Experimental Design

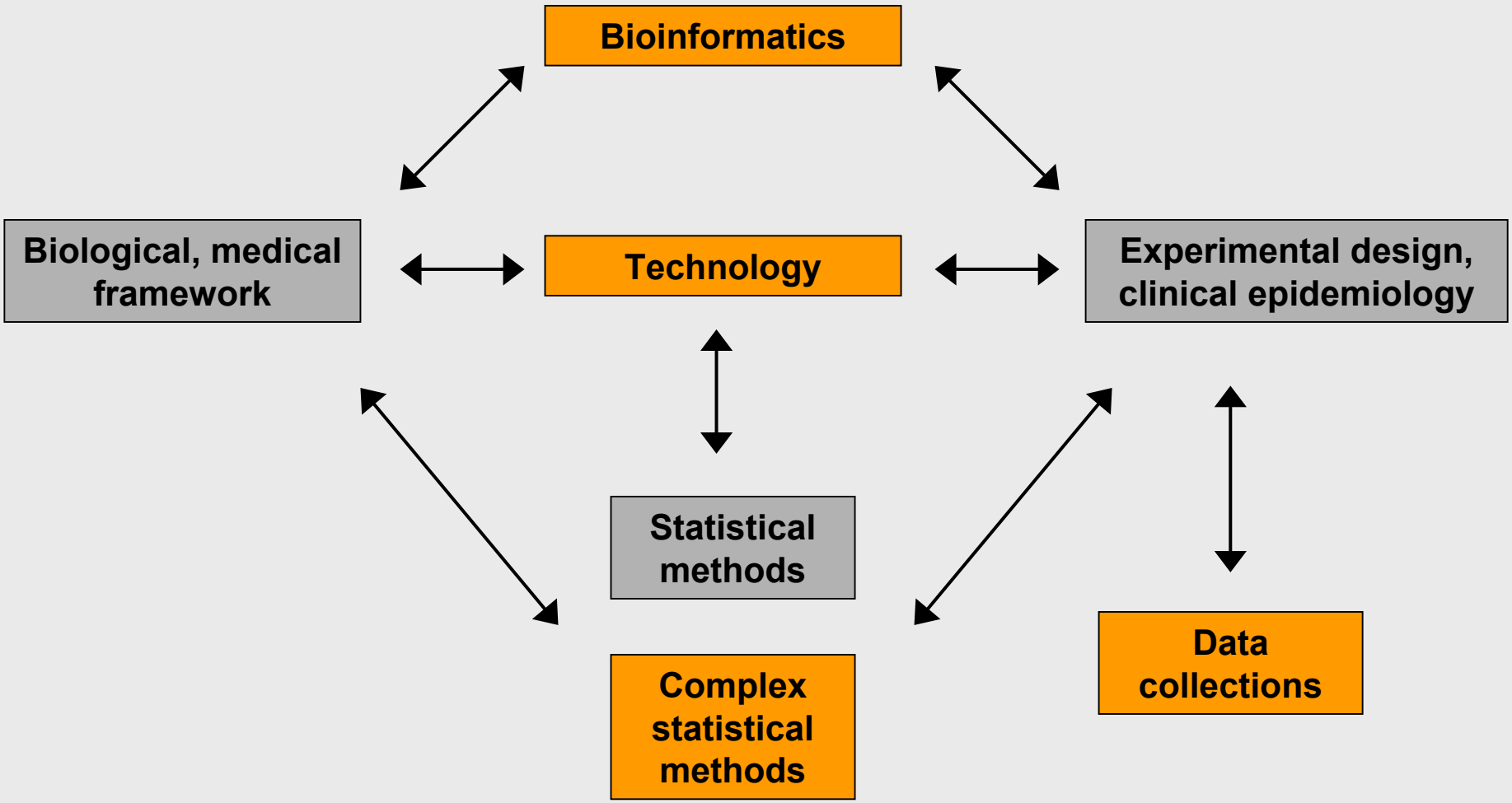
Interplay between Clinical Epidemiology, Bioinformatics, and Biostatistics

Ulrich Mansmann
Department of Medical Biometrics and Informatics
University of Heidelberg

Overview

- Introduction: Basics of experimental design
  Sample size calculation
  Pepe et al. (2003), Biometrics,
- The need of clinical epidemiology
  Chang JC et al. (2003), The Lancet, **362**:362-369
- Good statistical practice in genomic profiling studies
  Fredriksen, C.M. et al. (2003) J Cancer Res Clin Oncol 129: 263–271
- Discussion

Micro-array experiments



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 = Usable knowledge

Decisions on the experimental design influence the measurement model.

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)?

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)

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)

$$\Delta\text{logFC}_{obs} = \Delta\text{logFC}_{true} + e$$
$$e \sim N(0, \sigma^2 \cdot [1/n+1/m])$$

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***.

For two independent random variables X, Y and two real numbers a, b it holds:

$$Var(aX+bY)=a^2Var(X)+b^2Var(y)$$

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.

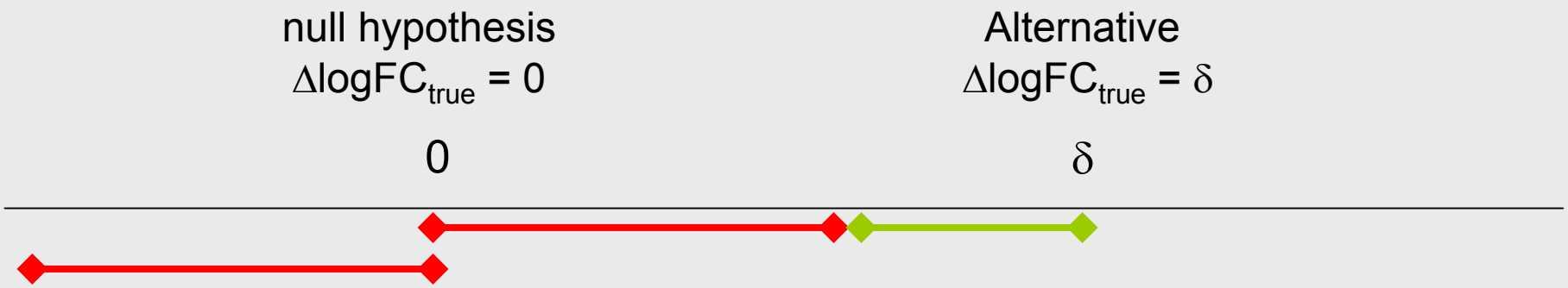
This involves the question: Could the presented data reasonable have come from the model if the hypothesis is correct?

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.

Controlling the power – sample size calculations

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



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

or

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

$$\sigma^2_{n,m} = \sigma^2 \cdot \left(\frac{1}{n} + \frac{1}{m}\right)$$

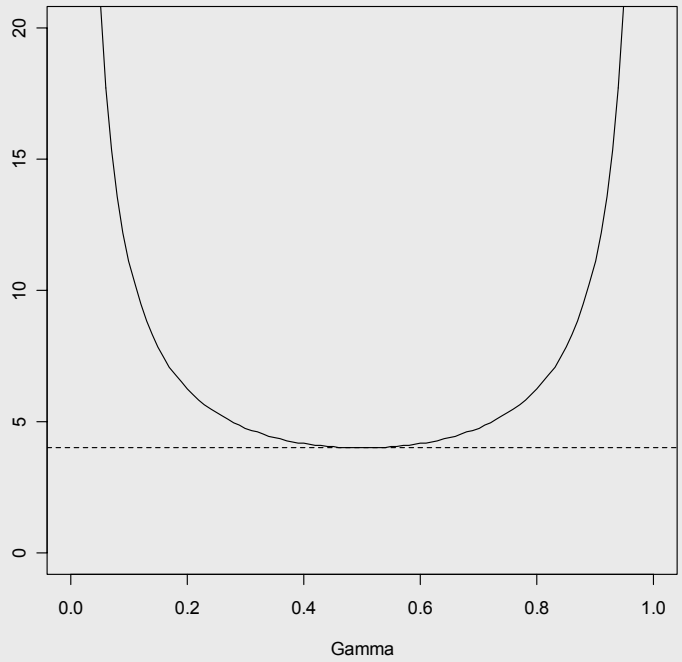
σ^2 - variation of gene expression

Controlling the power – sample size calculations

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

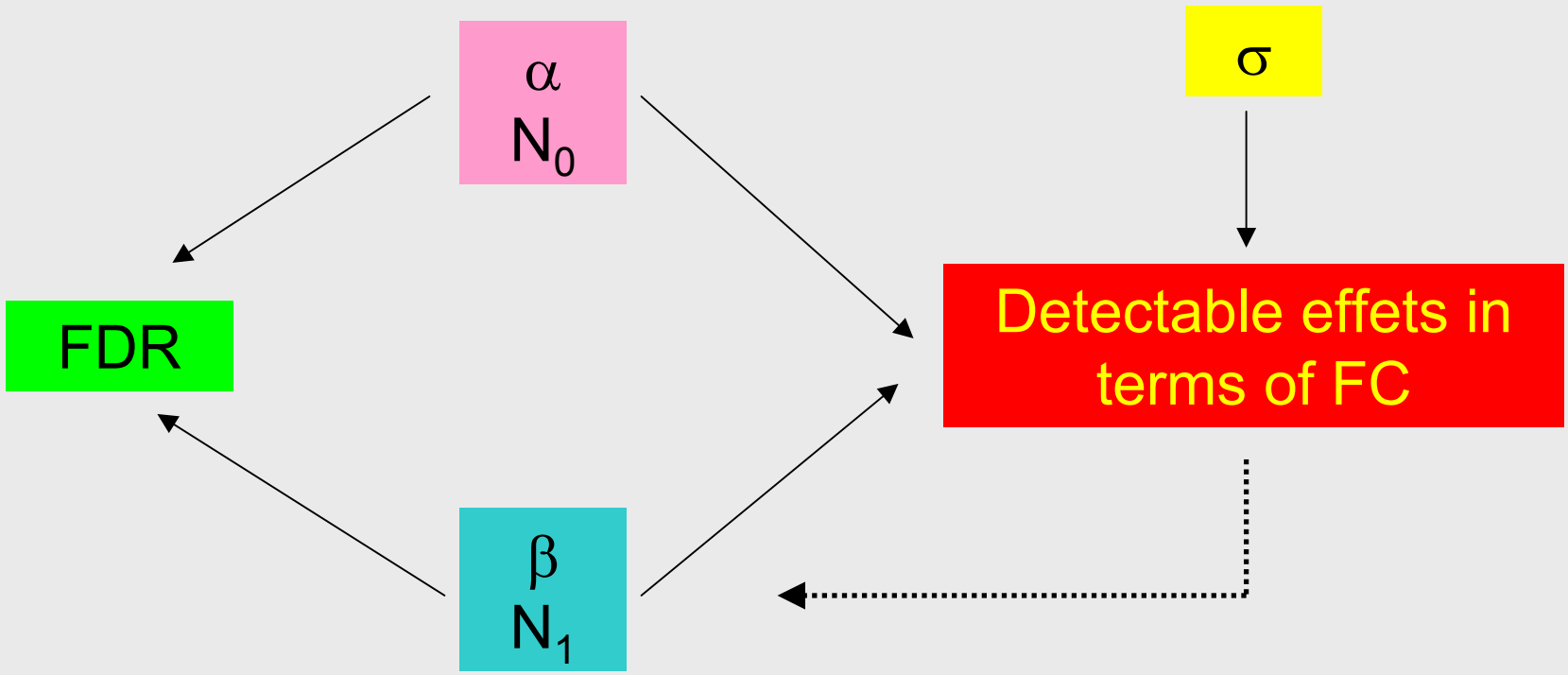
$n = N \cdot \gamma$ and $m = N \cdot (1-\gamma)$ with M – total size of experiment and $\gamma \in]0,1[$

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



The size of the experiment is minimal if γ = ½.

Differential gene expression: planning a simple experiment



FDR

FDR = E[FT/(FP+TP)] = E[FP/AP]	AP= FP+TP: all positives, sum of true (TP) and false positives (FP)
marginal FDR (mFDR):	$\alpha \cdot n_0 / (\alpha \cdot n_0 + [1-\beta] \cdot n_1)$
positive FDR (pFDR):	E[FP/AP | AP>0]
conditional FDR (cFDR)	E[FP/AP | AP=k]
empirical FDR (eFDR)	only accessible in a simulation
Independence model:	FP ~ Pois($\alpha \cdot n_0$), TP ~ Binom(n_1, 1-β)

Scenario - Sample Size

α = 0.0001, n_0 = 22100, β = 0.1, n_1 = 183

mFDR=0.013 FDR ~ 0.013
eFDR: Median =0.012, Q.95 = 0.030
cFDR.90 = 0.03

Sigma: 0.2 (Based on data from similar arrays and normalization)
FC = 2
arrays per group = 5 per group

Gene lists and sample size calculation

Why gene lists?

- Find genes that are differentially expressed between normal and diseased samples.
- Gene is related to protein product (or antibody of it) which may be detectable in blood, urine, ...
- Over-expressed genes are of higher interest: To detect the presence of a new aberrant protein is easier than detecting a decreased level of a normal protein.
- Many over-expressed genes may not be good markers for screening. Essays may be to difficult, genes may also be related to processes which also occur in other settings: inflammation, growth, ...
- Need of finding a sizeable number of overexpressed genes to arrive at a subset which might have the potential for screening.

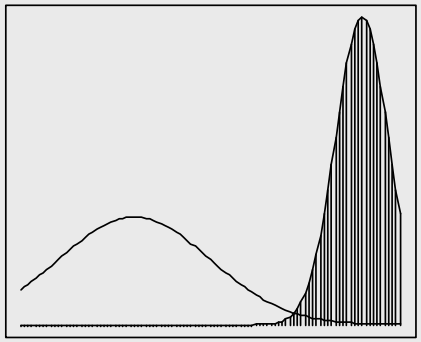
Properties of a *good* gene list

- Contain the relevant subset of genes out of a large pool of candidate genes.
- Contain a low number of false positive findings.
- The ranking of the candidates in the list should be stable and informative.
- Information with respect to a well defined purpose:
  i.e. in case of screening markers have to be highly specific.
- P-values versus selection probabilities: *Gene g ranked in the top k*
- Purpose is not to test *equal* versus *unequal* distribution.
  The issue is to rank genes according the extent of differential expression.

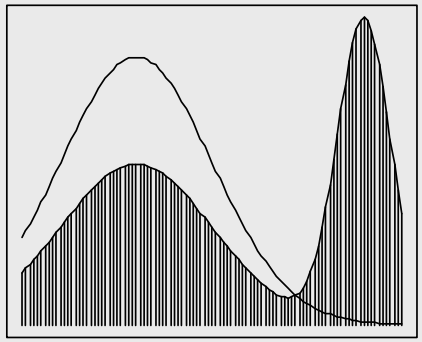
P-value versus discrimination measures

- Perform gene wise simple test or statistical procedures which allow for stratification and covariate adjustment
- Pepe et al. (2003) Biometrics propose for overexpression:
  AUC
  pAUC(t_0) with $t_0 = P[Y_c > u]$ - rate of false positives
  ROC(t_0)
- Example: Huang et al (2003) The Lancet 361:1590-1596.
  The data contains microarrays of 52 women with breast cancer of whom 34 did not experience a recurrence of the tumour during a 3 years time period.
  Human U95Av2 gene chip with 12625 probe sets

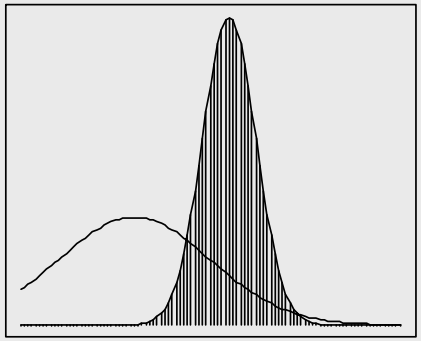
Scenario I

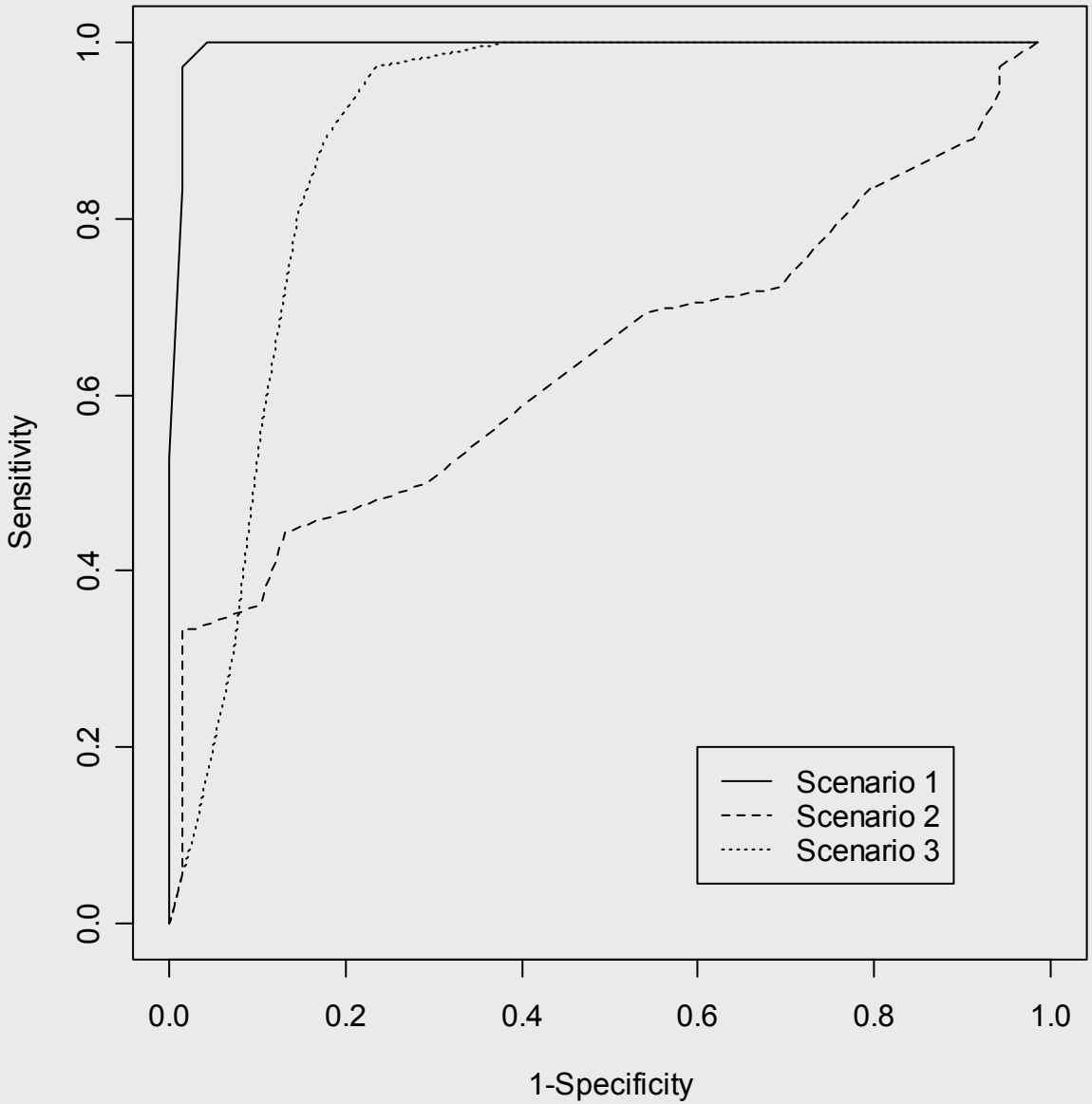


Scenario II



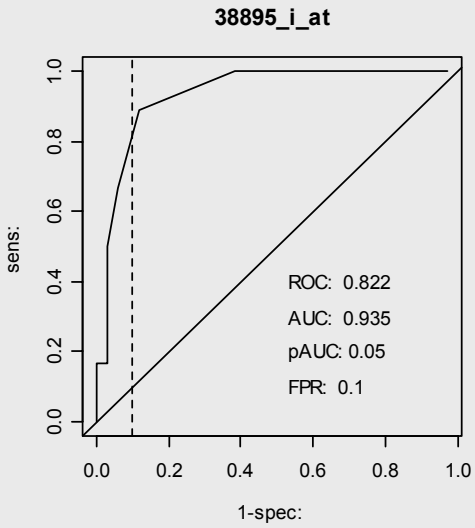
Scenario II

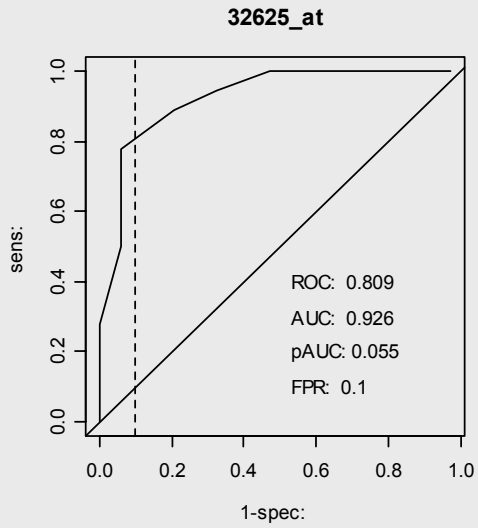


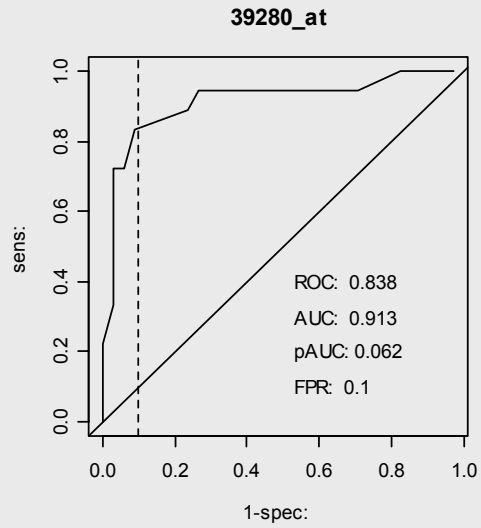

Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Scenario 1
Scenario 2
Scenario 3

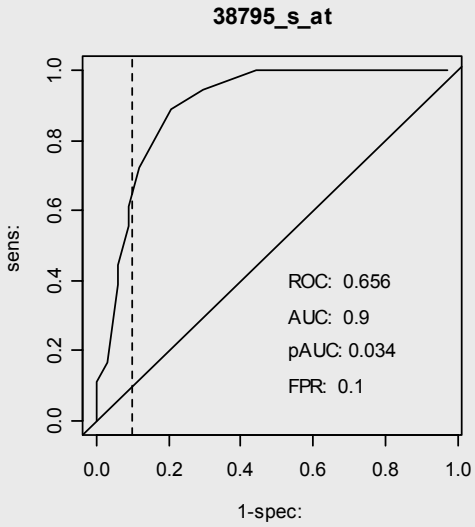
P-value and differential expression

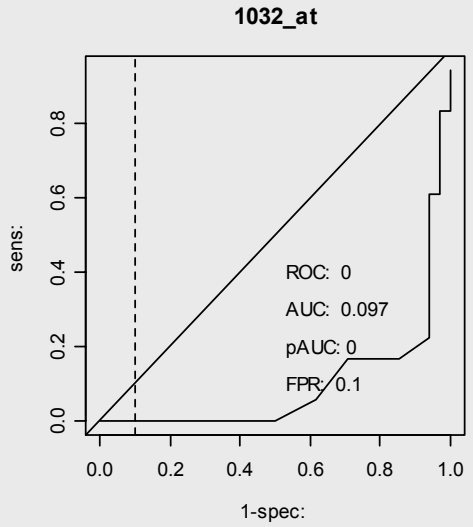
	rawp	Bonferroni	Holm	Hochberg	SidakSS	SidakSD	BH	BY
38895_i_at	0	0.0026	0.0026	0.0026	0.0026	0.0026	0.0026	0.0261
32625_at	0	0.0053	0.0053	0.0053	0.0053	0.0053	0.0027	0.0267
39280_at	0	0.0130	0.0130	0.0130	0.0129	0.0129	0.0031	0.0306
38795_s_at	0	0.0211	0.0211	0.0211	0.0209	0.0209	0.0031	0.0306
1032_at	0	0.0232	0.0232	0.0232	0.0230	0.0230	0.0031	0.0306
33706_at	0	0.0256	0.0256	0.0256	0.0252	0.0252	0.0031	0.0306
35222_at	0	0.0256	0.0256	0.0256	0.0252	0.0252	0.0031	0.0306
965_at	0	0.0256	0.0256	0.0256	0.0252	0.0252	0.0031	0.0306
35225_at	0	0.0340	0.0339	0.0339	0.0334	0.0334	0.0031	0.0306
39547_at	0	0.0373	0.0373	0.0373	0.0366	0.0366	0.0031	0.0306
31685_at	0	0.0410	0.0409	0.0409	0.0401	0.0401	0.0031	0.0306
33673_r_at	0	0.0410	0.0409	0.0409	0.0401	0.0401	0.0031	0.0306
34151_at	0	0.0450	0.0449	0.0449	0.0440	0.0439	0.0031	0.0306
518_at	0	0.0450	0.0449	0.0449	0.0440	0.0439	0.0031	0.0306
1608_at	0	0.0541	0.0541	0.0540	0.0527	0.0526	0.0031	0.0306
36195_at	0	0.0541	0.0541	0.0540	0.0527	0.0526	0.0031	0.0306
33650_at	0	0.0593	0.0593	0.0593	0.0576	0.0575	0.0031	0.0306
38189_s_at	0	0.0593	0.0593	0.0593	0.0576	0.0575	0.0031	0.0306
33290_at	0	0.0713	0.0712	0.0711	0.0688	0.0687	0.0031	0.0306
33558_at	0	0.0713	0.0712	0.0711	0.0688	0.0687	0.0031	0.0306

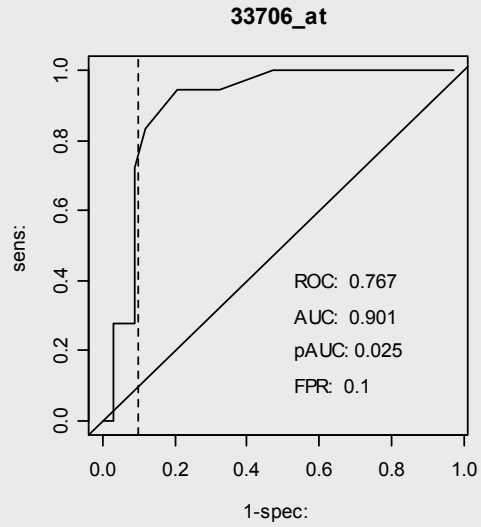

38895_i_at
sens:
1-spec:
ROC: 0.822
AUC: 0.935
pAUC: 0.05
FPR: 0.1


32625_at
sens:
1-spec:
ROC: 0.809
AUC: 0.926
pAUC: 0.055
FPR: 0.1


39280_at
sens:
1-spec:
ROC: 0.838
AUC: 0.913
pAUC: 0.062
FPR: 0.1


38795_s_at
sens:
1-spec:
ROC: 0.656
AUC: 0.9
pAUC: 0.034
FPR: 0.1


1032_at
sens:
1-spec:
ROC: 0
AUC: 0.097
pAUC: 0
FPR: 0.1


33706_at
sens:
1-spec:
ROC: 0.767
AUC: 0.901
pAUC: 0.025
FPR: 0.1

Probability for gene selection $P_g(k)$

$$P_g(k) = \text{Prob}[\text{Gene g ranked in the top k}]$$

Calculate $P_g(k)$ by bootstrap

- Sample n_C control arrays with replacement
  Sample n_D disease arrays with replacement
- Break randomly ties by adding a minuscule random noise (jitter)
  This makes bootstrap distribution of rank statistics more reflective of the actual distribution across different realisations of the experiment.
- Calculate the statistic of interest AUC_g, $ROC_g(t_0)$, $pAUC_g(t_0)$ and determine its rank for gene g. Notice if g is ranked in the top k genes.

Results for the Huang data on recurrence

Rank	Gene	AUC	Pg20	Gene	ROC	Pg20	Gene	pAUC	Pg20
1	38895_i_at	0.9355	1.00	39280_at	0.8378	1.00	39280_at	0.0621	1.00
2	32625_at	0.9257	1.00	35653_at	0.8333	1.00	39737_at	0.0611	0.99
3	39280_at	0.9134	1.00	35763_at	0.8333	0.99	32004_s_at	0.0601	0.96
4	35222_at	0.9011	0.99	38895_i_at	0.8222	0.98	34658_at	0.0587	0.93
5	33706_at	0.9011	0.99	32625_at	0.8089	0.97	41267_at	0.0578	0.91
6	38795_s_at	0.9003	0.98	33290_at	0.7889	0.96	32166_at	0.0565	0.87
7	33650_at	0.8954	0.97	32166_at	0.7778	0.92	33325_at	0.0556	0.83
8	36195_at	0.8938	0.95	33706_at	0.7667	0.91	36670_at	0.0554	0.79
9	965_at	0.8938	0.93	36670_at	0.7667	0.87	36195_at	0.0552	0.77
10	35225_at	0.8922	0.86	33237_at	0.7444	0.83	40823_s_at	0.0551	0.73
11	36471_f_at	0.8897	0.84	35116_at	0.7444	0.79	32625_at	0.0549	0.70
12	34151_at	0.8889	0.83	33925_at	0.737	0.76	33079_at	0.0546	0.66
13	40823_s_at	0.8864	0.79	33650_at	0.7333	0.76	35853_at	0.0546	0.63
14	33888_at	0.884	0.75	40823_s_at	0.7333	0.75	36731_g_at	0.0544	0.61
15	33290_at	0.8791	0.68	635_s_at	0.7311	0.73	36268_at	0.0539	0.58
16	518_at	0.8775	0.64	39082_at	0.7296	0.69	37134_f_at	0.0539	0.52
17	38189_s_at	0.8758	0.57	35853_at	0.7278	0.63	37903_at	0.0539	0.51
18	465_at	0.8725	0.57	34361_at	0.7267	0.58	41587_g_at	0.0531	0.48
19	35763_at	0.8709	0.53	32887_at	0.7222	0.55	35225_at	0.0529	0.41
20	845_at	0.8709	0.46	33307_at	0.7222	0.51	41397_at	0.0529	0.36

Aim – sample size calculation

- Gene expression experiments are expensive, therefore they tend to be small, this not necessarily bad.
- Task is to select informative genes from the pool of genes studied, the criterion for choosing sample sizes should be that they be large enough to ensure that informative genes have a high chance of being selected for further study.
- Resources exist such that the top-ranked k_0 (k_0 =100) genes will be considered for further study.
  Sample size is determined by the requirement that an informative gene ranking in the top k_1 (k_1=30) genes has a probability of 1-β of being ranked into the top k_0 in the experiment.
- Quantities of interest:
  $P_g(k_0|k_1)$ = P[Gene g ranked in the top k_0|true rank is in the top k_1]
  $P(k_0|k_1)$ = P[All true top k_1 genes ranked in the top k_0]

Computational Strategies

- Quantities of interest:
  $P_g(k_0|k_1)$ = P[Gene g ranked in the top k_0|true rank is in the top k_1]
  $P(k_0|k_1)$ = P[All true top k_1 genes ranked in the top k_0]
- Calculation of quantities of interest:
  Simulation study? But how to simulate a microarray experiment?
- Bootstrap:
  Use *pilot data* to bootstrap possible realisations of experiments.
- Simple Bootstrap:
  Bootstrap the pilot data and calculate $P_g(k_0|k_1)$ or $P(k_0|k_1)$.
- Bootstrap the bootstrap:
  Information on the distribution of the estimates of $P_g(k_0|k_1)$ or $P(k_0|k_1)$.
- van der Laan M, Bryan J (2001) Biostatistics 2:1-17

Results for the Huang data on recurrence

1000 Bootstrap samples were performed

$P_g(k_0 \mid k_1)$		k_0 = 100				
True Ranking	(k_1)	10	20	30	40	50
15	15	1.00	0.993	0.971	0.944	0.891
25	25	1.00	0.997	0.986	0.971	0.955
50	50	1.00	1.000	1.000	0.993	0.975

$P(k_0 \mid k_1)$		k_0 = 100				
True Ranking	(k_1)	10	20	30	40	50
15	15	0.988	0.866	0.711	0.332	0.163
25	25	0.999	0.955	0.871	0.517	0.254
50	50	1.000	1.000	0.965	0.781	0.552

The need of clinical epidemiology

Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer

Chang JC et al. (2003), The Lancet, **362**:362-369

- Neoadjuvant chemotherapy for locally advanced breast cancer
  Good response → breast cons. Surgery, assoc. with survival
  → prognostic value
- Need for a good response predictor
  → possibly of post-surgical relevance for adjuvant treatment for micrometastates
- 24 patients with locally advanced cancers, phase II — small sample size
  sensitive/resistant → tumour residual volume — relevant endpoint?
  12625 probe sets → 1628 with highest variance — prediction
  → 91 used for signature
  PPV 93%, NPV 83% — population

How to consider such results?
How to draw conclusions?

Careful design

- **Classifiers** derived from small series will be *overfitted* to the original dataset and may not have *general applicability*.
- **No test sample**, external sample (no clinical details) of six patients. **Cross-validation** or **leave one out** approach does not correct enough for the **overfitting** in small samples.
- No **stringent criteria** for assigning outcomes to samples**.**

What is a sensitive / resistant tumour?

Careful design

- Decision made at the end of the study on the basis of the observed median *relative residual volume* of disease.
  Tumour size - product of the two largest perpendicular diameters measured before and after → percentage of residual disease
  median of *relative residual volume* used to divide cancers into roughly equal groups - sensitive and resistant tumours.
- No pathological response considered, endpoint not correlated to survival.
- Measures based on tumour value may not have clinical relevance.

Sensitive and resistant tumours differ in tumour size and histology

	sensitive	resistant
Median perpendicular ∅ [mm]	80	30
invasive ductal carcinoma	5/11	11/13
invasive mammary carcinoma	6/11	2/13

1.) Size inversely related to response, responsive tumours have largest median diameter.

2.) Histology (exact Banard's Test p = 0.0528): Why do response rates in ductal carcinoma differ from invasive mammary carcinoma

Classifier might represent differences in size and histology rather than docetaxel sensitivity.

Classifier does not include genes that have previously been associated with taxane resistance.

Appropriate statistical framework

Computational reproducibility

- Van ‘t Veer LJ et al. (2002), Nature, **415**: 530-536
- Huang E et al. (2003), The Lancet, **361**:1590-1596

Tibshirani and Efron report: „We reanalysed the breast cancer data from van ,t Veer et al. ... Even with some help of the authors, we were unable to exactly reproduce this analysis“ (*Statistical applications in Genetics and Molecular Biology,* Vol.1 , Article 1, 2002)

Huang et al. Present 52 patients, 18 with tumour recurrence. The authors present a novel algorithm for classification which was at the time of the Lancet publication not described in the literature. From the description on the web page we could not replicated the calculation.

Cross-validation, the wrong and the right way

Consider a simple classifier for microarrays:

1.) Starting with all genes, find the 200 genes having the largest correlation with the class labels
2.) Carry out nearest-centroid classification using only these 200 genes

How do we estimate the test set performance of this classifier?

Wrong: Apply cross-validation to step 2
Right: Apply cross-validation to steps 1 and 2

It is easy to simulate realistic data with the class independent of the outcome.
The true test error: 50% (half/half distribution of class labels)
But *wrong* CV error estimate is zero.

I know at least 4 high profile papers where this error is made.

Horror example: Frederiksen CM et al. (2003)
J Cancer Res Clin Oncol, 129:263-271

Classification of Dukes'B and C colorectal cancers using expression arrays

Frederiksen CM, Knudsen SL, Orntoft TF (2003) J. Cancer Res Clin Oncol, 129:263-271

Chip: HuGeneFL (6800 probesets, ~5000 genes)

25 subjects, five sample from normal tissue, colon cancer tissue Duke‘s stage A, B, C, D

Results:
The data indicates that it is possible at least to classify Dukes‘B and C colorectal tumours with microarrays.

Stage	SAM (1)	SAM (2)	Covariance
0	100%	80%	80%
A	20%	40%	20%
B	60%	80%	80%
C	100%	100%	100%
D	0%	0%	20%

base: 5 subjects per group

Generate Data with no correlation between class labile and expression value

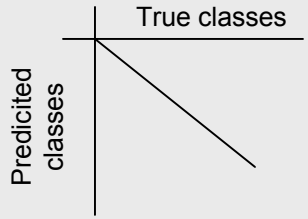
```
> aa.generate.data.rfc
function (anz.genes=5000,gr.names=aa.group.names)
{
        n<-anz.genes
        m<-length(gr.names)
        mat<-matrix(rnorm(n*m,0,1),ncol=m)
        colnames(mat)<-gr.names
        return(mat)
}
> aa.group.names
 [1] "O.1" "O.2" "O.3" "O.4" "O.5" "A.1" "A.2" "A.3" "A.4" "A.5" "B.1" "B.2"
[13] "B.3" "B.4" "B.5" "C.1" "C.2" "C.3" "C.4" "C.5" "D.1" "D.2" "D.3" "D.4"
[25] "D.5"
```

Preprocessing strategies

```
> aa.reduce.f.rfc
function (data=aa.data.5000,y.class=aa.gr.exp.25,anz.sel=400)
{
        require(multtest)
        yy<-y.class-1
        nn.1<-length(unique(yy))
        nn.2<-length(yy)
        f.res<-mt.teststat(data,yy,test="f",na=.mt.naNUM,nonpara="n")
        f.pvl<-1-pf(f.res,nn.1-1,nn.2-nn.1)
        f.adj<-mt.rawp2adjp(f.pvl)
        f.sel<-f.adj[[2]][1:anz.sel]
        return(data[f.sel,])
}

> aa.reduce.cor.rfc
function (data=aa.data.5000,y.class=aa.gr.exp.25,anz.sel=200)
{
        cor.res<-apply(data,1,cor,y=y.class)
        cor.srt<-order(cor.res)
        nn<-length(cor.res)
        ss<-c(1:anz.sel,(nn-anz.sel+1):nn)
        return(data[cor.srt[ss],])
}
```

Generate new data, make preprocessing and perform classification

```
> aa.cor.simulation.rfc
function (i=1,how.many.genes=5000,how.many.select=200)
{
        dd.full<-aa.generate.data.rfc(anz.genes=how.many.genes)
        dd.redu<-aa.reduce.cor.rfc(data=dd.full,anz.sel=how.many.select)
        return(aa.generate.class.rfc(data.train=dd.redu))
}

> aa.f.simulation.rfc
function (i=1,how.many.genes=5000,how.many.select=200)
{
        dd.full<-aa.generate.data.rfc(anz.genes=how.many.genes)
        dd.redu<-aa.reduce.f.rfc(data=dd.full,anz.sel=how.many.select)
        return(aa.generate.class.rfc(data.train=dd.redu))
}

> aa.generate.class.rfc
function (data.train=aa.data.cor,y.class=as.factor(aa.gr.exp.25),k.nn=3)
{
        require(class)
        y.knn<-knn.cv(t(data.train),y.class,k=k.nn)
        return(diag(table(y.knn,y.class)))
}
```



Perform the simulation

```
> aa.cor.simulation.run.rfc
function (anz.simul=10,h.m.g=5000,h.m.s=200)
{
        mm<-matrix(1:anz.simul,ncol=1)
        rr<-
apply(mm,1,aa.cor.simulation.rfc,how.many.genes=h.m.g,how.many.select=h.m.s)
        return(rr)
}


> aa.f.simulation.run.rfc
function (anz.simul=10,h.m.g=5000,h.m.s=200)
{
        mm<-matrix(1:anz.simul,ncol=1)
        rr<-
apply(mm,1,aa.f.simulation.rfc,how.many.genes=h.m.g,how.many.select=h.m.s)
        return(rr)
}
```

Evaluate simulation

```
aa.how.many.above.rfc
function (x,cut.point)
{
        return(sum(ifelse(x>=cut.point,1,0)))
}

table(apply(aa.cor.simulation.run.1.res,2, aa.how.many.above.rfc,cut.point=4))

 0  1  2  3  4
 7 38 83 60 12

> table(apply(aa.f.simulation.run.1.res,2, aa.how.many.above.rfc,cut.point=4))

  3   4   5
  2  31 167
```

Summary / Discussion

- Interpretability of results is determined by a series of methodological decisions made by the scientists who perform the experiment / study.
- Decisions on technology, experimental design, data analysis and validation of results need to be integrated into a comprehensive statistical framework. Because of the complexity and novelty of the problem there is only a little progress to achieve this integration.
- At the moment there is a huge transfer of classical biostatistical methods into the field of microarray experiments. It is necessary to adapt these methods to bioinformatic information.
- *If the collection, analysis and interpretation of the data are flawed then it may not only be a waste of a valuable resource - we could draw faulty conclusions and potentially risk our health and environment.*

DNA microarrays: Vital statistics (2003) Nature, 424:610-612