

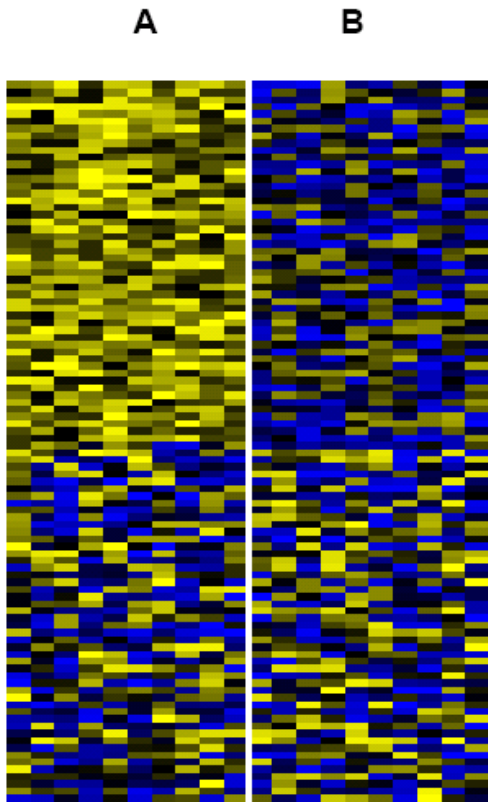
Screening

Methods Course: Gene Expression Data Analysis

-Day Three –

Rainer Spang

Comparing Conditions



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
U1-AS1	120.0
U1-AS2	120.0
U1-AS3	120.0
U1-AS4	120.0
U1-AS5	120.0
U1-AS6	120.0
U1-AS7	120.0
U1-AS8	120.0
U1-AS9	120.0
U1-AS10	120.0
U1-AS11	120.0
U1-AS12	120.0
U1-AS13	120.0
U1-AS14	120.0
U1-AS15	120.0
U1-AS16	120.0
U1-AS17	120.0
U1-AS18	120.0
U1-AS19	120.0
U1-AS20	120.0
U1-AS21	120.0
U1-AS22	120.0
U1-AS23	120.0
U1-AS24	120.0
U1-AS25	120.0
U1-AS26	120.0
U1-AS27	120.0
U1-AS28	120.0
U1-AS29	120.0
U1-AS30	120.0
U1-AS31	120.0
U1-AS32	120.0
U1-AS33	120.0
U1-AS34	120.0
U1-AS35	120.0
U1-AS36	120.0
U1-AS37	120.0
U1-AS38	120.0
U1-AS39	120.0
U1-AS40	120.0
U1-AS41	120.0
U1-AS42	120.0
U1-AS43	120.0
U1-AS44	120.0
U1-AS45	120.0
U1-AS46	120.0
U1-AS47	120.0
U1-AS48	120.0
U1-AS49	120.0
U1-AS50	120.0

B

Gene	Expression
U1-AS1	120.0
U1-AS2	120.0
U1-AS3	120.0
U1-AS4	120.0
U1-AS5	120.0
U1-AS6	120.0
U1-AS7	120.0
U1-AS8	120.0
U1-AS9	120.0
U1-AS10	120.0
U1-AS11	120.0
U1-AS12	120.0
U1-AS13	120.0
U1-AS14	120.0
U1-AS15	120.0
U1-AS16	120.0
U1-AS17	120.0
U1-AS18	120.0
U1-AS19	120.0
U1-AS20	120.0
U1-AS21	120.0
U1-AS22	120.0
U1-AS23	120.0
U1-AS24	120.0
U1-AS25	120.0
U1-AS26	120.0
U1-AS27	120.0
U1-AS28	120.0
U1-AS29	120.0
U1-AS30	120.0
U1-AS31	120.0
U1-AS32	120.0
U1-AS33	120.0
U1-AS34	120.0
U1-AS35	120.0
U1-AS36	120.0
U1-AS37	120.0
U1-AS38	120.0
U1-AS39	120.0
U1-AS40	120.0
U1-AS41	120.0
U1-AS42	120.0
U1-AS43	120.0
U1-AS44	120.0
U1-AS45	120.0
U1-AS46	120.0
U1-AS47	120.0
U1-AS48	120.0
U1-AS49	120.0
U1-AS50	120.0

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

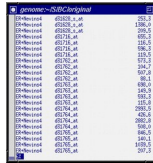
Is two-fold trust worthy?

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:

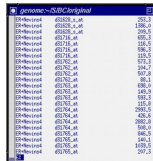
A



Gene	Expression Level
20250_s_at	120.0
20251_s_at	110.0
20252_s_at	200.0
20253_s_at	100.0
20254_s_at	100.0
20255_s_at	110.0
20256_s_at	100.0
20257_s_at	100.0
20258_s_at	100.0
20259_s_at	100.0
20260_s_at	100.0
20261_s_at	100.0
20262_s_at	100.0
20263_s_at	100.0
20264_s_at	100.0
20265_s_at	100.0
20266_s_at	100.0
20267_s_at	100.0
20268_s_at	100.0
20269_s_at	100.0
20270_s_at	100.0
20271_s_at	100.0
20272_s_at	100.0
20273_s_at	100.0
20274_s_at	100.0
20275_s_at	100.0
20276_s_at	100.0
20277_s_at	100.0
20278_s_at	100.0
20279_s_at	100.0
20280_s_at	100.0
20281_s_at	100.0
20282_s_at	100.0
20283_s_at	100.0
20284_s_at	100.0
20285_s_at	100.0
20286_s_at	100.0
20287_s_at	100.0
20288_s_at	100.0
20289_s_at	100.0
20290_s_at	100.0
20291_s_at	100.0
20292_s_at	100.0
20293_s_at	100.0
20294_s_at	100.0
20295_s_at	100.0
20296_s_at	100.0
20297_s_at	100.0
20298_s_at	100.0
20299_s_at	100.0
20300_s_at	100.0

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

B



Gene	Expression Level
20250_s_at	120.0
20251_s_at	110.0
20252_s_at	200.0
20253_s_at	100.0
20254_s_at	100.0
20255_s_at	110.0
20256_s_at	100.0
20257_s_at	100.0
20258_s_at	100.0
20259_s_at	100.0
20260_s_at	100.0
20261_s_at	100.0
20262_s_at	100.0
20263_s_at	100.0
20264_s_at	100.0
20265_s_at	100.0
20266_s_at	100.0
20267_s_at	100.0
20268_s_at	100.0
20269_s_at	100.0
20270_s_at	100.0
20271_s_at	100.0
20272_s_at	100.0
20273_s_at	100.0
20274_s_at	100.0
20275_s_at	100.0
20276_s_at	100.0
20277_s_at	100.0
20278_s_at	100.0
20279_s_at	100.0
20280_s_at	100.0
20281_s_at	100.0
20282_s_at	100.0
20283_s_at	100.0
20284_s_at	100.0
20285_s_at	100.0
20286_s_at	100.0
20287_s_at	100.0
20288_s_at	100.0
20289_s_at	100.0
20290_s_at	100.0
20291_s_at	100.0
20292_s_at	100.0
20293_s_at	100.0
20294_s_at	100.0
20295_s_at	100.0
20296_s_at	100.0
20297_s_at	100.0
20298_s_at	100.0
20299_s_at	100.0
20300_s_at	100.0

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

The information in the variability is crucial

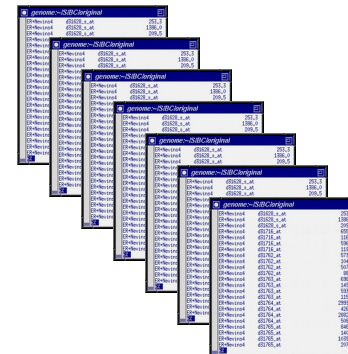
In addition to the differences in gene expression you also have a vital interest in its variability ... This information is needed to obtain meaningful lists of genes

Therefore: Invest 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:

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

Precision by Repetition

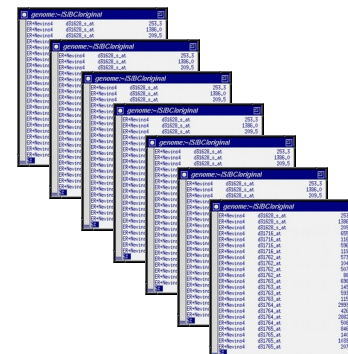
Repetitions lead to a more precise measurement of gene expression. Single expression measurements are very noisy, average expression across several repetitions is much less noisy

Therefore: Invest in repeated experiments!

A



B



The Additive Scale

Most statistics works on an additive scale

Biology works on a multiplicative scale

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*

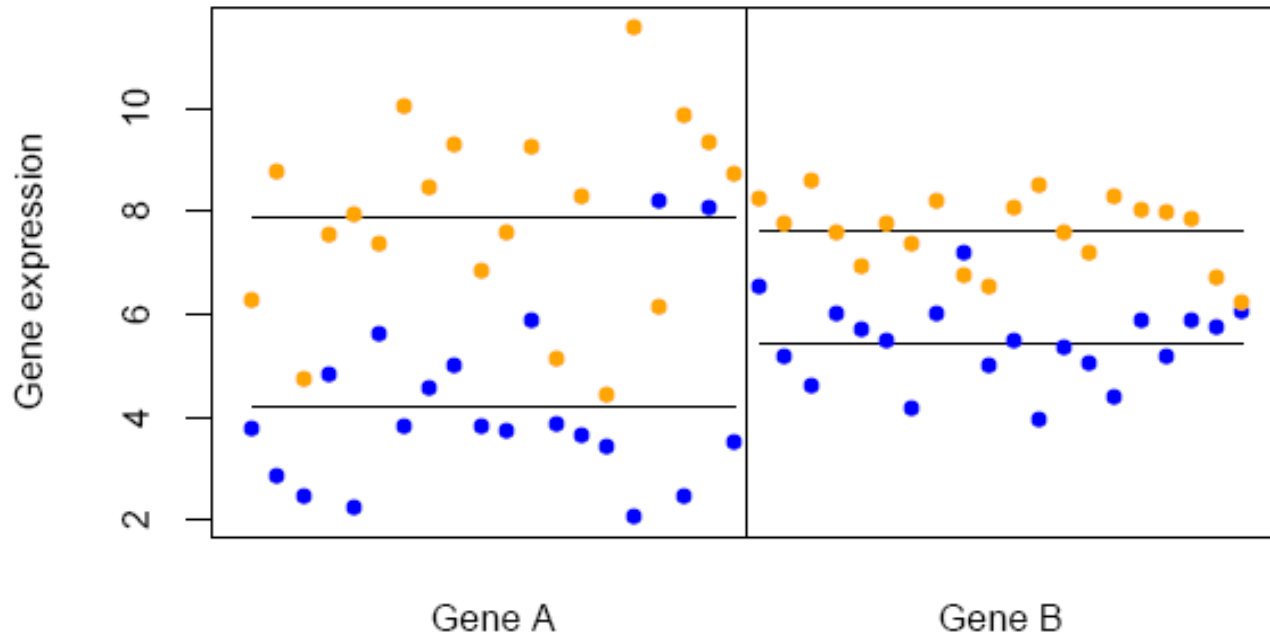
Ranking is not Testing

Ranking: Finding the right genes

Testing: Deciding whether genes are significant

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**

Which scores are there?

Fold Change & Log Ratios

You have transformed your data to additive scale!

Factors become differences:

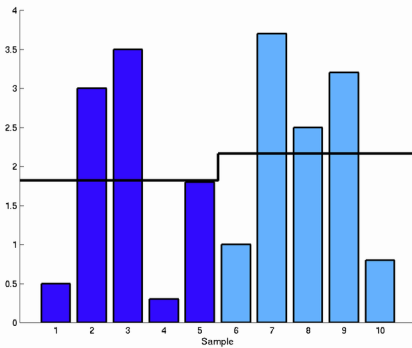
$$\log(a/b) = \log(a) - \log(b)$$

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

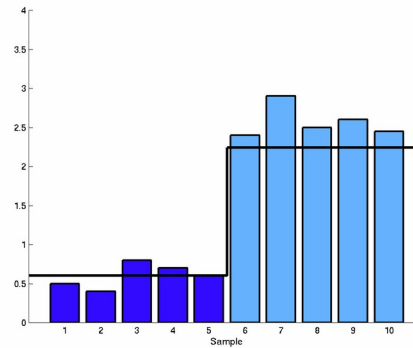
$$LR = \bar{X}_1 - \bar{X}_2$$

T-Score

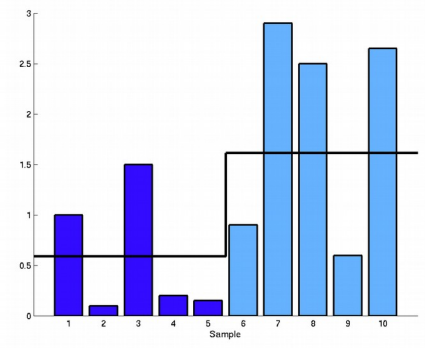
Idea: Take variances into account



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 an already small variance

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)

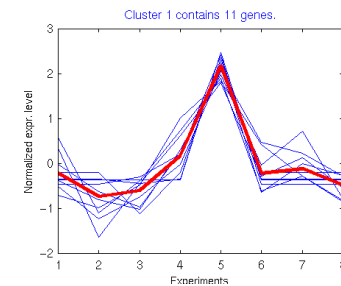
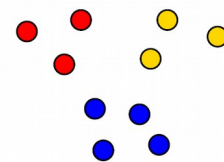
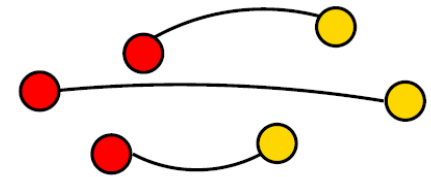
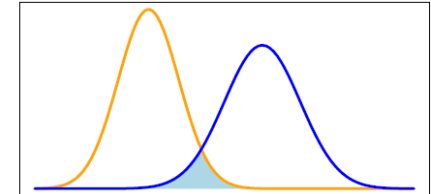
PAUc Score (separation)

paired t-Score (paired Data)

F-Score (more than 2 conditions)

Correlation to a reference gene

etc etc



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

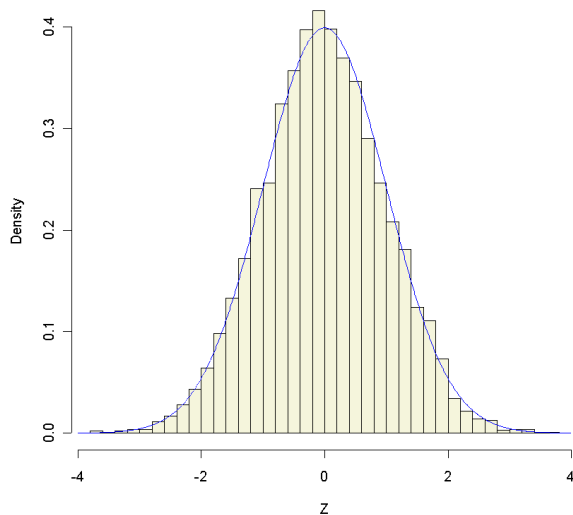
ALL vs AML (Golub et al.)

Which Score is the best one?

That depends on your
problem ...

*Measurement noise of expression differences is Gaussian for **all** genes ...*

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



Measurements are Gaussian

The average of Gaussians is Gaussian

The difference of Gaussians is Gaussian

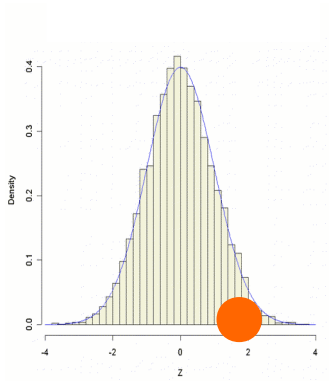
Some fold changes are over estimated and some are underestimated

***... but this changes after
sorting the fold changes !***

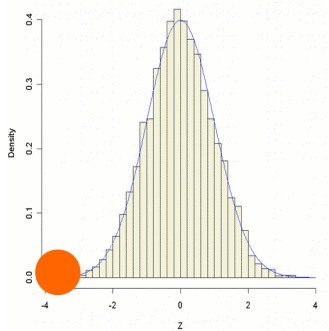
Gene 1	2.10 fold
Gene 2	2.08 fold
Gene 3	1.37 fold
Gene 4	5.91 fold
Gene 5	0.92 fold
Gene 6	2.85 fold

Rank 1	5.91 fold
Rank 2	2.85 fold
Rank 3	2.10 fold
Rank 4	2.08 fold
Rank 6	1.37 fold
Rank 7	0.92 fold

Estimation Errors

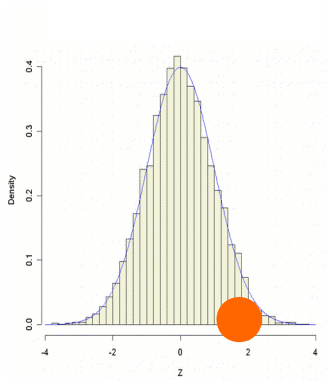


Genes, for which we **overestimate** the fold change ... **move up** in the ranking

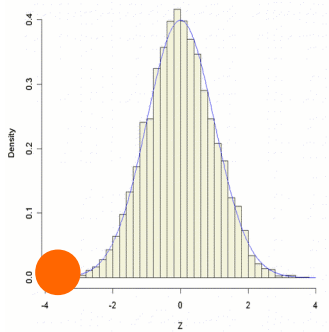


Genes, for which we **underestimate** the fold change ... **go down** in the ranking

Vice Versa

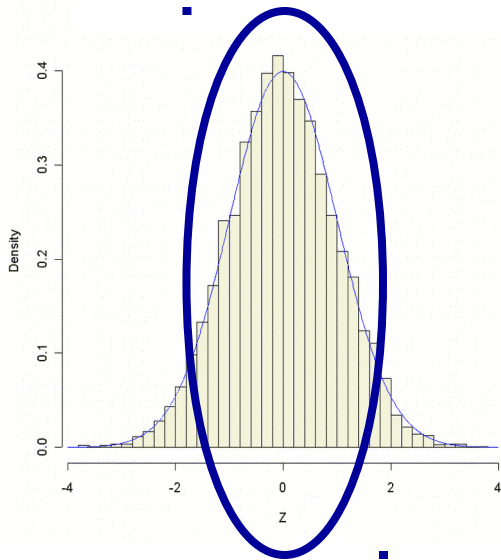


Genes high up in the ranking have most likely overestimated fold changes

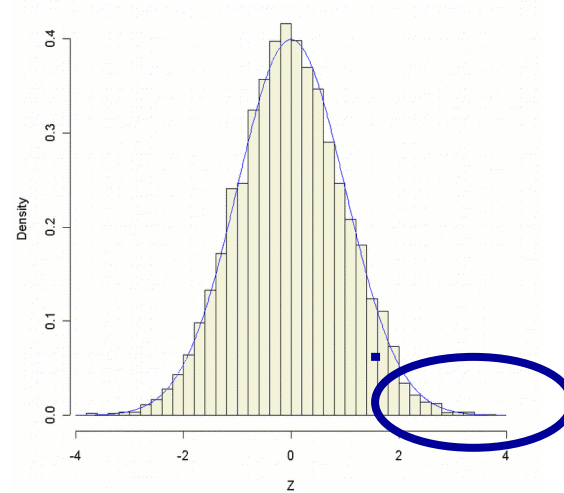


Genes far down in the ranking have most likely underestimated fold changes

The noise in rank 1



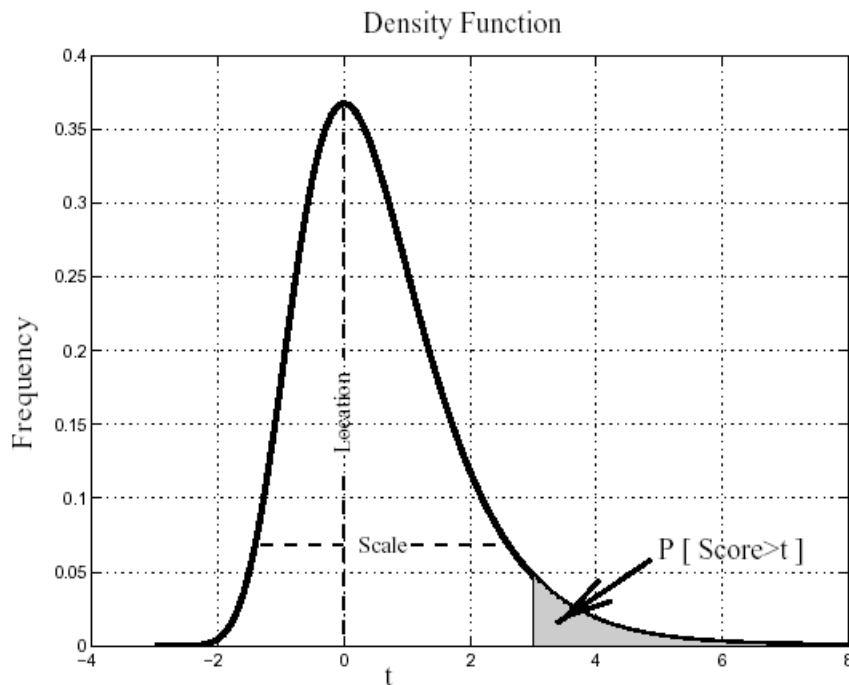
The noise for a randomly selected gene is centered around zero



The noise for the top ranking gene is centered around a positive offset

Extreme Value Distribution

$$\phi(t) = \theta^{-1} e^{-(t-\xi)/\theta} \exp\left(-e^{-\frac{t-\xi}{\theta}}\right)$$



The noise distribution is not only shifted to the right, it also changes its shape from a Gaussian to a Extreme Value Distribution

Outliers are much more frequent for this type of distribution

Screening Noise

Rank 1	5.91 fold
Rank 2	2.85 fold
Rank 3	2.10 fold
Rank 4	2.08 fold
Rank 6	1.37 fold
Rank 7	0.92 fold

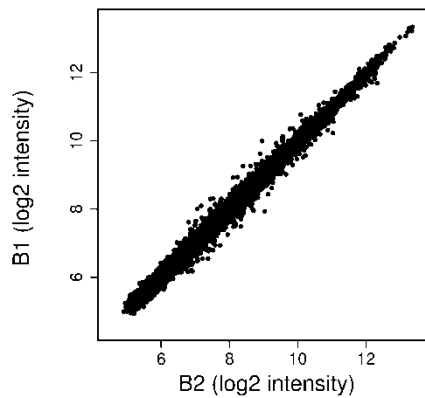
Screening for differentially expressed genes:

Increases noise

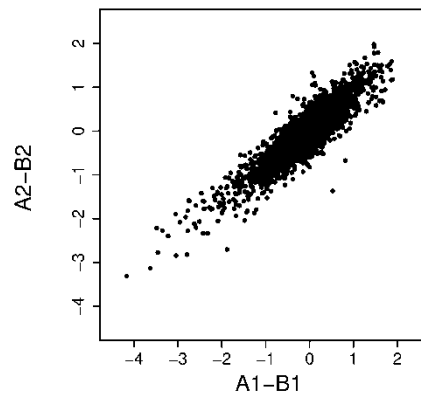
Yields biased fold changes

Increase the number of noise related outliers

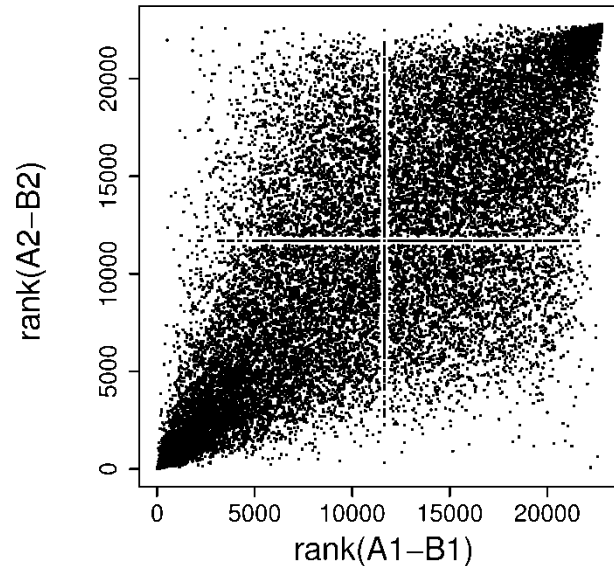
Reproducibility of Rankings



This reproducibility of absolute values translates to ...



... this reproducibility of expression differences, which translates to ...



... this reproducibility of the ranking of genes

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

Concept p-values:

We observe a score $s=1.27$

Can this be just a random fluctuation?

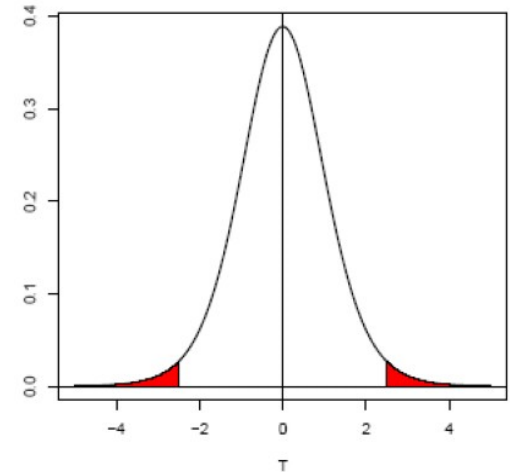
Assume: It is a random fluctuation

= 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

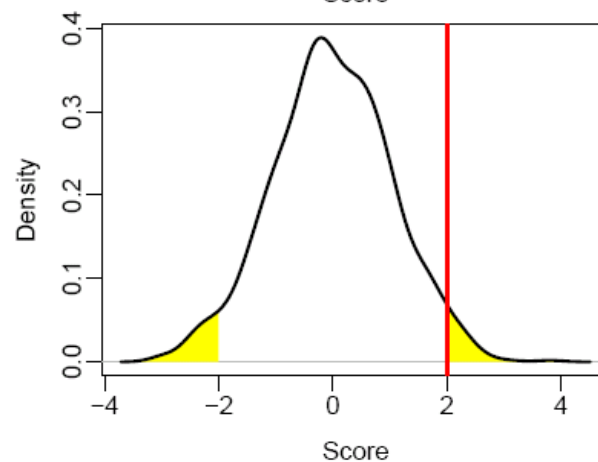
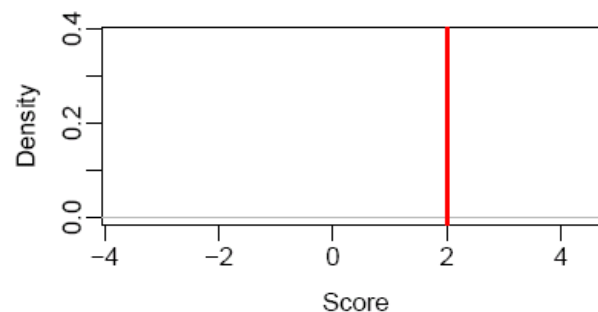
0	0	0	0	0	1	1	1	1	1
---	---	---	---	---	---	---	---	---	---

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
---	---	---	---	---	---	---	---	---	---



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!

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 no theoretical support for this

Controlling the error of the first kind

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

Nevertheless it can still happen and we call this event an **error of the first kind**

Why Most Published Research Findings Are False

John P. A. Ioannidis

PLoS Medicine 2005

Two ways to get 5 Nature publications ?

1. Good science and a little bit of luck
- 2 Fantasy and the error of the first kind

Make up 100 scientific hypotheses all of which should be incorrect but spectacular enough such that Nature would publish them, if one produced significant data to back them up

→ You can expect that about 5 projects will produce p -values < 0.05
Submit those to Nature

Limits of Statistical Significance

Significance tests cannot control the percentage of false published results

The proportion of false claims tested is driving this percentage

Choosing claims to test is not statistics but scientific practice

Corollary 5: The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.

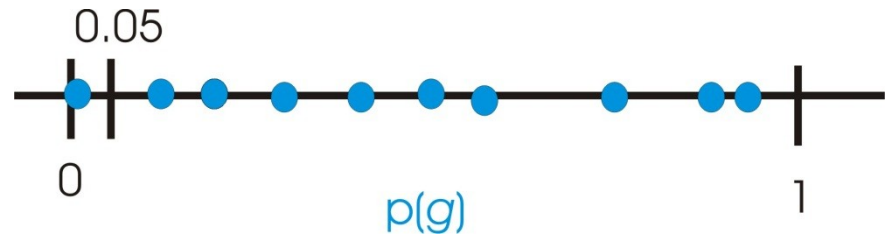
Corollary 6: The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

Multiple testing with only non-induced genes

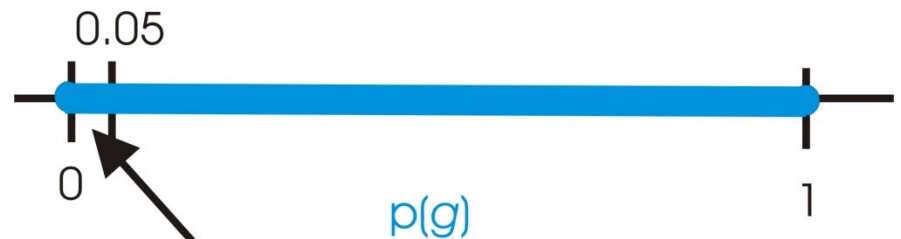
1 gene



10 genes



30,000 genes

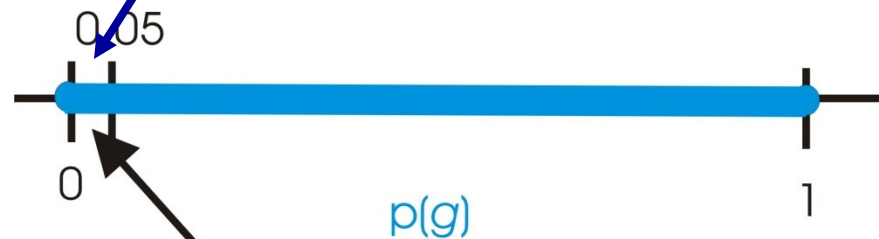


~1500 p-values

The Multiple Testing Problem

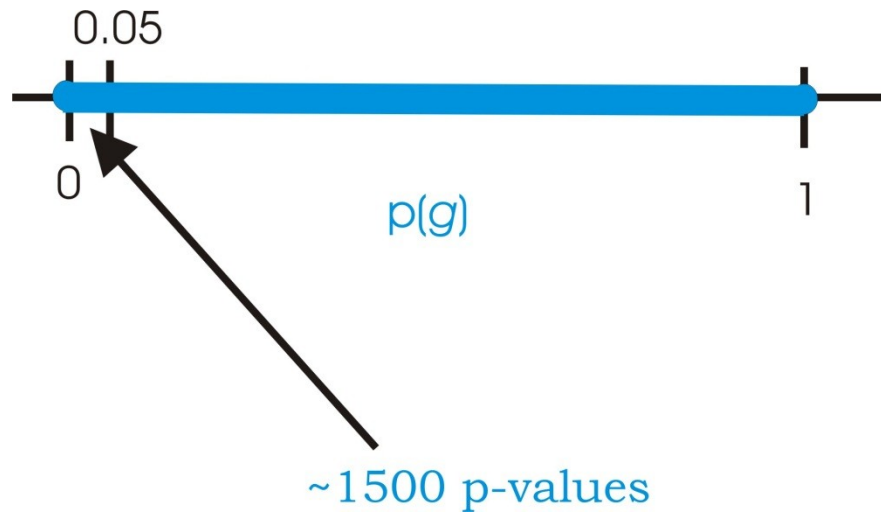


P-values are random numbers between 0 and 1. For only one such number it is unlikely to fall in this small interval, but if we have 30.000 such numbers many will be in there.



~1500 p-values

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.

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

No good idea

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

FDR vs. 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 a random fluctuation

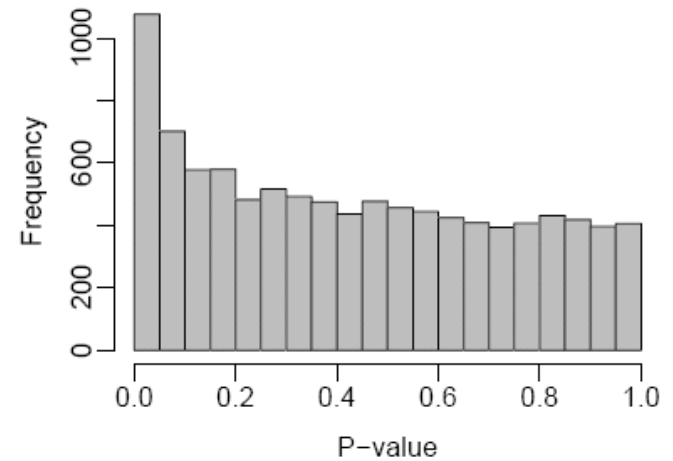
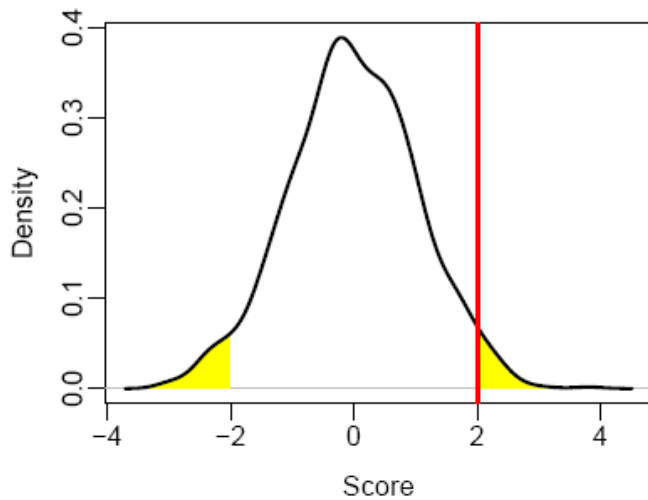
Conclusion: 4x change is significant

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.

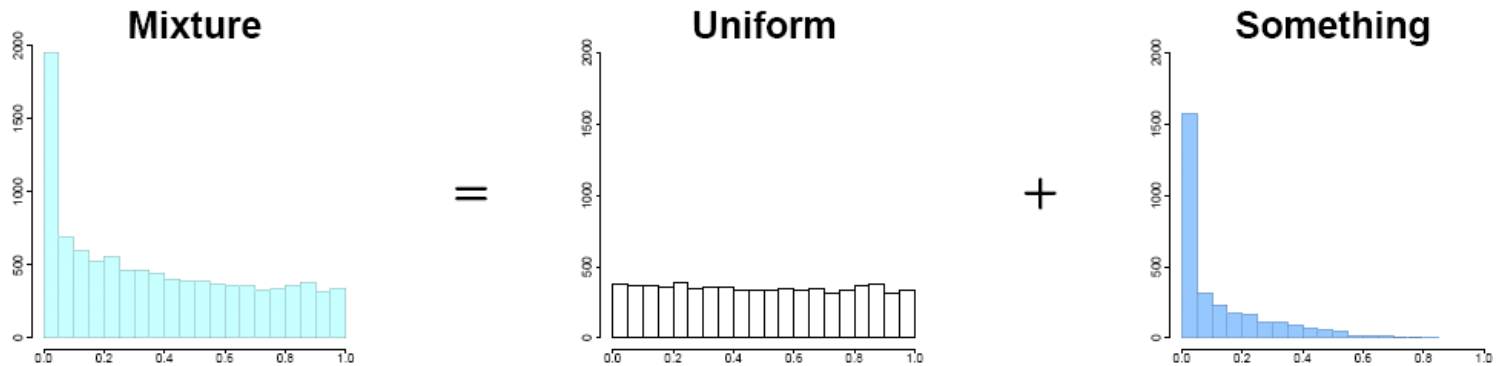
Conclusion: 4x change can be noise, but 150 genes on that level are too many to be explained just by random fluctuation.

In **p-value** 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

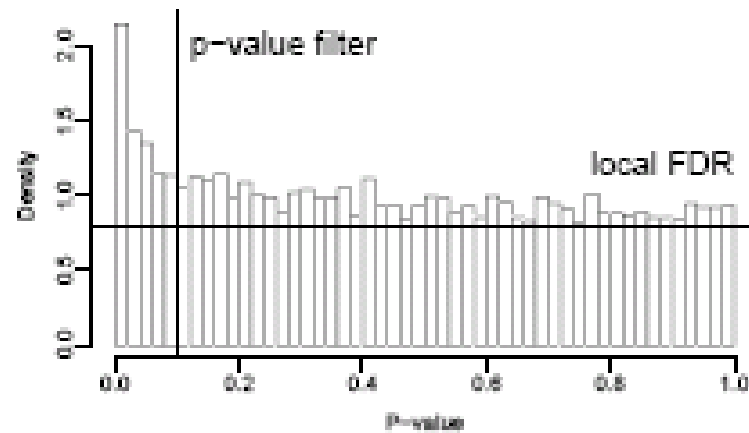


The mixture interpretation of the FDR



■

Horizontal vs. Vertical cutoffs

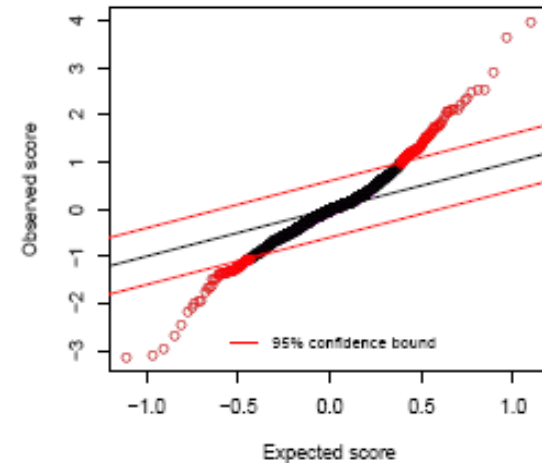


FWER: Vertical cutoff

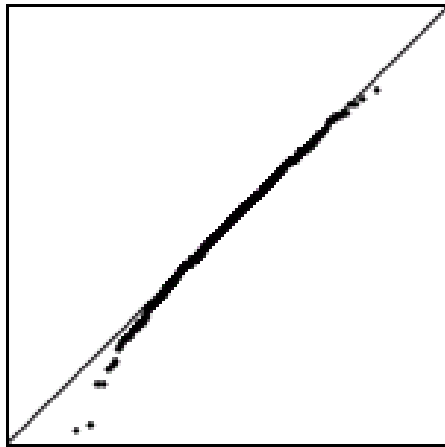
FDR: Horizontal cutoff

The typical plots

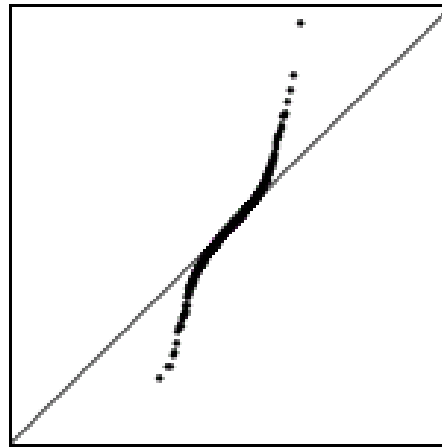
Expected random score vs observed scores: Deviations from the main diagonal are evidence for differentially expressed genes



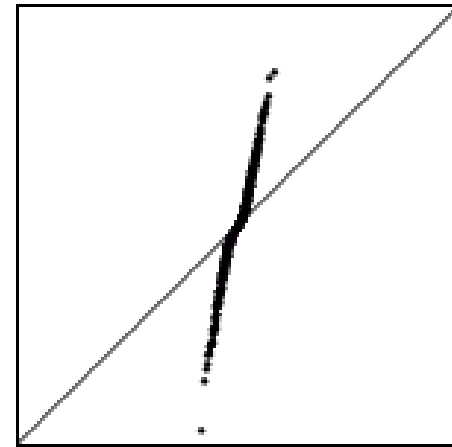
What you typically observe



**No differential
gene
expression**



**A lot of
differential
gene
expression**



**Global
changes in
gene
expression**

Finding the needle in the haystack

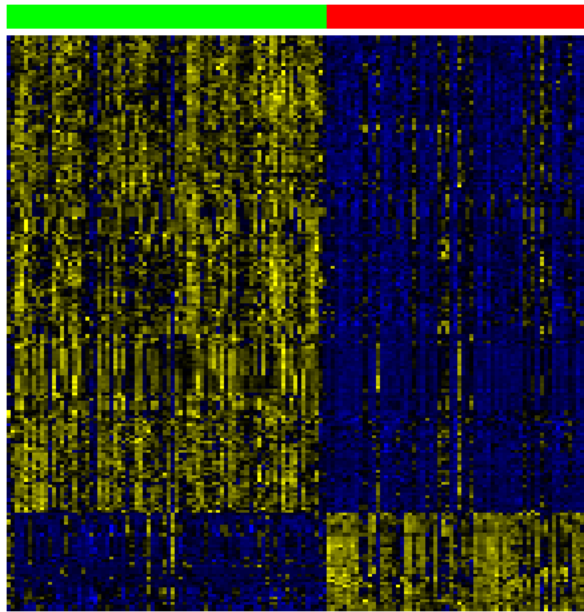
A common myth:

There are only a couple of genes that are truly different



The Avalanche

Aggressive lymphomas with and without a MYC-breakpoint



MYC-neg MYC-pos



Verbundprojekt maligne Lymphome

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

