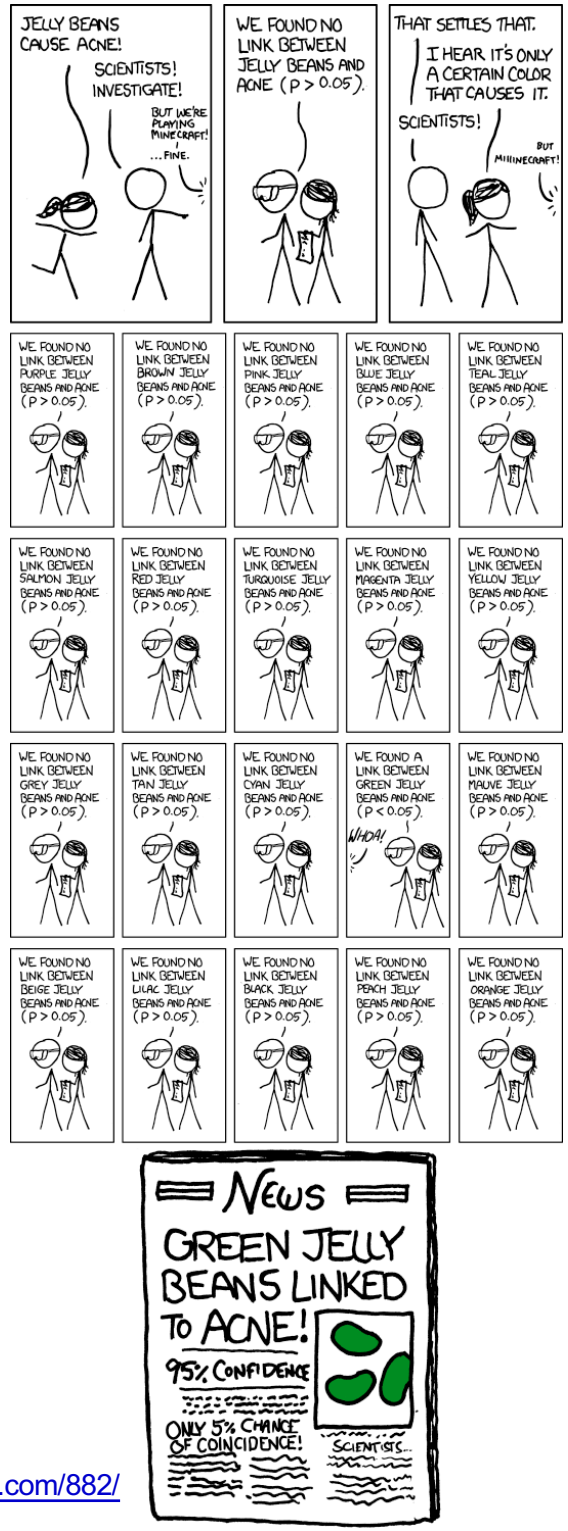
Statistics for bioinformatics

Multiple testing

Jacques van Helden

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.univ-amu.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)



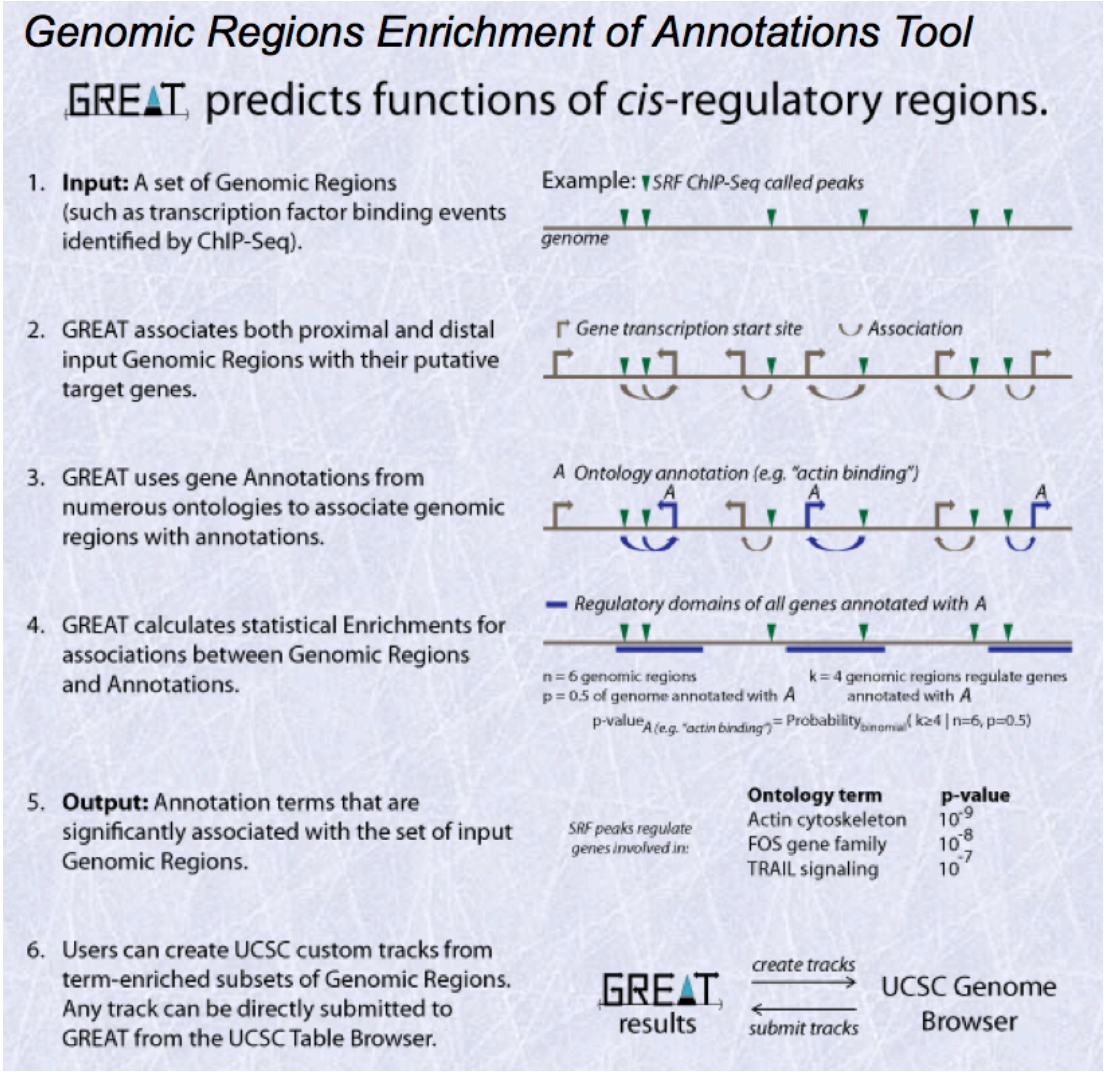
http://xkcd.com/882/

What is a P-value ?

- In the context of significance tests (e.g. detecting over-represented k-mers, or estimating the significance of BLAST matching scores), the P-value represents the probability to obtain by chance (under the null model) a value at least as distant from the expectation as the one observed.
  - $Pval = P(X \geq x_{obs})$
- The p-value is an estimation of the ***False Positive Risk*** *(***FPR***)*, which is interpreted as the probability to unduly consider a result as significant.
- The p-value is computed by measuring the left tail, the right tail or both of the theoretical distribution corresponding to the test statistics (e.g. binomial, chi-squared, Student, …).
- In the context of hypothesis testing, the P-value corresponds to the risk of first type error, which consists in rejecting the null hypothesis H_0 whereas it is true.
  - $Pval = P(RH_0|H_0)$
- The classical approach is to define *a priori* some threshold on the first error type. This threshold is conventionally represented by the Greek symbol α (alpha).
- In bioinformatics, we often take a slightly different approach: rather than applying a predefined cutoff, we sort all the tested objects (e.g. 25.000 genes) by increasing order of p-value, and evaluate the statistical significance and biological relevance of the top-ranking genes.

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.



- http://great.stanford.edu/

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 $p = E[F / m0]$
  - E[F] expected number of false positives (also called E-value) $E[F] = p * m0$
  - Pr(F >+ 1) family-wise error rate $FWER = 1 – (1 – p)^{m0}$
  - FDR False discovery rate $FDR = E[F/S] = E[F / (F + T)]$
  - Sp Specificity $Sp = (m0 – F) / m0$
  - Sn Sensitivity $Sn = T / m1$
- In practice
  - We never know the values of F, T, m0, m1, 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** *m0* and *m1* (see Storey and Tibshirani, 2003).

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

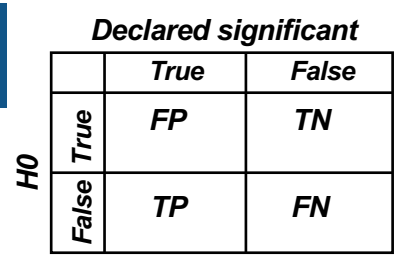
	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

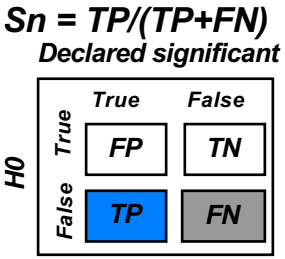
- Storey and Tibshirani. Statistical significance for genomewide studies. Proc Natl Acad Sci USA (2003) vol. 100 (16) pp. 9440-5

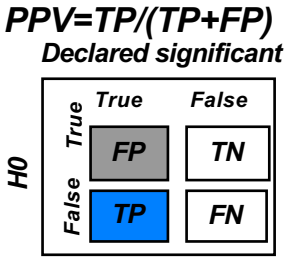
Validation statistics

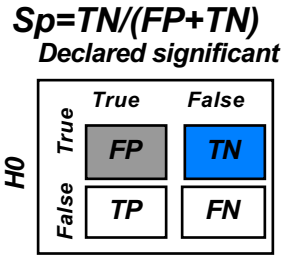
- Various statistics can be derived from the 4 elements of a contingency table.

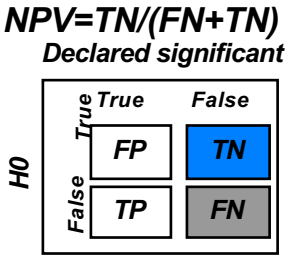
Abbrev	Name	Formula
TP	True positive	TP
FP	False positive	FP
FN	False negative	FN
TN	True negative	TN
KP	Known Positive	TP+FN
KN	Known Negative	TN+FP
PP	Predicted Positive	TP+FP
PN	Predicted Negative	FN+TN
N	Total	TP + FP + FN + TN
Prev	Prevalence	(TP + FN)/N
ODP	Overall Diagnostic Power	(FP + TN)/N
CCR	Correct Classification Rate	(TP + TN)/N
Sn	**Sensitivity**	**TP/(TP + FN)**
Sp	Specificity	TN/(FP + TN)
FPR	False Positive Rate	FP/(FP + TN)
FNR	False Negative Rate	FN/(TP + FN) = 1-Sn
PPV	**Positive Predictive Value**	**TP/(TP + FP)**
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)

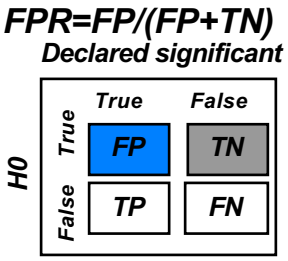


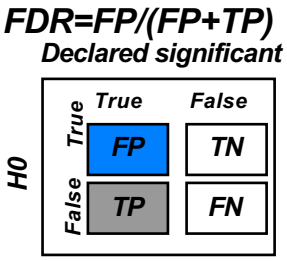


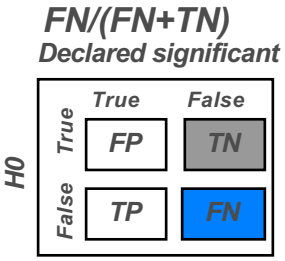


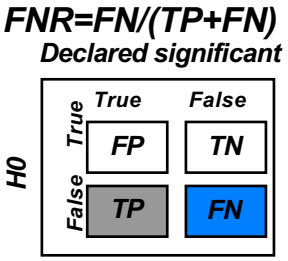








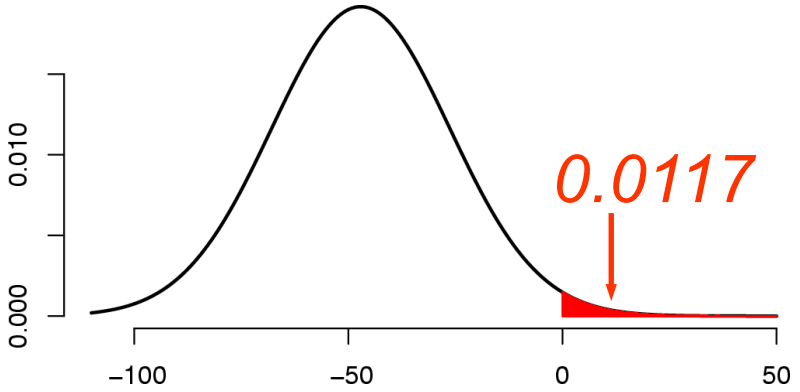




Multiple testing corrections

The problem of multiple testing

- Let us assume that the score of the alignment between two sequences has a p-value.
  - $P(X \geq x) = 0.0117$
- What would happen if we consider this p-value as significant, while scanning a database that contains *N=65,000,000* sequences (the size of Uniprot in 2016) ?
  - The risk of first type error will be challenged *N* type.
  - The significance threshold classically recommended for single tests ($\alpha = 001; \alpha = 0.001$) are this expected to return a large number of false positives.
- This situation of multiple testing is very frequent in bioinformatics.
  - An analysis to detect differentially expressed genes requires thousands of tests.
  - Genome-wide association studies (GWAS) are now routinely performed with SNP chips containing ~1 million SNPs.
  - Sequence similarity searches (e.g. BLAST against the whole Uniprot database) amount to evaluate billions of possible alignments between the query sequence and the millions 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" p-value threshold (e.g. $\alpha = 0.05$) by the number of tests (*N*).
- Notes
  - Bonferonni'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.
  - With the original Bonferoni correction, the displayed p-value is thus still the nominal p-value (i.e. the p-value attached to a single test). This is thus not properly speaking a correction on the p-value, but on the threshold.

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

where
α_n =Nominal p-value
N = Number of tests
α_b = Bonferonni-corrected threshold

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

$$Eval = N \cdot Pval$$

- If p=P(X >0)=0.0117 and the database contains N=200,000 entries, we expect to obtain $N*p = 2340$ 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 :
  - Instead of setting a threshold on the P-value, we should set a threshold on the E-value, which indicates the expected number of false positives for a given P-value.
  - ***Eval = Pval * N***
- **T*he E-value is not a probability***, but a positive Real number.
  - It can in principle take values higher than 1.
  - If we want to avoid false positives, this threshold should always be smaller than 1.
    - ***Threshold(Eval) ≤ 1***
- The fact to set a threshold ≤ 1 on the E-value is equivalent to Bonferroni's correction, which consists in adapting the threshold on the p-value.
- ***Threshold(Pval) $\leq 1/N$***
- The fact to set a threshold ≤ 1 on the E-value is equivalent to Bonferroni's correction, which consists in adapting the threshold on the p-value

Bonferroni-corrected P-value

- The so-called “Bonferroni-corrected” P-value is obtained by computing the E-value ($Eval = N \cdot Pval$) and truncating it to 1.

$$\boldsymbol{P_{Bonferroni}} = \begin{cases} \boldsymbol{N.Pval} & \textbf{if } \boldsymbol{Pval < 1/N} \\ \boldsymbol{1} & \textbf{otherwise} \end{cases}$$

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 positives among all the tests declared significant*** (the "discovered" cases).
  - FP number of false positives
  - TP number of true positives

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

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 ?

- TO BE COMPLETED
- See the practical about multiple testing correction on the supporting Web site.
  - http://pedagogix-tagc.univ-mrs.fr/courses/statistics_bioinformatics/

Fig 1 from Storey and Tibshirani, 2003)

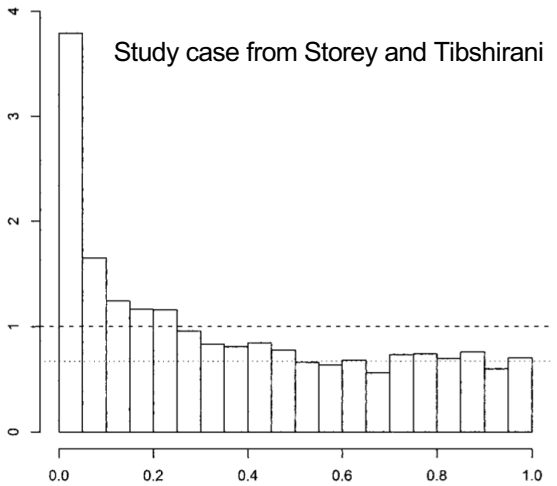
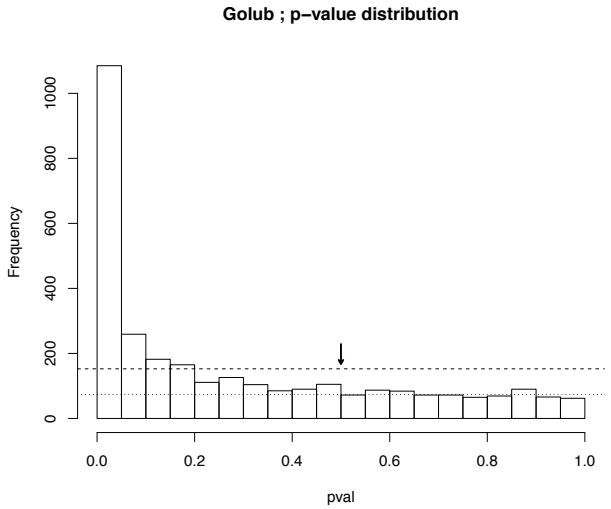
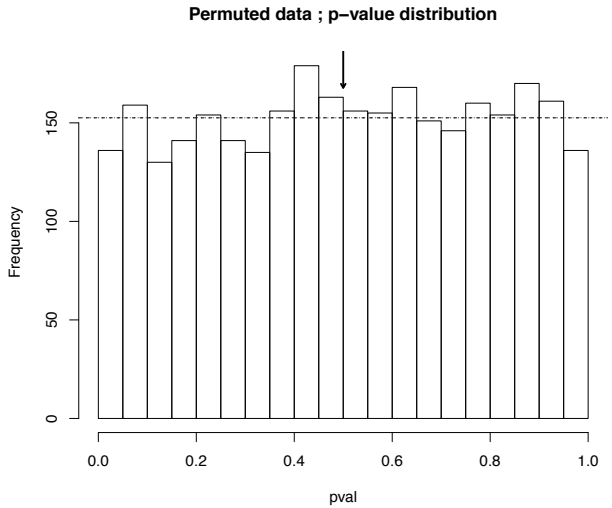


Fig. 1. A density histogram of the 3,170 *p* values from the Hedenfalk *et al.* (14) data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null *p* values.

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



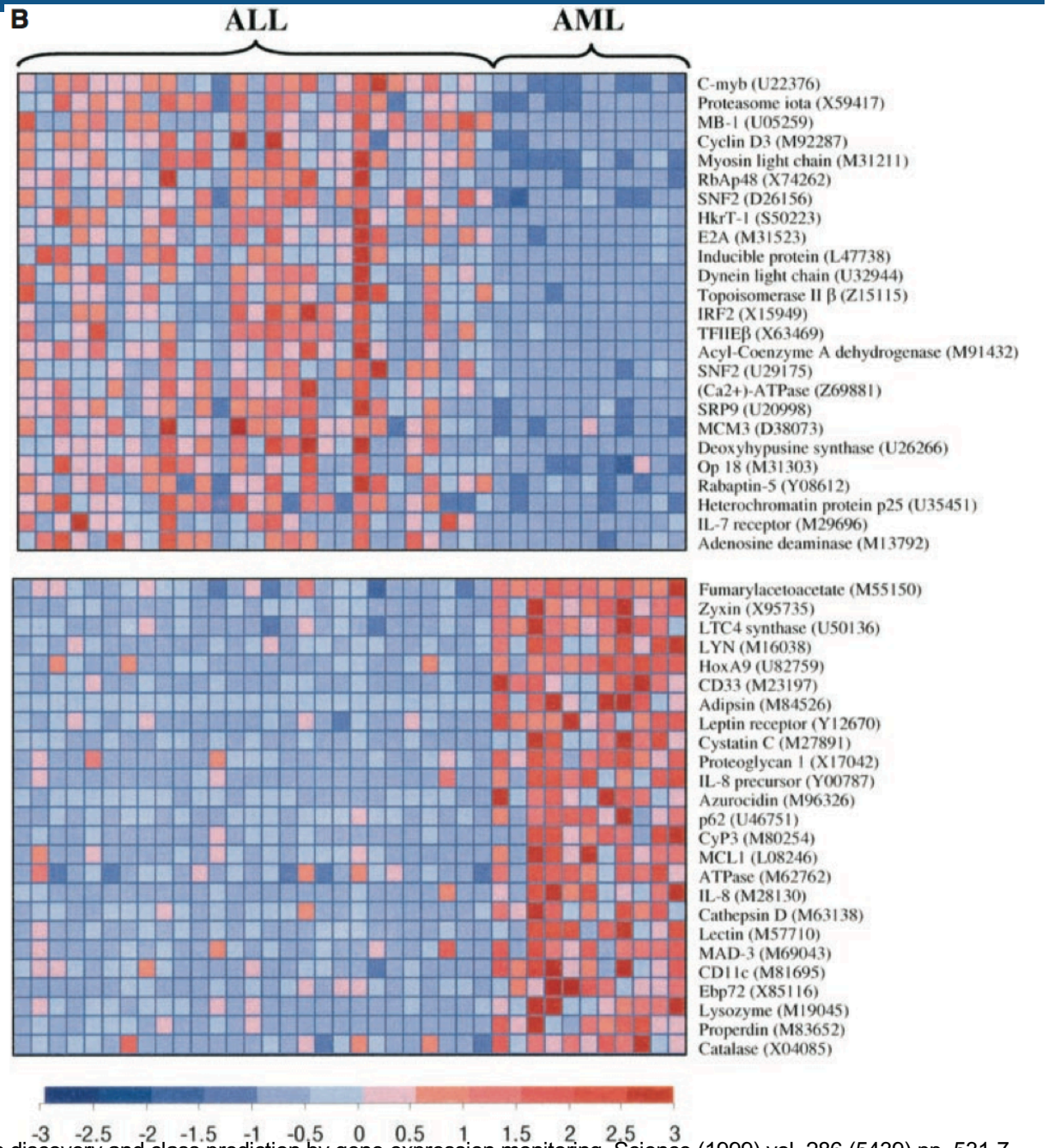
Negative control: permuted data from Golub et al. (1999)



- Storey and Tibshirani. Statistical significance for genomewide studies. Proc Natl Acad Sci USA (2003) vol. 100 (16) pp. 9440-5.

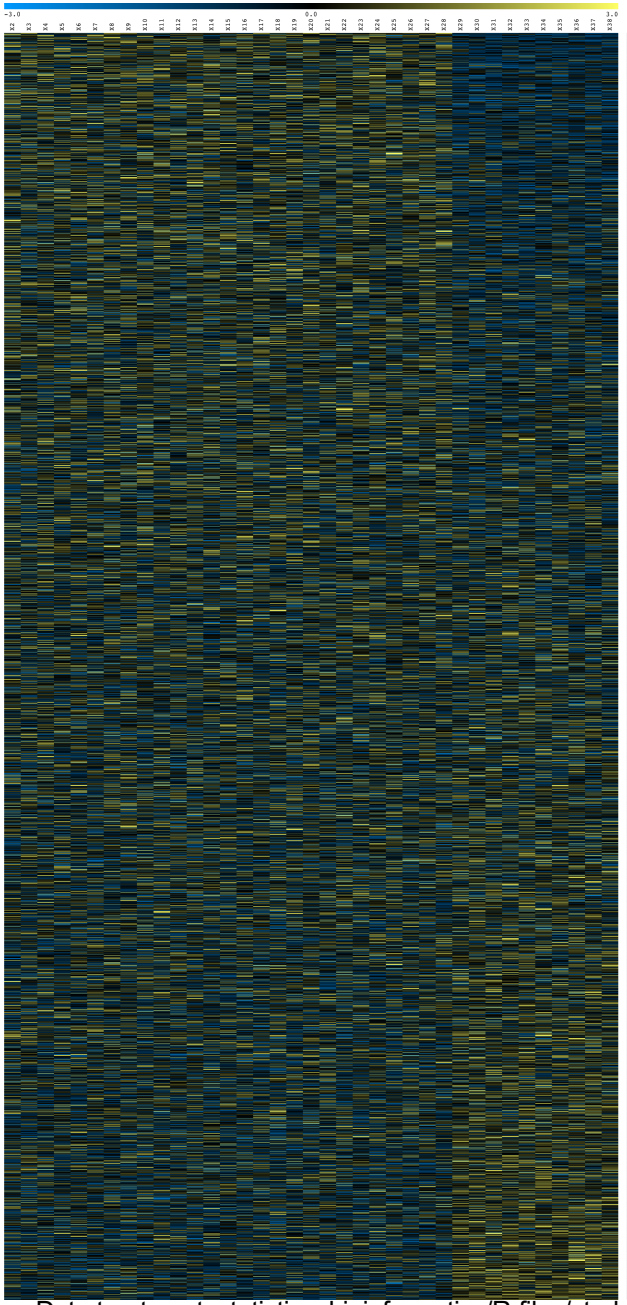
Study case for multitesting corrections

- 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

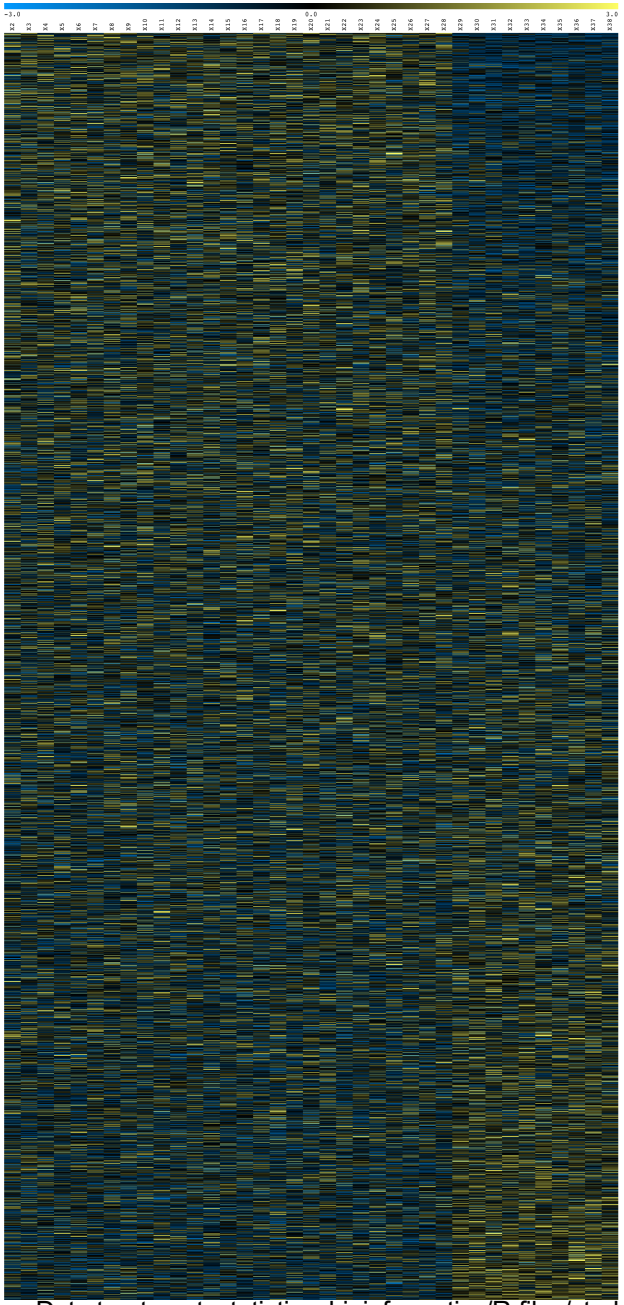
Some expression profiles sorted by difference between means



- 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 <= 0.01

- Data treatment: statistics_bioinformatics/R-files/student_test.R
- Image generated with TMeV

Some expression profiles sorted by difference between means

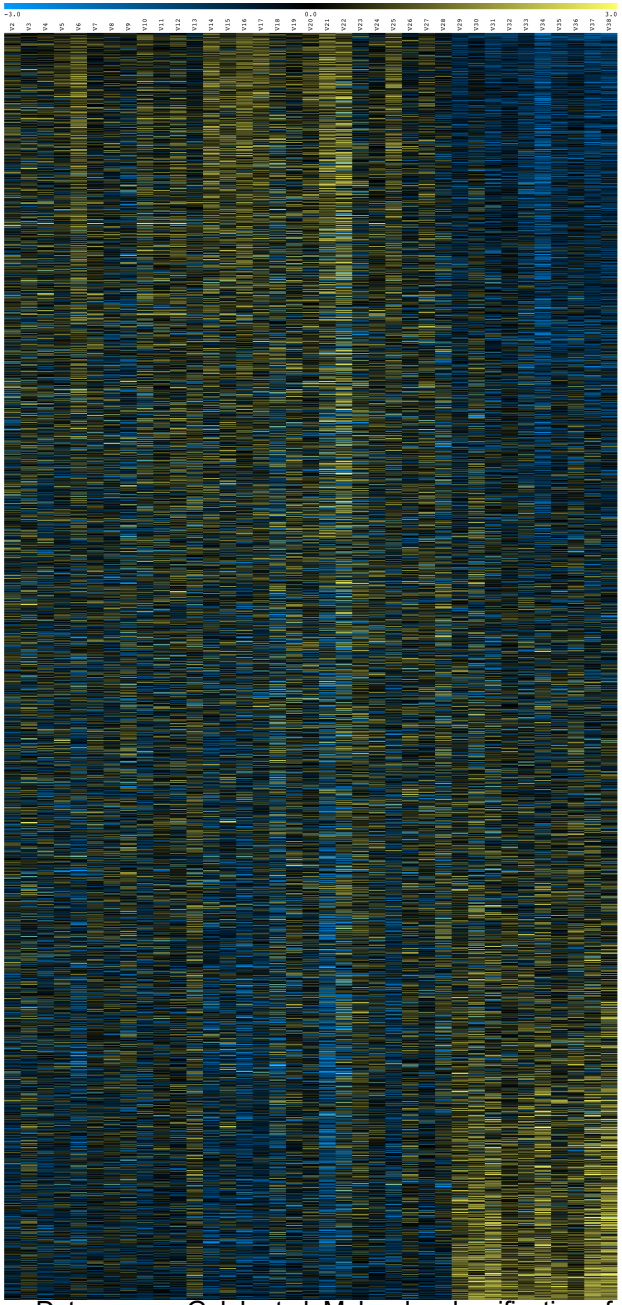


?

- 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



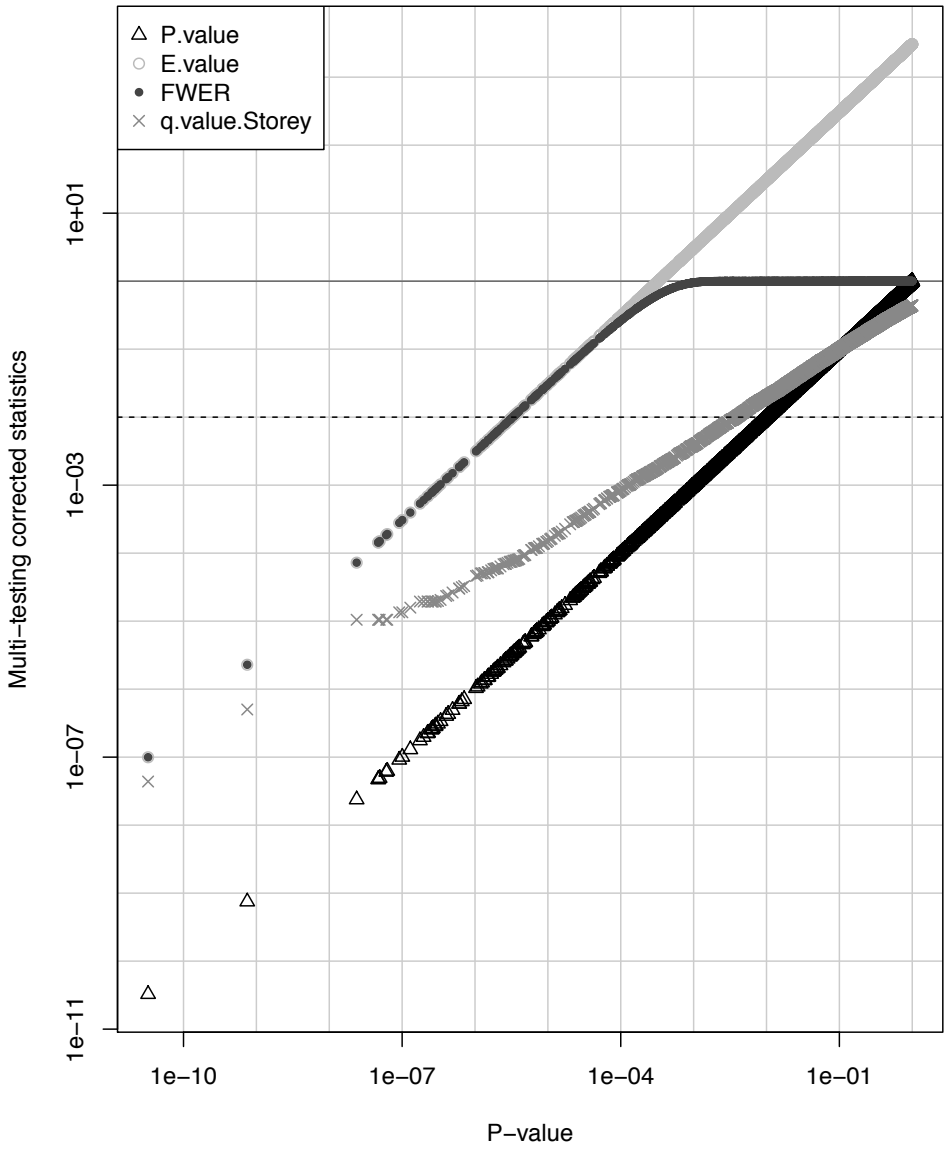
?

- 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

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

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



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

Criterion	genes	E(FP)	
Pval <= 0.01	673	Pval*N = 30	
Bonferoni (Pval <= 0.01/N) = 0.01	57	N*Pval/N	
Eval =Pval*N <= 0.01	57	Eval=0.01	
Eval <= 1	243	Eval = 1	
Qval (m0=N) <= 0.01	366	S*qval = 3.66	
Qval (Storey) <= 0.01	515	S*qval = 5.15