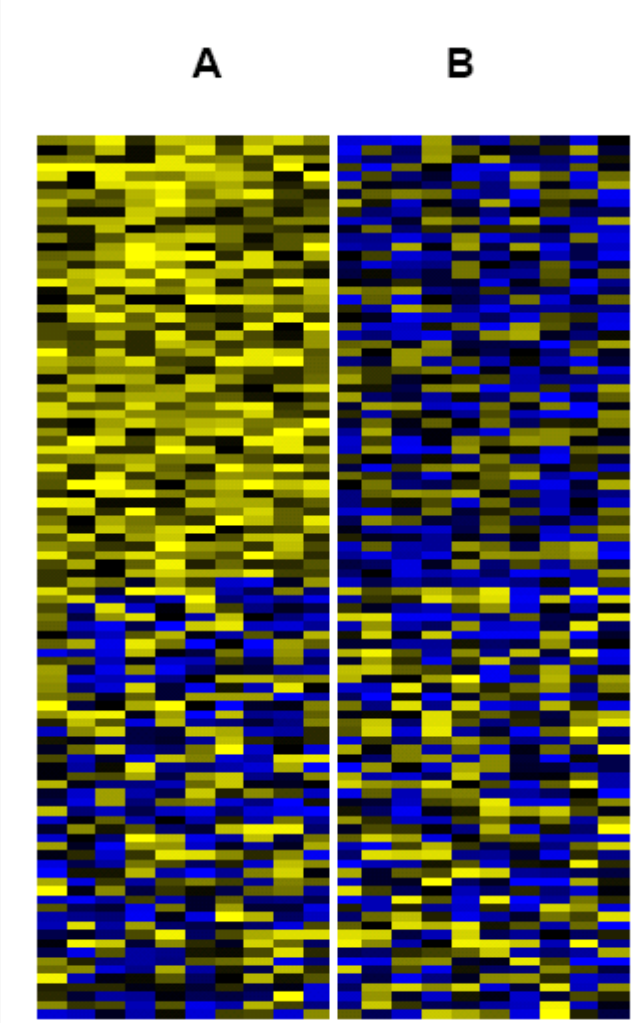
Differential Gene Expression

Rainer Spang

Courses in Practical DNA Microarray Analysis

NGFN
Nationales
Genomforschungsnetz



Two cell/tissue /disease types:

wild-type / mutant

control / treated

disease A / disease B

responding / non responding

etc. etc....

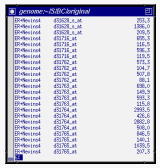
For every sample (cell line/patient) we have the expression levels of thousands of genes and the information whether it is A or B

Differential gene expression:

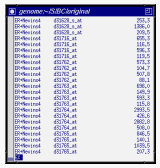
Which genes are differentially expressed in the two tissue type populations?

A cost efficient (cheap) experiment:

A



B



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

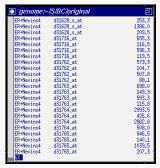
Is two-fold trust worthy?

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

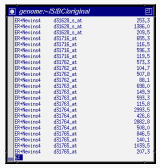
By no more than 10% than the answer is yes, by up to 500% then the answer is no.

A cost efficient (cheap) experiment II:

A



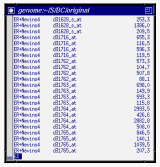
B



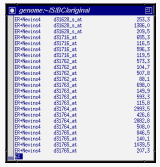
Is a three-fold induced gene more trust worthy than a two-fold induced gene?

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

A



B

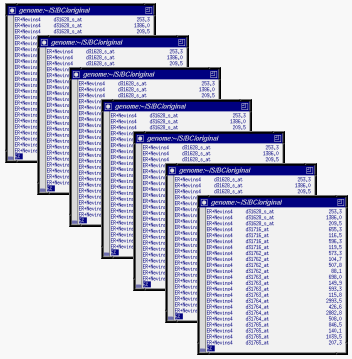


Conclusion: In addition to the differences in gene expression you also have a vital interest in its variability ... This information is needed to obtain meaningful lists of genes

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:

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

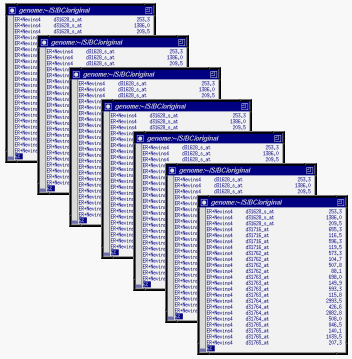
$$SE = \frac{1}{\sqrt{n}} SD$$

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

Ranking is not Testing

Ranking: Finding the right genes

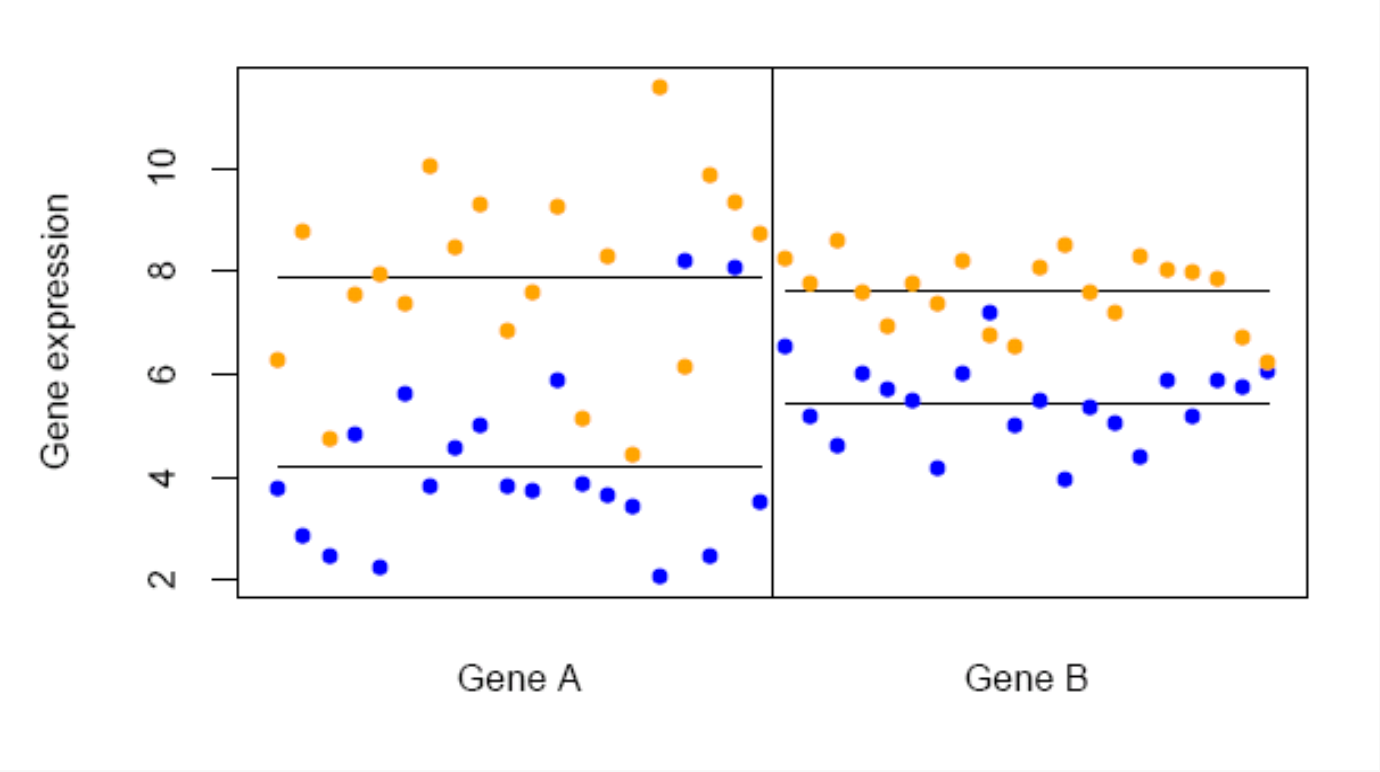
Testing: Deciding whether genes are significant

There is more then one way to rank

There is more then one way to test

The criteria for which ranking is best is different from the criteria which test is best … power is often no argument

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:

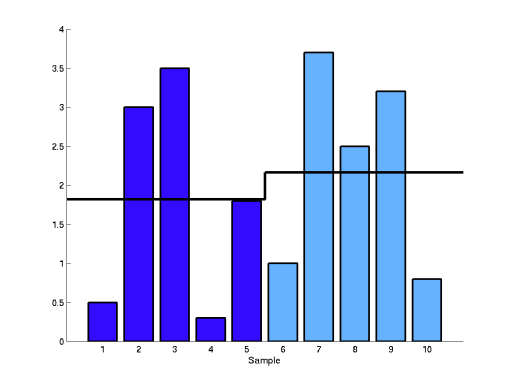
$$\log(a/b) = \log(a) - \log(b)$$

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

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

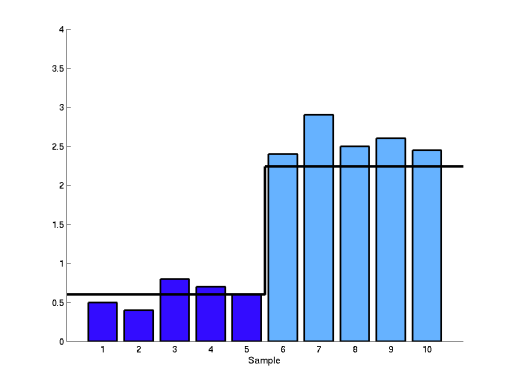
T-Score

Idea: Take variances into account



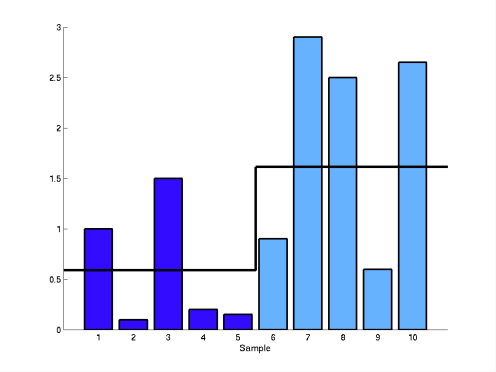
Change: low

Variance: high



Change: high

Variance: low



Change: high

Variance:high

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

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

Correction: Add a constant fudge factor s_0

→ Regularized T-score

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

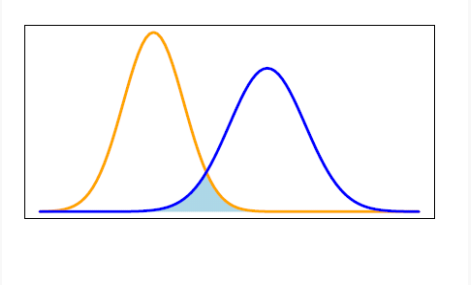
→Limma

→SAM

→Twighlight

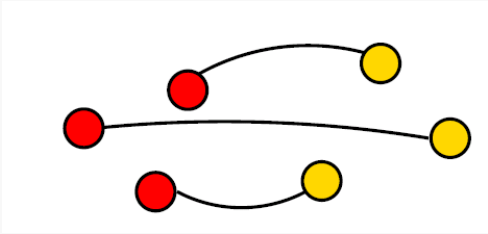
More Scores:

- Wilcoxon Score (robust)

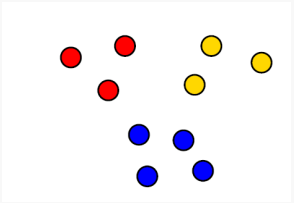


- PAUc Score (separation)

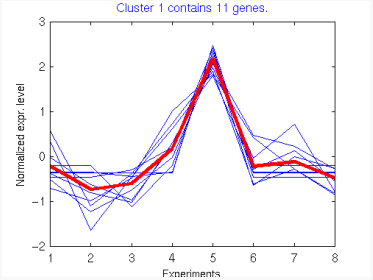
- paired t-Score (paired Data)



- F-Score (more then 2 conditions)



- Correlation to a reference gene



- etc etc

Different scores give different rankings

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

ALL vs AML (Golub et al.)

Which Score is the best one?

That depends on your problem ...

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?

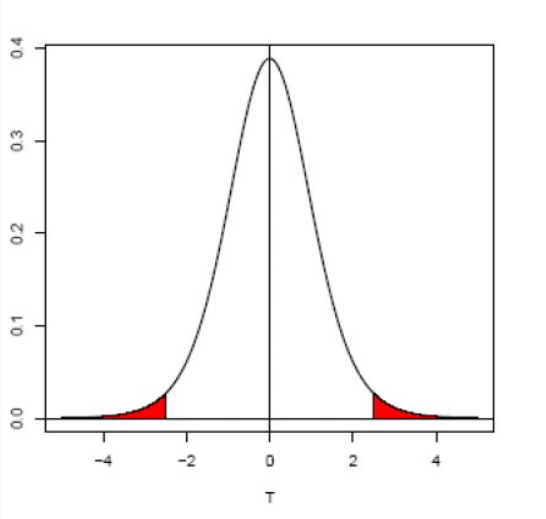
Assume: It is a random fluctuation

= 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)



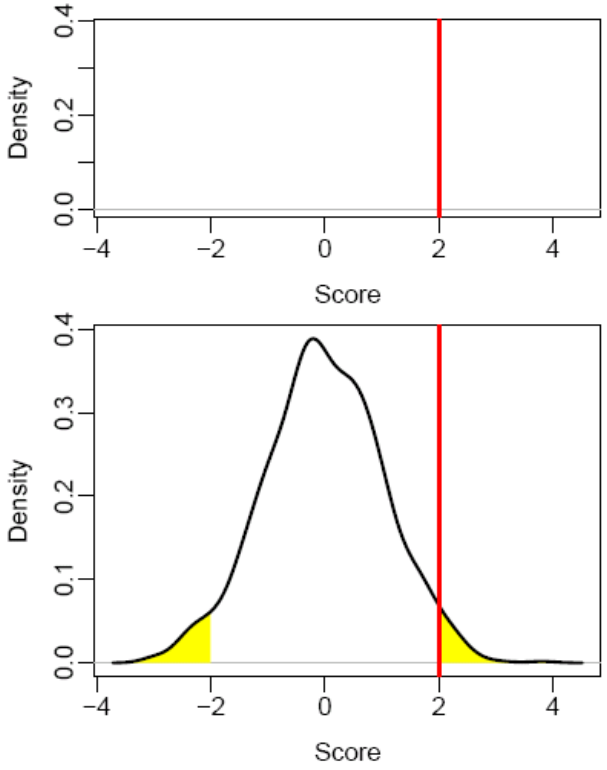
Permutations and empirical p-values

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:

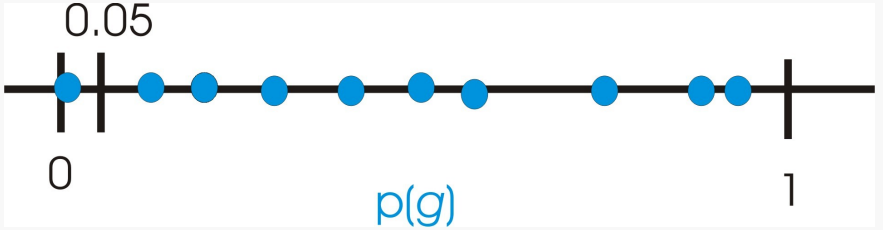
If the gene is not differentially expressed a small p-value is unlikely, hence we should be surprised by this observation.

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

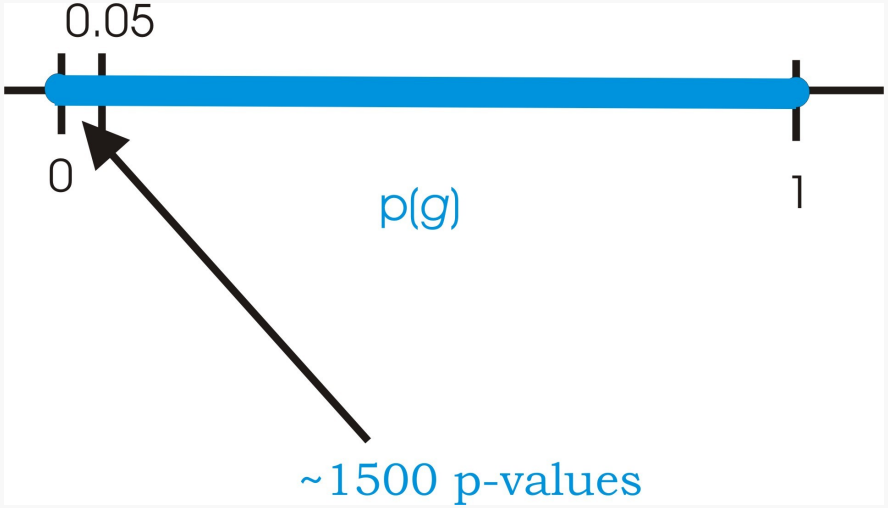
Multiple testing with only non-induced genes

1 gene

10 genes

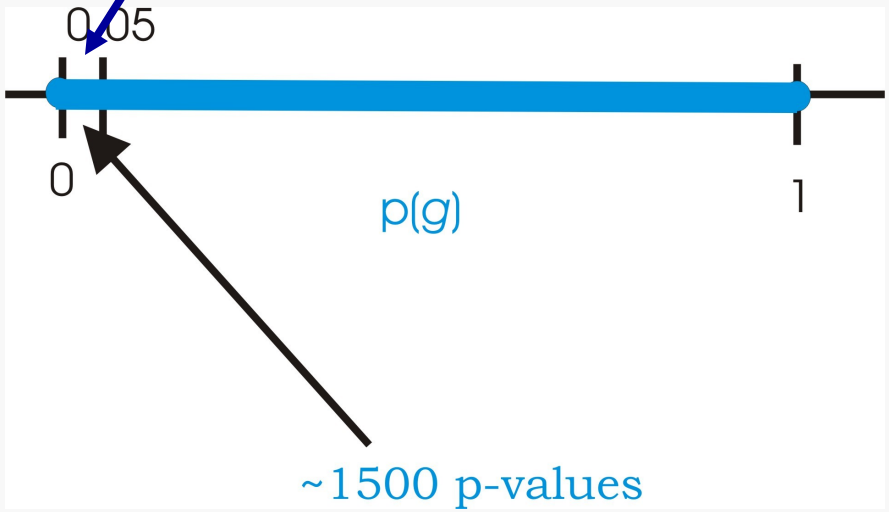


30,000 genes

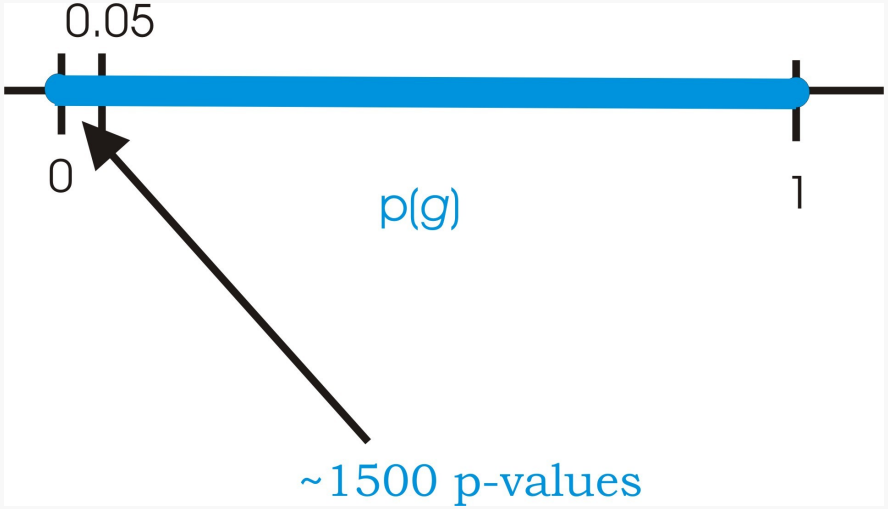


The Multiple Testing Problem

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.



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

1. Score genes and rank them
2. Choose a cutoff
3. 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

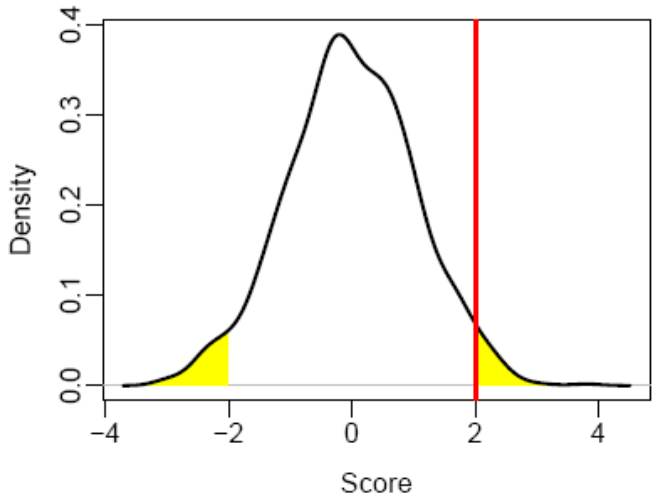
Conclusion: 4x change is significant

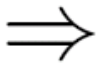
If a list of 150 genes with 4x change or more has a small estimated FDR this means that we have more genes on this level than would be expected by chance.

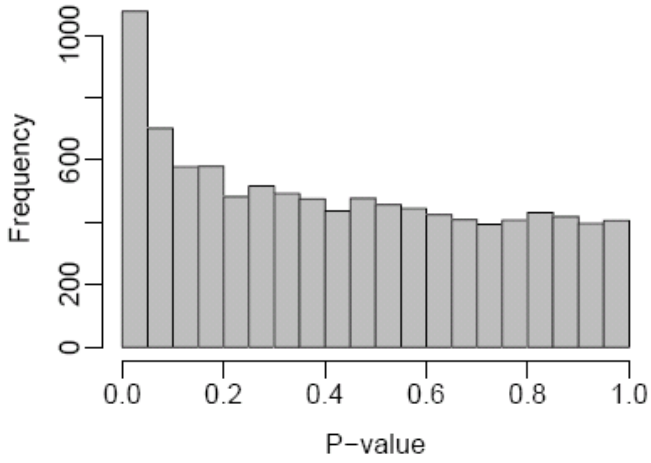
Conclusion: 4x change can be noise, but 150 genes on that level are too many to be explained just by random fluctuation.

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

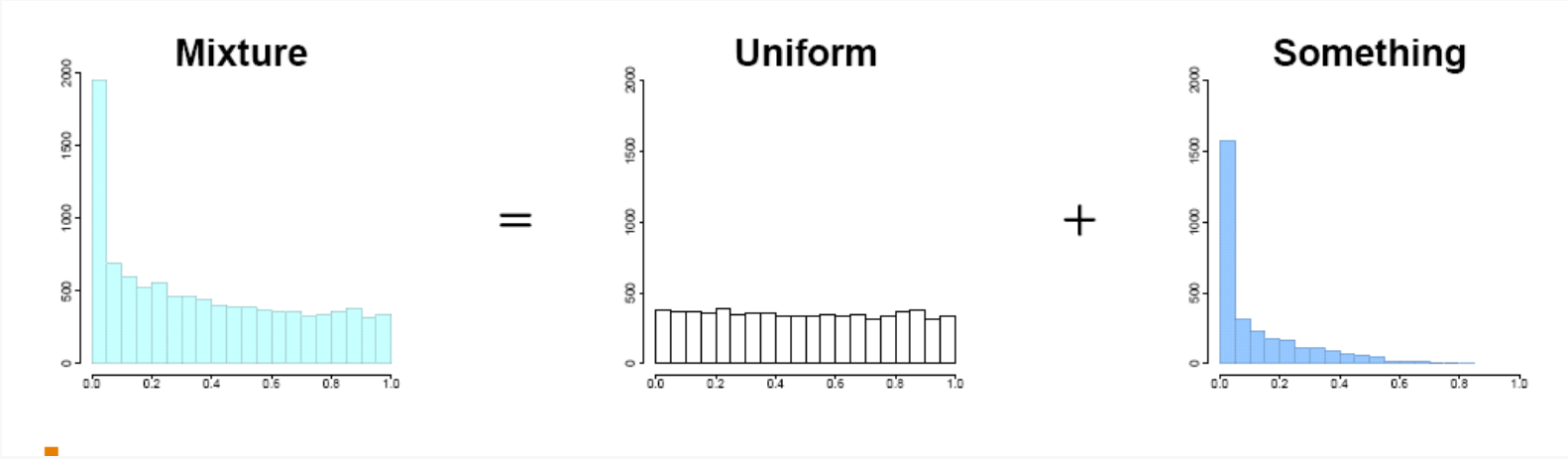
Histograms of the p-values of all genes on the array

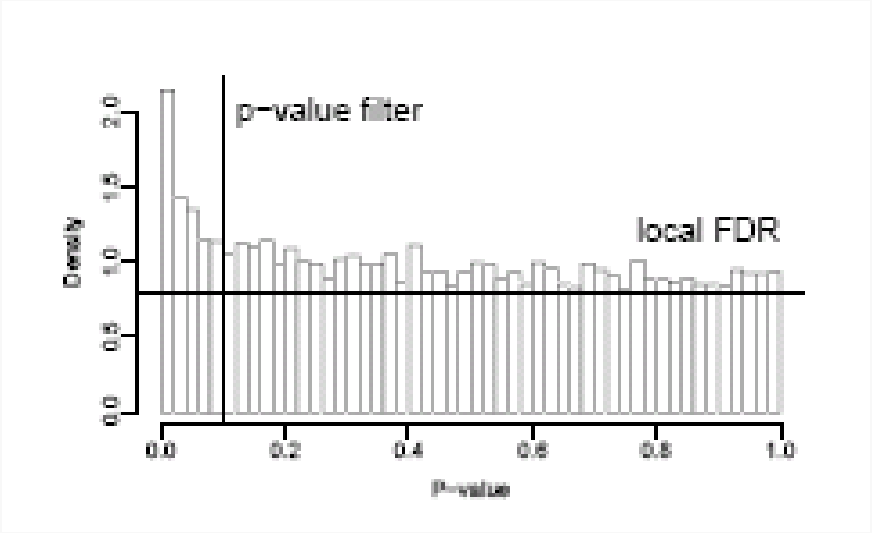






The mixture interpretation of the FDR

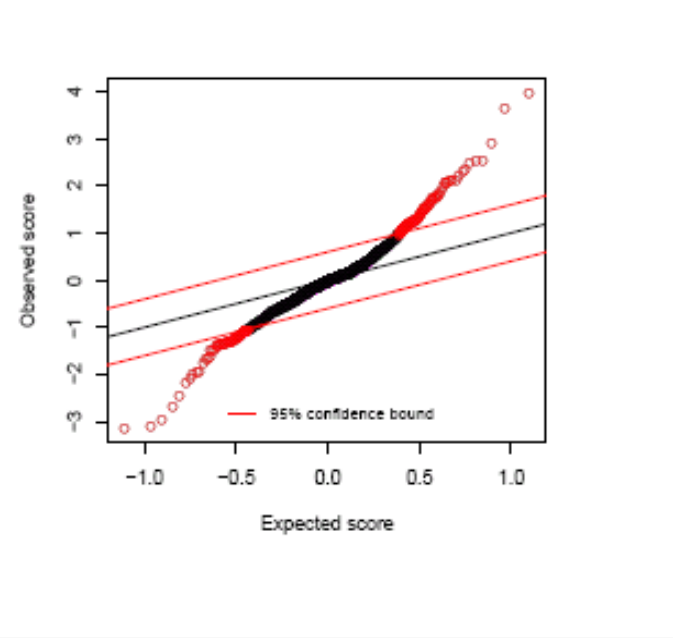




FWER: Vertical cutoff

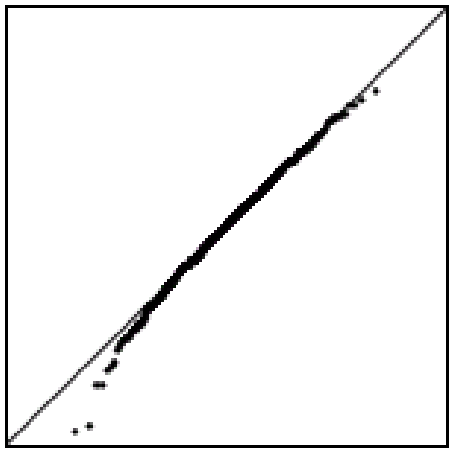
FDR: Horizontal cutoff

The typical plots

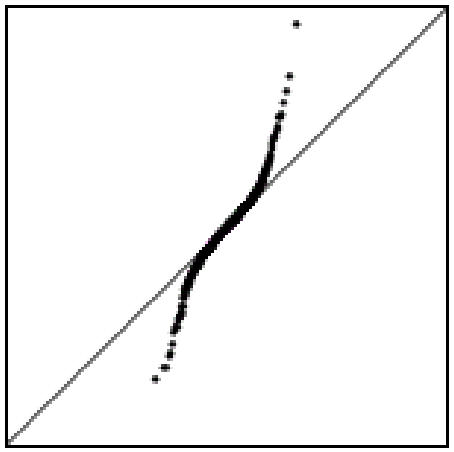


Expected random score vs observed scores: Deviations from the main diagonal are evidence for differentially expressed genes

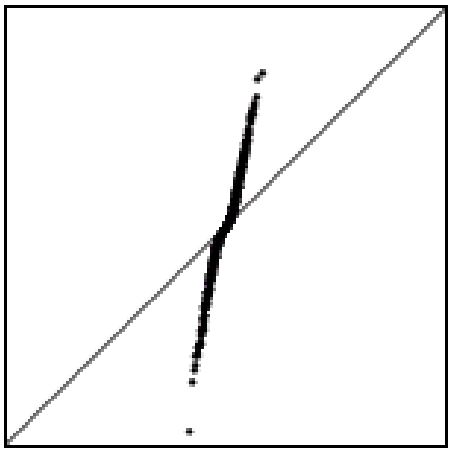
What you typically observe



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)
- 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