

## Statistics for bioinformatics

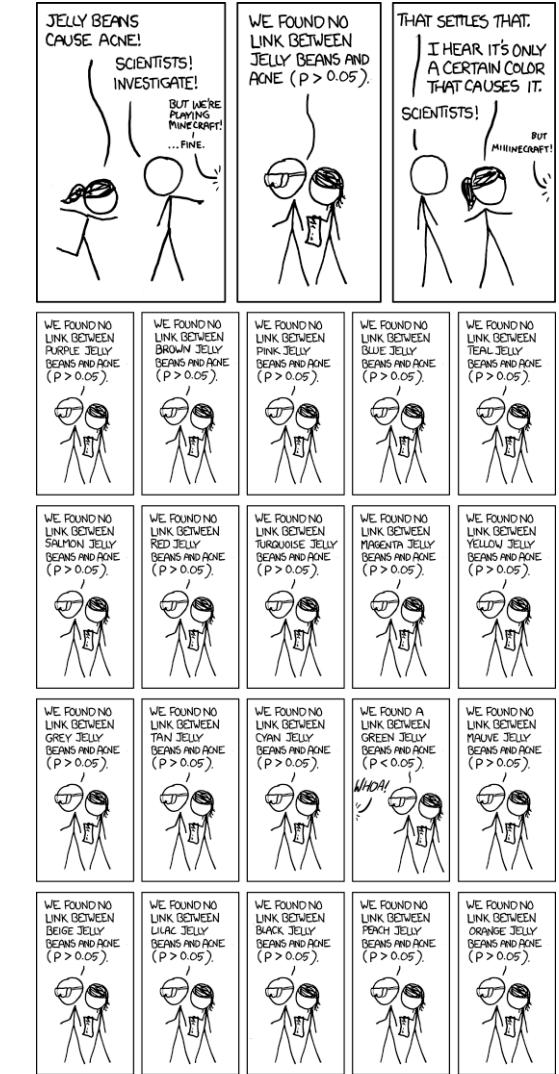
# Multiple testing

Jacques van Helden

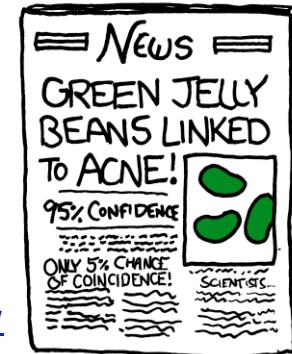
[Jacques.van-Helden@univ-amu.fr](mailto:Jacques.van-Helden@univ-amu.fr)

Aix-Marseille Université, France  
Technological Advances for Genomics and Clinics  
(TAGC, INSERM Unit U1090)  
<http://jacques.van-helden.perso.luminy.univmed.fr/>

FORMER ADDRESS (1999-2011)  
Université Libre de Bruxelles, Belgique  
Bioinformatique des Génomes et des Réseaux (BiGRe lab)  
<http://www.bigre.ulb.ac.be/>



<http://xkcd.com/882/>



## What is a P-value ?

- In the context of significance tests (e.g. detecting over-represented words, or estimating the significance of BLAST matching scores), the P-value represents the probability to generate by chance (under the background model) a value at least as distant from the expectation as the one we observe.
  - $P_{\text{val}} = P(X \geq \text{obs})$
- For the analyst, this P-value indicates the risk to consider something as significant whereas it is not, i.e. the **False Positive Risk (FPR)**.
- In the context of hypothesis testing, the concept of P-value is associated to the parameter alpha, the risk of first type error. The first type error consists in rejecting the null hypothesis  $H_0$  whereas it is true :  $P(RH_0|H_0)$ . This alpha risk is estimated by testing the significance of the observed statistics (e.g.  $\chi^2_{\text{obs}}$ ,  $t_{\text{obs}}$ ) according to the theoretical distribution.

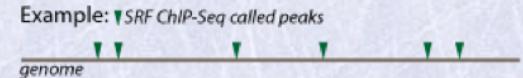
# Application example: GREAT - Genomic Regions Enrichment of Annotations Tool

- GREAT takes as input a set of genomic features (e.g. the peaks obtained from a ChIP-seq experiment).
- Identifies the set of genes matched by these features (genes are extended upstream and downstream to include regulatory regions).
- Assesses the enrichment of the set of target genes with each class of the Gene Ontology.
- One analysis involves several thousands of significant tests.

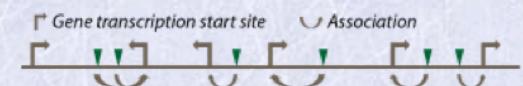
## Genomic Regions Enrichment of Annotations Tool

GREAT predicts functions of *cis*-regulatory regions.

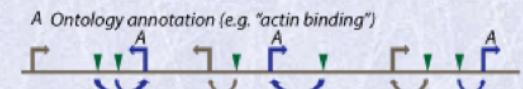
1. **Input:** A set of Genomic Regions (such as transcription factor binding events identified by ChIP-Seq).



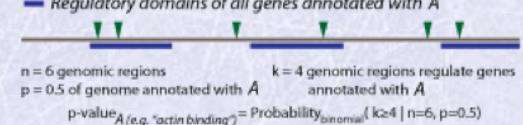
2. GREAT associates both proximal and distal input Genomic Regions with their putative target genes.



3. GREAT uses gene Annotations from numerous ontologies to associate genomic regions with annotations.



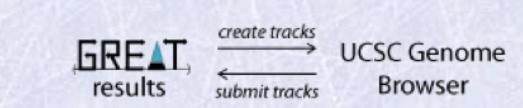
4. GREAT calculates statistical Enrichments for associations between Genomic Regions and Annotations.



5. **Output:** Annotation terms that are significantly associated with the set of input Genomic Regions.

Ontology term	p-value
Actin cytoskeleton	$10^{-9}$
FOS gene family	$10^{-8}$
TRAIL signaling	$10^{-7}$

6. Users can create UCSC custom tracks from term-enriched subsets of Genomic Regions. Any track can be directly submitted to GREAT from the UCSC Table Browser.



# Statistics

- Nomenclature
  - F number of false positives (FP)
  - T number of true positives (TP)
  - S number of tests called significant
  - $m_0$  number of truly null features
  - $m_1$  number of truly alternative features
  - m total number of features  $m = m_0 + m_1$
  - p threshold on p-value  $p = E[F / m_0]$
  - $E[F]$  expected number of false positives (also called E-value)  $E[F] = p * m_0$
  - $\Pr(F >+ 1)$  family-wise error rate  $FWER = 1 - (1 - p)^{m_0}$
  - FDR False discovery rate  $FDR = E[F/S] = E[F / (F + T)]$
  - Sp Specificity  $Sp = (m_0 - F) / m_0$
  - Sn Sensitivity  $Sn = T / m_1$
- In practice
  - We never know the values of F, T,  $m_0$ ,  $m_1$ , or any statistics derived from them.
  - The only observable numbers are the number of tests ( $m$ ), and the number of these declared significant ( $S$ ) or not ( $m-S$ ).
  - Some strategies have however been proposed to estimate  $m_0$  and  $m_1$  (see Storey and Tibshirani, 2003).

**Table 1. Possible outcomes from thresholding  $m$  features for significance**

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

# Validation statistics

- Various statistics can be derived from the 4 elements of a contingency table.

Abbrev	Name	Formula
TP	True positive	TP
FP	False positive	FP
FN	False negative	FN
TN	True negative	TN
KP	Known Positive	TP+FN
KN	Known Negative	TN+FP
PP	Predicted Positive	TP+FP
PN	Predicted Negative	FN+TN
N	Total	TP + FP + FN + TN
Prev	Prevalence	(TP + FN)/N
ODP	Overall Diagnostic Power	(FP + TN)/N
CCR	Correct Classification Rate	(TP + TN)/N
<b>Sn</b>	<b>Sensitivity</b>	<b>TP/(TP + FN)</b>
Sp	Specificity	TN/(FP + TN)
FPR	False Positive Rate	FP/(FP + TN)
FNR	False Negative Rate	FN/(TP + FN) = 1-Sn
<b>PPV</b>	<b>Positive Predictive Value</b>	<b>TP/(TP + FP)</b>
FDR	False Discovery Rate	FP/(FP+TP)
NPV	Negative Predictive Value	TN/(FN + TN)
Mis	Misclassification Rate	(FP + FN)/N
Odds	Odds-ratio	(TP + TN)/(FN + FP)
Kappa	Kappa	((TP + TN) - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))/(N - (((TP + FN)*(TP + FP) + (FP + TN)*(FN + TN))/N))
NMI	NMI n(s)	(1 - TP*log(TP) - FP*log(FP) - FN*log(FN) - TN*log(TN) + (TP+FP)*log(TP+FP) + (FN+TN)*log(FN+TN))/(N*log(N) - ((TP+FN)*log(TP+FN) + (FP+TN)*log(FP+TN)))
ACP	Average Conditional Probability	0.25*(Sn+ PPV + Sp + NPV)
MCC	Matthews correlation coefficient	(TP*TN - FP*FN) / sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]
Acc.a	Arithmetic accuracy	(Sn + PPV)/2
Acc.a2	Accuracy (alternative)	(Sn + Sp)/2
Acc.g	Geometric accuracy	sqrt(Sn*PPV)
Hit.noTN	A sort of hit rate without TN (to avoid the effect of their large number)	TP/(TP+FP+FN)

		Declared significant	
		True	False
True		FP	TN
False	True	TP	FN
False	False		

$$Sn = TP/(TP+FN)$$

Declared significant

		True	False
True		FP	TN
False		TP	FN
H0	False		
	True		

$$PPV=TP/(TP+FP)$$

Declared significant

		True	False
True		FP	TN
False		TP	FN
H0	False		
	True		

$$Sp=TN/(FP+TN)$$

Declared significant

		True	False
True		FP	TN
False		TP	FN
H0	False		
	True		

$$NPV=TN/(FN+TN)$$

Declared significant

		True	False
True		FP	TN
False		TP	FN
H0	False		
	True		

$$FPR=FP/(FP+TN)$$

Declared significant

		True	False
True		FP	TN
False		TP	FN
H0	False		
	True		

$$FDR=FP/(FP+TP)$$

Declared significant

		True	False
True		FP	TN
False		TP	FN
H0	False		
	True		

$$FNR=FN/(FN+TN)$$

Declared significant

		True	False
True		FP	TN
False		TP	FN
H0	False		
	True		

$$FNR=FN/(TP+FN)$$

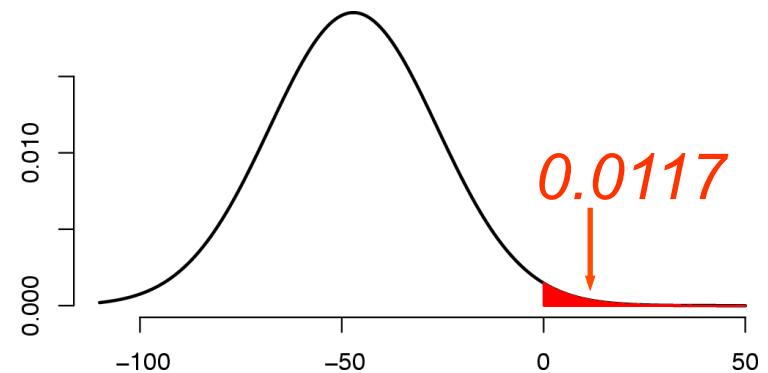
Declared significant

		True	False
True		FP	TN
False		TP	FN
H0	False		
	True		

## *Multiple testing corrections*

## The problem of multiple testing

- Let us assume that the score of the alignment between two sequences has a p-value
  - $P(X > 0) = 0.0117$
- What would happen if we consider this score as significant, while scanning a database that contains 200,000 sequences ?
- Let N be the number of tests
  - The risk of error (P-value) associated to each gene will thus be challenged N times.
  - The significance thresholds generally used for single testing ( $\alpha = 0.01, 0.001$ ) are thus likely to return many false positive.
- The situation of multiple testing is very frequent in bioinformatics
  - Assessing the significance of each gene on a chip represents thousands of simultaneous tests.
  - Genome-wide association studies (GWAS) are now routinely performed with SNP chips containing 600.000 SNPs.
  - Sequence similarity searches (e.g. BLAST a sequence against all known proteins) amount to compare a query sequences to billions of database entries.



## *Multiple testing correction : Bonferroni's rule*

- A first approach to correct for multiple tests is to apply Bonferroni's rule.
- Adapt the p-value threshold ("alpha risk") to the number of simultaneous tests.
- Bonferroni's rule recommends to lower the alpha risk below  $1/N$ , where  $N$  is the number of tests.
- A stringent application of Bonferroni's correction is to divide the "usual" alpha risk (e.g. 0.05) by the number of tests.
- Note: Bonferroni's correction is reputed to be too stringent. It performs an efficient control of the false positive rate, but at the cost of a loss of sensitivity. Alternative corrections attempt to increase the sensitivity.

$$\alpha_b = \frac{\alpha_n}{N} \leq \frac{1}{N}$$

where

$N$  is the number of tests

$\alpha_n$  is the usual threshold on the nominal p-value

$\alpha_b$  is the Bonferroni-corrected threshold on p-value

## *Multiple testing correction : from P-value to E-value*

- If  $p=P(X > 0)=0.0117$  and the database contains  $N=200,000$  entries, we expect to obtain  $N \cdot p = 2340$  false positives !
- We are in a situation of multi-testing : each analysis amounts to test  $N$  hypotheses.
- The E-value (expected value) allows to take this effect into account :
  - $\text{Eval} = \text{Pval} * N$
  - Instead of setting a threshold on the P-value, we should set a threshold on the E-value.
  - If we want to avoid false positive, this threshold should always be smaller than 1.
    - $\text{Threshold}(\text{Eval}) \leq 1$
- The fact to set a threshold  $\leq 1$  on the E-value is equivalent to Bonferroni's correction, which consists in adapting the threshold on the p-value.
  - $\text{Threshold}(\text{Pval}) \leq 1/N$

$$\text{Eval} = N \cdot \text{Pval}$$

## *Multiple testing correction : Family-wise Error Rate (FWER)*

- Another correction for multiple testing consists in estimating the Family-Wise Error Rate (FWER).
- The FWER is the probability to observe *at least* one false positive in the whole set of tests. This probability can be calculated quite easily from the P-value ( $Pval$ ).

$$FWER = 1 - (1 - Pval)^N$$

# False Discovery Rate (FDR)

- Yet another approach is to consider, for a given threshold on P-value, the False Discovery Rate (FDR), i.e. the proportion of false predictions within a set of tests declared significant.
  - FP            number of false positives
  - TP            number of true positives
- Difficulty: this measure involves an estimation of the number of true positives.
- We thus assume that the set of tests is subdivided in two subsets
  - $m_0$  elements ***under the null hypothesis ("truly null")***.
  - $m_1$  elements not under the null hypothesis.
- However, how could we evaluate the respective sizes of these subsets ? This would require to know a priori which elements are truly null or not.

$$FDR = FP / (FP + TP)$$

## Summary - Multi-testing corrections

$$\alpha_{Bonf} \leq \frac{1}{N}$$

- Bonferroni rule adapt significance threshold

$$Eval = N \cdot Pval$$

- E-value      expected number of false positives

$$FWER = 1 - (1 - Pval)^N$$

- FWER      Family-wise error rate:  
probability to observe  
at least one false positive

$$FDR = FP / (FP + TP)$$

- FDR      False discovery rate:  
estimated rate of false positives  
among the predictions

***The “q-value”  
(Storey and Tibshirani, 2003)***

# How can we estimate the $m_0 / m_1$ proportions ?

- Storey and Tibshirani proposed an elegant method to estimate the respective numbers of objects that are under the null hypothesis ( $m_0$ ) or not ( $m_1$ ).
- Principle
  - Draw the distribution of all p-values.
  - Under null hypothesis, this distribution should be uniform.
  - Generally, the upper part of the distribution essentially contains truly null objects.
  - Use this upper part to estimate the number of truly null objects in the whole set.
- See the practical about multiple testing correction on the supporting Web site.
  - [http://pedagogix-tagc.univ-mrs.fr/courses/statistics\\_bioinformatics/](http://pedagogix-tagc.univ-mrs.fr/courses/statistics_bioinformatics/)

Fig 1 from Storey and Tibshirani, 2003

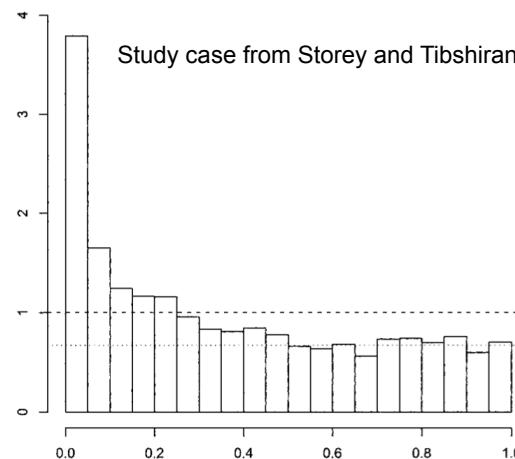
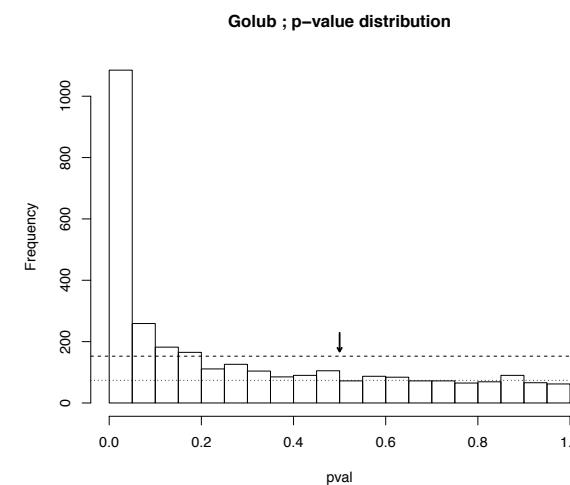
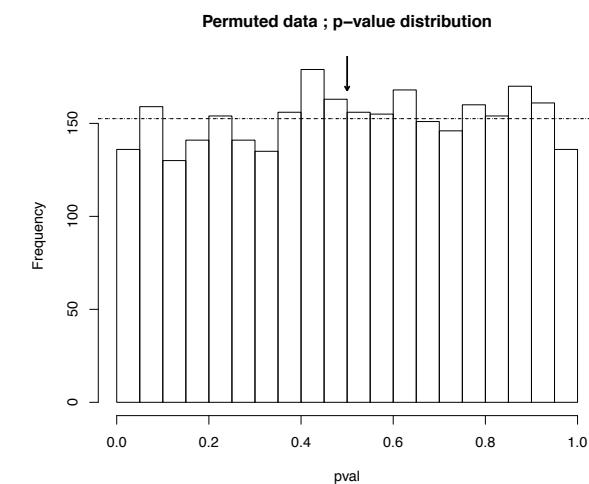


Fig. 1. A density histogram of the 3,170  $p$  values from the Hedenfalk et al. (14) data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null  $p$  values.

Application to another study case  
(ALL versus AML expression from Gobub et al., 1999)

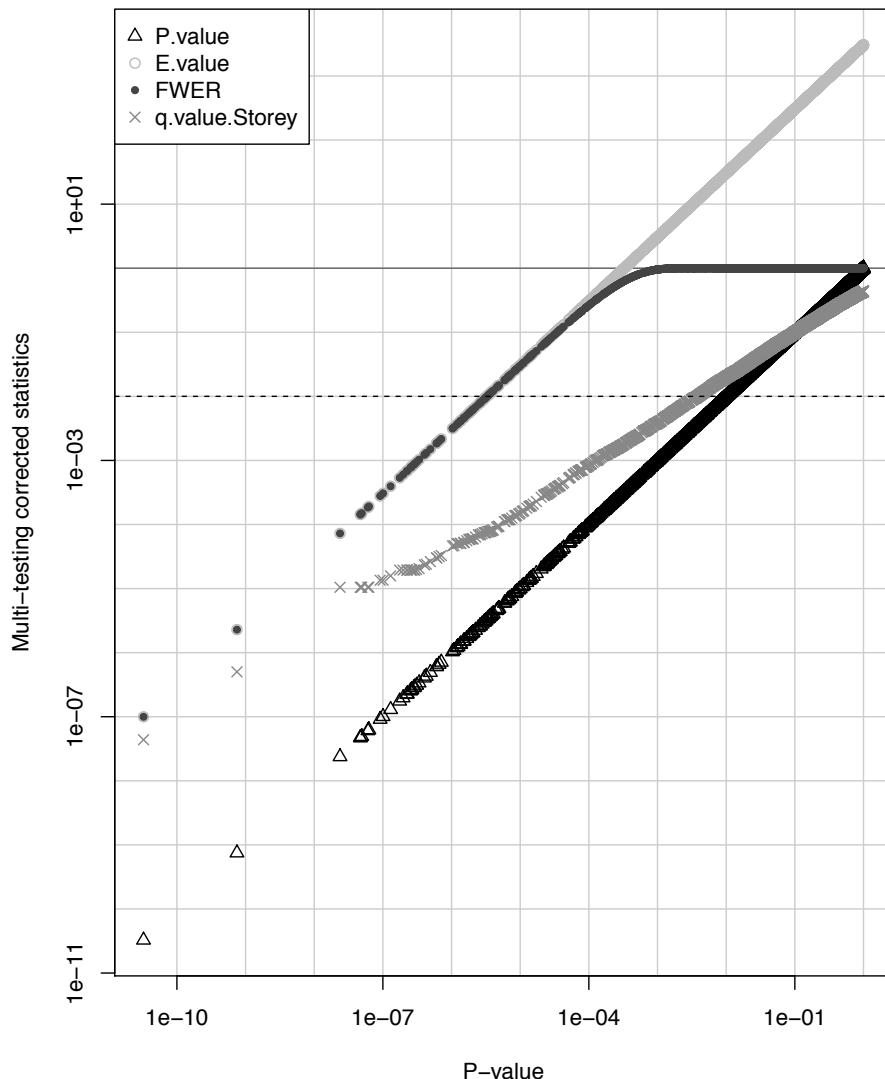


Negative control: permuted data from  
Golub et al. (1999)



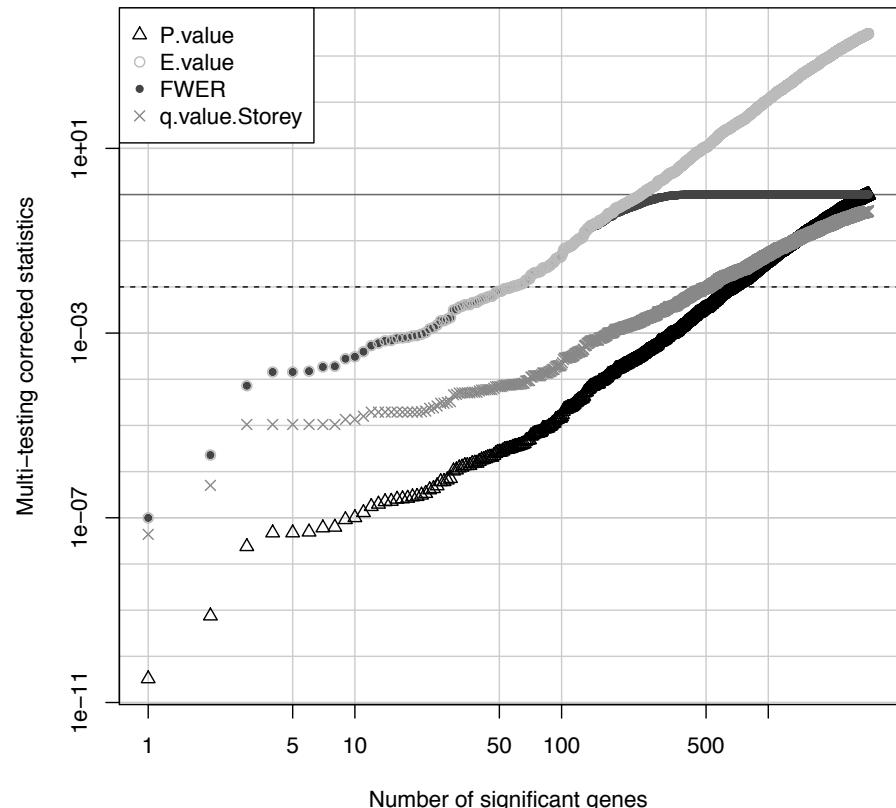
## Impact of multi-testing on the number of genes declared significant

- Left: un-corrected (P-value) versus multitemping-corrected statistics for controlling false positives.



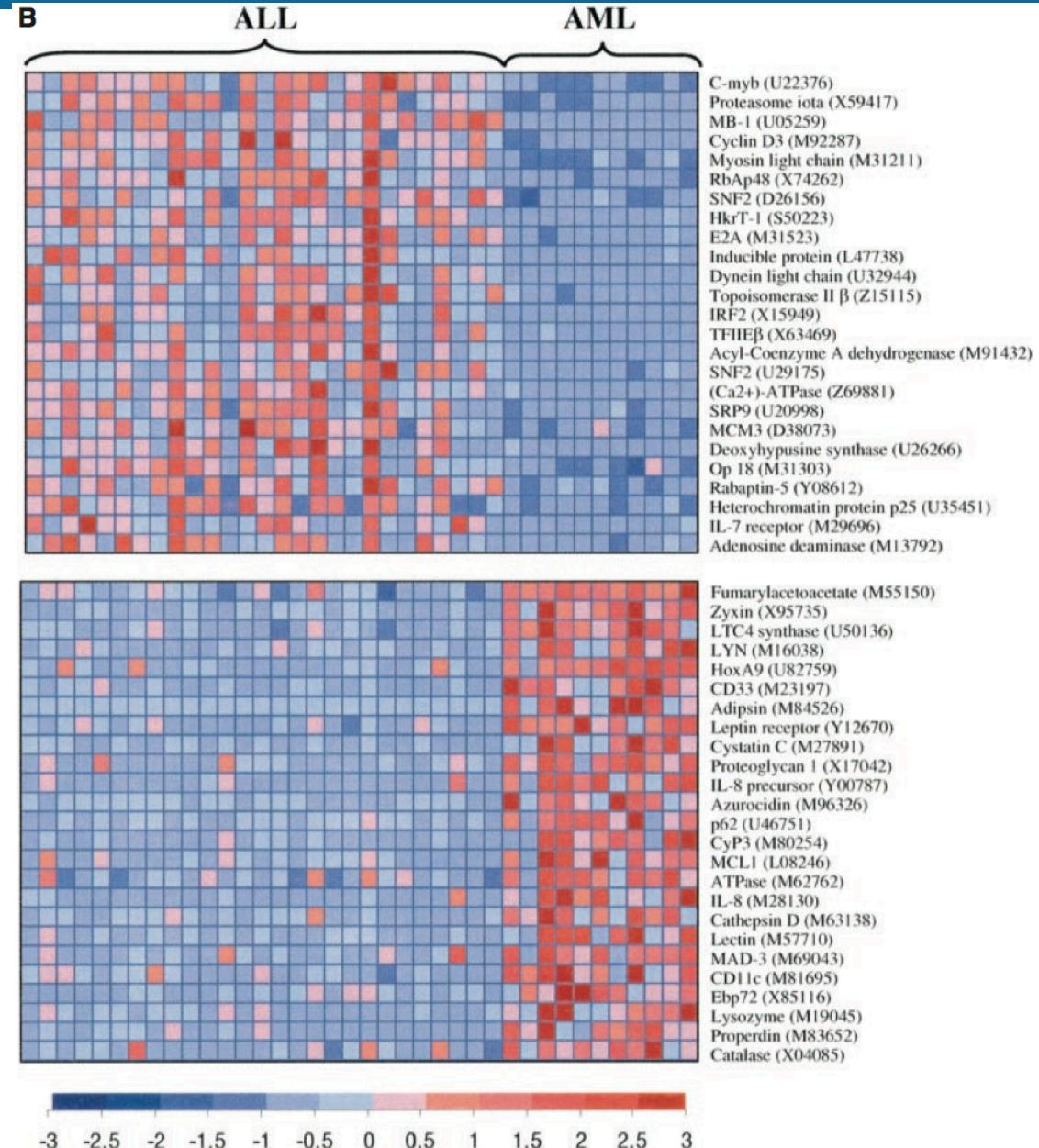
- Right: Number of genes declared significant as a function of the thresholds.

Criterion	genes	E(FP)	
Pval $\leq 0.01$	673		Pval*N = 30
Bonferroni (Pval $\leq 0.01/N$ ) = 0.01	57		$N^*Pval/N$
Eval = Pval*N $\leq 0.01$	57		Eval=0.01
Eval $\leq 1$	243		Eval = 1
Qval (m0=N) $\leq 0.01$	366		$S^*qval = 3.66$
Qval (Storey) $\leq 0.01$	515		$S^*qval = 5.15$



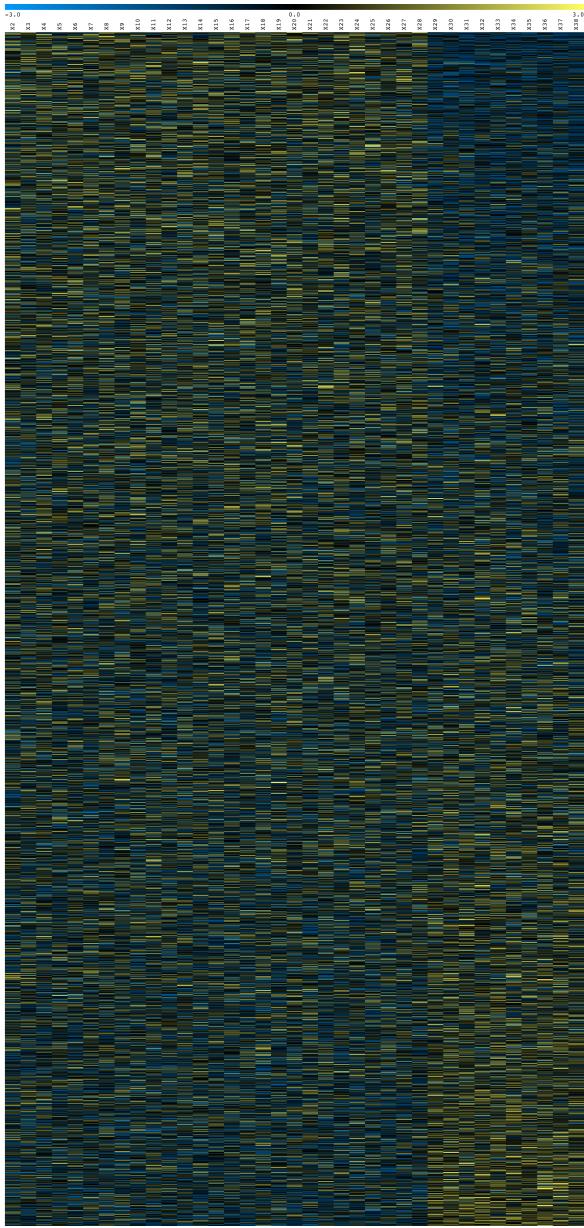
# Study case for multitempling corrections

- Historical article by Golub et al (1999).
- Authors used microarrays to measure the transcriptome of 38 samples
  - 27 Acute Lymphoblastic Leukemia
  - 11 Acute Myeloid Leukemia



- Data source: Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (1999) vol. 286 (5439) pp. 531-7

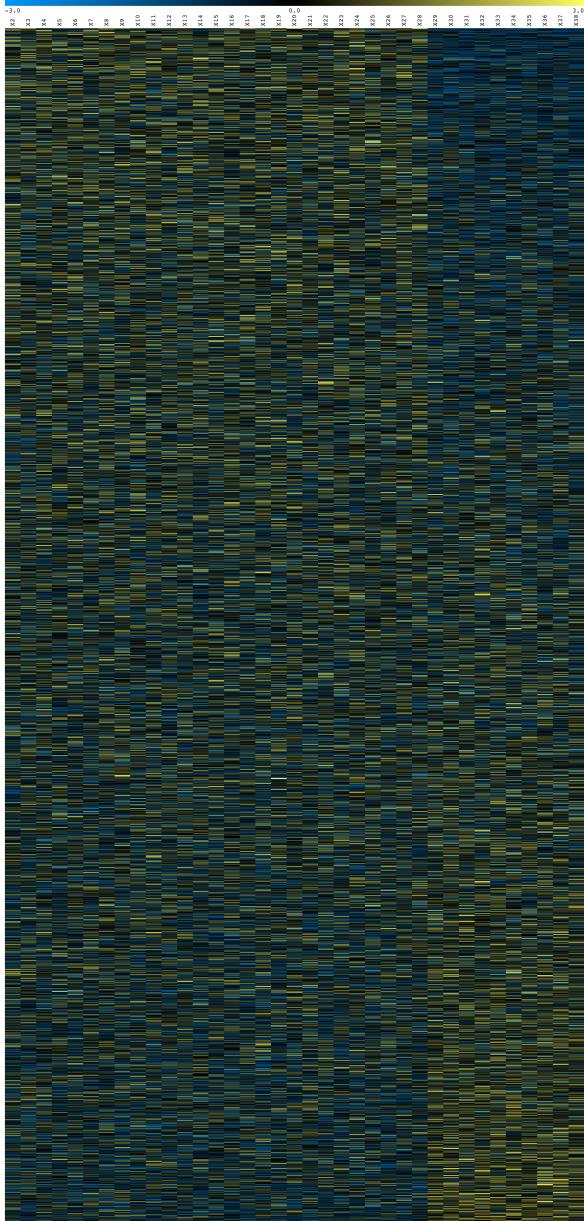
# Some expression profiles sorted by difference between means



- Data treatment: statistics\_bioinformatics/R-files/student\_test.R
- Image generated with TMev

- Let us assume that we have gene expression profiles sorted according to some criterion (e.g. the significance of a t-statistics).
- We would like to select the set of genes considered as significant (e.g. to establish a signature enabling to predict cancer type).
- **Question**
  - Where should we set the limit ?
- Classical approach: select all genes passing an a priori defined level of significance
  - E.g. P-value  $\leq 0.01$

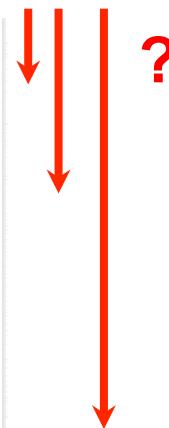
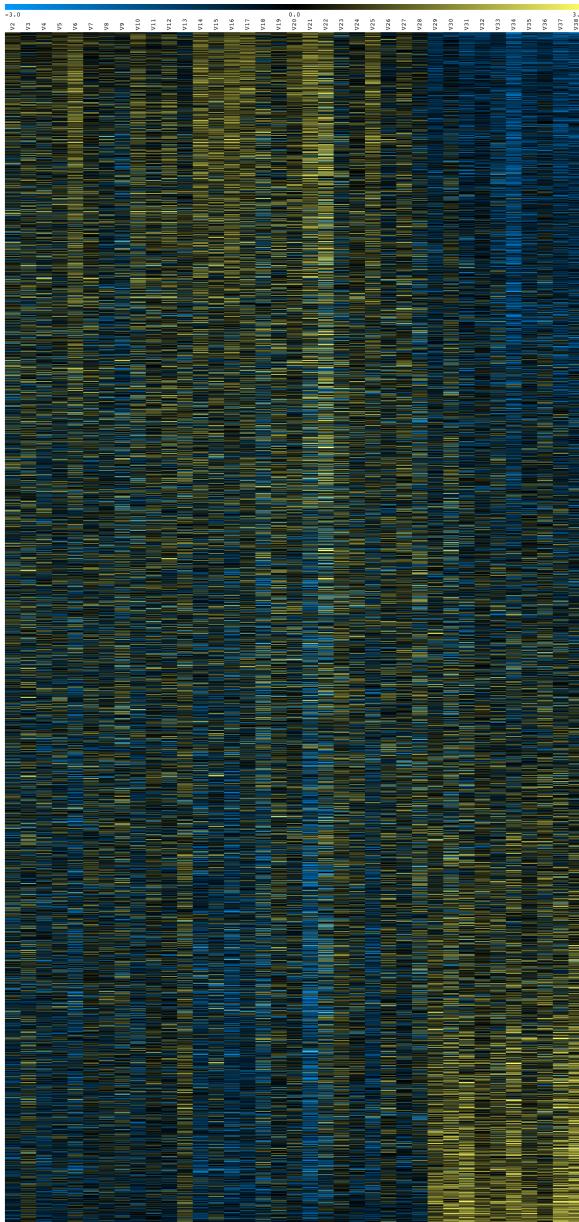
# *Some expression profiles sorted by difference between means*



- Surprise: this data set contains randomized data.
- There is thus not a single gene that should be considered as “truly differentially expressed”.

- Data treatment: statistics\_bioinformatics/R-files/student\_test.R
- Image generated with TMeV

# Golub's expression profiles sorted by difference between means



- Golub dataset (3051 genes) sorted by mean differences.
- The heat-map shows the chip-wise and gene-wise standardized z-scores, to highlight differences irrespective of gene-wise specific variance.
- Top: higher expression in ALL than AML.
- Top: higher expression in AML than ALL.
- Question
  - We want to select the ALL-specific genes to get a “signature” of ALL-type cancers.
  - Where should we set the limit ?

- Data source: Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (1999) vol. 286 (5439) pp. 531-7
- Data treatment: statistics\_bioinformatics/R-files/student\_test.R
- Image generated with TMeV