

---

# Differential Gene Expression

Rainer Spang

Courses in Practical DNA Microarray Analysis

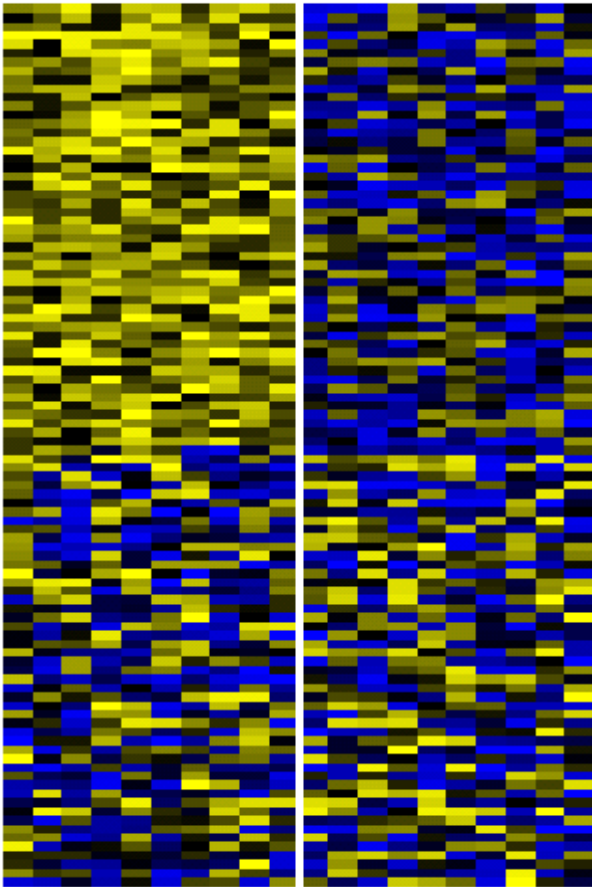
---



Nationales  
Genomforschungsnetz

A

B



## Two cell/tissue /disease types:

wild-type / mutant

control / treated

disease A / disease B

responding / non responding

etc. etc....

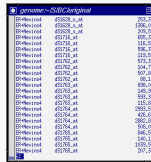
For every sample (cell line/patient) we have the expression levels of thousands of genes and the information whether it is A or B

## **Differential gene expression:**

**Which genes are differentially expressed in the two tissue type populations?**

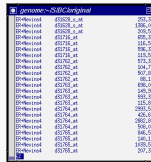
# A cost efficient (cheap) experiment:

A



Gene	Expression
gga00001	1.0
gga00002	1.0
gga00003	1.0
gga00004	1.0
gga00005	1.0
gga00006	1.0
gga00007	1.0
gga00008	1.0
gga00009	1.0
gga00010	1.0
gga00011	1.0
gga00012	1.0
gga00013	1.0
gga00014	1.0
gga00015	1.0
gga00016	1.0
gga00017	1.0
gga00018	1.0
gga00019	1.0
gga00020	1.0
gga00021	1.0
gga00022	1.0
gga00023	1.0
gga00024	1.0
gga00025	1.0
gga00026	1.0
gga00027	1.0
gga00028	1.0
gga00029	1.0
gga00030	1.0
gga00031	1.0
gga00032	1.0
gga00033	1.0
gga00034	1.0
gga00035	1.0
gga00036	1.0
gga00037	1.0
gga00038	1.0
gga00039	1.0
gga00040	1.0
gga00041	1.0
gga00042	1.0
gga00043	1.0
gga00044	1.0
gga00045	1.0
gga00046	1.0
gga00047	1.0
gga00048	1.0
gga00049	1.0
gga00050	1.0
gga00051	1.0
gga00052	1.0
gga00053	1.0
gga00054	1.0
gga00055	1.0
gga00056	1.0
gga00057	1.0
gga00058	1.0
gga00059	1.0
gga00060	1.0
gga00061	1.0
gga00062	1.0
gga00063	1.0
gga00064	1.0
gga00065	1.0
gga00066	1.0
gga00067	1.0
gga00068	1.0
gga00069	1.0
gga00070	1.0
gga00071	1.0
gga00072	1.0
gga00073	1.0
gga00074	1.0
gga00075	1.0
gga00076	1.0
gga00077	1.0
gga00078	1.0
gga00079	1.0
gga00080	1.0
gga00081	1.0
gga00082	1.0
gga00083	1.0
gga00084	1.0
gga00085	1.0
gga00086	1.0
gga00087	1.0
gga00088	1.0
gga00089	1.0
gga00090	1.0
gga00091	1.0
gga00092	1.0
gga00093	1.0
gga00094	1.0
gga00095	1.0
gga00096	1.0
gga00097	1.0
gga00098	1.0
gga00099	1.0
gga00100	1.0

B



Gene	Expression
gga00001	1.0
gga00002	1.0
gga00003	1.0
gga00004	1.0
gga00005	1.0
gga00006	1.0
gga00007	1.0
gga00008	1.0
gga00009	1.0
gga00010	1.0
gga00011	1.0
gga00012	1.0
gga00013	1.0
gga00014	1.0
gga00015	1.0
gga00016	1.0
gga00017	1.0
gga00018	1.0
gga00019	1.0
gga00020	1.0
gga00021	1.0
gga00022	1.0
gga00023	1.0
gga00024	1.0
gga00025	1.0
gga00026	1.0
gga00027	1.0
gga00028	1.0
gga00029	1.0
gga00030	1.0
gga00031	1.0
gga00032	1.0
gga00033	1.0
gga00034	1.0
gga00035	1.0
gga00036	1.0
gga00037	1.0
gga00038	1.0
gga00039	1.0
gga00040	1.0
gga00041	1.0
gga00042	1.0
gga00043	1.0
gga00044	1.0
gga00045	1.0
gga00046	1.0
gga00047	1.0
gga00048	1.0
gga00049	1.0
gga00050	1.0
gga00051	1.0
gga00052	1.0
gga00053	1.0
gga00054	1.0
gga00055	1.0
gga00056	1.0
gga00057	1.0
gga00058	1.0
gga00059	1.0
gga00060	1.0
gga00061	1.0
gga00062	1.0
gga00063	1.0
gga00064	1.0
gga00065	1.0
gga00066	1.0
gga00067	1.0
gga00068	1.0
gga00069	1.0
gga00070	1.0
gga00071	1.0
gga00072	1.0
gga00073	1.0
gga00074	1.0
gga00075	1.0
gga00076	1.0
gga00077	1.0
gga00078	1.0
gga00079	1.0
gga00080	1.0
gga00081	1.0
gga00082	1.0
gga00083	1.0
gga00084	1.0
gga00085	1.0
gga00086	1.0
gga00087	1.0
gga00088	1.0
gga00089	1.0
gga00090	1.0
gga00091	1.0
gga00092	1.0
gga00093	1.0
gga00094	1.0
gga00095	1.0
gga00096	1.0
gga00097	1.0
gga00098	1.0
gga00099	1.0
gga00100	1.0

We observe a gene with a two-fold higher expression in profile A than in profile B.

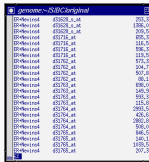
Is two-fold trust worthy?

Well, by how much can this gene change in group A and in group B?

By no more than 10% than the answer is yes, by up to 500% then the answer is no.

# A cost efficient (cheap) experiment II:

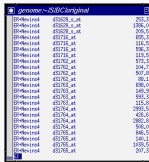
A



Gene	Expression
gmsm1	1.0
gmsm2	1.0
gmsm3	1.0
gmsm4	1.0
gmsm5	1.0
gmsm6	1.0
gmsm7	1.0
gmsm8	1.0
gmsm9	1.0
gmsm10	1.0
gmsm11	1.0
gmsm12	1.0
gmsm13	1.0
gmsm14	1.0
gmsm15	1.0
gmsm16	1.0
gmsm17	1.0
gmsm18	1.0
gmsm19	1.0
gmsm20	1.0
gmsm21	1.0
gmsm22	1.0
gmsm23	1.0
gmsm24	1.0
gmsm25	1.0
gmsm26	1.0
gmsm27	1.0
gmsm28	1.0
gmsm29	1.0
gmsm30	1.0
gmsm31	1.0
gmsm32	1.0
gmsm33	1.0
gmsm34	1.0
gmsm35	1.0
gmsm36	1.0
gmsm37	1.0
gmsm38	1.0
gmsm39	1.0
gmsm40	1.0
gmsm41	1.0
gmsm42	1.0
gmsm43	1.0
gmsm44	1.0
gmsm45	1.0
gmsm46	1.0
gmsm47	1.0
gmsm48	1.0
gmsm49	1.0
gmsm50	1.0
gmsm51	1.0
gmsm52	1.0
gmsm53	1.0
gmsm54	1.0
gmsm55	1.0
gmsm56	1.0
gmsm57	1.0
gmsm58	1.0
gmsm59	1.0
gmsm60	1.0
gmsm61	1.0
gmsm62	1.0
gmsm63	1.0
gmsm64	1.0
gmsm65	1.0
gmsm66	1.0
gmsm67	1.0
gmsm68	1.0
gmsm69	1.0
gmsm70	1.0
gmsm71	1.0
gmsm72	1.0
gmsm73	1.0
gmsm74	1.0
gmsm75	1.0
gmsm76	1.0
gmsm77	1.0
gmsm78	1.0
gmsm79	1.0
gmsm80	1.0
gmsm81	1.0
gmsm82	1.0
gmsm83	1.0
gmsm84	1.0
gmsm85	1.0
gmsm86	1.0
gmsm87	1.0
gmsm88	1.0
gmsm89	1.0
gmsm90	1.0
gmsm91	1.0
gmsm92	1.0
gmsm93	1.0
gmsm94	1.0
gmsm95	1.0
gmsm96	1.0
gmsm97	1.0
gmsm98	1.0
gmsm99	1.0
gmsm100	1.0

Is a three-fold induced gene more trust worthy than a two-fold induced gene?

B



Gene	Expression
gmsm1	1.0
gmsm2	1.0
gmsm3	1.0
gmsm4	1.0
gmsm5	1.0
gmsm6	1.0
gmsm7	1.0
gmsm8	1.0
gmsm9	1.0
gmsm10	1.0
gmsm11	1.0
gmsm12	1.0
gmsm13	1.0
gmsm14	1.0
gmsm15	1.0
gmsm16	1.0
gmsm17	1.0
gmsm18	1.0
gmsm19	1.0
gmsm20	1.0
gmsm21	1.0
gmsm22	1.0
gmsm23	1.0
gmsm24	1.0
gmsm25	1.0
gmsm26	1.0
gmsm27	1.0
gmsm28	1.0
gmsm29	1.0
gmsm30	1.0
gmsm31	1.0
gmsm32	1.0
gmsm33	1.0
gmsm34	1.0
gmsm35	1.0
gmsm36	1.0
gmsm37	1.0
gmsm38	1.0
gmsm39	1.0
gmsm40	1.0
gmsm41	1.0
gmsm42	1.0
gmsm43	1.0
gmsm44	1.0
gmsm45	1.0
gmsm46	1.0
gmsm47	1.0
gmsm48	1.0
gmsm49	1.0
gmsm50	1.0
gmsm51	1.0
gmsm52	1.0
gmsm53	1.0
gmsm54	1.0
gmsm55	1.0
gmsm56	1.0
gmsm57	1.0
gmsm58	1.0
gmsm59	1.0
gmsm60	1.0
gmsm61	1.0
gmsm62	1.0
gmsm63	1.0
gmsm64	1.0
gmsm65	1.0
gmsm66	1.0
gmsm67	1.0
gmsm68	1.0
gmsm69	1.0
gmsm70	1.0
gmsm71	1.0
gmsm72	1.0
gmsm73	1.0
gmsm74	1.0
gmsm75	1.0
gmsm76	1.0
gmsm77	1.0
gmsm78	1.0
gmsm79	1.0
gmsm80	1.0
gmsm81	1.0
gmsm82	1.0
gmsm83	1.0
gmsm84	1.0
gmsm85	1.0
gmsm86	1.0
gmsm87	1.0
gmsm88	1.0
gmsm89	1.0
gmsm90	1.0
gmsm91	1.0
gmsm92	1.0
gmsm93	1.0
gmsm94	1.0
gmsm95	1.0
gmsm96	1.0
gmsm97	1.0
gmsm98	1.0
gmsm99	1.0
gmsm100	1.0

Actually this depends on the within class variability of the two genes again, it can be the other way round.

A

Terminal window A displays a list of gene expression data. Each line represents a gene across two conditions (L and R). The columns are: Gene ID (e.g., D141500\_4), Log2(FPKM) L, Log2(FPKM) R, and a third column of values. The values range from approximately 10.0 to 20.0.

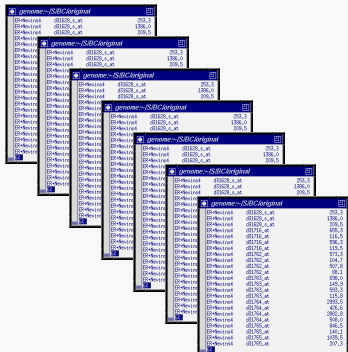
B

Terminal window B displays a list of gene expression data, similar to window A, but with a different set of values for the same genes across the L and R conditions.

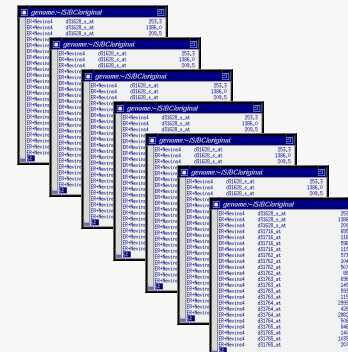
**Conclusion:** In addition to the differences in gene expression you also have a vital interest in its variability ... This information is needed to obtain meaningful lists of genes

**Therefore:** Invest money in repeated experiments!

A



B



# Standard Deviation and Standard Error

**Standard Deviation (SD):** Variability of the measurement

**Standard Error (SE):** Variability of the mean of several measurements

**n Replications**

**Normal Distributed Data:**

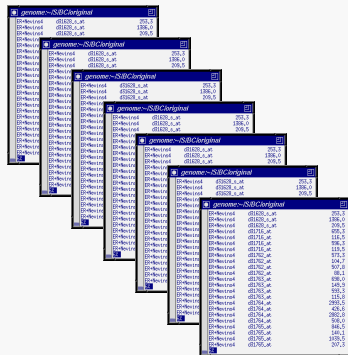
$$SE = \frac{1}{\sqrt{n}} SD$$

$$SE = \frac{1}{\sqrt{n}} SD$$

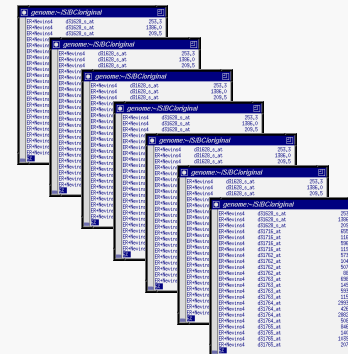
**Conclusion:** Repetitions lead to a more precise measurement of gene expression. Single expression measurements are very noisy, average expression across several repetitions is much less noisy

**Therefore:** Invest money in repeated experiments!

A



B





# The additive scale:

You will want to use the wealth of statistical theory to analyze your data

- Most **statistics works on an additive scale** (Significance of differences etc ...)
- **Gene expression works on a multiplicative scale** (fold changes ...)

**Conclusion:** Transform your data to the additive scale

- Simple way: take logs
- Better way: use variance stabilization

# Questions:

*Which genes are differentially expressed?*

→ **Ranking**

*Are these results „significant“*

→ **Statistical Analysis**

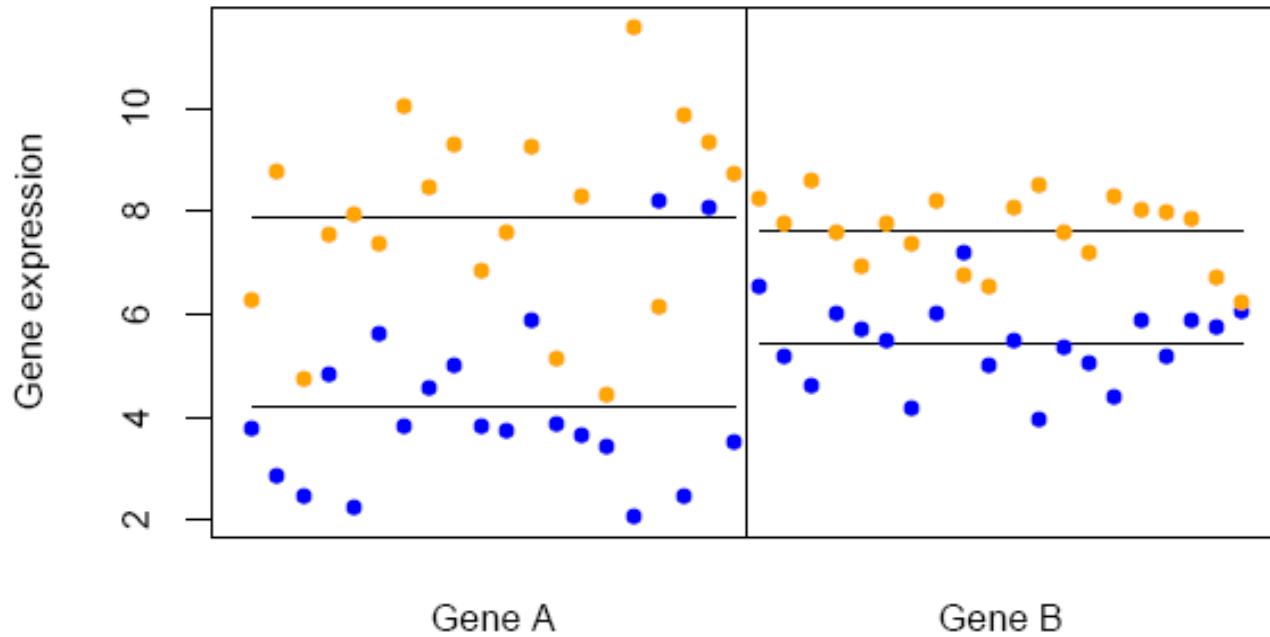
# Ranking:

**Problem:** Produce an ordered list of differentially expressed genes starting with the most up regulated gene and ending with the most down regulated gene

Ranking means finding the right genes  
... drawing our attention to them

In many applications it is the most important step

# Which gene is more differentially expressed?



# Ranking is **Scoring**

**You need to score differential  
gene expression**

**Different scores lead to different  
rankings**

**What scores are there?**

# Fold Change & Log Ratios

You have transformed your data to additive scale!

Factors become differences:

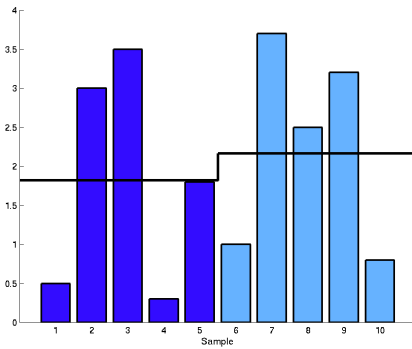
$$\log(a/b) = \log(a) - \log(b)$$

if you want to talk by fold change you compute the average expression in both groups and subtract them.

$$LR = \bar{X}_1 - \bar{X}_2$$

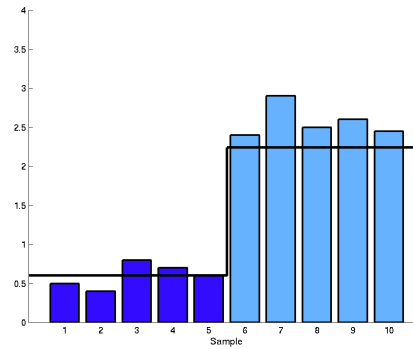
# T-Score

**Idea:** Take variances into account



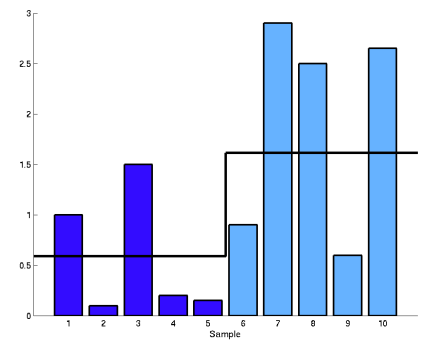
**Change:** low  
high

**Variance:** high  
**Variance:** high



**Change:** high

**Variance:** low



**Change:**

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

# Fudge Factors:

You need to estimate the variance from data

You might underestimate a already small variance  
(constantly expressed genes)

The denominator in T becomes really small

**Constantly expressed genes show up on top of the list**

**Fix:** Add a constant fudge factor  $s_0$

→ Regularized T-score

$$T_r = \frac{\bar{X}_1 - \bar{X}_2}{c(s + s_0)}$$

→ Limma

→ SAM

→ Twighlight



# Univariate Biomarkers

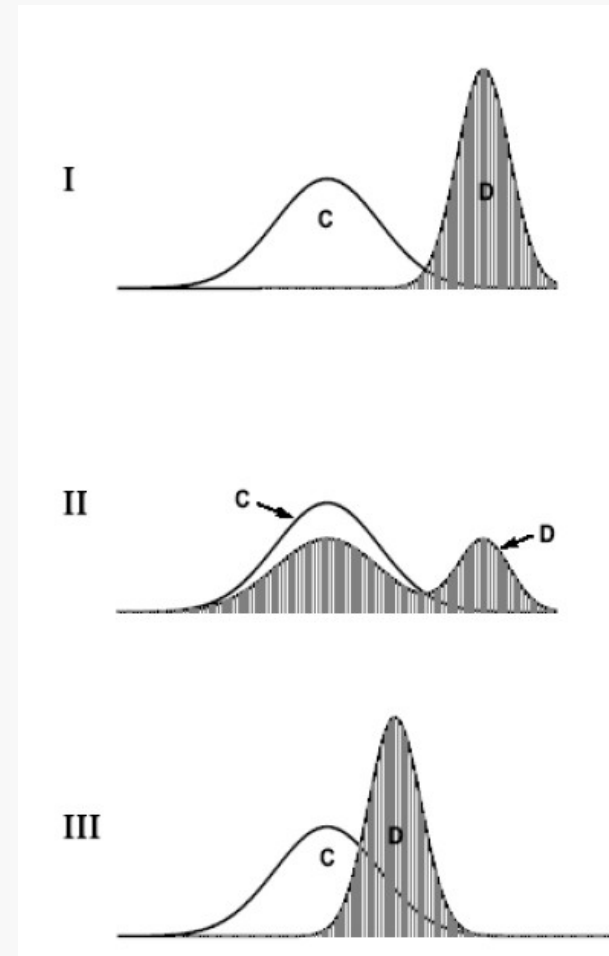
t-scores focus on the difference of population means

This does not imply good separation of the classes (II & III)

→ no good biomarkers

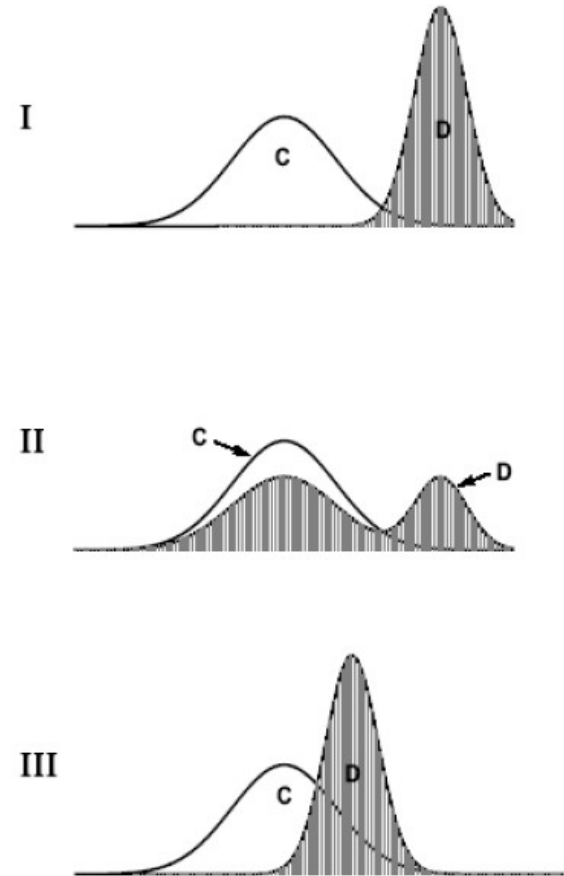
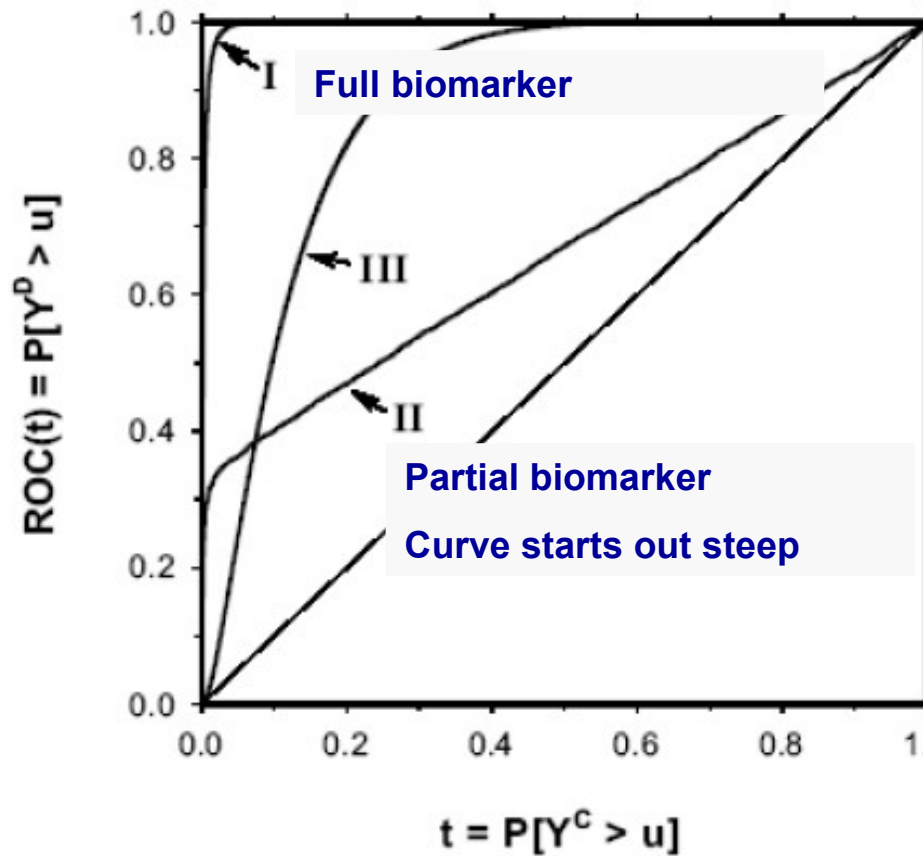
II identifies at least a subset of the members in the D class reliably

→ partial biomarker



Pepe et al., Biometrics 2003

# ROC-Curves

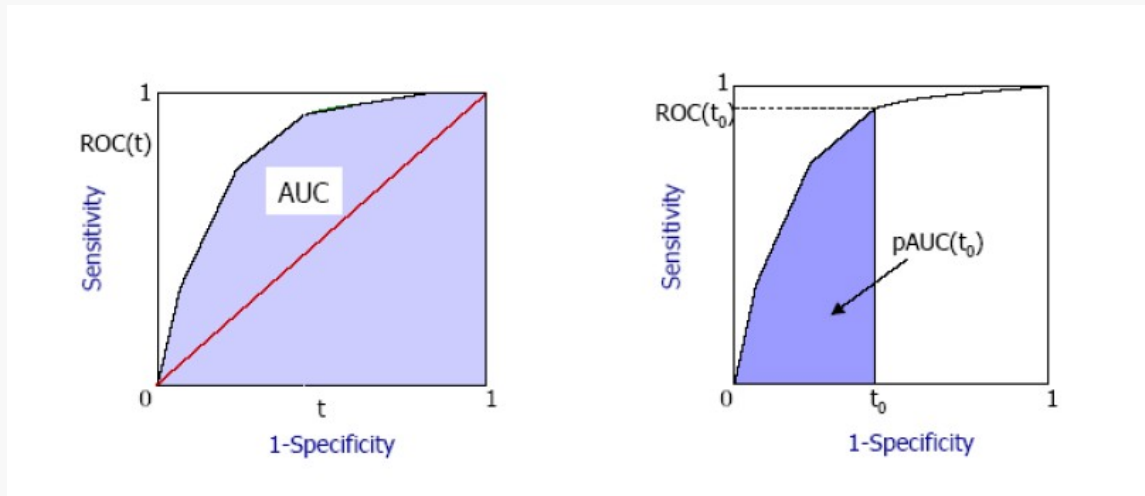


# (Partial) area under the curve

**AUC:** Score for univariate discrimination ability

→ Full biomarkers

**p(AUC)(t):** → Partial biomarkers (curve starts out steep)



**Multivariate biomarkers** → signatures

... see class tomorrow

# Confounders

You have compared two types of disease A and B and you have identified the 100 top scoring genes.

75% of the patients with disease A are man, while only 38% of patients with disease B are man.

**Problem:**

Are the observed expression differences disease or sex specific?

How can we score genes such that disease specific genes rank high?



# Linear models

$$y_i = a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + \varepsilon_i$$

= 1 if patient has  
disease A

= 0 otherwise

= 1 if patient has  
disease B

= 0 otherwise

= 1 if patient is  
male

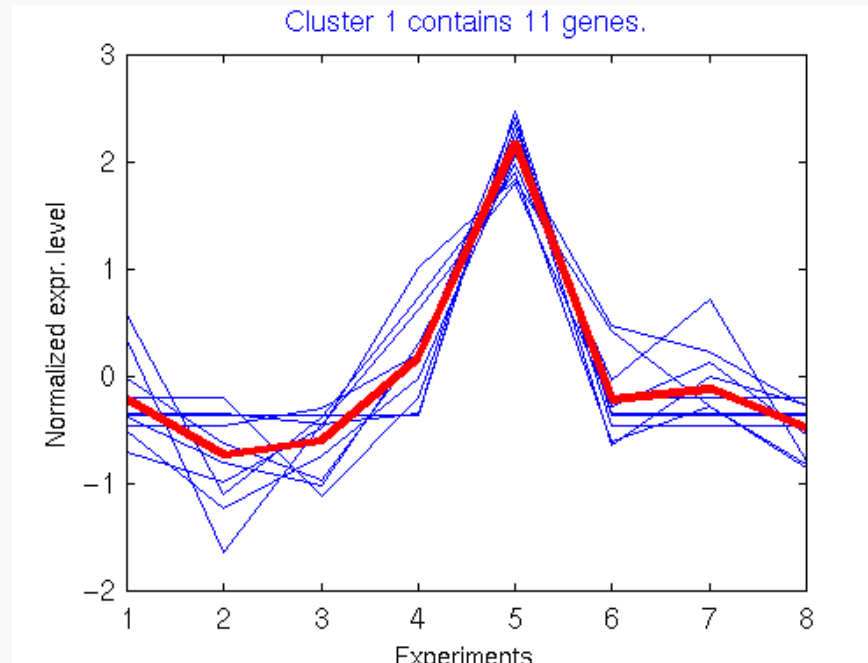
= 0 otherwise

$a_1$ - $a_2$  scores differences associated with the  
disease type independently from the gender

$a_1$ - $a_2$  is called a **contrast** and the matrix  $(x_{ik})$  the  
**design matrix** of the linear model

→ Limma package

# Correlation to a reference gene



**This is also a screening and testing problem and not a clustering problem**

# Different scores give different rankings

Gene	t-score	Limma	Fudge	Log ratio	Wilcoxon	pAUC
<i>MGST1</i>	1	1	3	21	5	27
<i>DF</i>	2	2	1	1	22	4
<i>CD33</i>	3	3	8	87	1	3
<i>CST3</i>	4	4	2	2	4	1
<i>TCF3</i>	5	5	11	58	3	5
<i>MLP</i>	6	7	22	118	8	28
<i>CSTA</i>	7	6	5	18	11	10
<i>CTSD</i>	8	8	27	144	7	12
<i>SPTAN1</i>	9	9	19	62	12	17
<i>CCND3</i>	10	11	17	51	10	6
<i>PSMA6</i>	20	18	24	63	21	30
<i>CD63</i>	30	30	46	120	29	158
<i>FCER1G</i>	40	38	23	29	49	164
<i>SPI1</i>	50	48	20	10	46	64
<i>LTC4S</i>	60	63	150	359	105	45

**ALL vs AML (Golub et al.)**

**Which Score is the best one?**

**That depends on your problem ...**



# Rankings are notoriously **unstable**

The scores of 30.000 genes typically form an almost continuous spectrum with little or no outliers.

The difference in score between genes that are several hundred ranks apart are so small that they can not be reproduced

*The ability of microarrays to reliably identify differentially expressed genes is low ...*



# Next Question:

**Ok, I chose a score and found a set of candidate genes**

**Can I trust the observed expression differences?**

**→ Statistical Analysis**

# P-Values

**Everyone knows that the p-value must be below 0.05**

**0.05 is a holy number both in medicine and biology**

**... what else should you know about p-values**

# Rumors

**If the gene is not differentially expressed the p-value is high**

**If the gene is differentially expressed the p-values is low**

**Both these statements are wrong!**

# The basic Idea behind **p-values**:

We observe a score  $s=1.27$

Can this be just a random fluctuation?

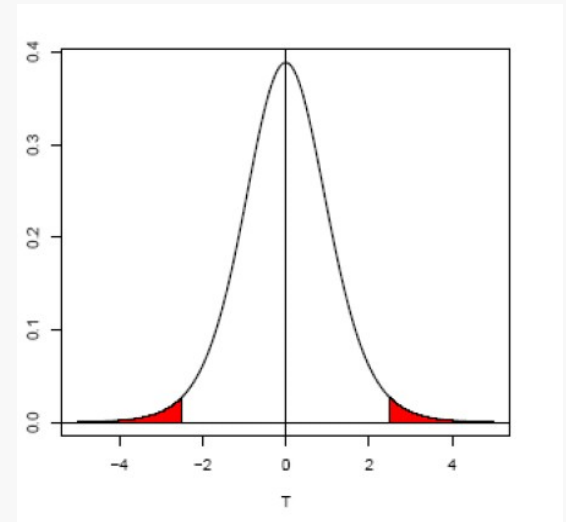
**Assume:** It is a random fluctuation

→ The gene is not differentially expressed

→ The null hypothesis holds

**Theory** gives us the distribution of the score under this assumption

**P-Value:** Probability that a random score is equal or higher to  $s=1.27$  in absolute value (two sided test)



# Permutations and empirical p-values

Target class labels

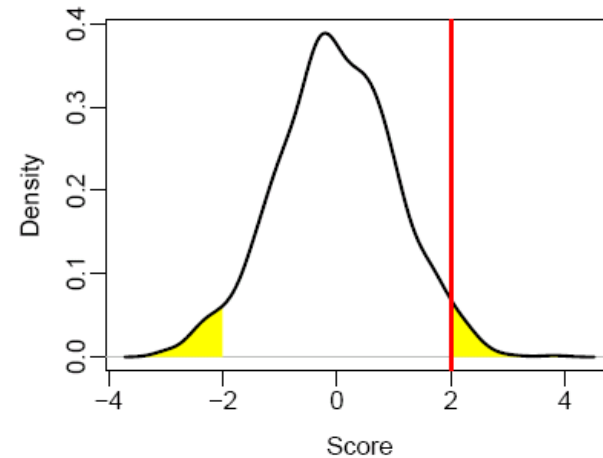
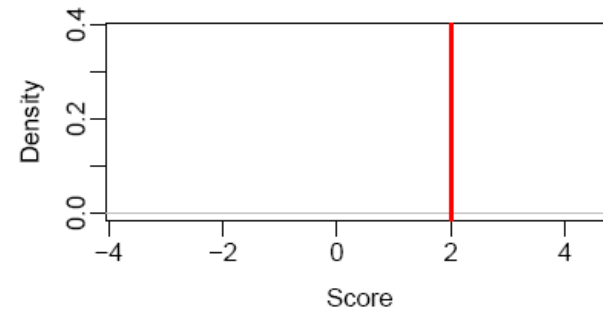
0	0	0	0	0	1	1	1	1	1
---	---	---	---	---	---	---	---	---	---

Permuted class labels

0	1	1	0	0	0	1	0	1	1
1	0	1	1	1	0	0	0	0	1
0	1	1	0	0	1	1	0	0	1

⋮

0	0	1	1	1	0	1	0	1	0
---	---	---	---	---	---	---	---	---	---



**If a gene is not differentially expressed:**

*The p-value is a random number between 0 and 1!*



**It is unlikely that such a number is below 0.05 (5% probability)**

**If a gene is differentially expressed:**

*The p-value has no meaning, since it was computed under the assumption that the gene is not differentially expressed.*

**We hope that it is small since the score is high, but there is absolutely no theoretical support for this**



## ***Testing only one gene:***

**If the gene is not differentially expressed a small p-value is unlikely, hence we should be surprised by this observation.**

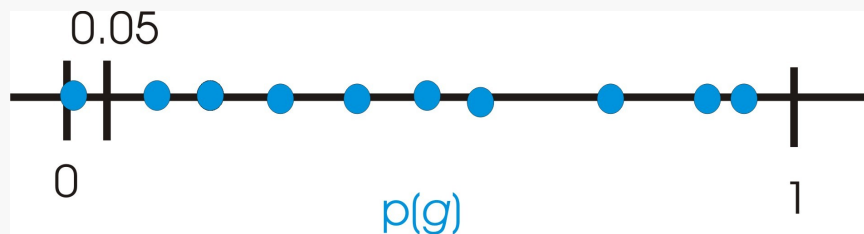
***If we make it a rule that we discard the gene if the p-values is above 0.05, it is unlikely that a random score will pass this filter***

# Multiple testing with only non-induced genes

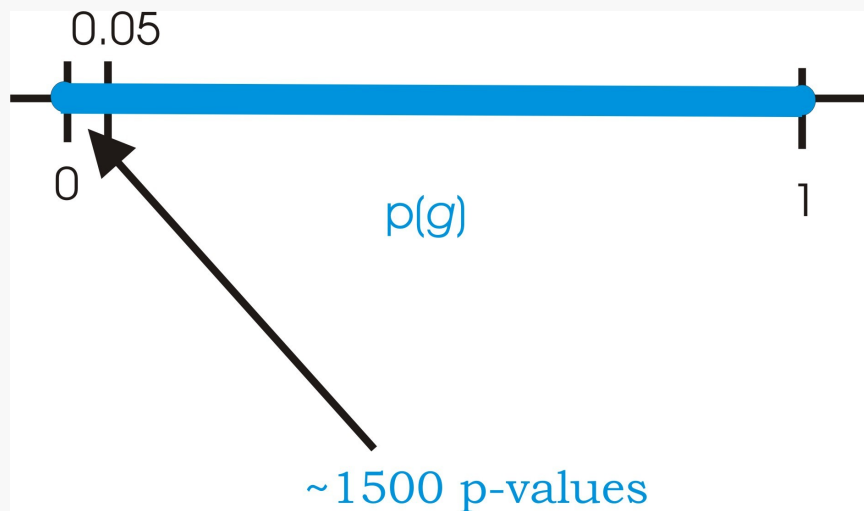
1 gene



10 genes



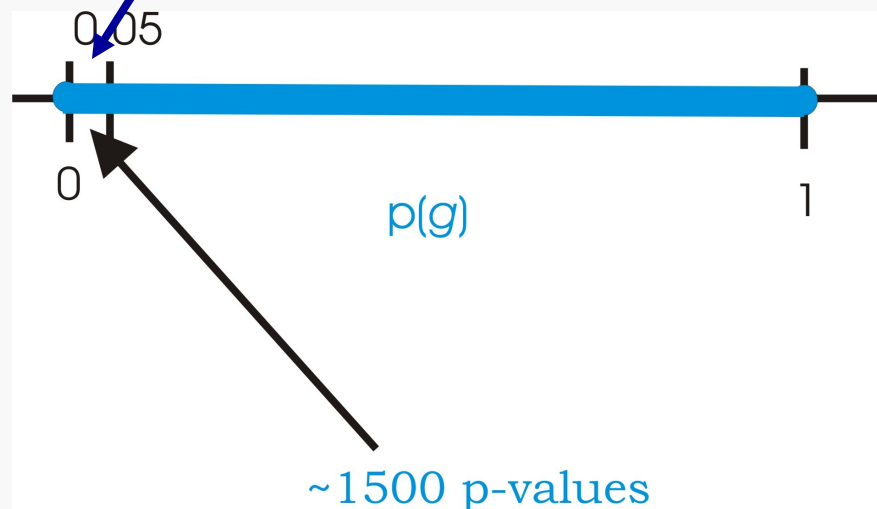
30,000 genes



# The Multiple Testing Problem



**P-values are random numbers between 0 and 1. For only one such number it is unlikely to fall in this small interval, but if we have 30.000 such numbers many will be in there.**



# Extreme value statistics

**Validation experiment:** Hybridize the same probe twice and score the differences

**Observation:** Some genes show 3 fold changes

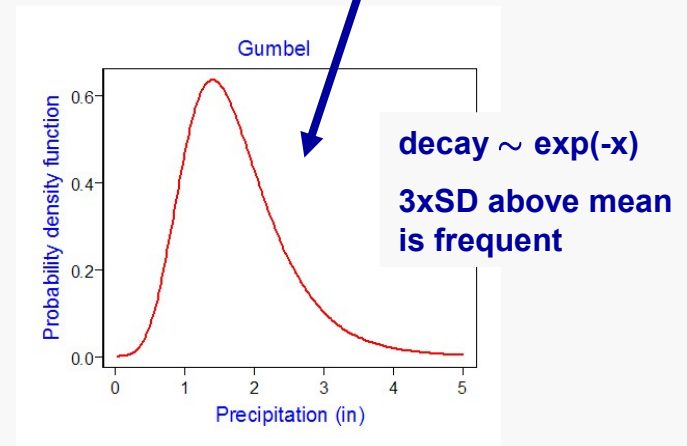
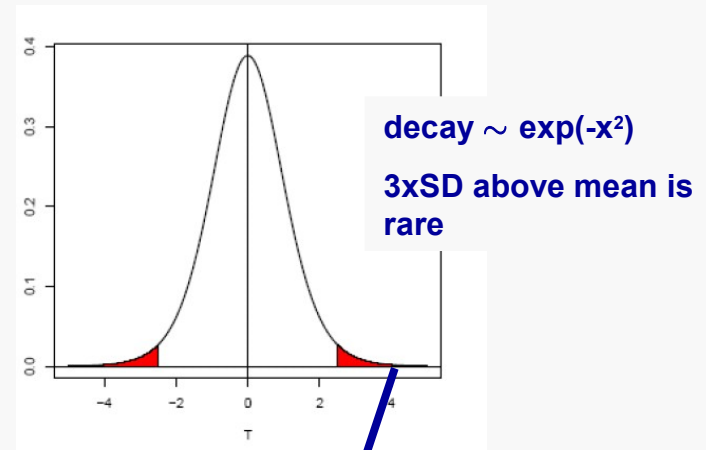
**Wrong conclusion:** Microarray experiments are not reproducible

A randomly selected gene is very reproducible ... the 3 fold change is „caused“ by looking at the genes with the highest score ... the ends of the ranked lists

... taking the maximum of 30.000 genes causes much more noise than the measurement. This is a general problem of a screening approach !!!

Distribution of a **random gene**

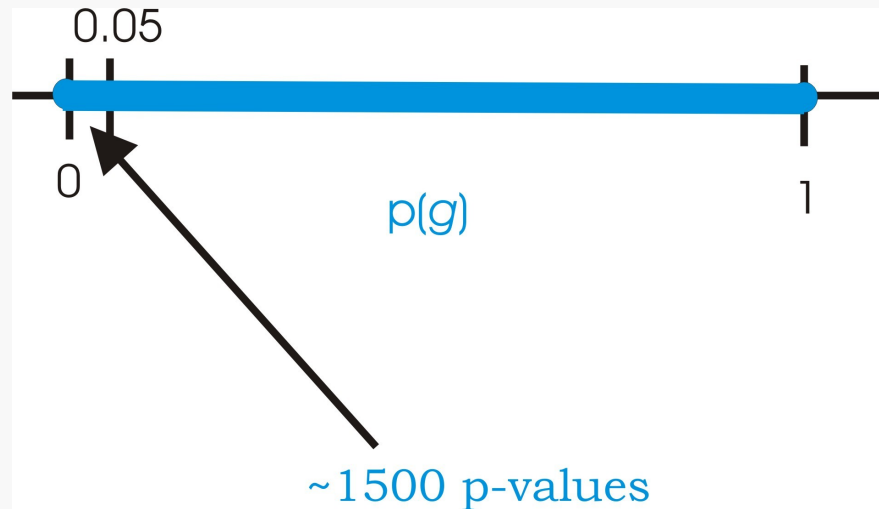
Normal Distribution



Distribution of the **maximal scoring gene**

Gumbel Distribution

# Controlling the family wise error rate (FWER)



If we want to avoid random numbers in this interval we need to make it smaller. The more numbers, the smaller. For 30.000 numbers very small.

This strategy is called: **Controlling the family wise error rate**

# How to control the FWER?

Note, that adjusting the interval border can also be done by adjusting the p-values and leaving the cut off at 0.05.

There are many ways to adjust p-values for multiple testing:

Bonferoni:  $p_{adj} = p N$

Better: Westfall and Young → Exercises

**In microarray studies controlling the FWER is not a good idea ... It is too conservative.**

**A different type of error measure became more popular**

**The False Discovery Rate**

***What is the idea?***

# The FDR

- **Score genes and rank them**
- **Choose a cutoff**
- **Loosely speaking:** The FDR is the best guess for the number of false positive genes that score above the cutoff



# The confusing literature:

There are many different definitions of the false discovery rate in the literature:

- Original: Benjamini-Hochberg
- Positive FDR
- Conditional FDR
- Local FDR

There is also a fundamental difference between **controlling** and **estimating** a FDR

**In microarray analysis it became popular to use estimated FDRs**

***Differences to p-values:***

**The FDR refers to a **list** of genes. The p-value refers to a single gene.**

**The p-value is based on the assumption that the gene is not differentially expressed, the FDR makes no such assumption.**

**P-values need to be corrected for multiplicity, FDRs not!**

# Another difference in concept:

If a 4x change has a small p-value, this means that 4x change is too high to be random fluctuation

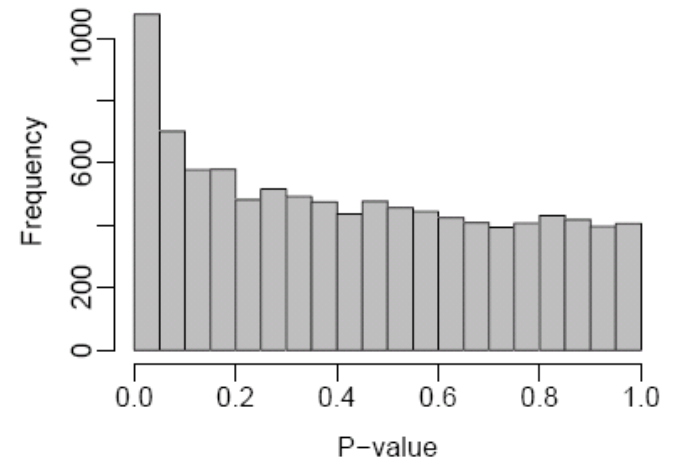
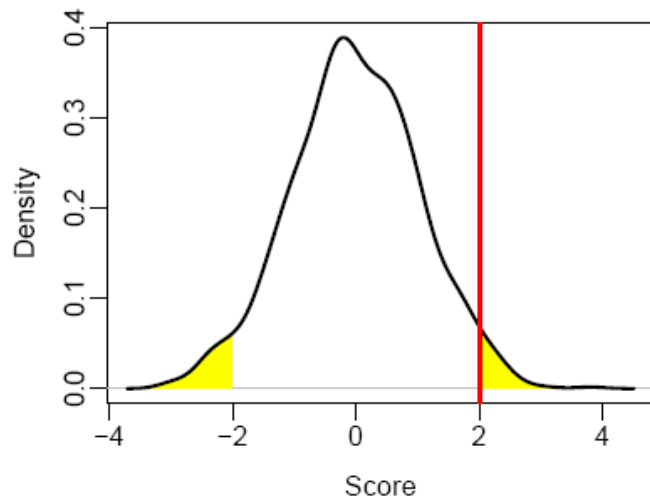
**Conclusion:** 4x change is significant

If a list of 150 genes with 4x change or more has a small estimated FDR this means that we have more genes on this level than would be expected by chance.

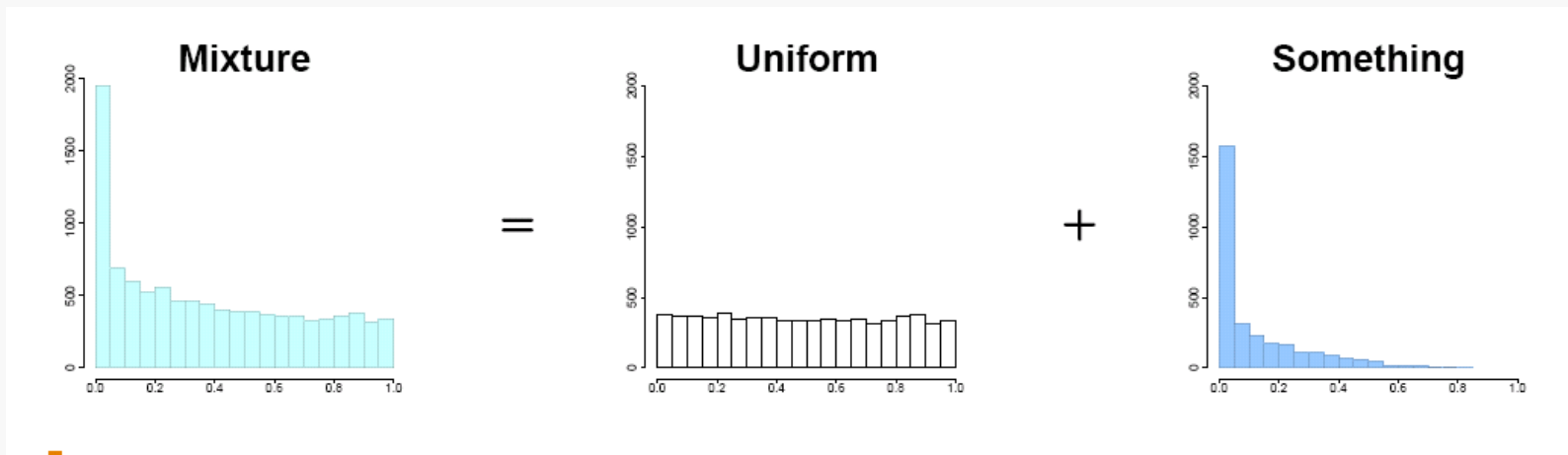
**Conclusion:** 4x change can be noise, but 150 genes on that level are too many to be explained just by random fluctuation.

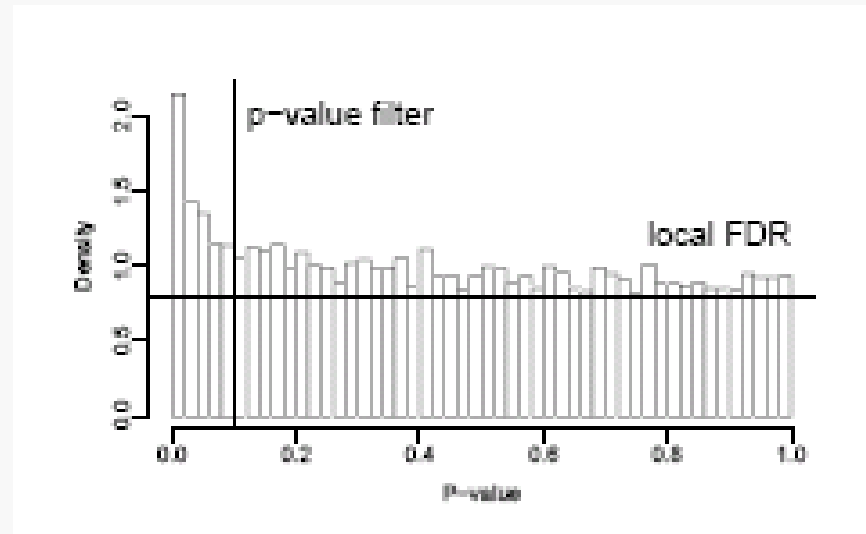
In **FWER** Analysis the fold change **4x** is significant, in **FDR** Analysis it is the number **150** that is significant.

# Histograms of the p-values of all genes on the array



# The mixture interpretation of the FDR

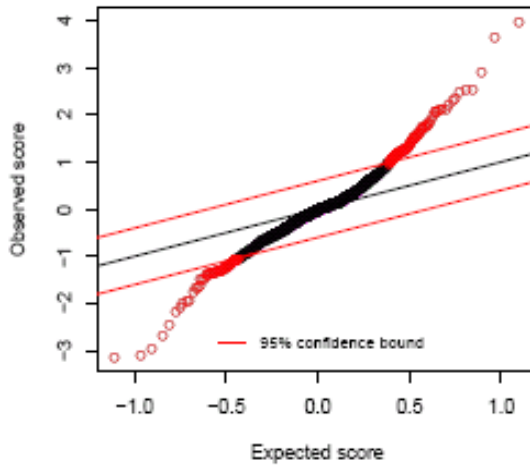




**FWER:** Vertical cutoff

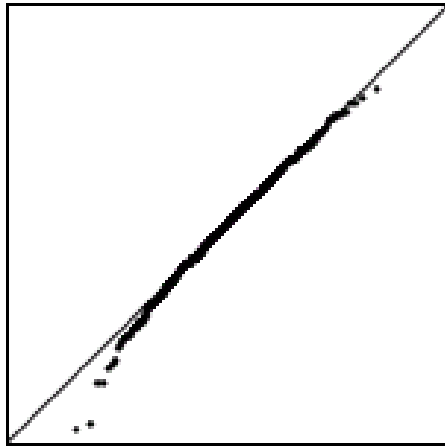
**FDR:** Horizontal cutoff

# The typical plots

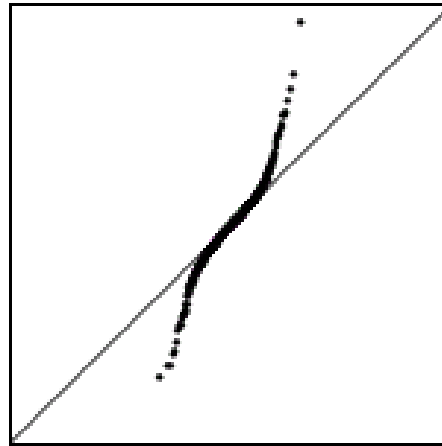


**Expected random score vs observed scores:  
Deviations from the main diagonal are  
evidence for differentially expressed genes**

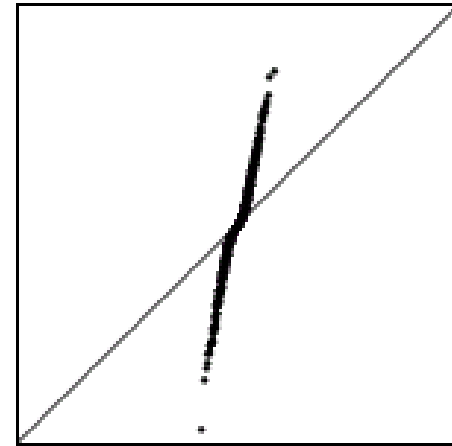
# What you typically observe



**No differential  
gene  
expression**



**A lot of  
differential  
gene  
expression**



**Global  
changes in  
gene  
expression**



# Summary

- Replications are useful, not only for statistical reasons (5-8 per leg)
- **Rankings are instable**
- **Screening increases the measurement noise**
- **Low FWER p-values will lead to many missed genes**
- **FDR (SAM) seems more appropriate**
- **Often there are many induced genes**
- **There are many open questions related to this type of intensive multiple tests**

# *Questions*



# Coffee

