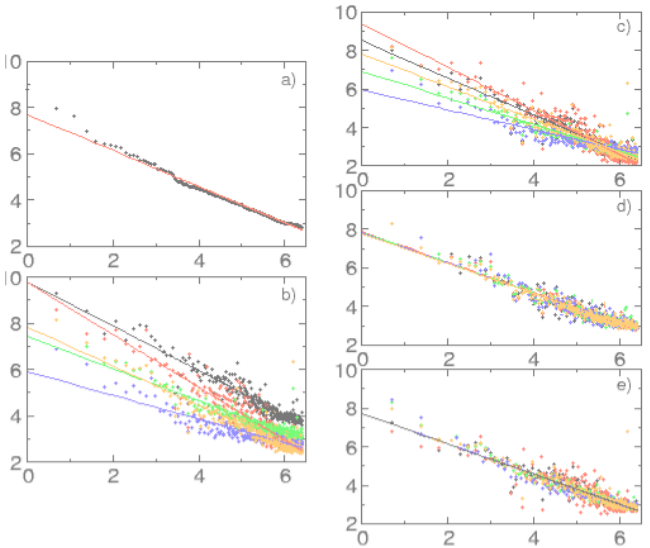
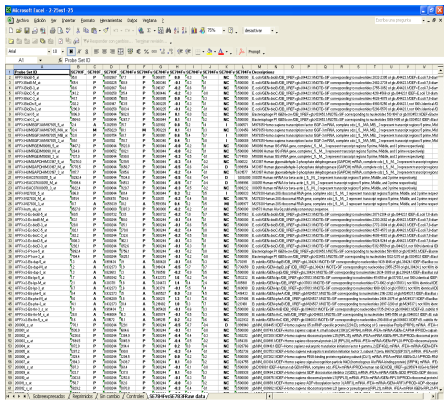
Course on Microarray Gene Expression Analysis

::: Differential Expression Analysis



Daniel Rico
drico@cnio.es

Bioinformatics Unit
CNIO



Upregulation

No Change

Downregulation

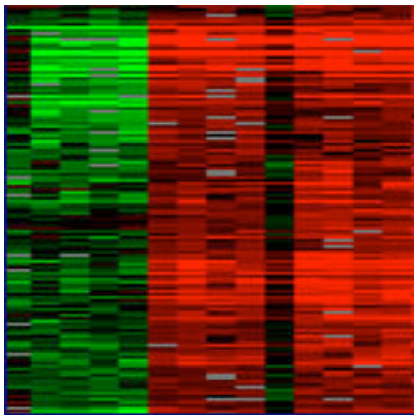
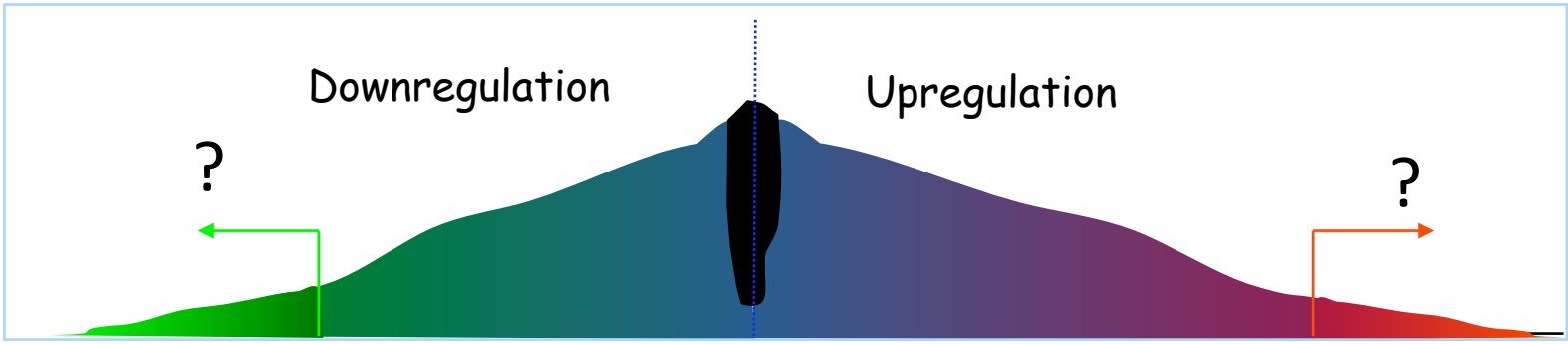
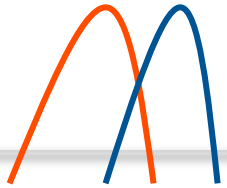


Image analysis comparison
(normalization and filtering)

Data analysis

Ratios (or not…)
Log_2 transform

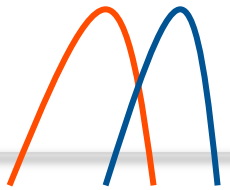




"To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of."

Ronald A. Fisher: Indian Statistical Congress, 1938, vol. 4, p. 17

ASK BEFORE DOING THE EXPERIMENTS!!!!!

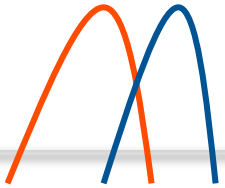


1. **Replication.** It allows the experimenter to obtain an estimate of the experimental error

2. **Randomization.** It requires the experimenter to use a random choice for every factor that is not of interest but might influence the outcome of the experiment. Such factors are called **nuisance factors**. Ex.: printing of replicate spots on the array.

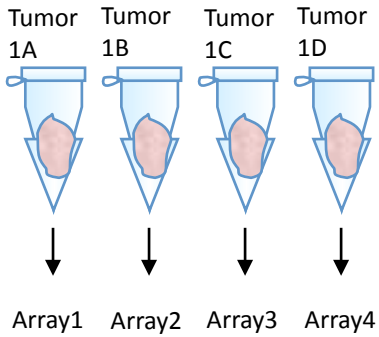
3. **Blocking:** method of createing homogeneous blocks of data in which in which the nuisance factor is kept constant and the factor of interest is allowed to vary. It is used to increase the accuracy with which the influence of the various factors is assessed in a given experiment. Ex.: the microarray slide itself.

"Block what you can, randomize what you cannot"

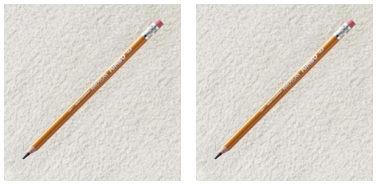


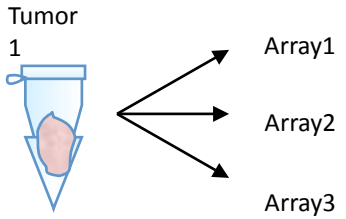
At least **5 replicates por clase** (biological!!!!!)

a) Biological replicates:

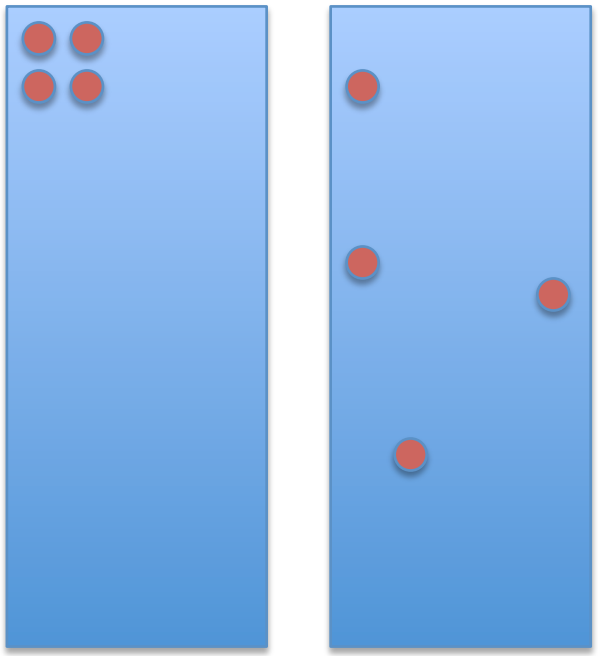


b) Technical replicates:

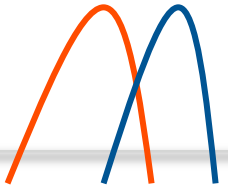




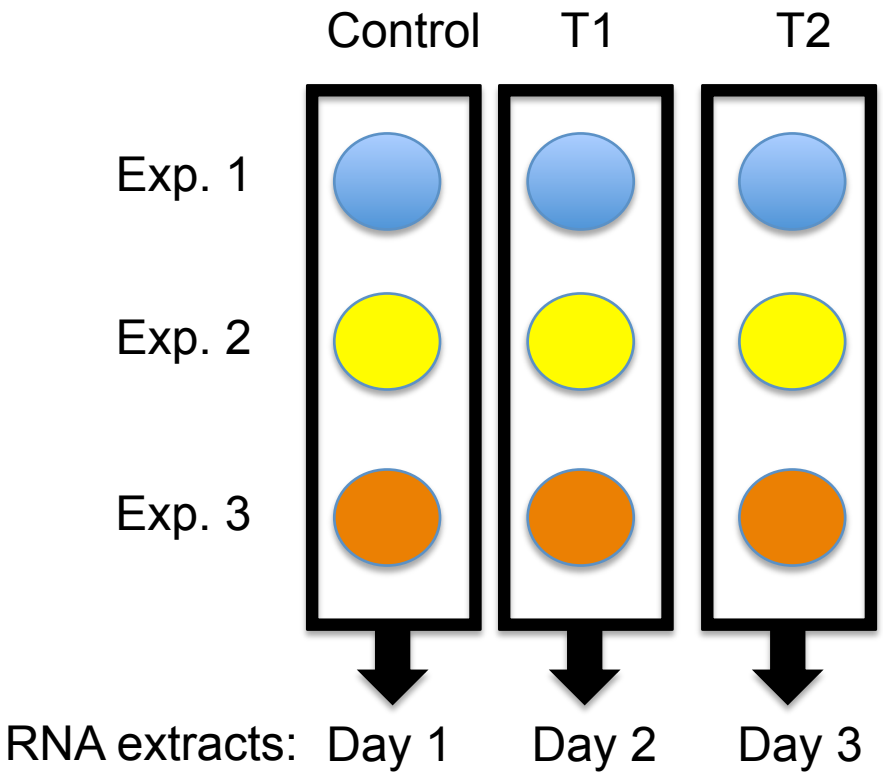
Each gene is spotted in quadruplicate: randomize position in the slide

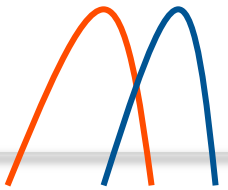


Not randomized Randomized

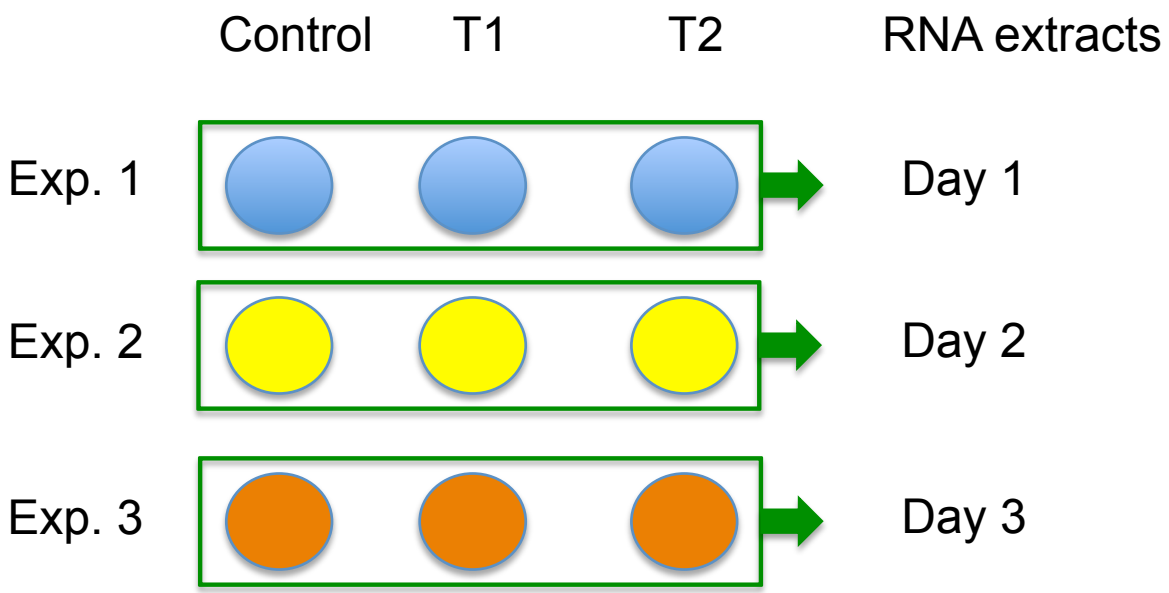


Treatment and RNA extraction days are confounded!!!

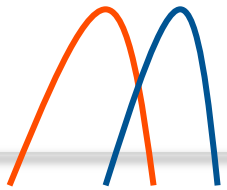




Make coherent blocks:



::: Fold change is NOT the way!

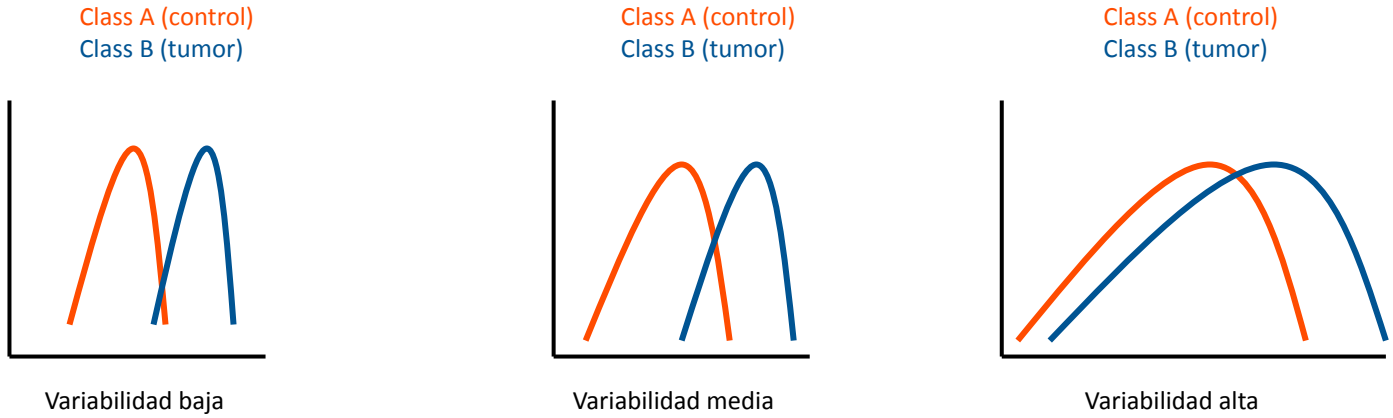


-Fold change: Expression ratio between 2 groups (ie. Tumor/control)

Differentially expressed genes (DEG) are selected if they pass a se *cut-off*

Ej. 2.5 (Schena et al), 3 (DeRisi)

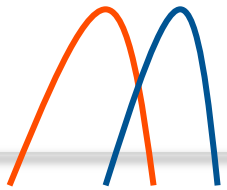
The statistical significance of a change depends on the variability and within group and between groups, and this variability (variance) differs greatly for each gene.



Fold change approach simply ignore this information (that you have!!!)

To test for significant changes, we must perform a statistical test for each gene to obtain a p-value.

::: Nine steps for hypothesis testing

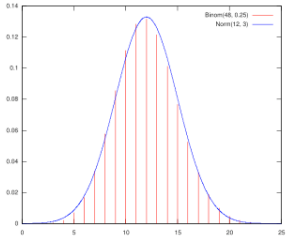


1. State the problem.
2. State the null and alternative hypothesis.
3. Choose the level of significance.
4. Find the appropriate statistical model and test stastistic.
5. Calculate the appropriate test statistic.
6. Determine the p-value of the test statistic (the prob. of it occurring by chance).
7. Compare the p-value with the chosen significance level.
8. Reject or do not reject Ho based on the test above.
9. Answer the question in step 1.

:::Parametric and non parametric methods

Parametric methods

- Assume that the data follow normal distribution.



N(μ=12,σ=3)

T test.
Test difference in means between 2 independent populations with equal variances. Welch T-test for unequal variances.

Paired T test.
T test for paired data (blocks of 2 elements).
Example: Treatment in right arm, left arm as control

ANOVA.
Analysis of variance, for more than 2 populations.

Non parametric methods

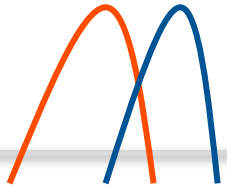
- Appropriate when normality cannont be assumed.
- More robust (less sensitive to outliers).
- Less sensitive than parametric methods to detect significant changes.
- They order the data by expression, and use the rank to test.

Ex. Gene 63; 4 treatments and 5 controls; rank 1,2,3 ,4,5,6,7,8,9

Mann-Whitney test.
Test for differences in medians between two independent populations.

Wilcoxon Signed Rank test.
Non-parametric test equivalent to the paired T test for paired samples (test if median of paired differences is zero)

Kruskal-Wallis.
Non-parametric test equivalent to ANOVA for more than 2 populations.

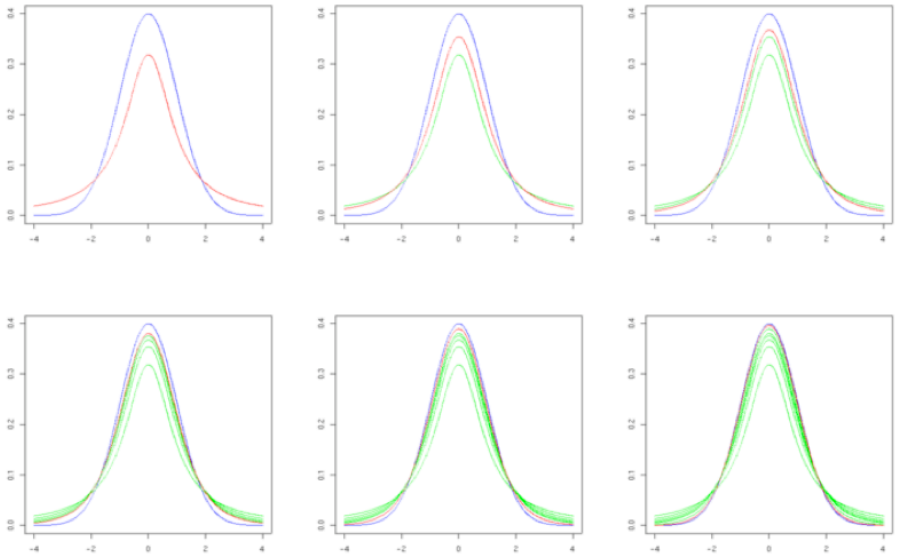


A t-test is any statistical hypothesis test in which the test statistic has a Student's t distribution if the null hypothesis is true. It is applied when the population is assumed to be normally distributed but the sample sizes are small enough that the statistic on which inference is based is not normally distributed because it relies on an uncertain estimate of standard deviation rather than on a precisely known value.

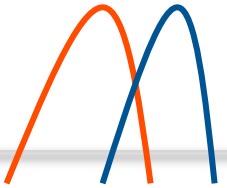
The overall shape of the probability density function of the t-distribution resembles the bell shape of a normally distributed variable with mean 0 and variance 1, except that it is a bit lower and wider. As the number of degrees of freedom grows, the t-distribution approaches the normal distribution with mean 0 and variance 1.

The following images show the density of the t-distribution for increasing values of v. The normal distribution is shown as a blue line for comparison.; Note that the t-distribution (red line) becomes closer to the normal distribution as v increases. For v = 30 the t-distribution is almost the same as the normal distribution.

Density of the *t*-distribution (red and green) for 1, 2, 3, 5, 10, and 30 df compared to normal distribution (blue)



http://en.wikipedia.org/wiki/Student%27s_t-test

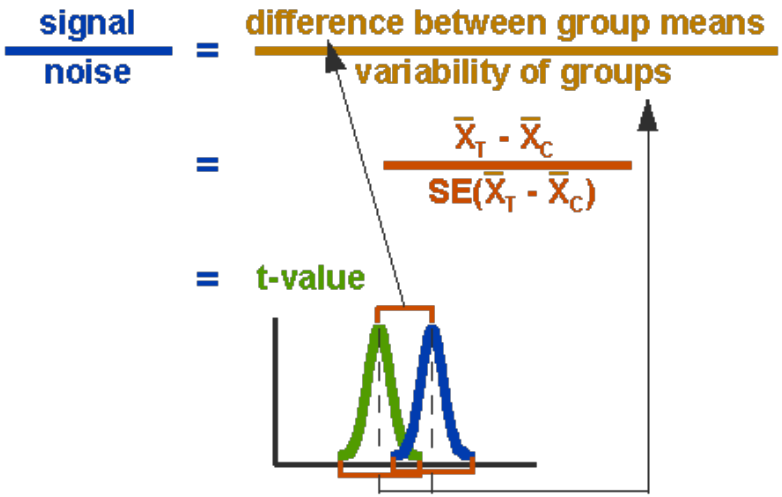


Difference between group means

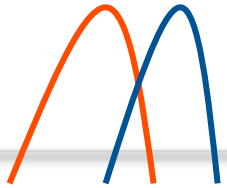
Test statistic → $t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Pooled standard deviation

http://en.wikipedia.org/wiki/Student%27s_t-test



http://www.socialresearchmethods.net/kb/stat_t.php



$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where:

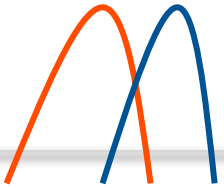
Pooled Standard Deviation

$$S_{X_1X_2} = \sqrt{\frac{(n_1-1)S_{X1}^2 + (n_2-1)S_{X2}^2}{n_1+n_2-2}}.$$

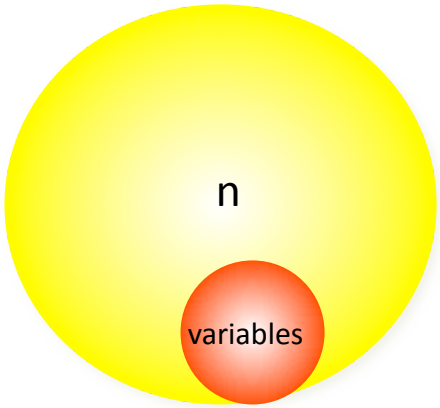
http://en.wikipedia.org/wiki/Pooled_standard_deviation

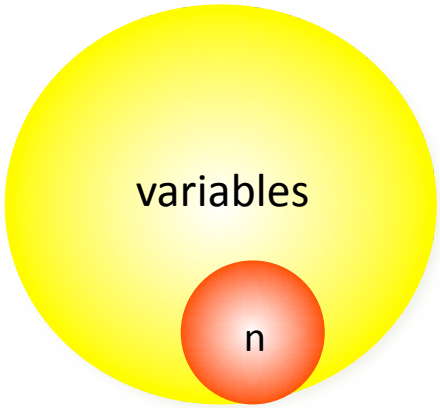
1. Open the file T_test_with_Excel.xls
2. Observe the expression data for the gene AC002378 in controls (C) and tumors (T).
3. See the formula for the “pooled SD” (Standard Deviation).
4. Calculate the t value for the difference between C and T averages (use formula above). *Hints: n1 is 6, n2 is 6, square root in Excel is: SQRT().*
5. Use the function TDIST() to calculate the p-value (probility of observing this value of t by chance. *Hint: degrees of freedom for a T test are: n1 + n2 – 2.*

::: Probems in identifying DEGs with microarrays



Classic statistical analysis





Statistical analysis in microarray scenario

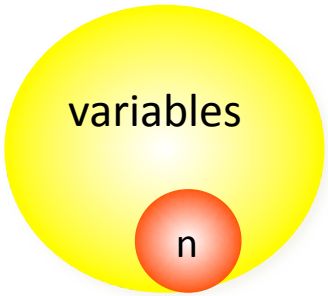
Basic Problem: many genes/tests with few replicates (per sample)
Statisticians prefer: many replicates/observations/data with few genes/conditions/tests to investigate eg by t-test

$$t = \frac{(\bar{x}_T - \bar{x}_c)}{s / \sqrt{1/n_1 + 1/n_2}}$$

Given mean expression values $\bar{x}_T$, $\bar{x}_C$ and variance $\mathbf{s^2}$
Variances/sd's of expression values **poorly** estimated for each gene *i*,

there exist methods to obtain better estimate of variance
eg estimate of $s^{*2} = B\, s_i^2 + (1\text{-}B)\, s^2$

B = weighting (0<B<1)
s_i^2= variance of gene i , s^2= variance of all (other) genes (10000 ?)



LIMMA (Linear Models for Microarray Data, Gordon Smyth 2004):

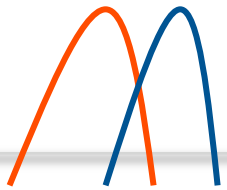
$$t = \frac{(\bar{x}_T - \bar{x}_c)}{s^* / \sqrt{1/n_1 + 1/n_2}}$$

Find $\mathbf{s_i}$ for each gene i – find average s_i over genes = s.

Adequate for small sample sizes (n).
Better estimation of variance, borrowing information from other genes.
Gives less false positives than standard ttest
Allows paired analysis, co-variates and ANOVA (R and Asterias-Pomelo II)

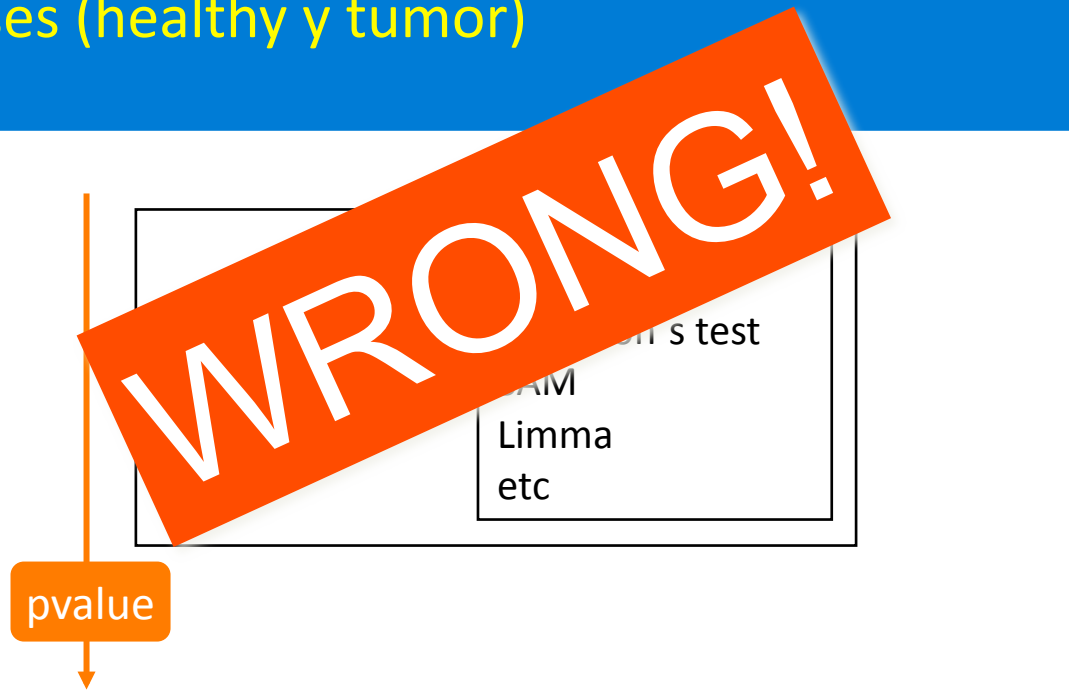
"Assumes normality but performs well generally" (Kim 2006)

SAM (Statistical Analysis of Microarrays, Tusher 2001): another good alternative based on permutations, but need more replicates



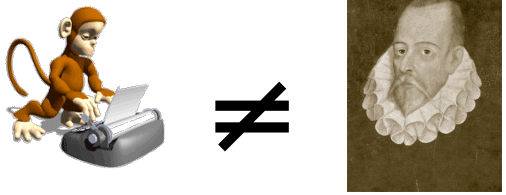
Example

20 normalized arrays
1000 genes
2 classes (healthy y tumor)



Differentially expressed genes between classes

::: Multiple testing: is a monkey able to write a sentence of “El Quijote”?



¿  =  ?

We run into the multiple testing problem:
We are not testing one hypotheses, but many hypotheses one for each gene.

Suppose:

1) 10 independent genes.
So, we have 10 null hypotheses, one for each gene.

2) No significant differences in gene expression between 2 classes (H_0 is true). Thus, the probability that a particular test (say, for gene 3) is declared significant at level 0.05 is exactly 0.05...Good
(Prob of reject H_0 in 1 test if H_0 is true = 0.05)

3) However, the probability of declaring at least one of the 10 hypotheses false (i.e. rejecting at least one, or finding at least one result significant) is:

Source: R. Díaz Uriarte

Pr(at least one null rejected) = 1 - Pr(all p>0.05) =

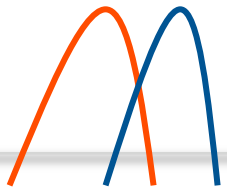
$1 - Pr(1-0.05)^{10} = 1 - 0.95^{10} = 0.401$

The more genes, the more serious is the problem.

In our example....1000 genes...
imagine the number of false positives that we would get without pvalues adjustment...

In summary, without control for multiple testing we would end up rejecting the null much more often than we should.

Source: R. Díaz Uriarte



1. Open a new spreadsheet in Excel.

2. Use the function rand() to generate random numbers between 0 and 1.

3. Generate a random matrix of 6 columns and 100 rows. Select the matrix and “Paste special” the values in another sheet.

4. Considering that the first 3 columns are controls and the other 3 are treatments, calculate a p-value with ttest(). Assume equal variances and select two tails. We will choose the level of significance to be 0.05.

5. Order the data by p-value. How many “genes” would be significantly expressed?

6. And if you extend the random matrix to 10,000 rows?

We want to calculate the number of H_0 that we have declared false (False positives)

We must adjust p-values for multiple testing… How??

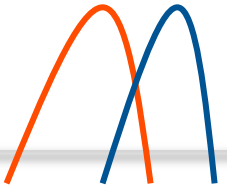
Control of FWER (prob. at list 1 false positive, conservative methods)

Bonferroni
Holm´s Bonferroni Step-Down
Westfall & Young permutation

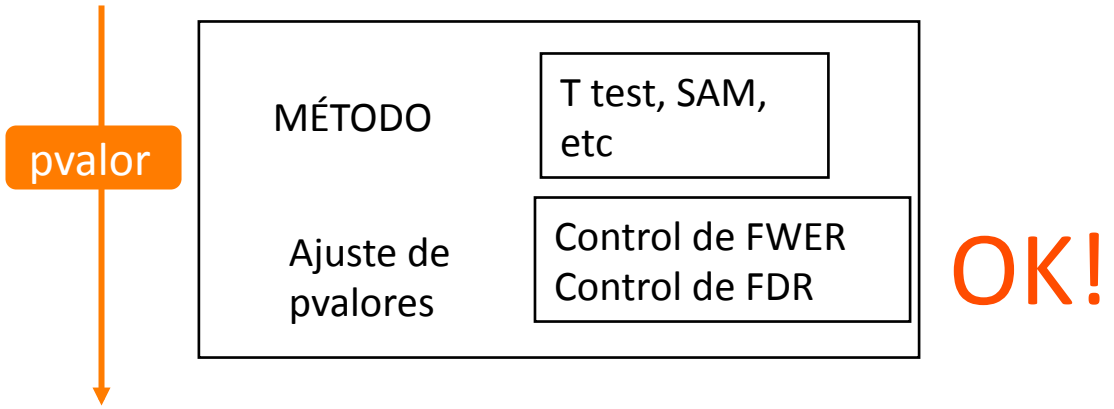
Control of FDR (rate of false positives in the results liberal methods)

Benjamini & Hochberg
Benjamini & Yekutieli

FWER: Type I Family Wise Error Rate
FDR: False Discovery Rate



20 normalized arrays
1000 genes
2 classes (healthy y tumor)



Differentially expressed genes between classes

FWER: Type I Family Wise Error Rate
FDR: False Discovery Rate

Error measures

- FamilyWise Error Rate (FWER). The FWER is defined as the probability of at least one Type I error (false positive): $FWER = \mathbb{P}(V > 0)$.
- False discovery rate (FDR). The FDR (Benjamini & Hochberg 1995) is the expected proportion of Type I errors among the rejected hypotheses, including cases where no hypotheses are significant:
  $FDR = \mathbb{E}\{\frac{V}{R}|R > 0\}\mathbb{P}(R > 0)$.

EXAMPLE: multiple-testing results.

We must used the FDR adjusted p-values!

Public tools:

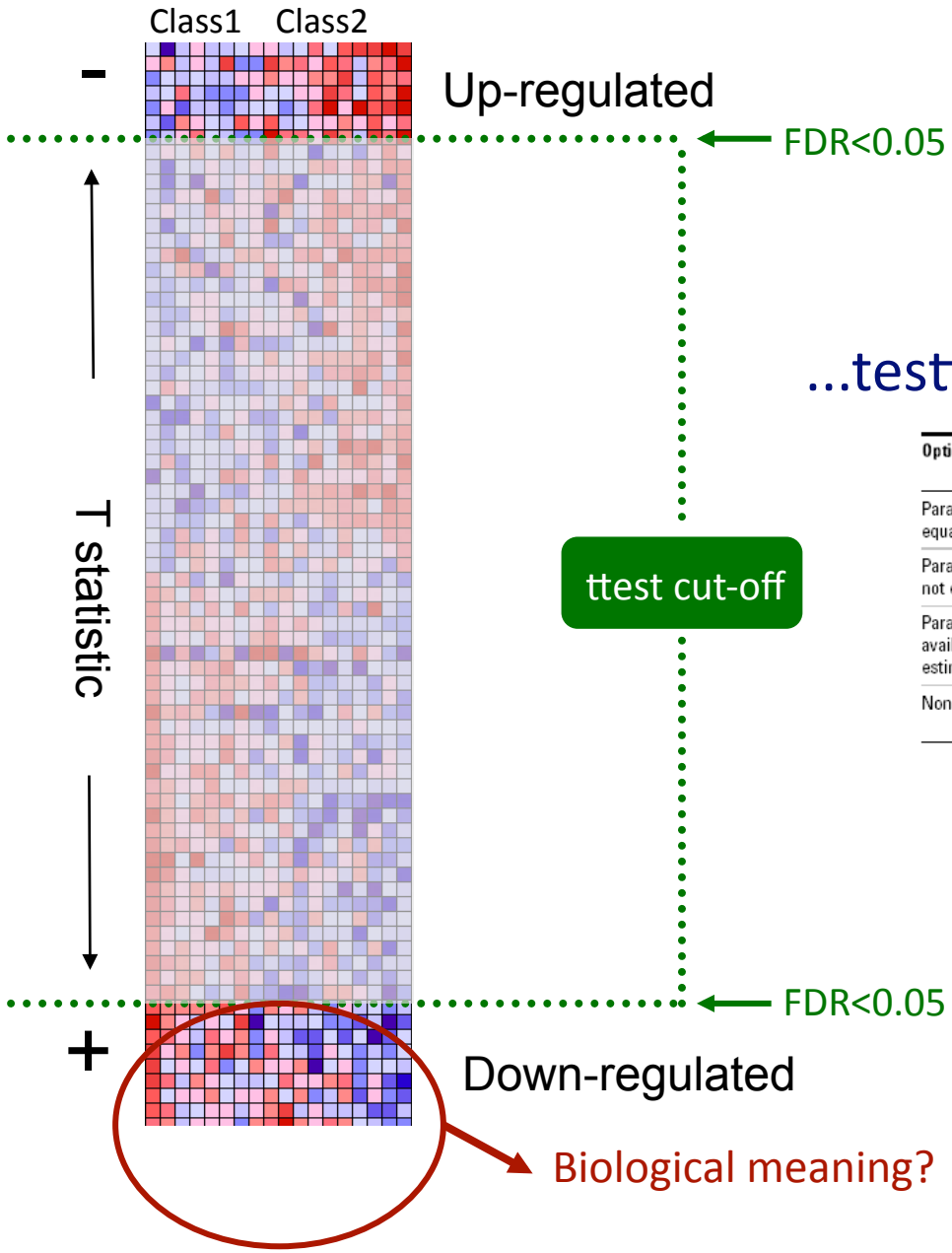
Asterias – POMELO II

GEPAS - TRex

Chosen test type: Limma t-test
Permutations used: Non permutation method
Covariables used: Selected test does not a low additio al covariables
Class labels: Click_here

	Gene Name	Row numbe	unadj.p	FDR_indep	Obs_stat	abs(Obs_stat)
1	H3024E09_	4941	1.0	1.0	-0.0	0.0
2	H3012E07_	4731	0.9991436	0.9992247	-0.001086	0.001086
3	Nup35	9513	0.9990753	0.9992247	0.001173	0.001173
4	Galc	4290	0.9985561	0.9991792	0.001831	0.001831
5	H3106C10_	6138	0.9987465	0.9991792	0.00159	0.00159
6	H4022B06_	7290	0.9989359	0.9991792	0.001349	0.001349
7	H4047C07_	7667	0.9986275	0.9991792	-0.001741	0.001741
8	Impa1	8457	0.9987782	0.9991792	0.001549	0.001549
9	Lrmp	8792	0.9988963	0.9991792	0.0014	0.0014
10	1110008P08Rik	84	0.9984148	0.9991688	-0.00201	0.00201
11	H3127G01_	6458	0.998439	0.9991688	-0.00198	0.00198
12	H4057B01_	7796	0.9982229	0.9991148	-0.002254	0.002254
13	Dapk1	3525	0.9978189	0.9988542	0.002766	0.002766
14	H3148B10_	6781	0.9978814	0.9988542	-0.002687	0.002687
15	Micl	9027	0.9977	0.9988542	-0.002917	0.002917
16	Rasd1	10387	0.9977588	0.9988542	0.002842	0.002842
17	AW146154	1673	0.9975489	0.9988459	-0.003108	0.003108
18	Lrrfip2	8803	0.9974447	0.9988228	-0.00324	0.00324
19	H3133C12_	6536	0.99714	0.9985987	0.003627	0.003627
20	Hmgb1	8234	0.996823	0.9983625	0.004029	0.004029
21	Agl	1834	0.9961663	0.9978287	0.004862	0.004862
22	H3037D05_	5129	0.9962091	0.9978287	-0.004807	0.004807
23	Srp9	11238	0.9959925	0.9977739	-0.005082	0.005082
24	H3076D09_	5745	0.995262	0.9971232	0.006008	0.006008
25	4930506D23Rik	1008	0.9951111	0.9970724	0.0062	0.0062
26	H3028B09_	4994	0.9951304	0.9970724	-0.006175	0.006175
27	Prcc	10071	0.9949421	0.9970724	0.006414	0.006414
28	Prkg1	10109	0.9950367	0.9970724	-0.006294	0.006294
29	H3004H10_	4603	0.9948015	0.9970672	0.006592	0.006592
30	H4059F04_	7836	0.9946651	0.9970116	0.006765	0.006765
31	Adam10	1785	0.9941076	0.9965338	0.007472	0.007472
32	1110067D22Rik	153	0.9938699	0.9964473	0.007774	0.007774

Statistical analysis-DEG



...testing genes independently...

Options	Specific test name: Analyzing 2 groups	Specific test name : Analyzing more than 2 groups
Parametric (variances equal)	Student's T-test	ANOVA
Parametric (variances not equal)	Welch t-test	Welch ANOVA
Parametric (use all available error estimate)	Welch t-test using error model variances	Welch ANOVA using error model variances
Nonparametric	Wilcoxon-Mann-Whitney test	Kruskal-Wallis test

FatiGO

Gene Set Enrichment Analysis - GSEA -