Detection of regulated genes in microarray data

Vincent Detours
IRIBHM – Université Libre de Bruxelles

vdetours@ulb.ac.be

A typical experiment: the data

The samples

- 12 PTC samples
- 12 patient-matched adjacent healthy tissues

The platform & processing

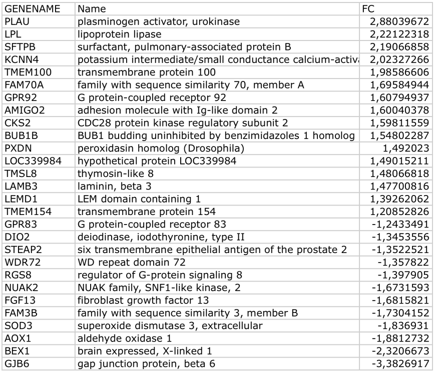
- Affymetrix® chips, 20,000 genes
- MAS 5.0 normalization

A typical experiment: gene selection

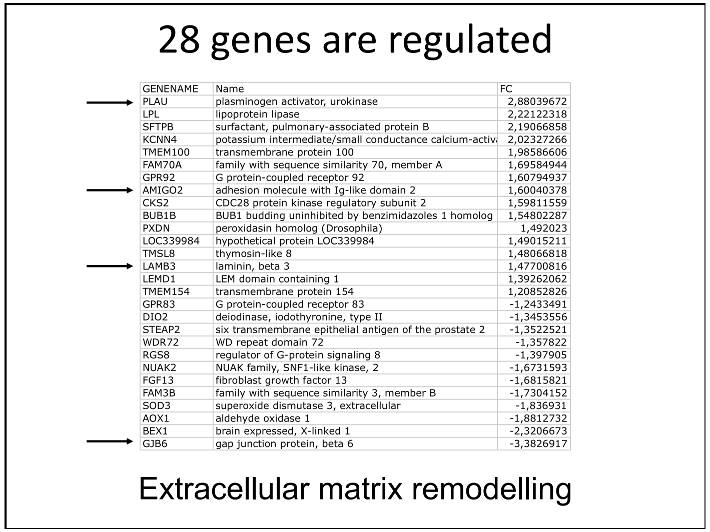
Up-regulated genes are selected as follows

1. Compute expression ratio of tumor vs. healthy tissue for each patient
2. Select genes with fold-change >2 in at least 8/12=2/3 of the patients

28 genes are regulated



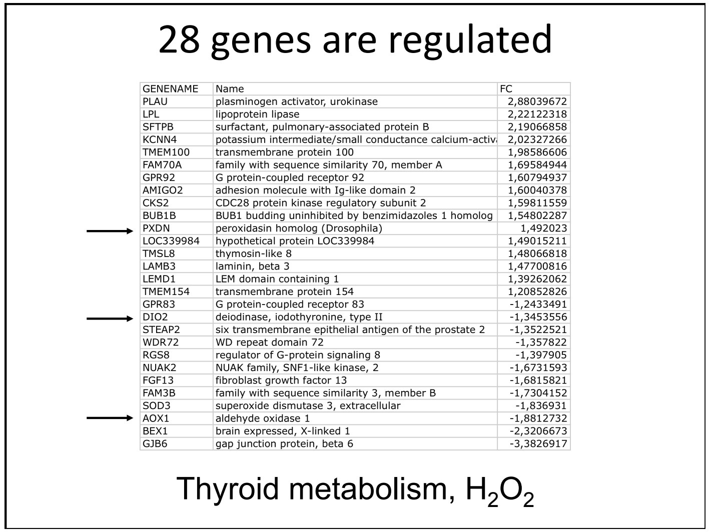
GENENAME	Name	FC
PLAU	plasminogen activator, urokinase	2,88039672
LPL	lipoprotein lipase	2,22122318
SFTPB	surfactant, pulmonary-associated protein B	2,19066858
KCNN4	potassium intermediate/small conductance calcium-activ	2,02327266
TMEM100	transmembrane protein 100	1,98586606
FAM70A	family with sequence similarity 70, member A	1,69584944
GPR92	G protein-coupled receptor 92	1,60794937
AMIGO2	adhesion molecule with Ig-like domain 2	1,60040378
CKS2	CDC28 protein kinase regulatory subunit 2	1,59811559
BUB1B	BUB1 budding uninhibited by benzimidazoles 1 homolog	1,54802287
PXDN	peroxidasin homolog (Drosophila)	1,492023
LOC339984	hypothetical protein LOC339984	1,49015211
TMSL8	thymosin-like 8	1,48066818
LAMB3	laminin, beta 3	1,47700816
LEMD1	LEM domain containing 1	1,39262062
TMEM154	transmembrane protein 154	1,20852826
GPR83	G protein-coupled receptor 83	-1,2433491
DIO2	deiodinase, iodothyronine, type II	-1,3453556
STEAP2	six transmembrane epithelial antigen of the prostate 2	-1,3522521
WDR72	WD repeat domain 72	-1,357822
RGS8	regulator of G-protein signaling 8	-1,397905
NUAK2	NUAK family, SNF1-like kinase, 2	-1,6731593
FGF13	fibroblast growth factor 13	-1,6815821
FAM3B	family with sequence similarity 3, member B	-1,7304152
SOD3	superoxide dismutase 3, extracellular	-1,836931
AOX1	aldehyde oxidase 1	-1,8812732
BEX1	brain expressed, X-linked 1	-2,3206673
GJB6	gap junction protein, beta 6	-3,3826917



28 genes are regulated

	GENENAME	Name	FC
→	PLAU	plasminogen activator, urokinase	2,88039672
	LPL	lipoprotein lipase	2,22122318
	SFTPB	surfactant, pulmonary-associated protein B	2,19066858
	KCNN4	potassium intermediate/small conductance calcium-activa	2,02327266
	TMEM100	transmembrane protein 100	1,98586606
	FAM70A	family with sequence similarity 70, member A	1,69584944
	GPR92	G protein-coupled receptor 92	1,60794937
→	AMIGO2	adhesion molecule with Ig-like domain 2	1,60040378
	CKS2	CDC28 protein kinase regulatory subunit 2	1,59811559
	BUB1B	BUB1 budding uninhibited by benzimidazoles 1 homolog	1,54802287
	PXDN	peroxidasin homolog (Drosophila)	1,492023
	LOC339984	hypothetical protein LOC339984	1,49015211
	TMSL8	thymosin-like 8	1,48066818
→	LAMB3	laminin, beta 3	1,47700816
	LEMD1	LEM domain containing 1	1,39262062
	TMEM154	transmembrane protein 154	1,20852826
	GPR83	G protein-coupled receptor 83	-1,2433491
	DIO2	deiodinase, iodothyronine, type II	-1,3453556
	STEAP2	six transmembrane epithelial antigen of the prostate 2	-1,3522521
	WDR72	WD repeat domain 72	-1,357822
	RGS8	regulator of G-protein signaling 8	-1,397905
	NUAK2	NUAK family, SNF1-like kinase, 2	-1,6731593
	FGF13	fibroblast growth factor 13	-1,6815821
	FAM3B	family with sequence similarity 3, member B	-1,7304152
	SOD3	superoxide dismutase 3, extracellular	-1,836931
	AOX1	aldehyde oxidase 1	-1,8812732
	BEX1	brain expressed, X-linked 1	-2,3206673
→	GJB6	gap junction protein, beta 6	-3,3826917

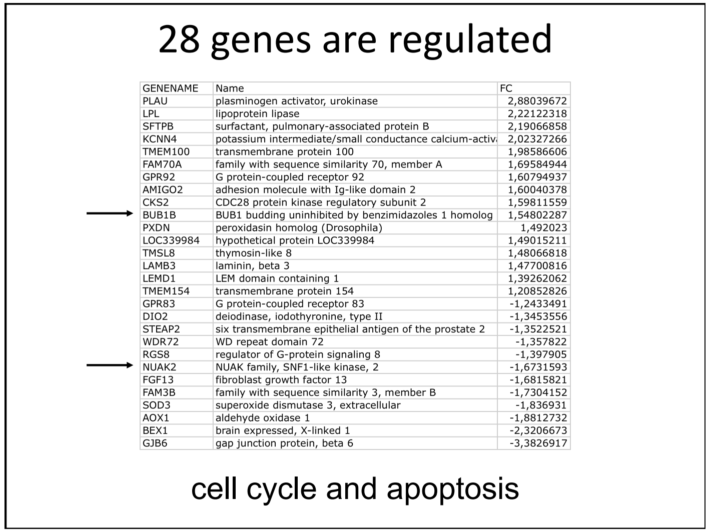
Extracellular matrix remodelling



28 genes are regulated

	GENENAME	Name	FC
	PLAU	plasminogen activator, urokinase	2,88039672
	LPL	lipoprotein lipase	2,22122318
	SFTPB	surfactant, pulmonary-associated protein B	2,19066858
	KCNN4	potassium intermediate/small conductance calcium-activa	2,02327266
	TMEM100	transmembrane protein 100	1,98586606
	FAM70A	family with sequence similarity 70, member A	1,69584944
	GPR92	G protein-coupled receptor 92	1,60794937
	AMIGO2	adhesion molecule with Ig-like domain 2	1,60040378
	CKS2	CDC28 protein kinase regulatory subunit 2	1,59811559
	BUB1B	BUB1 budding uninhibited by benzimidazoles 1 homolog	1,54802287
→	PXDN	peroxidasin homolog (Drosophila)	1,492023
	LOC339984	hypothetical protein LOC339984	1,49015211
	TMSL8	thymosin-like 8	1,48066818
	LAMB3	laminin, beta 3	1,47700816
	LEMD1	LEM domain containing 1	1,39262062
	TMEM154	transmembrane protein 154	1,20852826
	GPR83	G protein-coupled receptor 83	-1,2433491
→	DIO2	deiodinase, iodothyronine, type II	-1,3453556
	STEAP2	six transmembrane epithelial antigen of the prostate 2	-1,3522521
	WDR72	WD repeat domain 72	-1,357822
	RGS8	regulator of G-protein signaling 8	-1,397905
	NUAK2	NUAK family, SNF1-like kinase, 2	-1,6731593
	FGF13	fibroblast growth factor 13	-1,6815821
	FAM3B	family with sequence similarity 3, member B	-1,7304152
	SOD3	superoxide dismutase 3, extracellular	-1,836931
→	AOX1	aldehyde oxidase 1	-1,8812732
	BEX1	brain expressed, X-linked 1	-2,3206673
	GJB6	gap junction protein, beta 6	-3,3826917

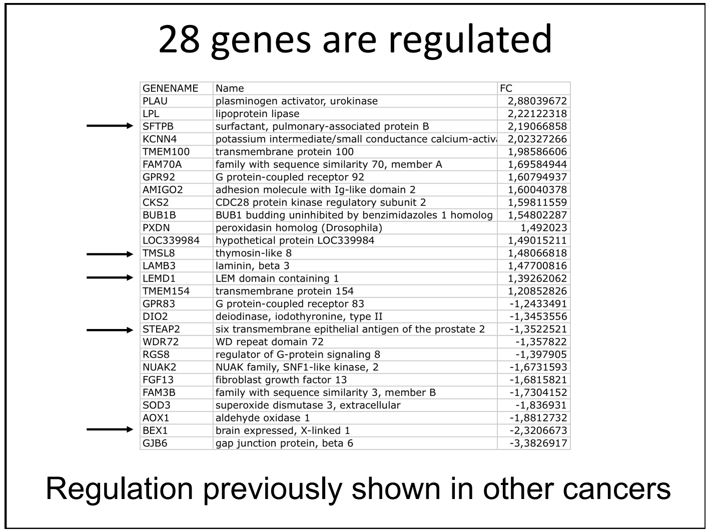
Thyroid metabolism, H_2O_2



28 genes are regulated

	GENENAME	Name	FC
	PLAU	plasminogen activator, urokinase	2,88039672
	LPL	lipoprotein lipase	2,22122318
	SFTPB	surfactant, pulmonary-associated protein B	2,19066858
	KCNN4	potassium intermediate/small conductance calcium-activa	2,02327266
	TMEM100	transmembrane protein 100	1,98586606
	FAM70A	family with sequence similarity 70, member A	1,69584944
	GPR92	G protein-coupled receptor 92	1,60794937
	AMIGO2	adhesion molecule with Ig-like domain 2	1,60040378
	CKS2	CDC28 protein kinase regulatory subunit 2	1,59811559
→	BUB1B	BUB1 budding uninhibited by benzimidazoles 1 homolog	1,54802287
	PXDN	peroxidasin homolog (Drosophila)	1,492023
	LOC339984	hypothetical protein LOC339984	1,49015211
	TMSL8	thymosin-like 8	1,48066818
	LAMB3	laminin, beta 3	1,47700816
	LEMD1	LEM domain containing 1	1,39262062
	TMEM154	transmembrane protein 154	1,20852826
	GPR83	G protein-coupled receptor 83	-1,2433491
	DIO2	deiodinase, iodothyronine, type II	-1,3453556
	STEAP2	six transmembrane epithelial antigen of the prostate 2	-1,3522521
	WDR72	WD repeat domain 72	-1,357822
	RGS8	regulator of G-protein signaling 8	-1,397905
→	NUAK2	NUAK family, SNF1-like kinase, 2	-1,6731593
	FGF13	fibroblast growth factor 13	-1,6815821
	FAM3B	family with sequence similarity 3, member B	-1,7304152
	SOD3	superoxide dismutase 3, extracellular	-1,836931
	AOX1	aldehyde oxidase 1	-1,8812732
	BEX1	brain expressed, X-linked 1	-2,3206673
	GJB6	gap junction protein, beta 6	-3,3826917

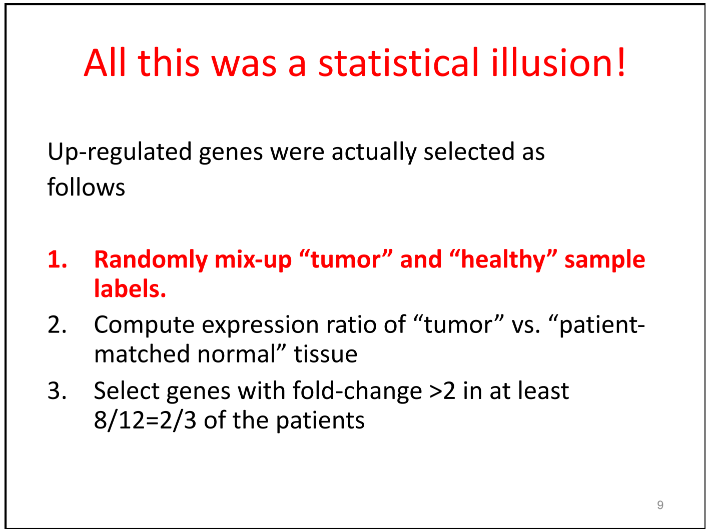
cell cycle and apoptosis



28 genes are regulated

	GENENAME	Name	FC
	PLAU	plasminogen activator, urokinase	2,88039672
	LPL	lipoprotein lipase	2,22122318
→	SFTPB	surfactant, pulmonary-associated protein B	2,19066858
	KCNN4	potassium intermediate/small conductance calcium-activa	2,02327266
	TMEM100	transmembrane protein 100	1,98586606
	FAM70A	family with sequence similarity 70, member A	1,69584944
	GPR92	G protein-coupled receptor 92	1,60794937
	AMIGO2	adhesion molecule with Ig-like domain 2	1,60040378
	CKS2	CDC28 protein kinase regulatory subunit 2	1,59811559
	BUB1B	BUB1 budding uninhibited by benzimidazoles 1 homolog	1,54802287
	PXDN	peroxidasin homolog (Drosophila)	1,492023
	LOC339984	hypothetical protein LOC339984	1,49015211
→	TMSL8	thymosin-like 8	1,48066818
	LAMB3	laminin, beta 3	1,47700816
→	LEMD1	LEM domain containing 1	1,39262062
	TMEM154	transmembrane protein 154	1,20852826
	GPR83	G protein-coupled receptor 83	-1,2433491
	DIO2	deiodinase, iodothyronine, type II	-1,3453556
→	STEAP2	six transmembrane epithelial antigen of the prostate 2	-1,3522521
	WDR72	WD repeat domain 72	-1,357822
	RGS8	regulator of G-protein signaling 8	-1,397905
	NUAK2	NUAK family, SNF1-like kinase, 2	-1,6731593
	FGF13	fibroblast growth factor 13	-1,6815821
	FAM3B	family with sequence similarity 3, member B	-1,7304152
	SOD3	superoxide dismutase 3, extracellular	-1,836931
	AOX1	aldehyde oxidase 1	-1,8812732
→	BEX1	brain expressed, X-linked 1	-2,3206673
	GJB6	gap junction protein, beta 6	-3,3826917

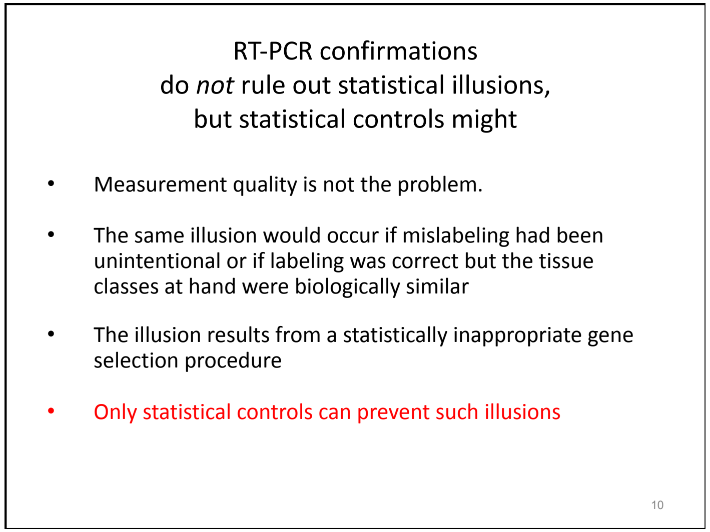
Regulation previously shown in other cancers



All this was a statistical illusion!

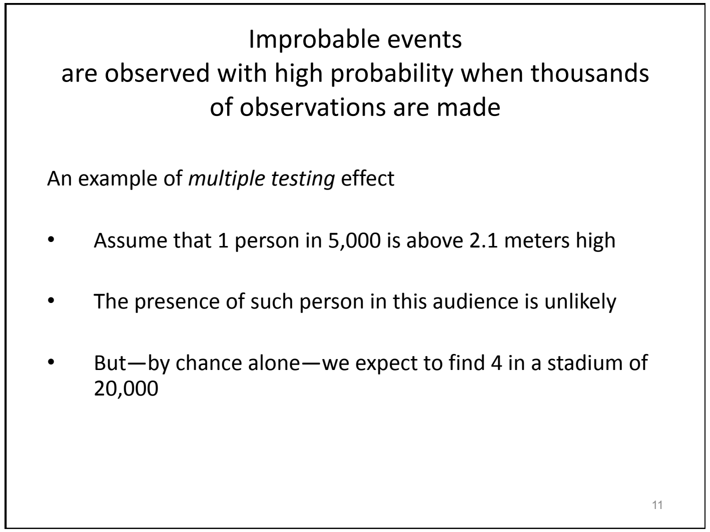
Up-regulated genes were actually selected as follows

1. **Randomly mix-up "tumor" and "healthy" sample labels.**
2. Compute expression ratio of "tumor" vs. "patient-matched normal" tissue
3. Select genes with fold-change >2 in at least 8/12=2/3 of the patients



RT-PCR confirmations do *not* rule out statistical illusions, but statistical controls might

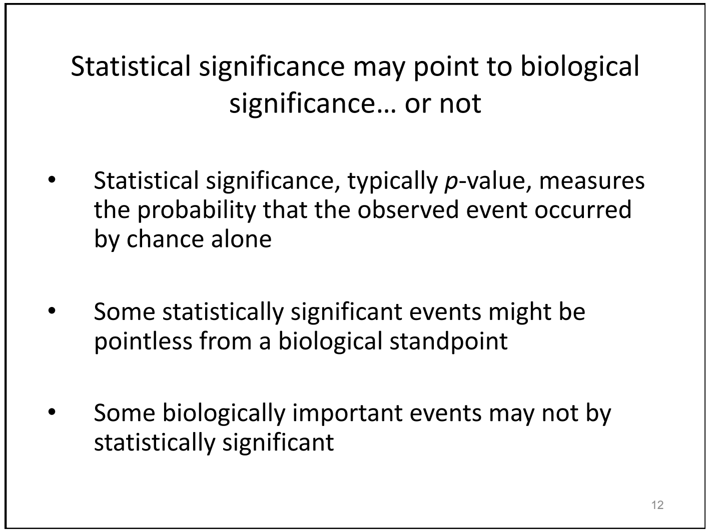
- Measurement quality is not the problem.
- The same illusion would occur if mislabeling had been unintentional or if labeling was correct but the tissue classes at hand were biologically similar
- The illusion results from a statistically inappropriate gene selection procedure
- Only statistical controls can prevent such illusions



Improbable events are observed with high probability when thousands of observations are made

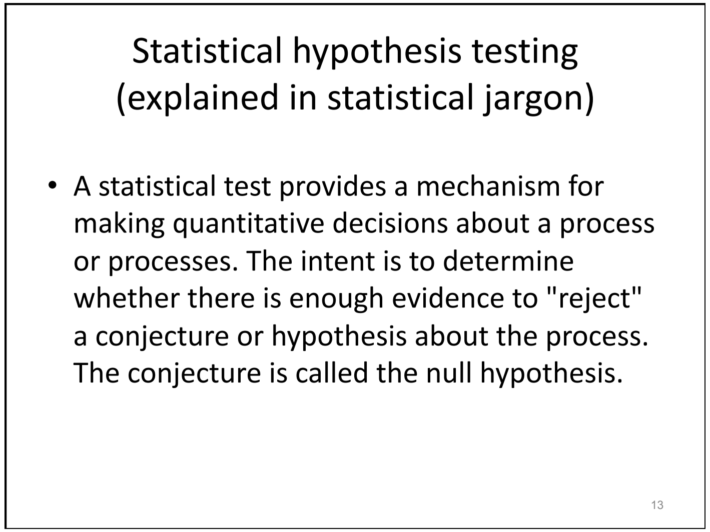
An example of *multiple testing* effect

- Assume that 1 person in 5,000 is above 2.1 meters high
- The presence of such person in this audience is unlikely
- But—by chance alone—we expect to find 4 in a stadium of 20,000



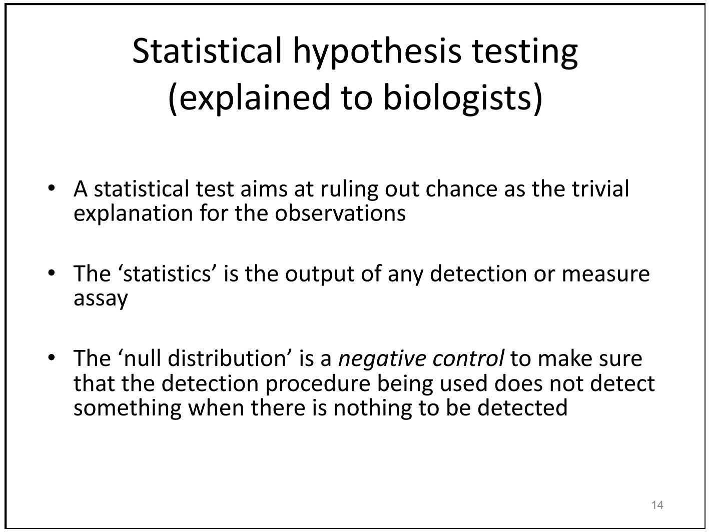
Statistical significance may point to biological significance… or not

- Statistical significance, typically *p*-value, measures the probability that the observed event occurred by chance alone
- Some statistically significant events might be pointless from a biological standpoint
- Some biologically important events may not by statistically significant



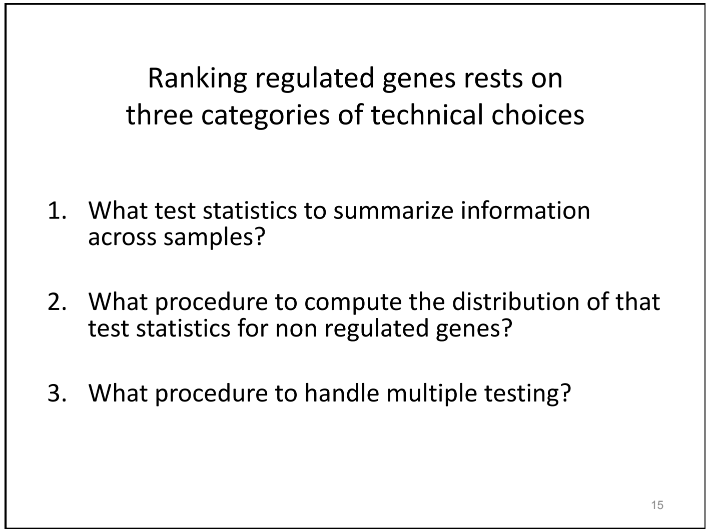
Statistical hypothesis testing (explained in statistical jargon)

- A statistical test provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process. The conjecture is called the null hypothesis.



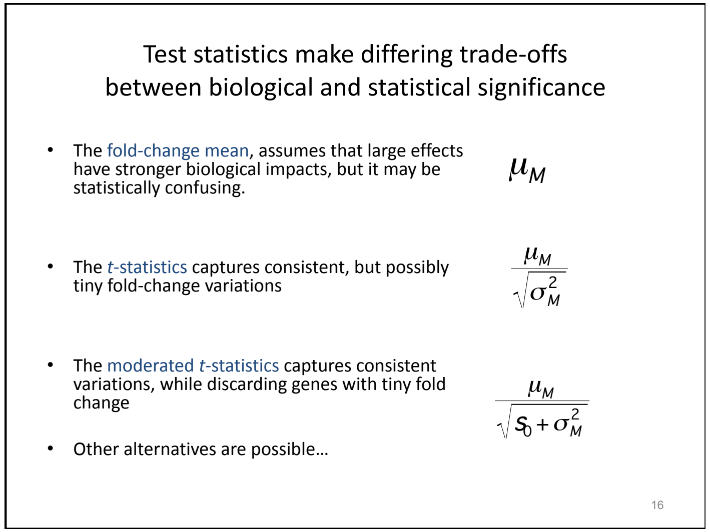
Statistical hypothesis testing (explained to biologists)

- A statistical test aims at ruling out chance as the trivial explanation for the observations
- The ‘statistics’ is the output of any detection or measure assay
- The ‘null distribution’ is a *negative control* to make sure that the detection procedure being used does not detect something when there is nothing to be detected



Ranking regulated genes rests on three categories of technical choices

1. What test statistics to summarize information across samples?
2. What procedure to compute the distribution of that test statistics for non regulated genes?
3. What procedure to handle multiple testing?



Test statistics make differing trade-offs between biological and statistical significance

- The fold-change mean, assumes that large effects have stronger biological impacts, but it may be statistically confusing. μ_M
- The *t*-statistics captures consistent, but possibly tiny fold-change variations $\frac{\mu_M}{\sqrt{\sigma_M^2}}$
- The moderated *t*-statistics captures consistent variations, while discarding genes with tiny fold change $\frac{\mu_M}{\sqrt{S_0+\sigma_M^2}}$
- Other alternatives are possible…

The null distribution is the negative control of statisticians

- The null distribution is the distribution of the test statistics when *no* gene is regulated
- Under certain hypothesis the *t*- and other classical statistics have a null distribution with a known mathematical form
- The null distribution may be estimated by repeatedly computing the statistics over randomly mixed-up data

p-values may be adjusted for multiple testing in various ways

- *p*-value (no adjustment): proportion of false positives among all genes

➢ Bonferroni correction:
significance_threshold = *0.05* / (number of tests)

➢ Family-wise error rate (FWER): probability that one or more genes considered regulated are false positives

➢ False discovery rate (FDR, *q*-value): proportion of false positives among genes considered regulated

Technical options may be combined in various ways

Significance analysis of microarrays (SAM, Tusher *et al*., 2001, *PNAS* 98, p5116) uses

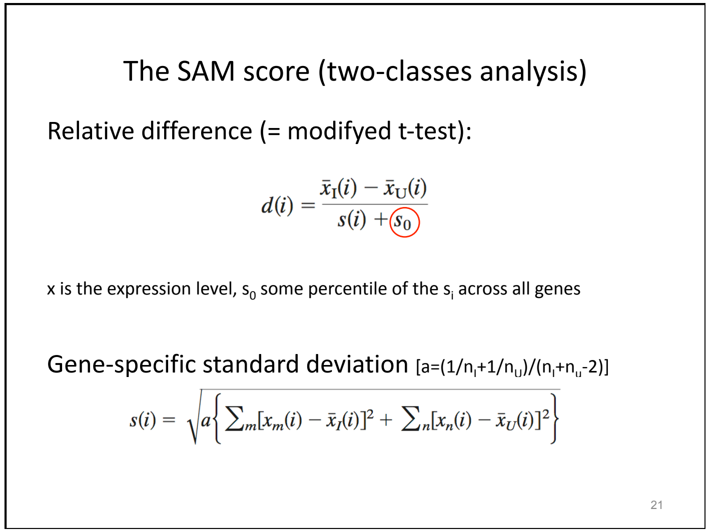
- Moderated *t*-statistics
- Permutation-based estimate of the null distribution
- Control of the false discovery rate

Significance Analysis of Microarrays

SAM is among the most widely used gene selection procedure (R, and Excel add-on)

It is flexible

- one class (deviation from no differential expression)
- paired/unpaired two-classes comparison
- F-test, i.e. *n* classes comparison
- Genes correlated with a continuous variable
- Genes associated with survival (Cox analysis)



The SAM score (two-classes analysis)

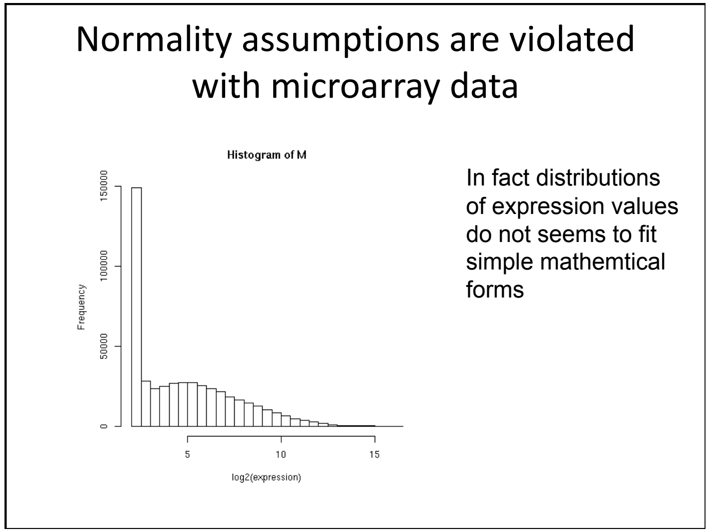
Relative difference (= modifyed t-test):

$$d(i) = \frac{\bar{x}_{\mathrm{I}}(i) - \bar{x}_{\mathrm{U}}(i)}{s(i) + \boxed{s_0}}$$

x is the expression level, s_0 some percentile of the s_i across all genes

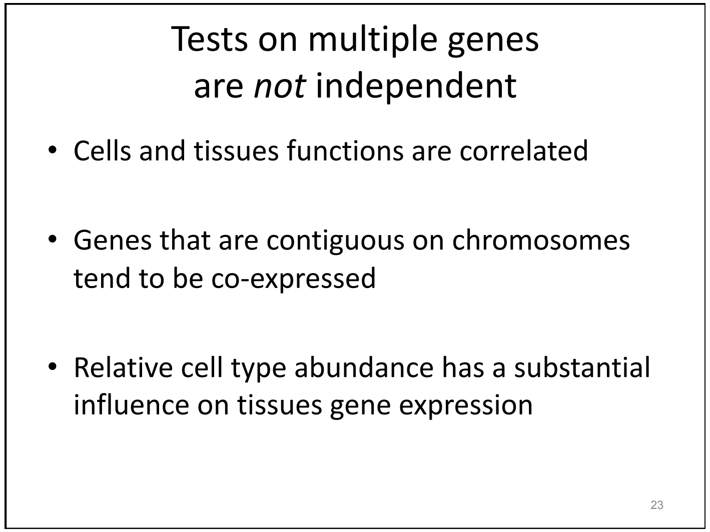
Gene-specific standard deviation $[a=(1/n_I+1/n_U)/(n_I+n_u-2)]$

$$s(i) = \sqrt{a\left\{\sum_m[x_m(i) - \bar{x}_I(i)]^2 + \sum_n[x_n(i) - \bar{x}_U(i)]^2\right\}}$$



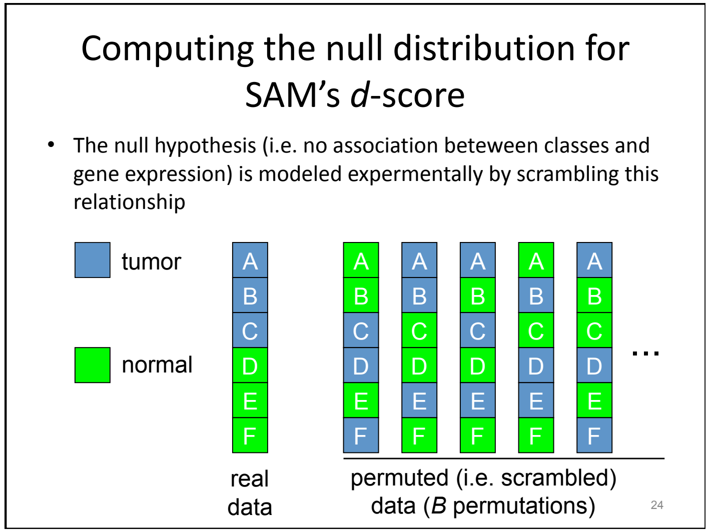
Normality assumptions are violated with microarray data

In fact distributions of expression values do not seems to fit simple mathemtical forms



Tests on multiple genes are *not* independent

- Cells and tissues functions are correlated
- Genes that are contiguous on chromosomes tend to be co-expressed
- Relative cell type abundance has a substantial influence on tissues gene expression



Computing the null distribution for SAM's *d*-score

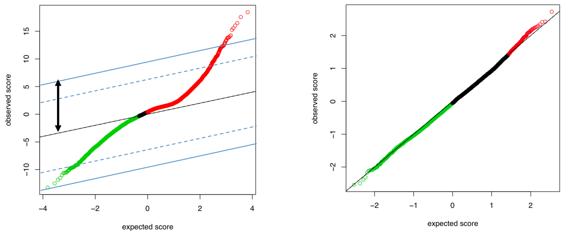
- The null hypothesis (i.e. no association beteween classes and gene expression) is modeled expermentally by scrambling this relationship

🟦 tumor

🟩 normal

real data	permuted (i.e. scrambled) data (*B* permutations)					
🟦 A	🟩 A	🟦 A	🟦 A	🟩 A	🟦 A	
🟦 B	🟩 B	🟦 B	🟩 B	🟦 B	🟩 B	
🟦 C	🟦 C	🟩 C	🟦 C	🟩 C	🟩 C	…
🟩 D	🟦 D	🟩 D	🟦 D	🟦 D	🟦 D	
🟩 E	🟩 E	🟦 E	🟦 E	🟦 E	🟩 E	
🟩 F	🟦 F	🟩 F	🟩 F	🟩 F	🟦 F	

Adjusting the False Discovery Rate (FDR)



False positives are more of a problem than false negatives in microarray studies

- Most studies yield list of several hundred regulated genes
- there are ususally enough positive genes to get biological insights, and too much for experimental follow up
- So far, the issue of power has been paid little attention to
- There is a tool to compute the power of SAM analyses (http://www-stat.stanford.edu/~tibs/SAM/)

The number of samples needed to reliably detect differential expression is computable

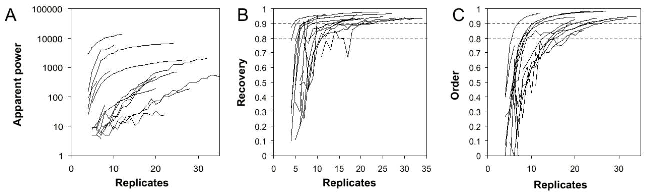


Fig. 3. Summary of results. Each line represents results for one data set shown in Table 1, at an FDR of 0.05. Not all of the 16 data sets are illustrated on these graphs, because some failed to meet criteria at this FDR (see our web site for more results). The plots are of the median values for all trials. Error bars are omitted for clarity. The dashed lines in (B) and (C) indicate the 0.8 and 0.9 levels. (**A**) Plot of the number of genes selected (apparent power, the size of *Ssel*). Note that the scale is logarithmic. (**B**) Recovery stability. (**C**) Order stability. Values below zero are not shown. Larger versions of this and the other figures are available as supplementary data.

From Pavlidis *et al*., (2001), *Bioinformatics* 19, 1620–7

The *q*-value

- Storey & Tibshirani (2003, PNAS 100, p9440) propose a formal procedure to adjust *p*-values for multiple testing
- The *q*-value of a test is the fraction of false positive among all the tests with statistics as or more extreme than this test
- The procedure takes *p*-values as input (available as R package)

The *q*-value calculation, roughly

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

FDR(t) = (# false positive)/(# called significant)

= E(F(*t*) / [F(*t*) + T(*t*)])

≈ E(F(*t*)) / E(F(*t*) + T(*t*)) (because m is large)

≈ $m_0 t$ / (# p_i<*t*) (because *p*-values are uniformly distributed under true H_0)

t is the *p*-value threshold (also called α)

m_0 can be estimated from the distribution of *p*-values

The trick: distribution of *p*-values contains information about the fraction of positive tests

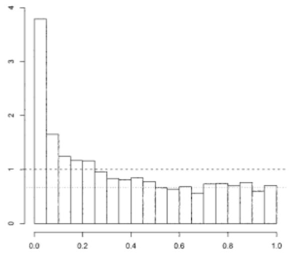


Fig. 1. A density histogram of the 3,170 *p* values from the Hedenfalk *et al.* (14) data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null *p* values.

The trick: distribution of *p*-values contains information about the fraction of positive tests

proportion of null *p*-values

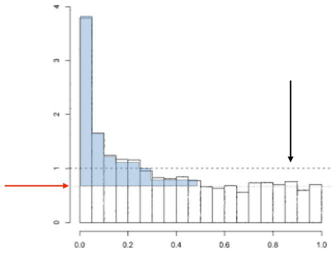


Fig. 1. A density histogram of the 3,170 *p* values from the Hedenfalk *et al.* (14) data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null *p* values.