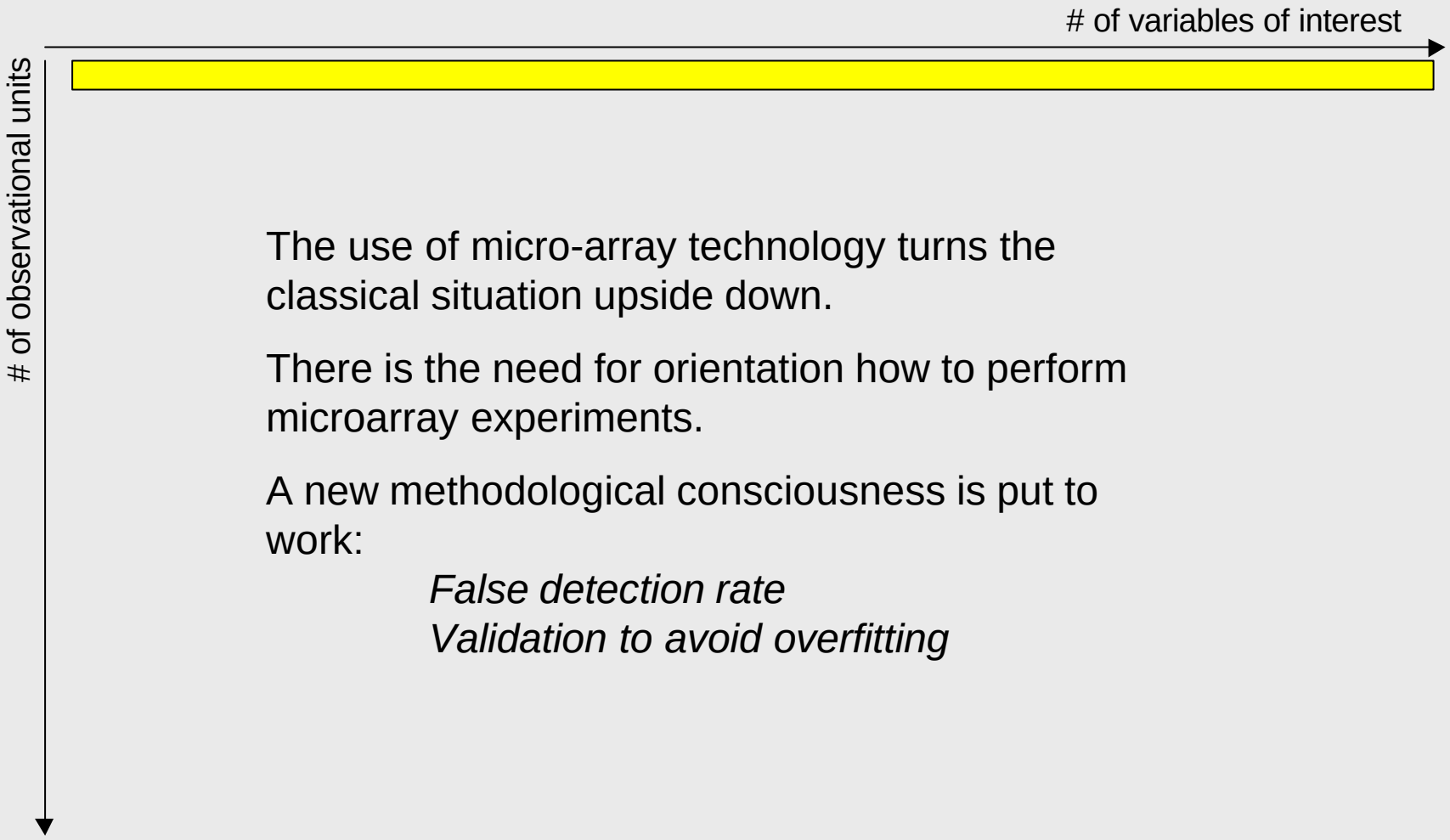
Information dilemma: too many or too few?



Classical situation of a clinical research project:

Statistical methods,
principles of clinical epidemiology and
principles of experimental design
allow to give a confirmatory answer, if results of the study describe reality or are caused by random fluctuations.

Information dilemma: too many or too few?



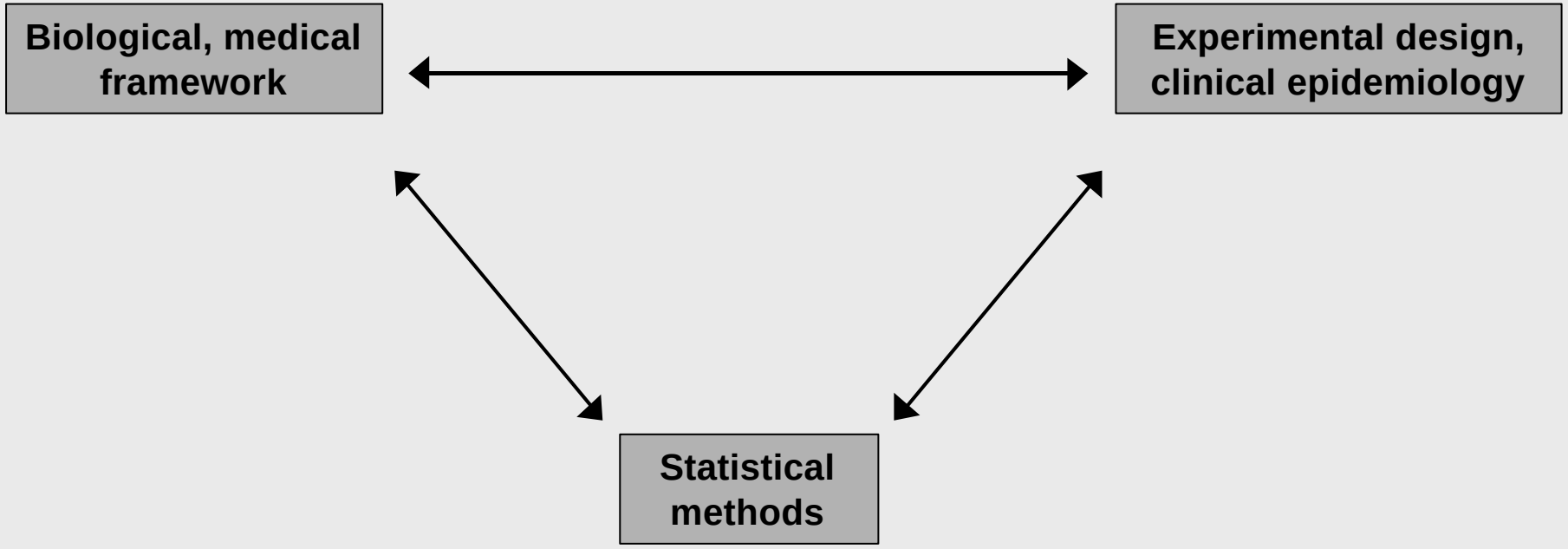
The use of micro-array technology turns the classical situation upside down.

There is the need for orientation how to perform microarray experiments.

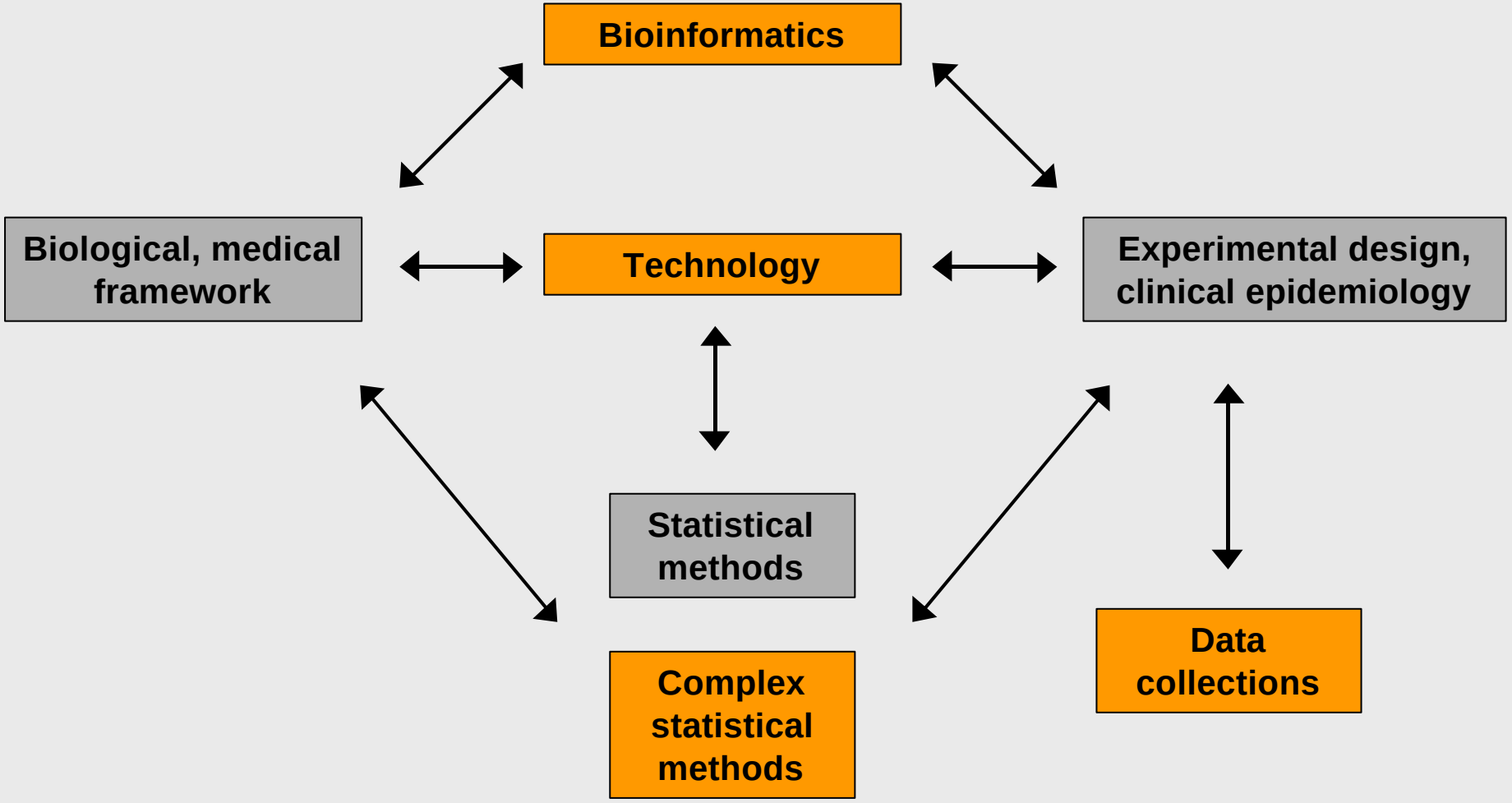
A new methodological consciousness is put to work:

False detection rate
Validation to avoid overfitting

Biometrical practice

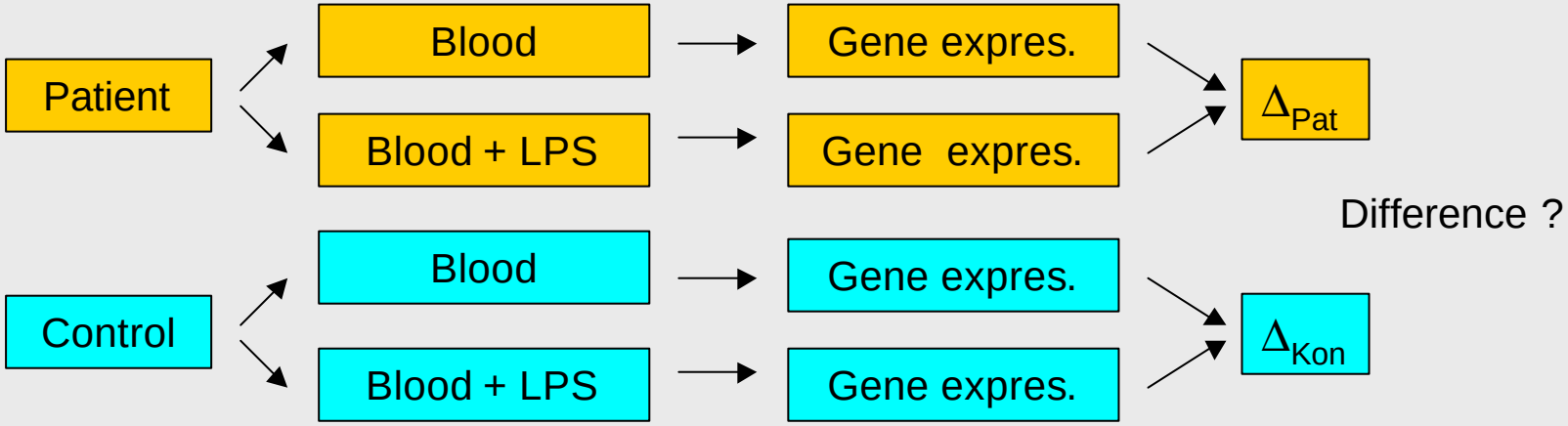


Micro-array experiments



Example LPS: *The setting*

Problem: Differential reaction on LPS stimulation in peripheral blood of stroke patients and controls?



Sample size has to be chosen with respect to financial restrictions
Peripheral blood is a special tissue, possible confounder

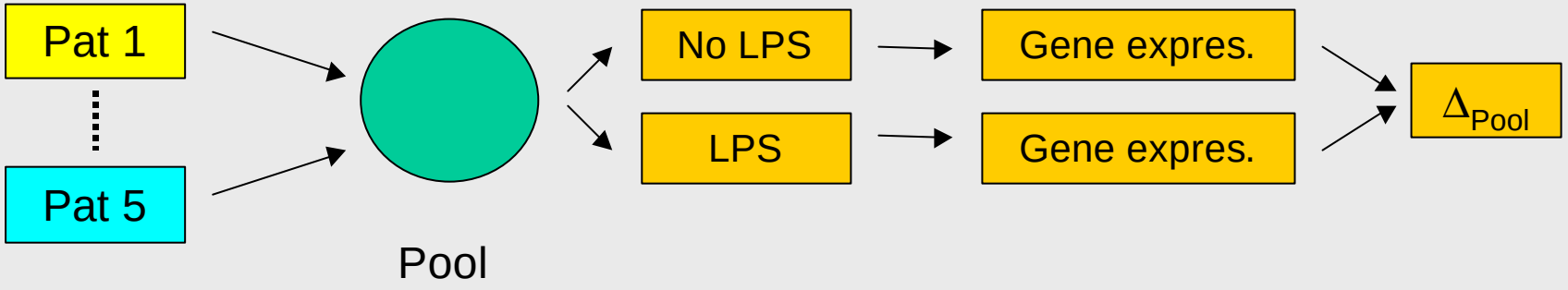
PNAS, 100:1896-1901

Chosen technology: Affymetrix (22283 genes)

Example LPS: *Design - Pooling*

Assume a linear model for appropriately transformed gene expression:

$y_{Pat,Gen}$ = transformed abundance + confounder effect + biol. var. + techn. var.



Correction for confounding - if composition of pools is homogeneous over possible confounder

Reduction of biological variability: σ_{biol}

No reduction of technological / array specific variability: σ_{tech}

Reduction of arrays is determined by $\Psi = \sigma_{tech} / \sigma_{biol}$.

Example LPS: *Design - Gene exclusion*

Array used codes for ~ 18000 genes

Do we have good rules to reduce the set of interesting genes?
How can we introduce a hierarchy into the gene list without manipulating the result of our analysis?

Possible solutions:

Bioinformatics:	Integration of pathway information into the analysis	
Statistics:	Use of genes with high inter-array variability - set cut-point	
	Meta-genes (West et al.) -	predefine # of meta-genes define cluster strategy

Example LPS: *Design - Gene exclusion*

Array used codes for ~ 18000 genes

Do we have good rules to reduce the set of interesting genes?
How can we introduce a hierarchy into the gene list without manipulating the result of our analysis?

Possible solutions:

Bioinformatics: Integration of pathway information into the analysis

Only possible for small problems

Statistics: Use of genes with high inter-array variability - set cut-point

Meta-genes (West et al.) - predefine # of meta-genes
define cluster strategy

Example LPS: *Design - Gene exclusion*

Array used codes for ~ 18000 genes

Do we have good rules to reduce the set of interesting genes?
How can we introduce a hierarchy into the gene list without manipulating the result of our analysis?

Possible solutions:

Bioinformatics: Integration of pathway information into the analysis

Only possible for small problems

Statistics: Use of genes with high inter-array variability - set cut-point

Mostly heuristic procedures / Kropf et al.

Meta-genes (West et al.) - predefine # of meta-genes
define cluster strategy

Example LPS: *Design - Gene exclusion*

Array used codes for ~ 18000 genes

Do we have good rules to reduce the set of interesting genes?
How can we introduce a hierarchy into the gene list without manipulating the result of our analysis?

Possible solutions:

Bioinformatics: Integration of pathway information into the analysis

Only possible for small problems

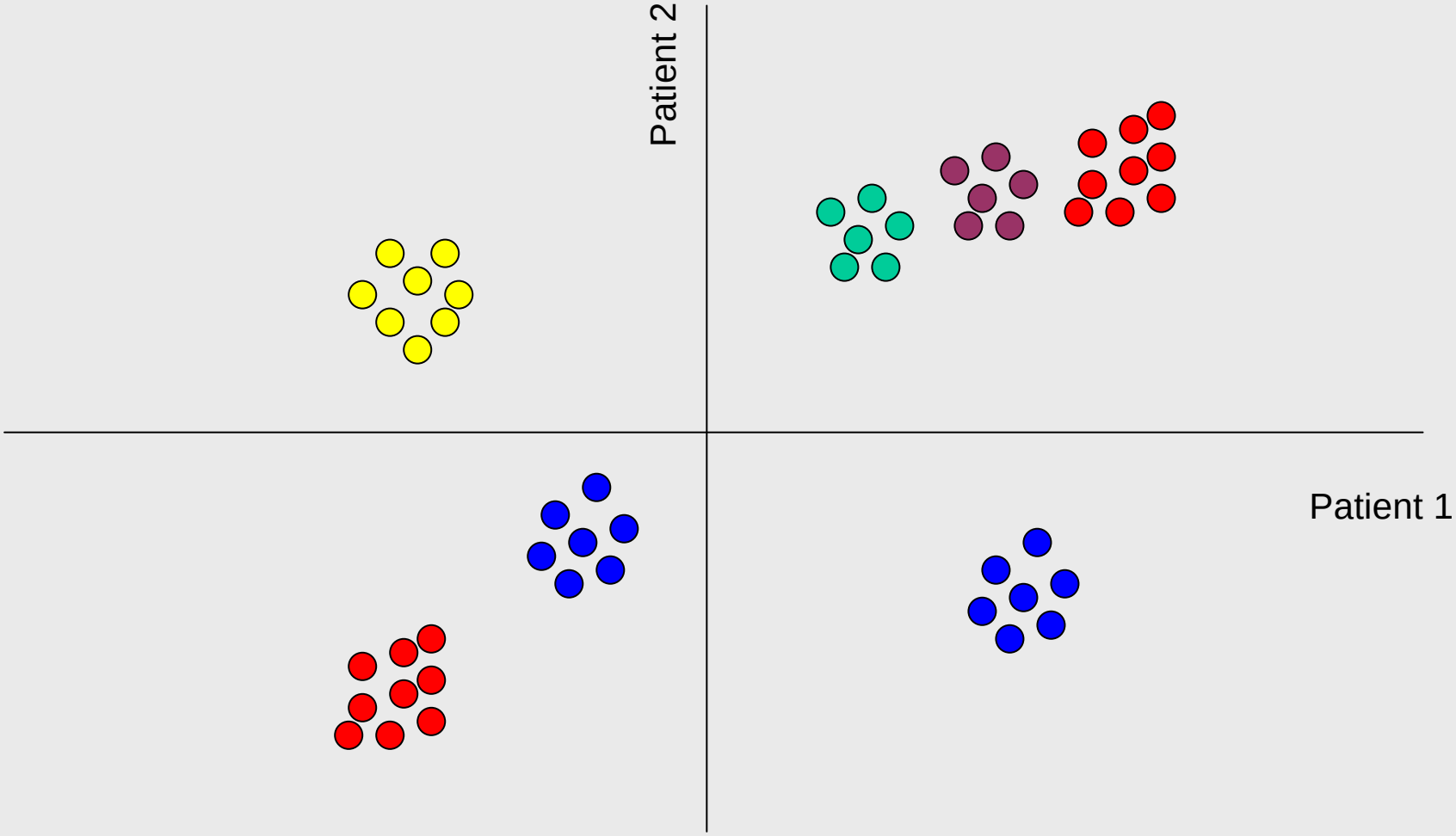
Statistics: Use of genes with high inter-array variability - set cut-point

Mostly heuristic procedures / Kropf et al.

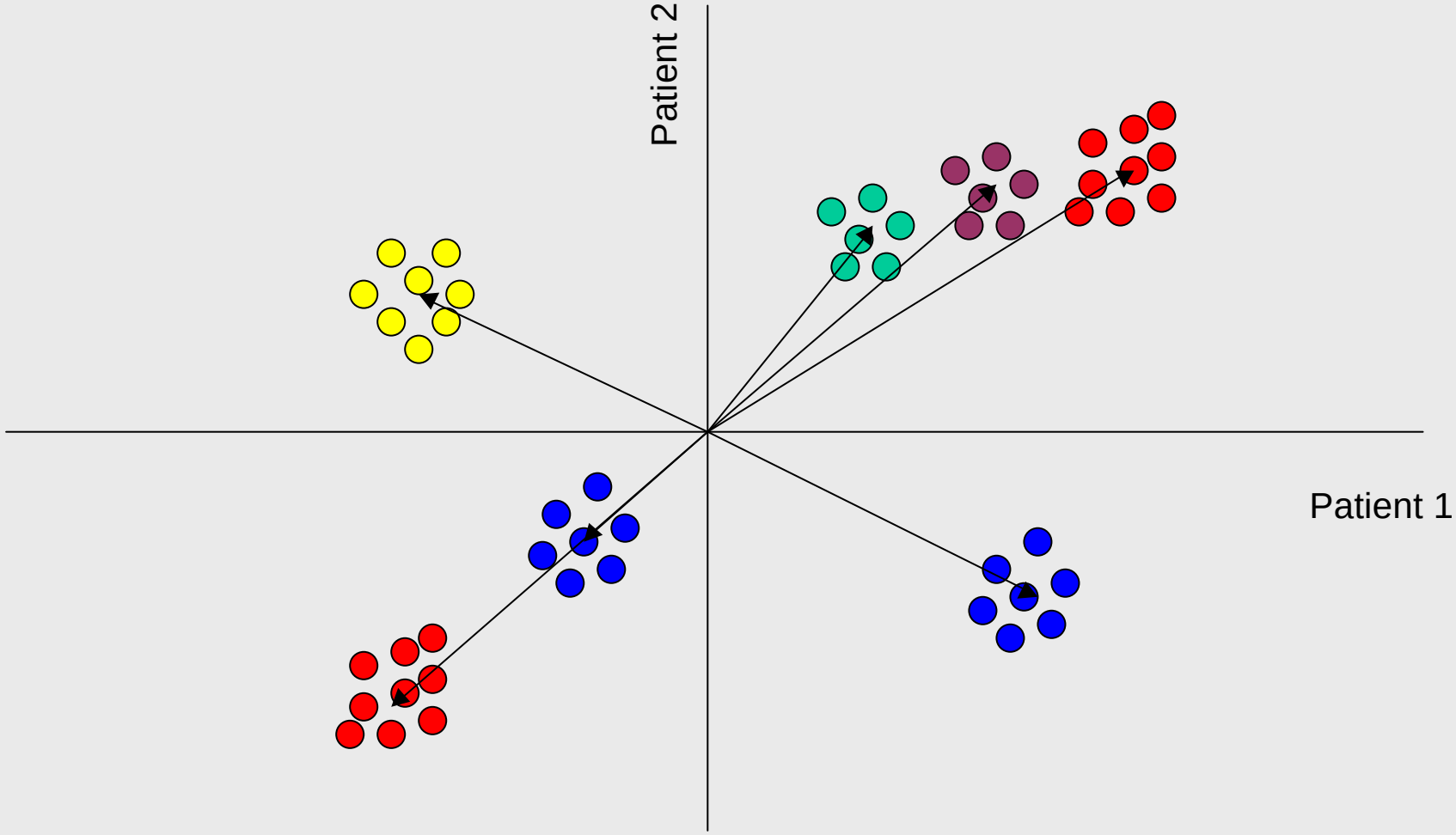
Meta-genes (West et al.) - predefine # of meta-genes
define cluster strategy

Not well evaluated

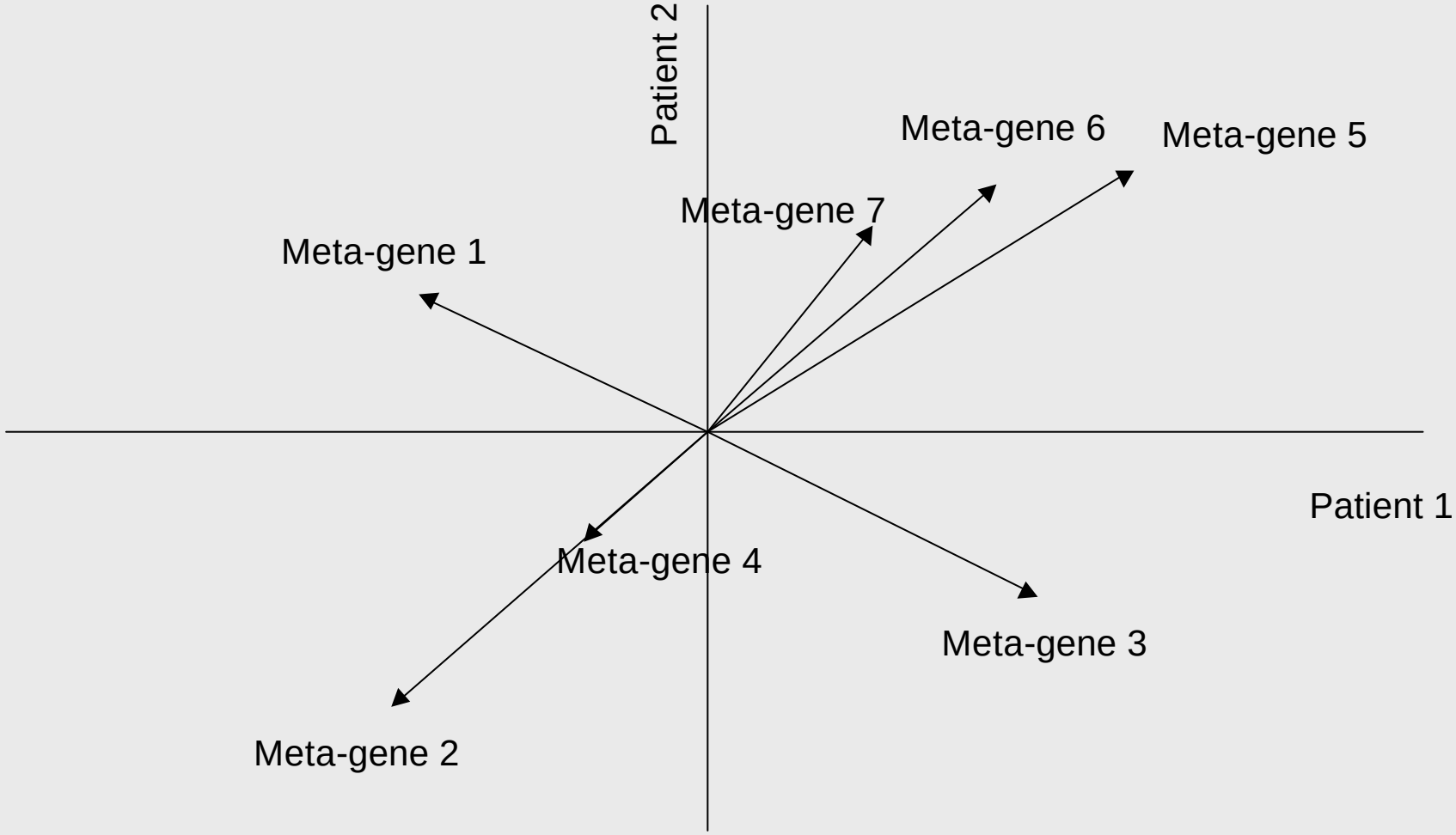
Meta-genes



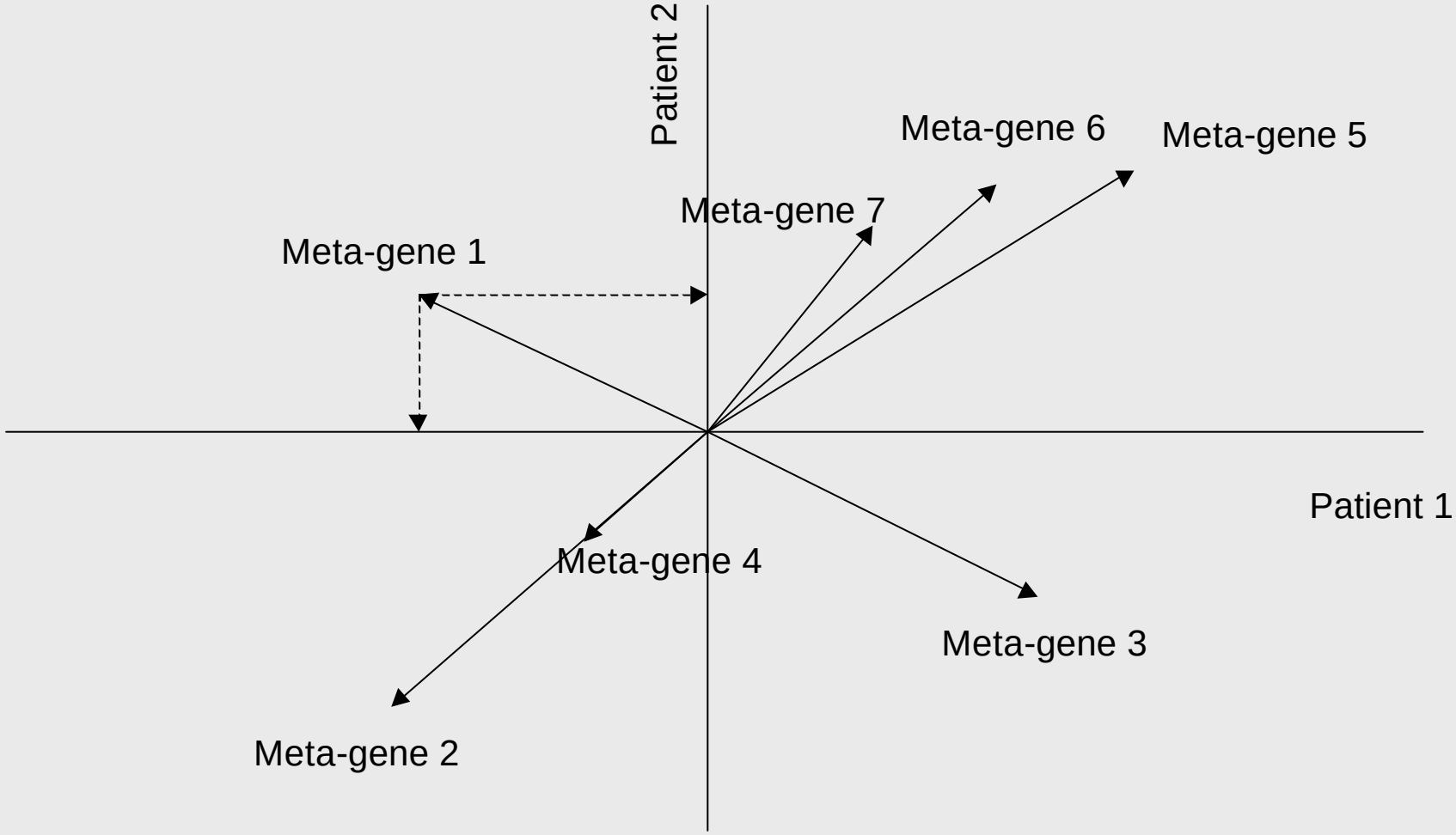
Meta-genes



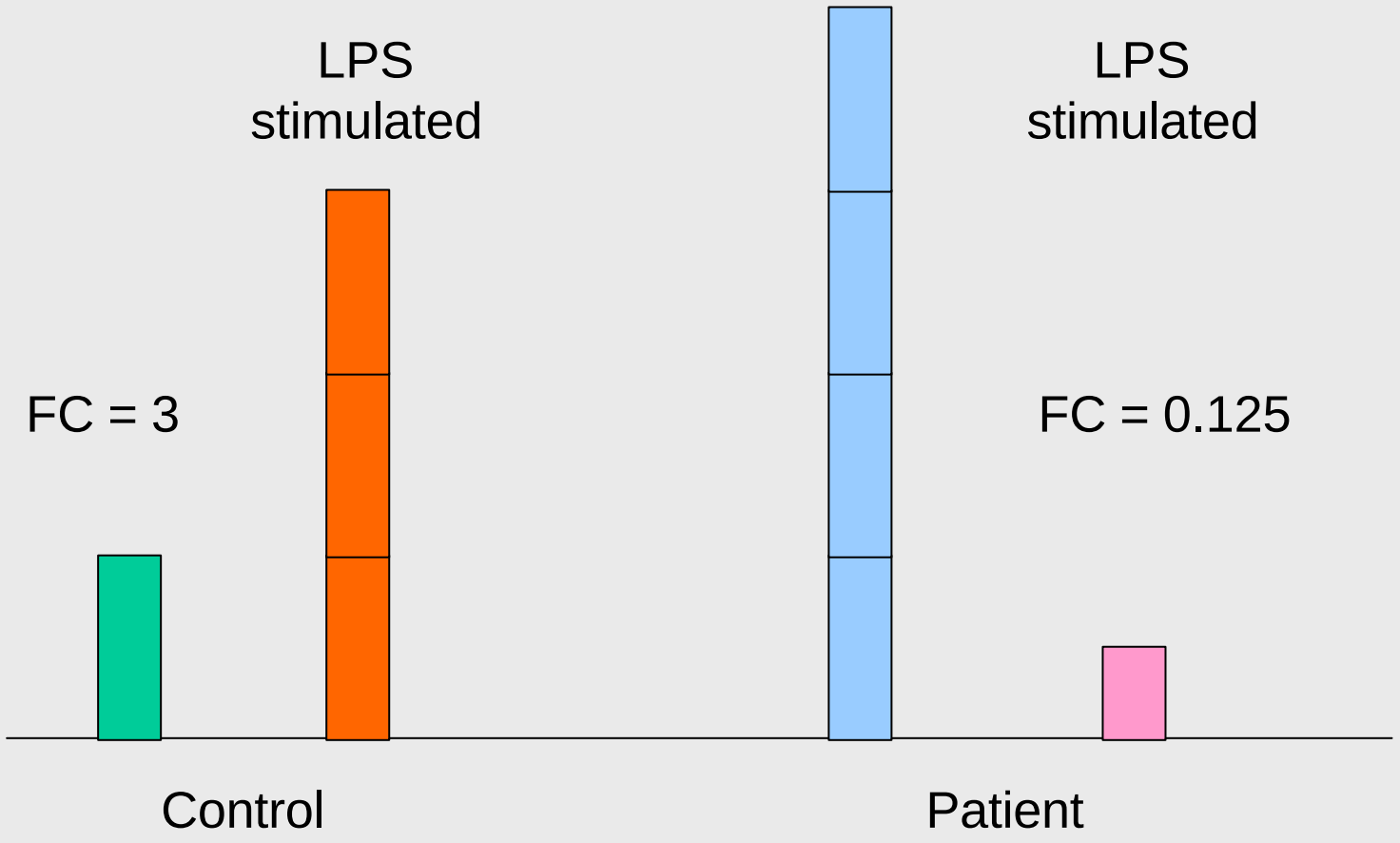
Meta-genes


Patient 2
Meta-gene 6
Meta-gene 5
Meta-gene 7
Meta-gene 1
Patient 1
Meta-gene 4
Meta-gene 3
Meta-gene 2

Meta-genes


Patient 2
Meta-gene 6
Meta-gene 5
Meta-gene 7
Meta-gene 1
Patient 1
Meta-gene 4
Meta-gene 3
Meta-gene 2

Example LPS: *Differential reaction DR*



Differential reaction (DR): log(0.125 / 3) = log(0.125) - log(3) = - 3.18

$DR = \Delta_{Pat} - \Delta_{Kon}$

Example LPS: *Data*

Pool (5 subjects)	Group	Sex distribution (Male:Female)	mean age
1	Control	1:4	60.8
2	Control	1:4	65.4
3	Control	2:3	61.6
4	Patient	4:1	64.4
5	Patient	5:0	66.2
6	Patient	3:2	74.4

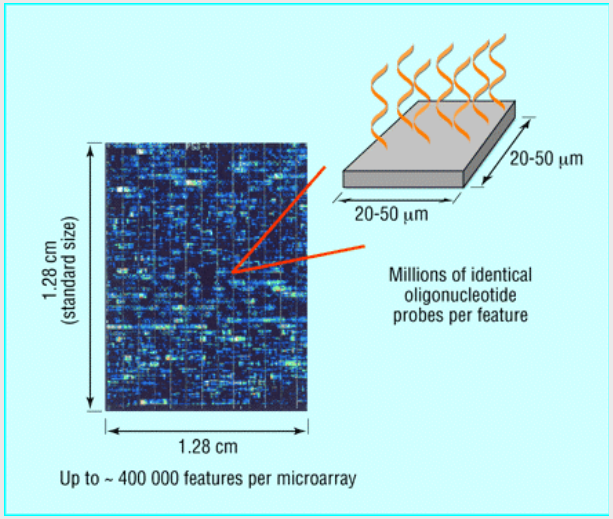
Example LPS: *Expression Summaries*

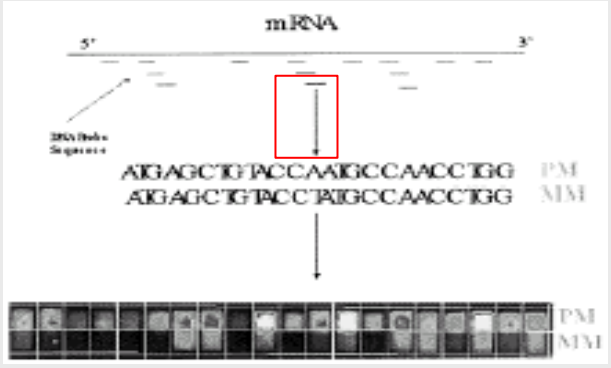
Quantification of expression

MSA5: *Tukey bi-weight* signal of PM/MM, which is log-transformed

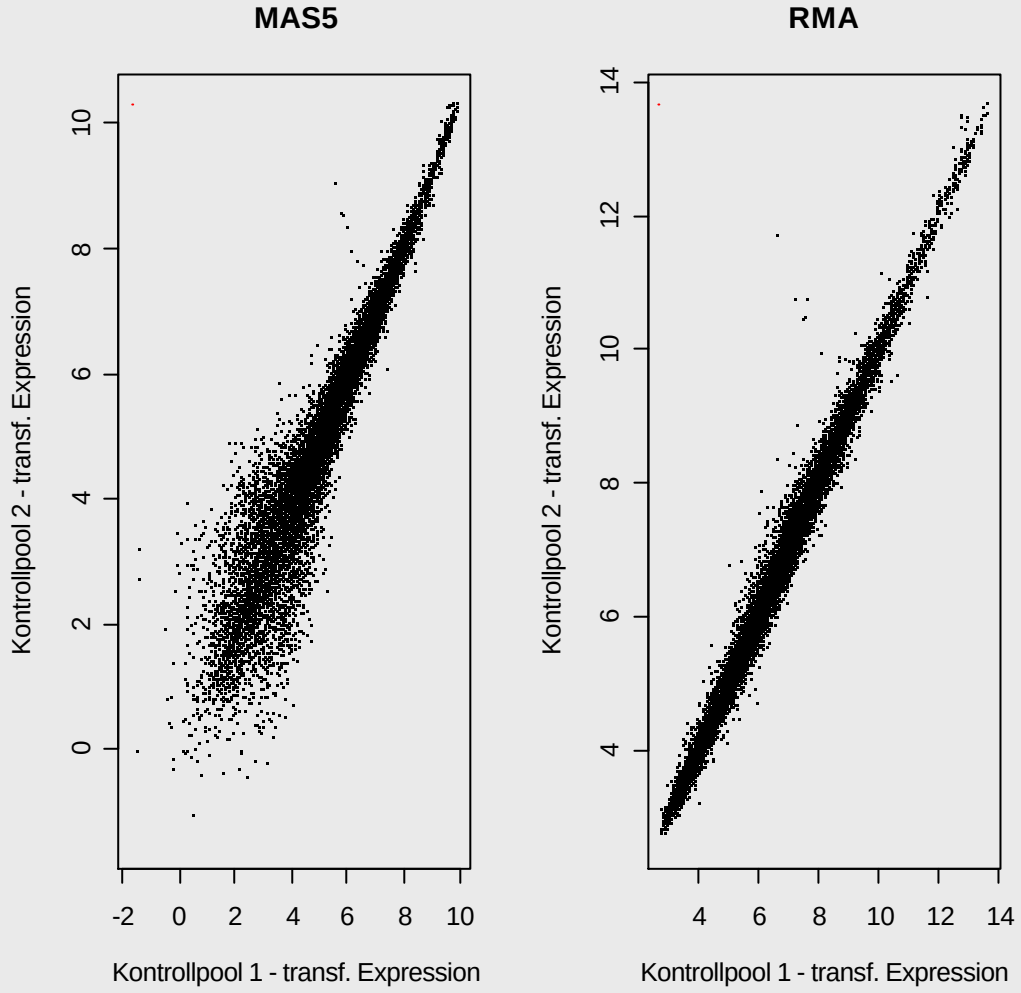
RMA: linear additive model for log(PM), *Median polish* to aggregate over probes

VSN: arsinh - transformation for PM values
Rock - Blythe model for expression



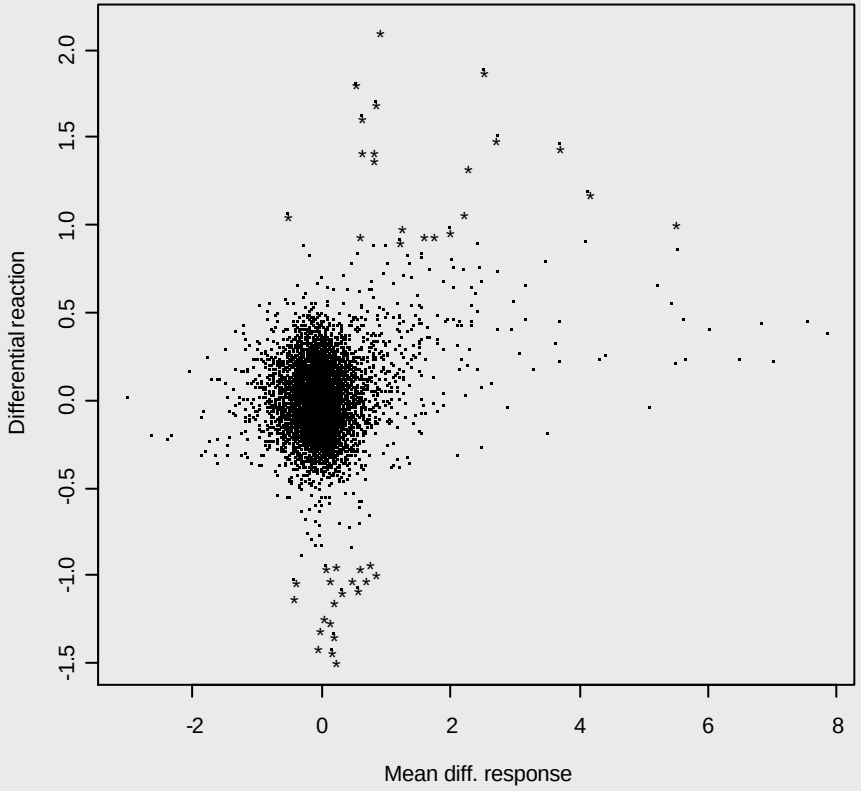


Example LPS: *Expression Summaries*

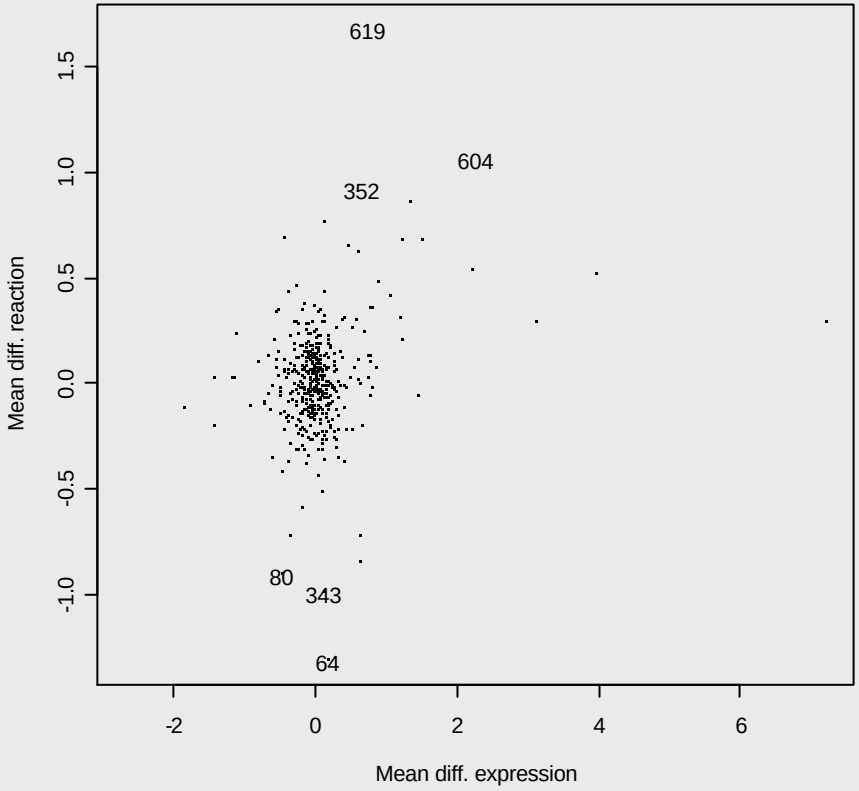


Example LPS: *First* l*ook on the data*

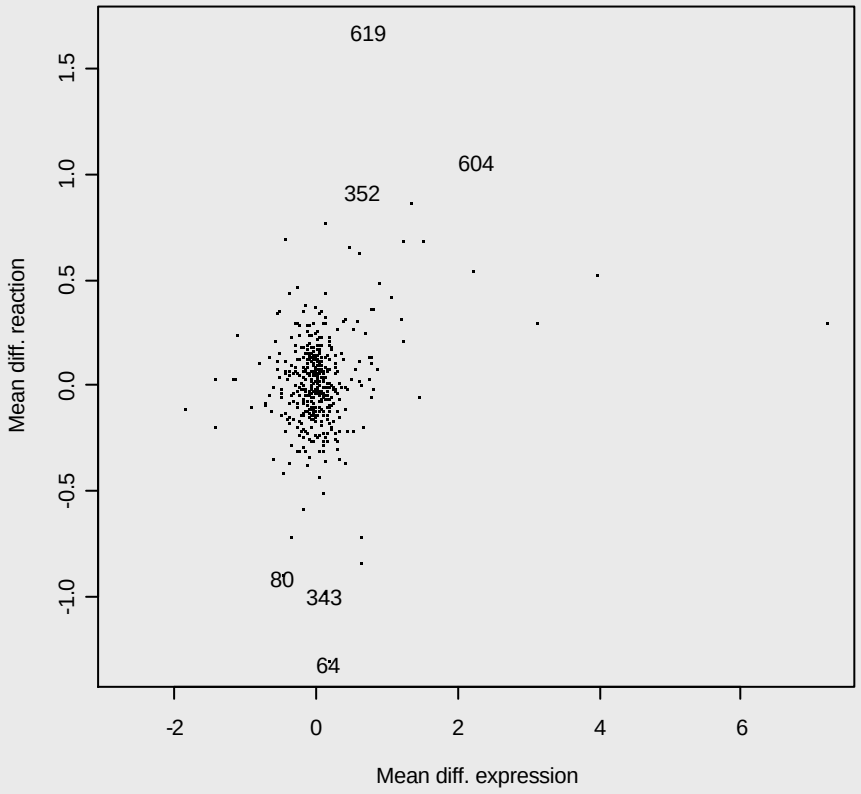
All 22253 genes



1000 meta-genes



Example LPS: *Metagenes*

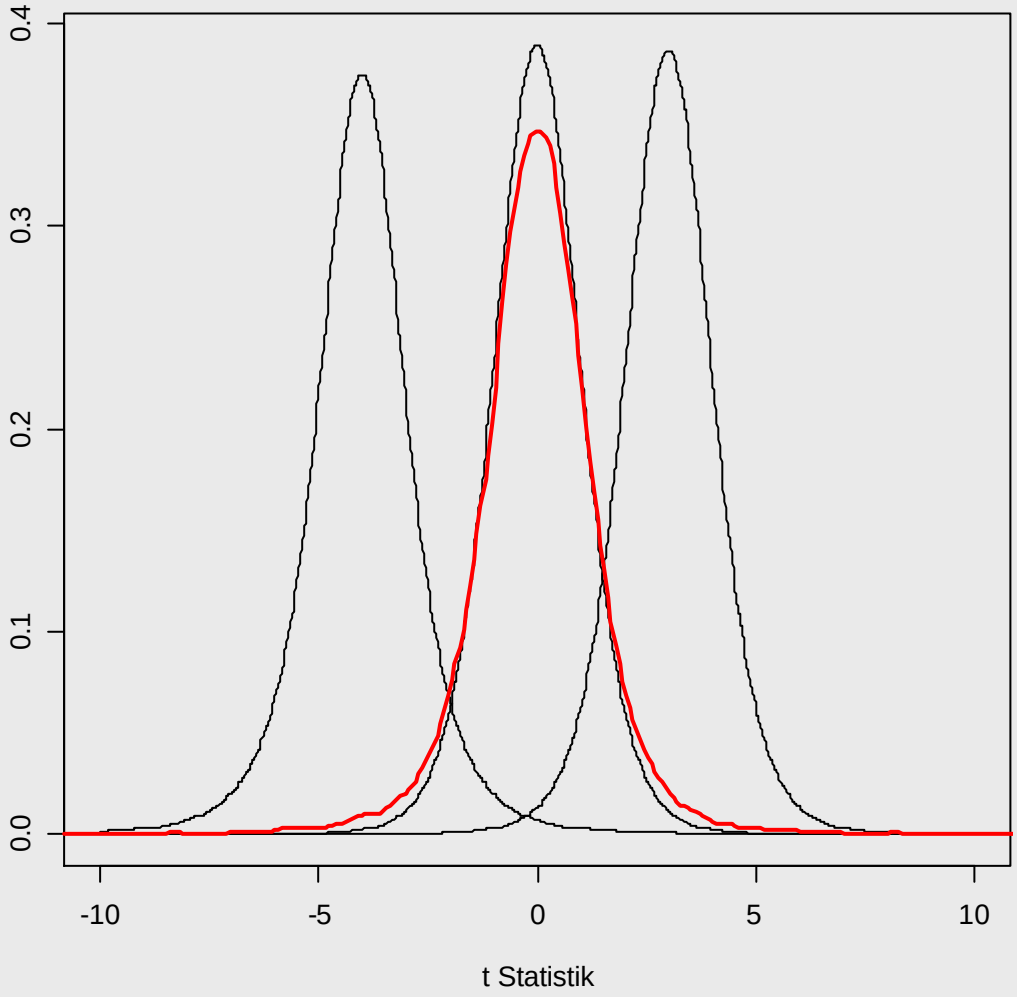



```
> meta.gene.rma.summary
$metagene.64
"201167_x_at" "204270_at"   "213606_s_at“
"219273_at"   "220557_s_at" "34478_at"

$metagene.80
"207425_s_at" "216234_s_at" "216629_at"

$metagene.343
"200935_at"   "201556_s_at" "205179_s_at" "207824_s_at“
"211790_s_at“ "214792_x_at" "217793_at"   "218600_at"

$metagene.352
"201625_s_at" "201627_s_at" "207387_s_at" "210692_s_at“
"211139_s_at“ "222061_at"

$metagene.604
"204747_at"   "205569_at"   "205660_at"   "210163_at"
"210797_s_at" "211122_s_at" "217502_at"

$metagene.619
"AFFX-HUMRGE/M10098_3_at"  "AFFX-HUMRGE/M10098_5_at“
"AFFX-HUMRGE/M10098_M_at"  "AFFX-r2-Hs18SrRNA-3_s_at"
"AFFX-r2-Hs18SrRNA-5_at"   "AFFX-r2-Hs18SrRNA-M_x_at"
```

Example LPS: *Metagenes - multiple testing*

```
> round(meta.gene.mult.test.rma[[1]][1:30,],5)

          rawp Bonferroni    Holm Hochberg SidakSS SidakSD      BH      BY
 [1,] 0.00001    0.00932 0.00932  0.00932 0.00928 0.00928 0.00504 0.03772
 [2,] 0.00001    0.01008 0.01007  0.01007 0.01003 0.01002 0.00504 0.03772
 [3,] 0.00002    0.01573 0.01570  0.01570 0.01560 0.01557 0.00524 0.03924
 [4,] 0.00004    0.04341 0.04328  0.04328 0.04248 0.04235 0.00941 0.07042
 [5,] 0.00005    0.04821 0.04802  0.04802 0.04707 0.04689 0.00941 0.07042
 [6,] 0.00006    0.05645 0.05617  0.05617 0.05489 0.05462 0.00941 0.07042
 [7,] 0.00008    0.07559 0.07513  0.07513 0.07280 0.07238 0.01080 0.08083
 [8,] 0.00012    0.12024 0.11940  0.11940 0.11330 0.11255 0.01236 0.09255
 [9,] 0.00012    0.12398 0.12299  0.12299 0.11661 0.11573 0.01236 0.09255
[10,] 0.00013    0.12696 0.12581  0.12581 0.11924 0.11823 0.01236 0.09255
[11,] 0.00014    0.13600 0.13464  0.13464 0.12716 0.12598 0.01236 0.09255
```

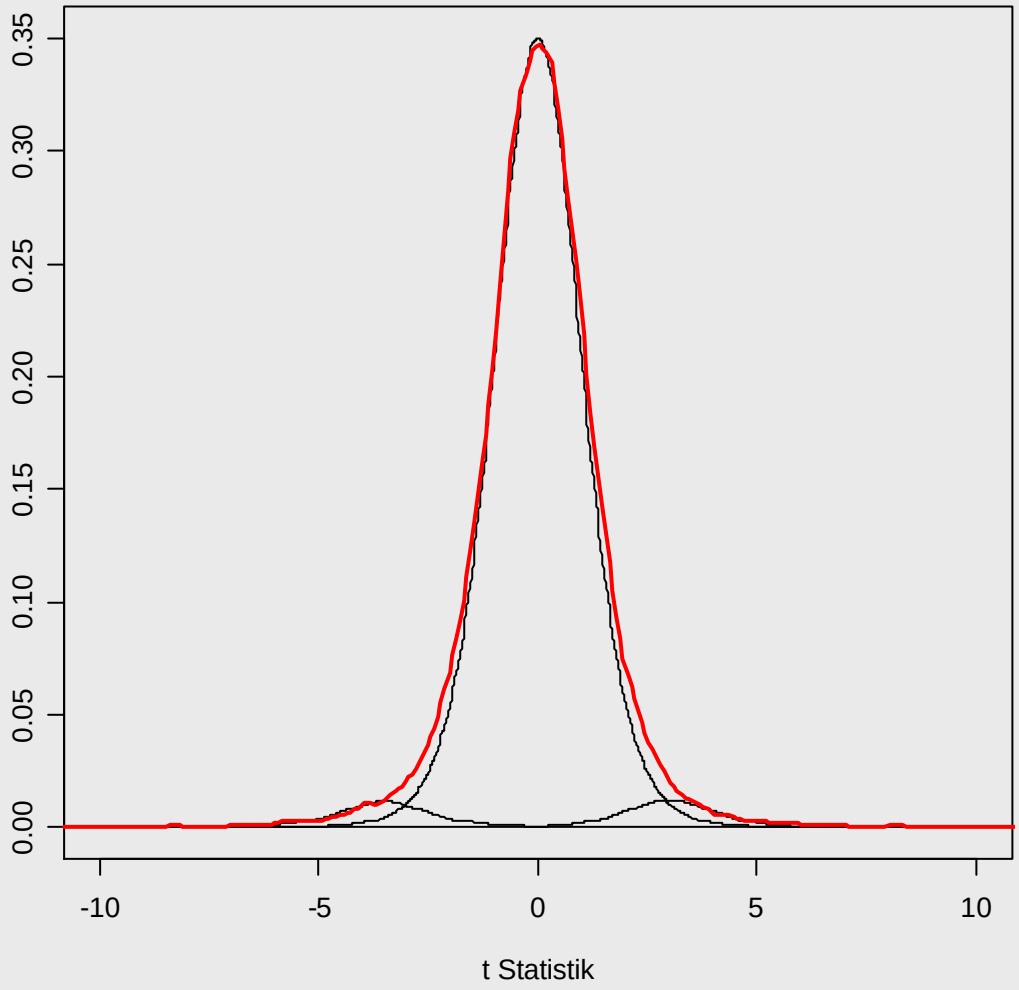
Example LPS: *Predictive analysis*



observed t-statistics

Mixture components
- without DR
- negative DR
- positive DR

Example LPS: *Predictive analysis*

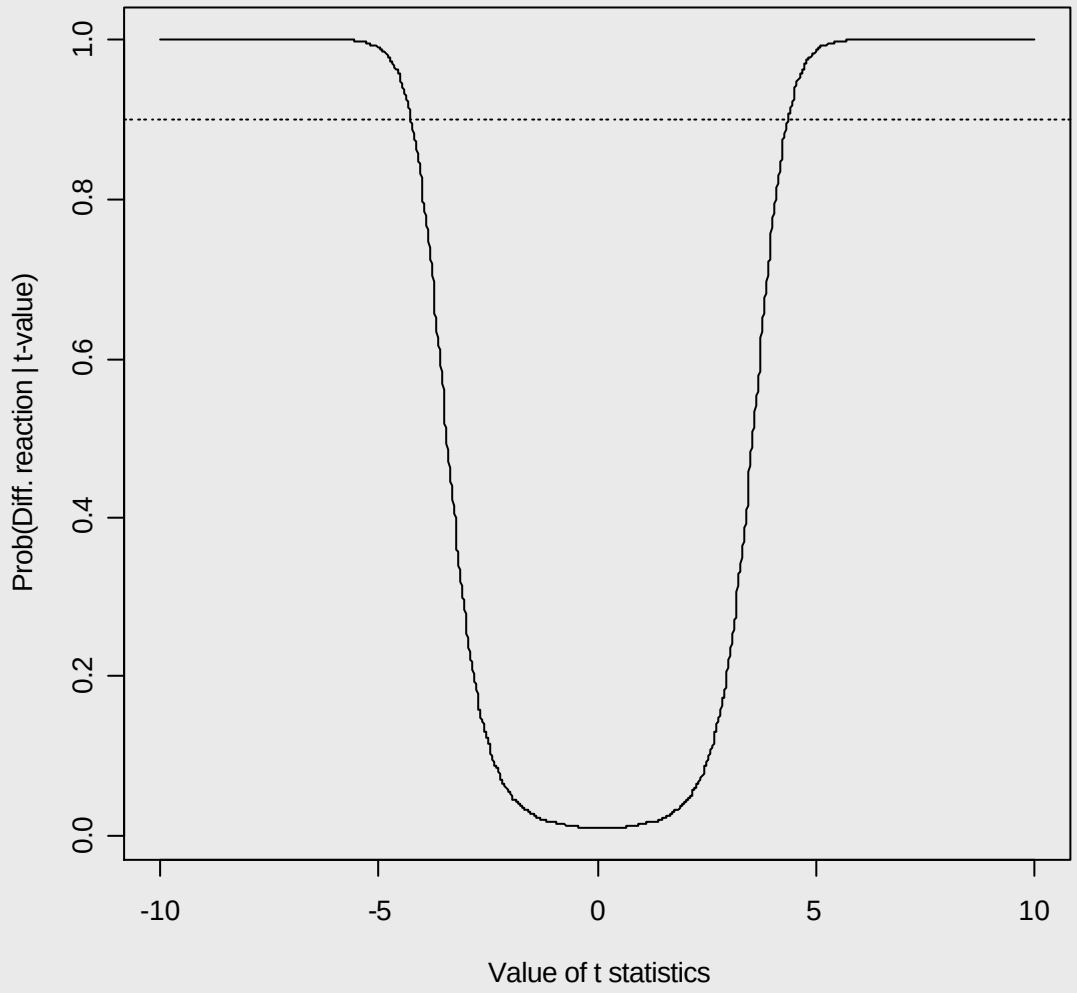


observed t-statistics

Mixture components
- without DR
- negative DR
- positive DR

Inference of mixture components by EM-algorithm or Gibbs-sampler

Example LPS: *Predictive analysis*

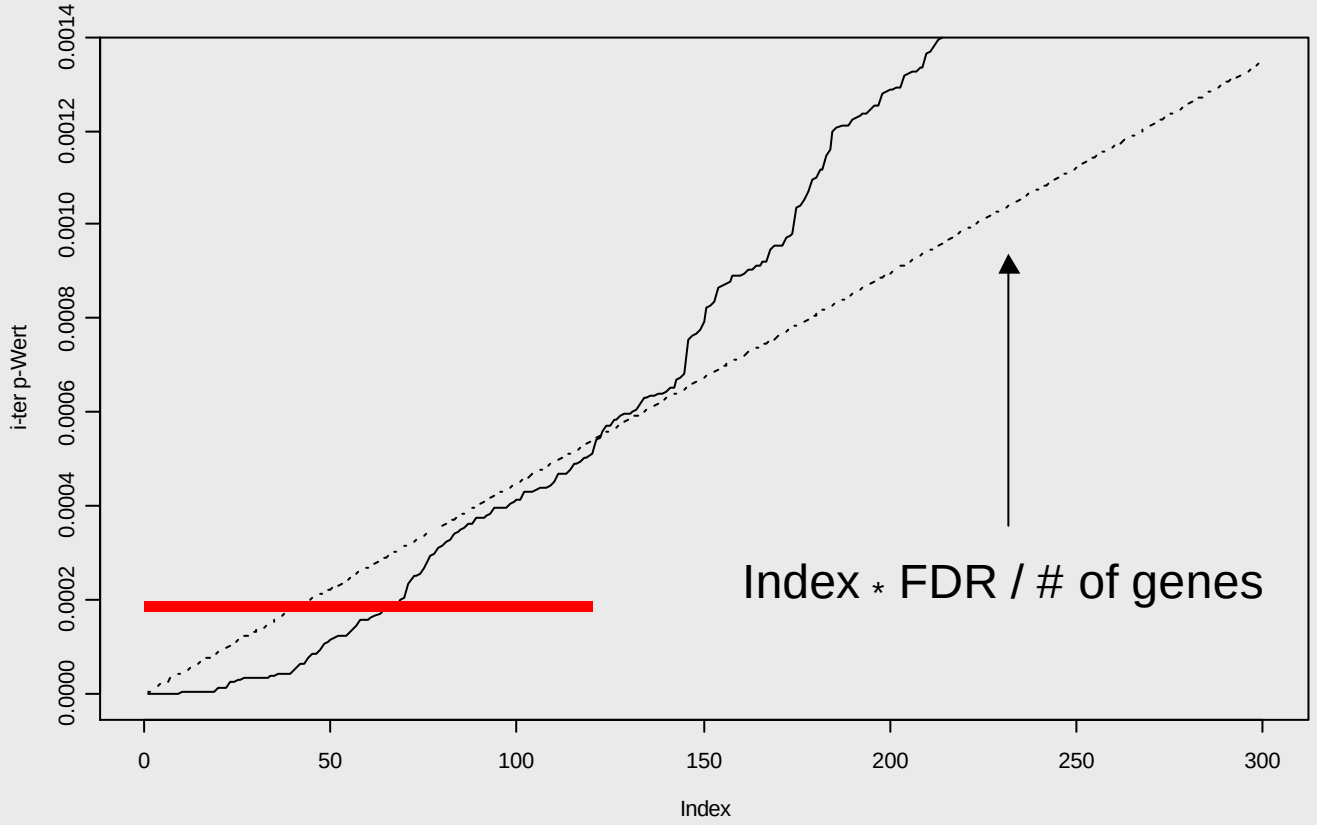


$P_{negative} = 0.0018$

$P_{positive} = 0.0031$

Example LPS: *Adjusted p-values*

Predictive analysis is closely related with frequentist test-theory: Procedure by Benjamini Hochberg (Efron, Storey, Tibshirani, 2001)



FDR: 0.1

BH:
red area contains in mean at most 10% false positive decisions.

Example LPS: *Interpretation*

Genes with differential reaction (RMA)

BH-procedure: # Genes: 122
PA with PPV>0.99: # Genes: 98 + 31

Genes with differential reaction (MAS5)

BH-procedure: # Genes: 42
PA with PPV>0.99: # Genes: 62

Set of genes common to RMA and MAS5 result: 27

Example LPS: *Interpretation*

- **Confounder, Covariates, Population variability** **-**
  Pools are unbalanced between cases and controls
- **Sample size calculation** **+**
  Setting allows with high chance to detect absolute effects above 2.
  cDNA arrays may have resulted in a more efficient analysis.
- **Generality** **+/-**
  Few pools may not give a representative sample of the patient group of interest.
- **Interpretability** **-**
  Inhomogeneities with respect to sex and age make it difficult to interpret DR as related to the disease.
- **Artificial assumptions** **-**
  Assumption of a linear model for confounder effects allows to assume an effect measurement fully attributable to the disease. Use of cDNA arrays would have automatically eliminated the confounder effects.

The most simple measurement model in microarray experiments

Situation: m arrays (Affimetrix) from *control* population
n arrays (Affimetrix) from population with
special condition /treatment

Observation of interest: Mean difference of log-transformed gene expression (ΔlogFC)

$$\Delta\log FC_{obs} = \Delta\log FC_{true} + e$$

$$e \sim N(0, \sigma^2 \cdot [1/n+1/m])$$

In an experiment with 5 arrays per population and the same variance for the expression of a gene of interest, the above formula implies that the variance of the ΔlogFC is only 40% (1/5+1/5 = 2/5 = 0.4) of the variability of a single measurement – ***taming of uncertainty***.

Separate out differences between the conditions / treatments from the uncontrolled variation that is assumed to be present.

Is $\Delta \log FC_{true} \neq 0$? – How to decide?

Special Decision rules: Statistical Tests

- When the probability model for the mechanism generating the observed data is known, hypotheses about the model can be tested.
- This involves the question: Could the presented data reasonable have come from the model if the hypothesis is correct?
- Usually a decision must be made on the basis of the available data, and some degree of uncertainty is tolerated about the correctness of that decision.
- These four components: data, model, hypothesis, and decision are basic to the statistical problem of hypothesis testing.

Quality of decision

	True state of gene	
Decision	Gene ***is*** diff. expr.	Gene ***is not*** diff. expr.
Gene ***is*** diff. expr.	**OK**	**false positive decision** happens with probability α
Gene ***is not*** diff. expr.	**false negative decision** happens with probability β	**OK**

Two sources of error: False positive rate α
False negative rate β

Power of a test: Ability to detect a difference if there is a true difference
Power – true positive rate or Power = 1 - β

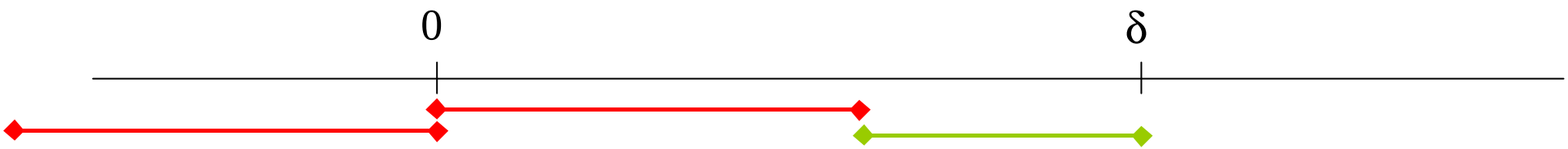
The Statistical test

- Question of interest (*Alterative*): Is the gene G differentially expressed between two cell populations?
- Answer the question via a ***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.*
  The test statistics T implies a *prob. distribution* **P** to quantify its variability when 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 the probability under the null hypothesis of an observation which is more extreme as the observation given by the data: $\mathbf{P}(T \geq t_{obs}) = p$.
- A criteria is needed to asses *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 $|t_{obs}| > t^*$ where t^* is the $1-\alpha$ or $1-\alpha/2$ quantile of the prob. distribution P. t^* is also called the *critical value.*
  The conditions $p < \mathbf{a}$ and $t_{obs} > t^*$ are equivalent. If $p < \mathbf{a}$ or $t_{obs} > t^*$ the null hypothesis will be rejected.
- If $p \geq \mathbf{a}$ or $t_{obs} \geq t^*$ the null hypothesis can not be rejected – this does not mean that it is true
  Absence of evidence for a difference is no evidence for an absence of difference.

Controlling the power – sample size calculations

The test should produce a significant result (level α) with a power of $1-\beta$ if $\Delta\text{logFC}_{true} = \delta$

null hypothesis: $\Delta\text{logFC}_{true} = 0$ alternative: $\Delta\text{logFC}_{true} = \delta$



$z_{1-\alpha/2}\ \sigma_{n,m}$ $z_{1-\beta}\ \sigma_{n,m}$ $\sigma^2_{n,m} = \sigma^2 \cdot (1/n+1/m)$

The above requirement is fulfilled if: $\delta = (z_{1-\alpha/2} + z_{1-\beta}) \cdot \sigma_{n,m}$

or

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

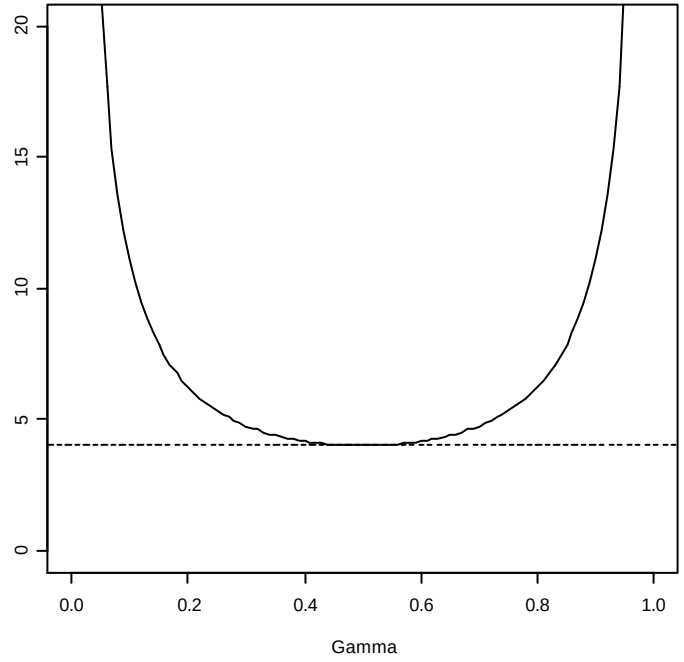
Controlling the power – sample size calculations

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

n = N·γ and m = N·(1-γ) with M – 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 g = ½.



Sample size calculation for a microarray experiment I

	Truth		
Test result	diff. expr. (H_1)	not diff. expr. (H_0)	
diff. expr.	D_1	D_0	D
not diff. expr.	U_1	U_0	U
Number of genes on array	G_1	G_0	G

$\alpha_0 = E[D_0]/G_0$ $\beta_1 = E[U_1]/G_1$ $FDR=E[D_0/D]$

E: expectation / mean number

family type I error probability: $\alpha_F = P[D_0>0]$

family type II error probability: $\beta_F = P[U_1>0]$

Sample size calculation for a microarray experiment II

Independent genes	Dependent Genes
$P[D_0=0] = (1-\alpha_0)^{G_0} = 1-\alpha_F$ $D_0 \sim \text{Binomial}(G_0, \alpha_0)$ $E[D_0] = G_0 \cdot \alpha_0$ Poissonapprox.: $E[D_0] \sim -\ln(1-\alpha_F)$ $P[U_1=0] = (1-\beta_1)^{G_1} = 1-\beta_F$ $E[U_1] = G_1 \cdot (1-\beta_1)$	Bonferroni: $\alpha_0 = \alpha_F / G_0$ No direct link between the probability for D_0 and α_F. $1-\beta_F \geq \max\{0, 1- G_1 \cdot \beta_1\}$ No direct link between the probability for U_1 and β_F.

Sample size calculation for a microarray experiment III

for an array with 33000 independent genes

What are useful α_0 and β_1?

$\alpha_F = 0.8$ $E[D_0] = - \ln(1-0.8) = 1.61 = \lambda$ $P(\text{exactly k false pos.}) = \exp(-\lambda) \cdot \lambda^k / (k!)$

false pos.	0	1	2	3	4	5
Prob.	0.200	0.322	0.259	0.139	0.056	0.018

P(at least six false positives) = 0.0062

32500 unexpressed genes: $\alpha_0 = 1.61/32500 = 0.0000495$

500 expressed genes, set $E[D_1] = 450$ $1-\beta_1 = 450/500 = 0.9$ $\beta_1 = 0.1$ $1-\beta_F = (1-\beta_1)^{G1} < 10^{-23}$

E[FDR] = 0.0035 95% quantile of FDR: 0.0089 (calculated by simulation)

Sample size calculation for a microarray experiment IV

In order to complete the sample size calculation for a microarray experiment, information on σ^2 is needed.

The size of the experiment, N, needed to detect a $\Delta\text{logFC}_{\text{true}}$ of δ on a significance level α and with power $1-\beta$ is:

$$N = 4 \cdot \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \cdot \sigma^2}{\delta^2}$$

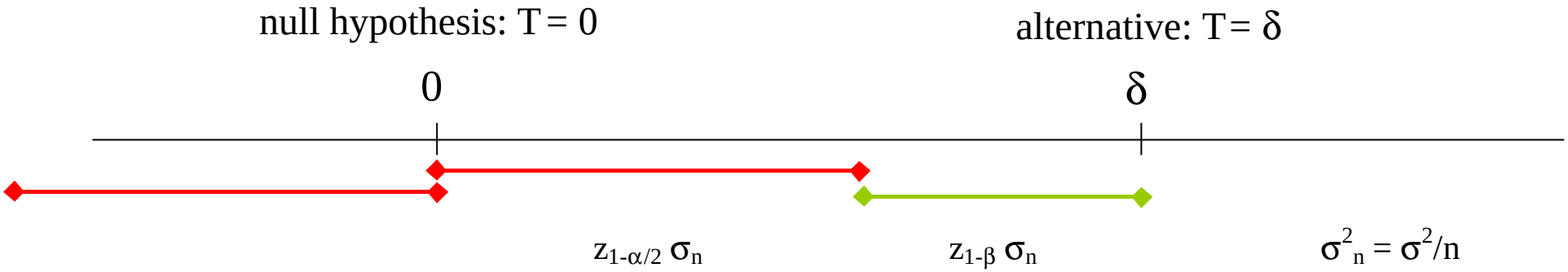
In a similar set of experiments σ^2 for a set of 20 VSN transformed arrays was between 1.55 and 1.85. One may choose the value $\sigma^2 = 2$.

δ	log(1.5)	log(2)	log(3)	log(5)	log(10)
N ($\sigma^2 = 2$)	1388	476	190	88	44
N ($\sigma^2 = 1$)	694	238	96	44	22

Sample size with $\alpha = 0.0000495$, $\beta = 0.1$

Sample size formula for a one group test

The test should produce a significant result (level α) with a power of $1-\beta$ if $T = \delta$



The above requirement is fulfilled if: $\delta = (z_{1-\alpha/2} + z_{1-\beta}) \cdot \sigma_n$

or

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

Measurement model for cDNA arrays

Gene expression under condition A – intensity of red colour,
Gene expression under condition B – intensity of green colour

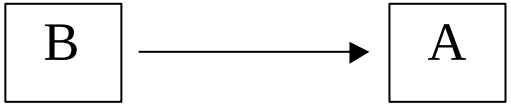
$$\text{Measurement: } m_{A/B} = \text{Log}_2\left(\frac{I_{red,A}}{I_{green,B}}\right) = \gamma_{A/B} + \delta + e$$

$\gamma_{A/B}$ – log-transformed true fold change of gene of condition A with respect to condition B
δ - dye effect, e – measurement error with E[e] = 0 and Var(e) = σ^2

Measurement $m_{A/B}$ is used to estimate unknown $\gamma_{A/B}$

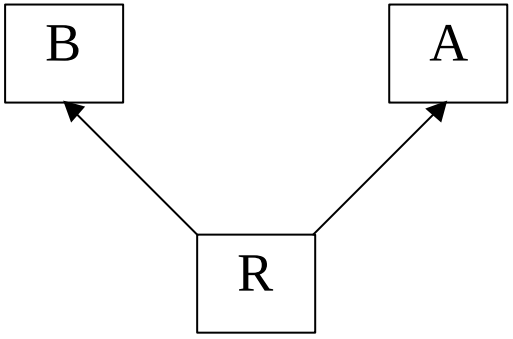
- *Vertices* mRNA samples
- *Edges* hybridization
- *Direction* Dye assignment

  Green ⟶ Red

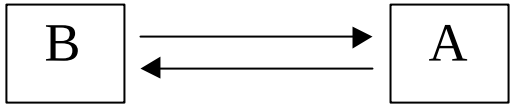


Estimation of *log fold change* $g_{A/B}$

Reference Design



Dye swap design



Estimate of $\gamma_{A/B}$

$$g^{R}_{A/B} = m_{A/R} - m_{B/R}$$

$$g^{DS}_{A/B} = (m_{A/B} - m_{B/A})/2$$

Variability of estimate

$$Var(g^{R}_{A/B}) = 2 \cdot \sigma^2$$

$$Var(g^{DS}_{A/B}) = 0.5 \cdot \sigma^2$$

Sample Size increases proportional to the variance of the measurement!

2x2 factorial experiments I

treatment / condition	Wild type	Mutation
before treatment	β	$\beta+\mu$
after treatment	$\beta+\tau$	$\beta+\tau+\mu+\psi$

β - baseline effect; τ - effect of treatment; μ - effect of mutation
ψ - differential effect on treatment between WT and MUT

treatment effect on gene expr. in WT cells: $\Delta^{WT} = (\beta + \tau) - \beta = \tau$
treatment effect on gene expr. in MUT cells: $\Delta^{MUT} = (\beta + \tau + \mu + \psi) - (\beta + \mu) = \tau + \psi$

differential treatment effect: $\Delta^{MUT} \neq \Delta^{WT}$ or $\psi \neq 0$

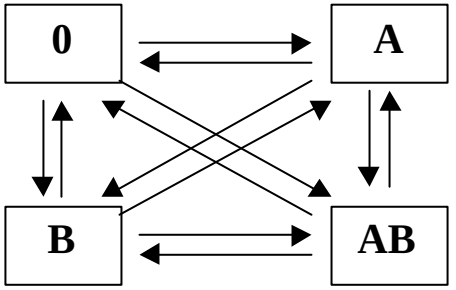
How many cDNA arrays are needed to show $\psi \neq 0$ with significance α and power $1-\beta$ if $|\psi| > \ln(5)$?

2x2 factorial experiments II

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

There are four possible condition / treatment combinations:

AB: treatment applied to MUT cells
A: treatment applied to WT cells
B: no treatment applied to MUT cells
0: no treatment applied to WT cells



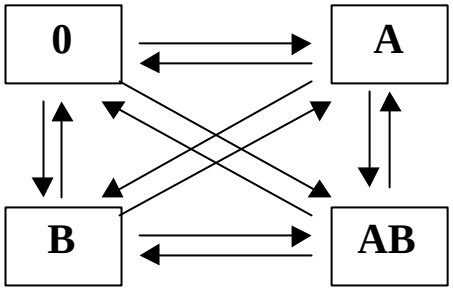
Design with 12 slides

2x2 factorial experiments III

Array	Measurement
$m_{A/0}$	$\gamma_{A/0} + \delta + e = \tau + \delta + e$
$m_{0/A}$	$-\gamma_{A/0} + \delta + e = -\tau + \delta + e$
$m_{B/0}$	$\gamma_{B/0} + \delta + e = \mu + \delta + e$
$m_{0/B}$	$-\gamma_{B/0} + \delta + e = -\mu + \delta + e$
$m_{AB/0}$	$\gamma_{AB/0} + \delta + e = \mu + \tau + \psi + \delta + e$
$m_{0/AB}$	$-\gamma_{AB/0} + \delta + e = -(\mu + \tau + \psi) + \delta + e$
$m_{AB/A}$	$\gamma_{AB/A} + \delta + e = \mu + \psi + \delta + e$
$m_{A/AB}$	$-\gamma_{AB/A} + \delta + e = -(\mu + \psi) + \delta + e$
$m_{AB/B}$	$\gamma_{AB/B} + \delta + e = \mu + \psi + \delta + e$
$m_{B/AB}$	$-\gamma_{AB/B} + \delta + e = -(\mu + \psi) + \delta + e$
$m_{A/B}$	$\gamma_{A/B} + \delta + e = \tau - \mu + \delta + e$
$m_{B/A}$	$-\gamma_{A/B} + \delta + e = -(\tau - \mu) + \delta + e$

- Each measurement has variance σ^2
- Parameter β is confounded with the dye effect

Regression analysis



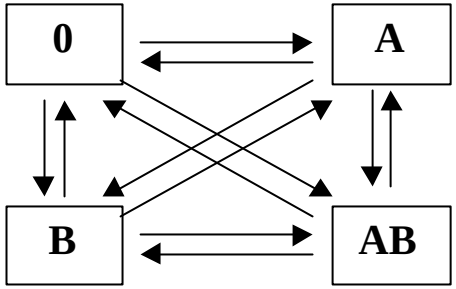
$$E\begin{pmatrix} M_{A/0} \\ M_{0/A} \\ M_{B/0} \\ M_{0/B} \\ M_{AB/0} \\ M_{0/AB} \\ M_{AB/A} \\ M_{A/AB} \\ M_{AB/B} \\ M_{B/AB} \\ M_{B/A} \\ M_{A/B} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & -1 & -1 \\ 1 & 1 & 0 & 1 \\ 1 & -1 & 0 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & -1 & 0 \end{pmatrix} \bullet \begin{pmatrix} \delta \\ \tau \\ \mu \\ \psi \end{pmatrix}$$

- For parameter $\theta = (\delta, \tau, \mu, \psi)$ 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:

- 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 IV



➢ total.2.by.2.design.mat

	delta	alpha	beta	psi
A/0	1	1	0	0
0/A	1	-1	0	0
B/0	1	0	1	0
0/B	1	0	-1	0
AB/0	1	1	1	1
0/AB	1	-1	-1	-1
AB/A	1	0	1	1
A/AB	1	0	-1	-1
AB/B	1	1	0	1
B/AB	1	-1	0	-1
B/A	1	-1	1	0
A/B	1	1	-1	0

```
> precision.2.by.2.rfc(x.mat)
$inv.mat

          tau     mu   psi
  tau   0.250  0.125 -0.25
   mu   0.125  0.250 -0.25
  psi  -0.250 -0.250  0.50

$effects
  tau    mu   psi   tau-mu
 0.25  0.25  0.50  0.25
```

$$\mathrm{Var}(A\text{-}B) = \mathrm{Var}(A) + \mathrm{Var}(B) - 2\cdot\mathrm{Cov}(A,B)$$

Sample size for differential treatment effect (DTE) in a 2 x 2 factorial designs I

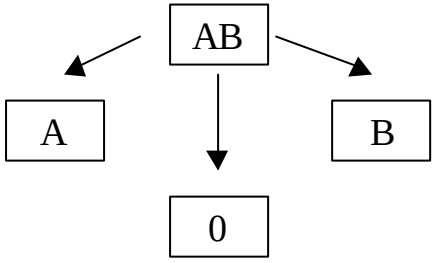
- Array has 20.000 genes: 19500 without DTE, 500 with DTE
- $\alpha_F = 0.9$, using Bonferroni adjustment: $\alpha = 0.9/20.000 = 0.0000462$
- Mean number of correct positives is set to 450: $1-\beta = 0.9$
- $\sigma^2 = 0.7$, taken from similar experiments
- A total dye swap design (12 arrays) estimates ψ with precision $\sigma^2/2 = 0.35$

$$N = [4.074 + 1.282]^2 \cdot 0.35 / \ln(5)^2 = 3.876$$

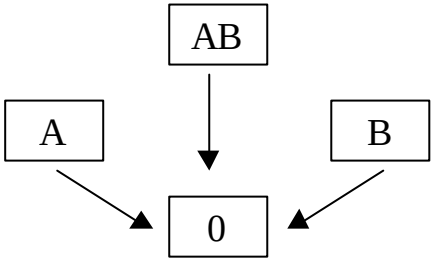
- The experiment would need in total 4 x 12 = 48 arrays
- Is there a chance to get the same result cheaper?

2 x 2 factorial designs V

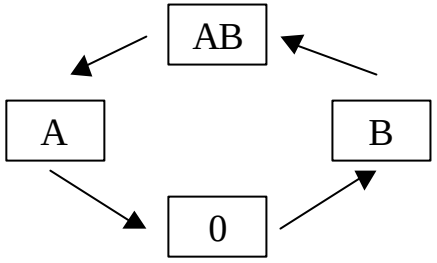
Design I
Common ref.



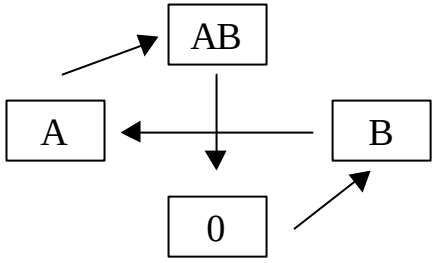
Design II
Common ref.



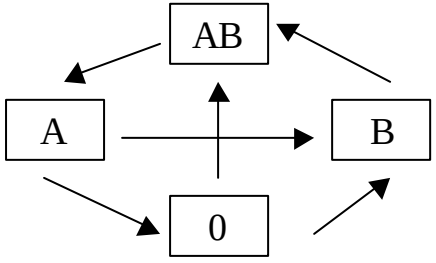
Design III
Connected



Design IV
Connected



Design V
All-pairs



Scaled variances of estimated effects

	D.I	D.II	D.III	D.IV	D.V	D.tot
tau	2	1	0.75	1.00	0.5	0.25
mu	2	1	0.75	0.75	0.5	0.25
psi	**3**	**3**	**1.00**	**2.00**	**1.0**	**0.50**
# chips	**3**	**3**	**4**	**4**	**6**	**12**

Sample size for differential treatment effect (DTE) in a 2 x 2 factorial designs II

Is there a chance to get the same result cheaper?

- Using total dye swap design, the experiment would need in total 4 x 12 = 48 arrays
- Using Design III, the effect of interest is estimated with doubled variance (4 $\rightarrow$ 8) but by using a design which need only 4 arrays (12 $\rightarrow$ 4).
- This reduces the number of arrays needed from 48 to 32.

Experimental Design - Conclusions

- Designs for *time course* experiments
- In addition to experimental constraints, design decisions should be guided by knowledge of which effects are of greater interest to the investigator.
- The unrealistic planning based on independent genes may be put into a more realistic framework by using simulation studies – speak to your bio – statistician/informatician
- How to collect and present *experience* from performed microarray experiments on which to base assumptions for planing (σ^2)?
- Further reading:

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

Lee MLT, Whitmore GA (2002), Power and sample size for DNA microarray studies, Stat. in Med., 21:3543-3570