Microarray Data Analysis

Finding Differential Expressed Genes

國立台灣大學資訊所

Course: 生物資訊與計算分子生物學

2007/11/06

吳漢銘
hmwu@stat.sinica.edu.tw
http://idv.sinica.edu.tw/hmwu

中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica

Outlines

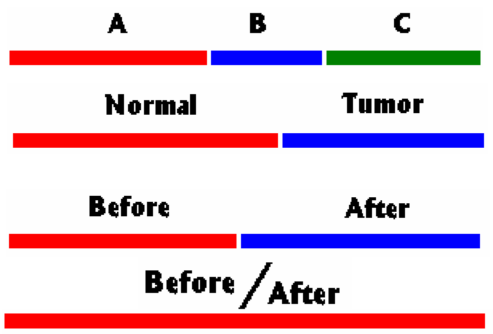
- Introduction: Three Demo Data Sets

- Hypothesis Testing
  - Parametric and Non-parametric Statistics
- Multiple Testing Corrections
- Post Hoc Tests

- Bootstrap Analysis
- Permutation Test

- Significance Analysis of Microarrays (SAM)
- Software: affylmGUI

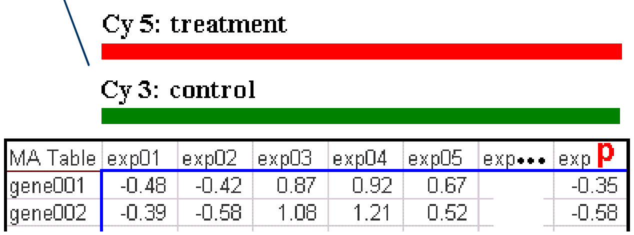
Finding Differentially Expressed Genes



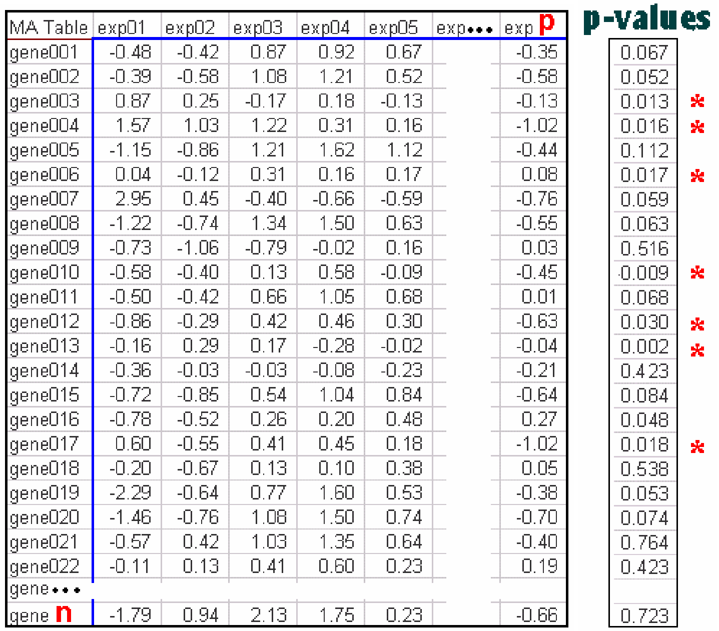
More than two samples

Two-sample (independent)

Paired-sample (dependent)

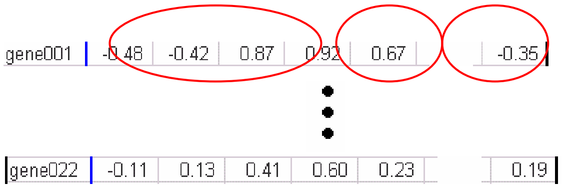

Cy 5: treatment
Cy 3: control

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58



MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p	p-values	
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35	0.067	
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58	0.052	
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13	0.013	*
gene004	1.57	1.03	1.22	0.31	0.16		-1.02	0.016	*
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44	0.112	
gene006	0.04	-0.12	0.31	0.16	0.17		0.08	0.017	*
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76	0.059	
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55	0.063	
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03	0.516	
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45	0.009	*
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01	0.068	
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63	0.030	*
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04	0.002	*
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21	0.423	
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64	0.084	
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27	0.048	
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02	0.018	*
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05	0.538	
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38	0.053	
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70	0.074	
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40	0.764	
gene022	-0.11	0.13	0.41	0.60	0.23		0.19	0.423	
gene•••									
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66	0.723	

Microarray Data Matrix



gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
⋮							
gene022	-0.11	0.13	0.41	0.60	0.23		0.19

- Select a statistic which will rank the genes in order of evidence for differential expression, from strongest to weakest evidence.

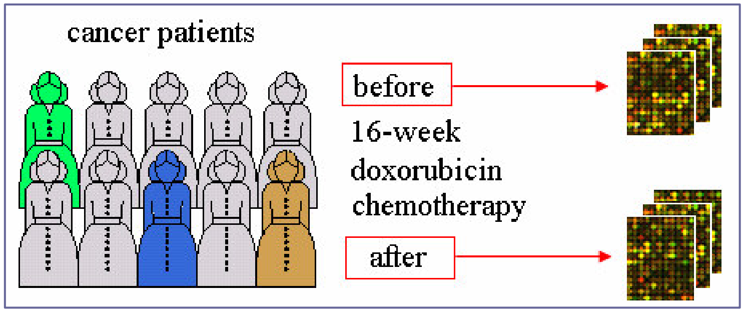
(Primary Importance): only a limited number of genes can be followed up in a typical biological study.

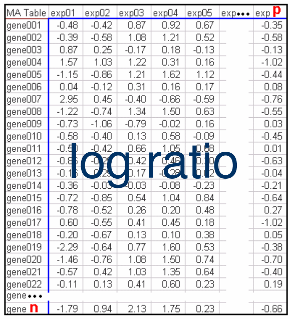
- Choose a critical-value for the ranking statistic above which any value is considered to be significant.

Example 1: Breast Cancer Dataset

cDNA microarrays

- Samples are taken from 20 breast cancer patients, before and after a 16 week course of doxorubicin chemotherapy, and analyzed using microarray. There are 9216 genes.
- **Paired data**: there are two measurements from each patient, one before treatment and one after treatment.
- These two measurements relate to one another, we are interested in the difference between the two measurements (the log ratio) to determine whether a gene has been up-regulated or down-regulated in breast cancer following that treatment.





MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01
gene012	-0.86	-0.29	0.42	0.46	0.10		-0.63
gene013	-0.16	0.25	0.17	-0.28	0.02		-0.04
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40
gene022	-0.11	0.13	0.41	0.60	0.23		0.19
gene•••							
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66

log ratio

9216 x 20

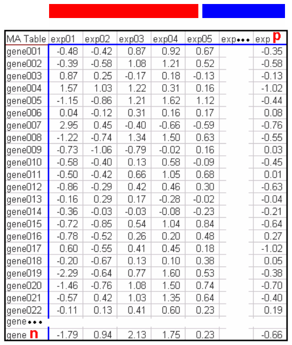
Perou CM, et al, (2000), Molecular portraits of human breast tumours. Nature 406:747-752.
Stanford Microarray Database: http://genome-www.stanford.edu/breast_cancer/molecularportraits/

Example 2: Leukemia Dataset

- Bone marrow samples are taken from
  - 27 patients suffering from acute lymphoblastic leukemia (ALL，急性淋巴細胞白血病) and
  - 11 patients suffering from acute myeloid leukemia (AML，急性骨髓性白血病) and analyzed using Affymetrix arrays.
  - There are 7070 genes.

- **Unpaired data**: there are two groups of patients (ALL, AML).

- We wish to identify the genes that are up- or down-regulated in ALL relative to AML. (i.e., to see if a gene is differentially expressed between the two groups.)



MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40
gene022	-0.11	0.13	0.41	0.60	0.23		0.19
gene•••							
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66

7070 x (27+11)

Golub, T.R et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531--537.

Cancer Genomics Program at Whitehead Institute for Genome Research

http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

Example 3: Small Round Blue Cell Tumors (SRBCT) Dataset

cDNA microarrays

- There are four types of small round blue cell tumors of childhood:
  - Neuroblastoma (NB) (12),
  - Non-Hodgkin lymphoma (NHL) (8),
  - Rhabdomyosarcoma (RMS) (20) and
  - Ewing tumours (EWS) (23).
  - Sixty-three samples from these tumours have been hybridized to microarray.
- We want to identify genes that are differentially expressed in one or more of these four groups.

More on SRBCT:
http://www.thedoctorsdoctor.com/diseases/small_round_blue_cell_tumor.htm

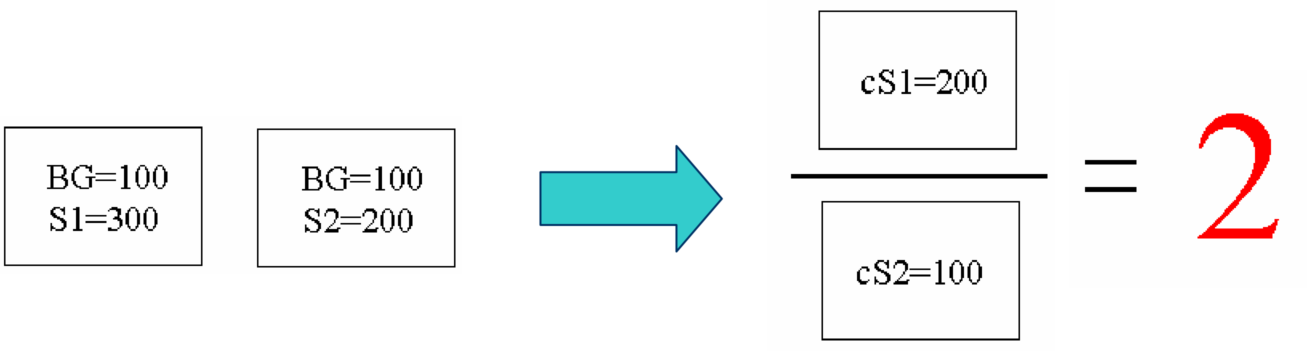
Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C, Peterson C and Meltzer P. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 2001, 7:673-679
Stanford Microarray Database

Fold-Change Method

Calculate the expression ratio in control and experimental cases and to rank order the genes. Chose a threshold, for example at least 2-fold up or down regulation, and selected those genes whose average differential expression is greater than that threshold.

Problems: it is an arbitrary threshold.

- In some experiments, no genes (or few gene) will meet this criterion.
- In other experiments, thousands of genes regulated.



BG=100
S1=300

BG=100
S2=200

cS1=200

cS2=100

= 2

- s2 close to BG, the difference could represent noise.
- It is more credible that a gene is regulated 2-fold with 10000, 5000 units)

- The average fold ratio does not take into account the extent to which the measurements of differential gene expression vary between the individuals being studied.

- The average fold ratio does not take into account the number of patients in the study, which statisticians refer to as the sample size.

Fold-Change Method (conti.)

Define which genes are significantly regulated might be to choose 5% of genes that have the largest expression ratios.

Problems:

- It applies no measure of the extent to which a gene has a different mean expression level in the control and experimental groups.
- Possible that no genes in an experiment have statistically significantly different gene expression.

Hypothesis Testing

A *hypothesis test* is a procedure for determining if an assertion about a characteristic of a population is reasonable.

Example

someone says that the average price of a gallon of regular unleaded gas in Massachusetts is $2.5.

How would you decide whether this statement is true?

- find out what every gas station in the state was charging and how many gallons they were selling at that price.
- find out the price of gas at a small number of randomly chosen stations around the state and compare the average price to $2.5.

- Of course, the average price you get will probably not be exactly $2.5 due to variability in price from one station to the next.

Suppose your average price was $2.23. Is this three cent difference a result of chance variability, or is the original assertion incorrect?

A **hypothesis test** can provide an answer.

CITGO
2.23 9/10 REGULAR UNLEADED
2.33 9/10 MID GRADE UNLEADED
2.43 9/10 PREMIUM UNLEADED

Terminology

- The ***null hypothesis:***
  - H**0**: $\mu = 2.5$. (the average price of a gallon of gas is $2.5)
- The ***alternative hypothesis:***
  - H**1**: $\mu > 2.5$. (gas prices were actually higher)
  - H**1**: $\mu < 2.5$.
  - H**1**: μ != 2.5.
- The ***significance level (alpha)***
  - Alpha is related to the degree of certainty you require in order to reject the null hypothesis in favor of the alternative.
  - Decide in advance to reject the null hypothesis if the probability of observing your sampled result is less than the significance level.
  - Alpha = 0.05: the probability of incorrectly rejecting the null hypothesis when it is actually true is 5%.
  - If you need more protection from this error, then choose a lower value of alpha .

Example

H_0: No differential expressed.

H_0: There is no difference in the mean gene expression in the group tested.

H_0: The gene will have equal means across every group.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ (…= μ_n)

The *p*-values

- *p* is the probability of observing your data under the assumption that the null hypothesis is true.
- *p* is the probability that you will be in error if you reject the null hypothesis.
- *p* represents the probability of **false positives** (Reject $\mathbf{H_0}$ | $\mathbf{H_0}$ true).

p=0.03 indicates that you would have only a 3% chance of drawing the sample being tested if the null hypothesis was actually true.

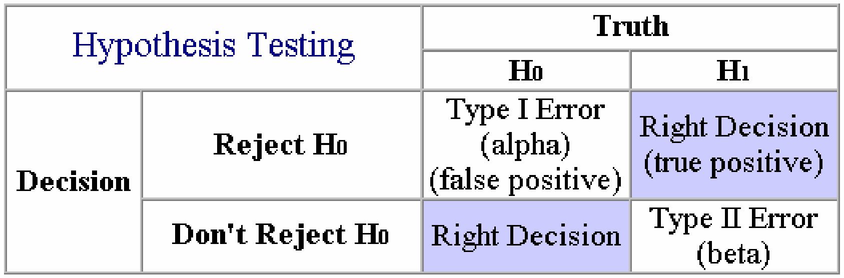
Decision Rule

- Reject H_0 if *P* is less than alpha.
- $P < 0.05$ commonly used. (Reject $\mathbf{H_0}$**,** the test is significant)
- The lower the p-value, the more significant the difference between the groups.

P is *not* the probability that the null hypothesis is true!

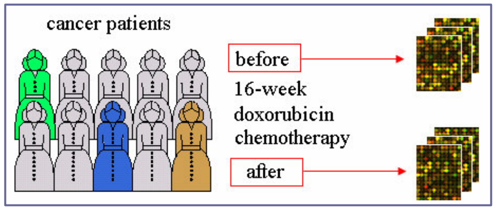
$$\text{Power} = 1 - \beta .$$

Type I Error (alpha): calling genes as differentially expressed when they are NOT
Type II Error: NOT calling genes as differentially expressed when they ARE

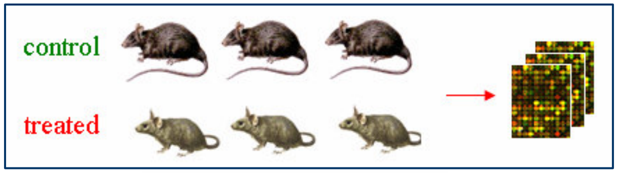


Hypothesis Testing		Truth	
		H_0	H_1
Decision	Reject H_0	Type I Error (alpha) (false positive)	Right Decision (true positive)
	Don't Reject H_0	Right Decision	Type II Error (beta)

Hypothesis Testing



Dependent samples



Independent samples

Comparison	Two Groups		More than two Groups
Hypothesis Testing	Paired data	Unpaired data	Complex data
Parametric (variance equal)	One sample t-test	Two-sample t-test	One-Way Analysis of Variance (ANOVA)
Parametric (variance not equal)	Welch t-test		Welch ANOVA
Non-Parametric (無母數檢定)	Wilcoxon Signed-Rank Test	Wilcoxon Rank-Sum Test (Mann-Whitney U Test)	Kruskal-Wallis Test

Steps of Hypothesis Testing

1. Determine the null and alternative hypothesis, using mathematical expressions if applicable.

2. Select a significance level (alpha).

3. Take a random sample from the population of interest.

4. Calculate a test statistic from the sample that provides information about the null hypothesis.

5. Decision

Hypothesis Testing:

two-sided z-test & p-value

H$_0$: μ=35 null hypothesis

H$_1$: $\mu \neq$35 alternative hypothesis (μ>35; μ<35)
one-sided

α signifcant level : =0.05

test statistic $z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

Reject H$_0$ if $|z| > z_{0.05}$

$$H_0 : \mu = m$$
$$H_1 : \mu \neq m$$
$$\alpha = P_{H_0}(|Z| > z_{\alpha/2})$$

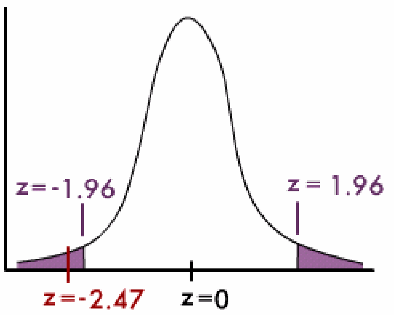
Sample Data: =33.6
test statistic: z=-2.47

$(1-\alpha)100\%$ Confidence Interval:

$$P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha$$

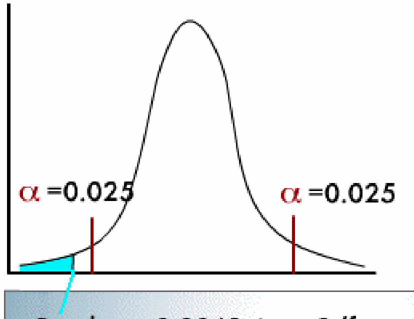
$$\text{p-value} = P_{H_0}(|Z| > z_0),\ z_0 = \frac{\bar{X}-m}{\sigma/\sqrt{n}}$$

The Classical Approach



Conclusion: since the z value of the test statistic (-2.47) is less than the critical value of z= -1.96, we reject the null hypothesis.

The *P*-Value Approach



P -value = 0.0068 times 2 (for a 2-sided test) = 0.0136

Conclusion: since the *P* -value of 0.0136 is less than the significance level of α=0.05, we reject the null hypothesis.

If A Result is Statistically Significant

There are two possible explanations:

- The populations are identical, so there really is no difference.
  - By chance, you obtained larger values in one group and smaller values in the other.
  - Finding a statistically significant result when the populations are identical is called making a Type I error (false positives).
  - If you define statistically significant to mean "P<0.05", then you'll make a Type I error in 5% of experiments where there really is no difference.

OR

- The populations really are different, so your conclusion is correct.
  - The difference may be large enough to be scientifically interesting.
  - Or it may be tiny and trivial.

Hypothesis Tests on Microarray Data

- The null hypothesis is that there is no biological effect.
  - For a gene in Breast Cancer Dataset, it would be that this gene is not differentially expressed following doxorubicin chemotherapy.
  - For a gene in Leukemia Dataset, it would be that this gene is not differentially expressed between ALL and AML patients.
- If the null hypothesis were true, then the variability in the data does not represent the biological effect under study, but instead results from difference between individuals or measurement error.
- The smaller the p-value, the less likely it is that the observed data have occurred by chance, and the more significant the result.
- *p=0.01* would mean there is a 1% chance of observing at least this level of differential gene expression by random chance.
- We then select differentially expressed genes not on the basis of their fold ratio, but on the basis of their p-value.

H_0: no differential expressed.

- The test is significant
  = Reject **H_0**
- False Positive
  = (Reject **H_0** | **H_0** true)
  = concluding that a gene is differentially expressed when in fact it is not.

- **A p-value=0.05 indicates that you would have only a 5% chance of drawing the sample being tested if the null hypothesis was actually true.**
- **The p-value is the smallest level of significance at which a null hypothesis may be rejected**

Statistical Power

- Question: What if I do a t-test on a pair of samples and fail to reject the null hypothesis--does this mean that there is no significant difference?
- Answer: Maybe yes, maybe no.

- For two-sample t-test, power is the probability of rejecting the hypothesis that the means are equal when they are in fact not equal. P(RH0 | not H0)
  Power is one minus the probability of Type-II error.

- The power of the test depends upon the sample size, the magnitudes of the variances, the alpha level, and the actual difference between the two population means.
- Usually you would only consider the power of a test when you failed to reject the null hypothesis.

- High power is desirable (0.7 to 1.0). High power means that there is a high probability of rejecting the null hypothesis when the null hypothesis is false.

- This is a critical measure of precision in hypothesis testing and needs to be considered with care.

One Sample t-test

The One-Sample t-test compares the mean score of a sample to a known value. Usually, the known value is a population mean.

Assumption: the variable is normally distributed.

One sample t-test

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$ (two-tailed).

μ: population mean.

α: significant level (e.g., 0.05).

Test Statistic:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \quad t_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$\bar{X}$: sample mean.

S: sample standard deviation.

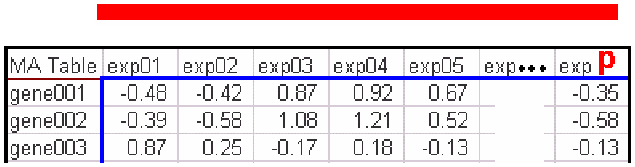
n: number of observations in the sample.

- Reject H_0 if $|t_0| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ: $\bar{X} - t_{\alpha/2} S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2} S/\sqrt{n}$
- p-value $= P_{H_0}(|\mathbf{T}| > t_0)$, $\mathbf{T} \sim t_{n-1}$.

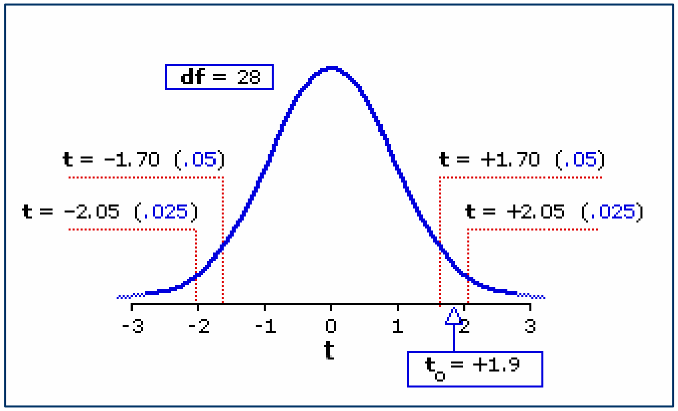
Question

■ whether a gene is differentially expressed for a condition with respect to baseline expression?

■ H_0: $\mu = 0$ (log ratio)



MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13



Two Sample t-test

Paired Sample t-test

$H_0 : \mu_d = \mu_0$
$H_1 : \mu_d \neq \mu_0$ (two-tailed).
μ_d: mean of population differences.
α: significant level (e.g., 0.05).
Test Statistic:

$$T_d = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}, \quad t_d = \frac{\bar{d} - \mu_0}{S_d/\sqrt{n}}$$

$\bar{d}$: average of sample differences.

S_d: standard deviation of sample difference

n: number of pairs.

- Reject H_0 if $|t_d| > t_{\alpha/2, n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ_d: $\bar{d} - t_{\alpha/2} S/\sqrt{n} \leq \mu_d < \bar{d} + t_{\alpha/2} S/\sqrt{n}$
- p-$value = P_{H_0}(|\mathbf{T}| > t_d)$, $\mathbf{T} \sim t_{n-1}$.

Two Sample t-test (Unpaired)

$H_0 : \mu_x - \mu_y = \mu_0$
$H_0 : \mu_x - \mu_y \neq \mu_0$

α: significant level (e.g., 0.05).

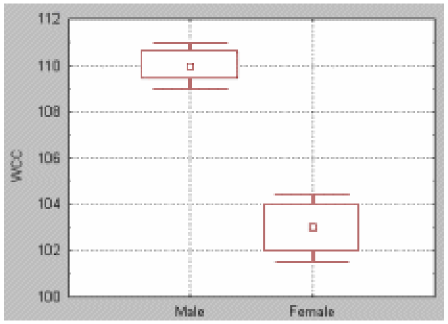
Test Statistic:

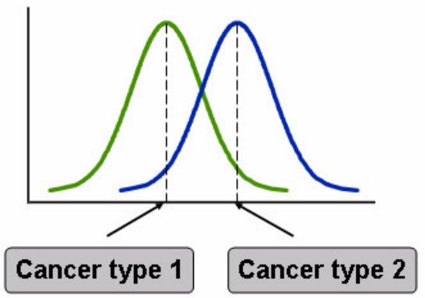
$$t_0 = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

for homogeneous variances:
$df = n + m - 2$

for heterogeneous variances:
adjusted df

Reject H_0 if $|t_0| > t_{\alpha/2,\, df}$





Paired t-test

Applied to a gene From Breast Cancer Data

- The gene acetyl-Coenzyme A acetyltransferase 2 (ACAT2) is on the microarray used for the breast cancer data.
- We can use a paired t-test to determine whether or not the gene is differentially expressed following doxoruicin chemotherapy.
- The samples from before and after chemotherapy have been hybridized on separate arrays, with a reference sample in the other channel.
  - Normalize the data.
  - Because this is a reference sample experiment, we calculate the log ratio of the experimental sample relative to the reference sample for before and after treatment in each patient.
  - Calculate a single log ratio for each patient that represents the difference in gene expression due to treatment by subtracting the log ratio for the gene before treatment from the log ratio of the gene after treatment.
  - Perform the t-test. t=3.22 compare to t(19).
  - The p-value for a two-tailed one sample t-test is 0.0045, which is significant at a 1% confidence level.
- Conclude: this gene has been significantly down-regulated following chemotherapy at the 1% level.

Unpaired t-test

Applied to a Gene From Leukemia Dataset

- The gene metallothionein IB is on the Affymetrix array used for the leukemia data.
  - To identify whether or not this gene is differentially expressed between the AML and ALL patients.
  - To identify genes which are up- or down-regulation in AML relative to ALL.
- Steps
  - the data is log transformed.
  - t=-3.4177, p=0.0016
- Conclude that the expression of metallothionein IB is significantly higher in AML than in ALL at the 1% level.

Assumptions of t-test

- The distribution of the data being tested is normal.
  - For paired t-test, it is the distribution of the subtracted data that must be normal.
  - For unpaired t-test, the distribution of both data sets must be normal.
- **Plots**: Histogram, Density Plot, QQplot,…
- **Test for Normality**: Jarque-Bera test, Lilliefors test, Kolmogorov-Smirnov test.

- Homogeneous: the variances of the two population are equal.
- Test for equality of the two variances: Variance ratio F-test.

Note:

◆ If the two populations are symmetric, and if the variances are equal, then the *t* test may be used.

◆ If the two populations are symmetric, and the variances are not equal, then use the two-sample unequal variance t-test or Welch's *t* test.

One-Way ANOVA

Using Analysis of Variance, which can be considered to be a generalization of the *t*-test, when

- compare more than two groups (e.g., *drug 1*, *drug 2*, and *placebo*), or
- compare groups created by more than one independent variable while controlling for the separate influence of each of them (e.g., *Gender*, *type of Drug*, and *size of Dose*).

- For two group comparisons, ANOVA will give results identical to a *t*-test.
- One-way ANOVA compares groups using one parameter.

- We can test the following:
  - Are all the means from more than two populations equal?
  - Are all the means from more than two treatments on one population equal? (This is equivalent to asking whether the treatments have any overall effect.)

This comparison is performed for each gene.

One-Way ANOVA (conti.)

Assumptions

- The subjects are sampled randomly.
- The groups are independent.
- The population variances are homogenous.
- The population distribution is normal in shape.

As with t tests, violation of homogeneity is particularly a problem when we have quite different sample sizes.

Homogeneity of variance test

- Bartlett's test (1937)
- Levene's test (Levene 1960)
- O'Brien (1979)

ANOVA Table

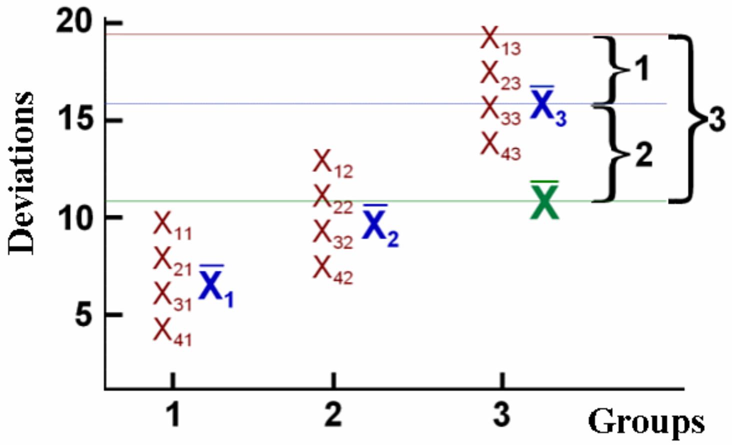
Groups

1	2	… j …	k
X_{11}	X_{12}	$\cdots$ X_{1j} $\cdots$	X_{1k}
X_{21}	X_{22}	$\cdots$ X_{2j} $\cdots$	X_{2k}
		$\cdots$	
X_{i1}	X_{i2}	$\cdots$ X_{ij} $\cdots$	X_{ik}
$\vdots$			$X_{n_k k}$
	$X_{n_2 2}$	$\cdots$ $\vdots$ $\cdots$	
$X_{n_1 1}$		$X_{n_j j}$	

$$T_j = \sum_{i=1}^{n_j} X_{ij} \qquad \bar{X}_j = \frac{T_j}{n_j}$$

$$T = \sum_{j=1}^{k} T_j \qquad \bar{X} = \frac{T}{N}$$

$$S^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} \frac{(X_{ij}-\bar{X})^2}{N-1}$$



$$(X_{ij}-\bar{X}) = (X_{ij}-\bar{X}_j) + (\bar{X}_j - \bar{X})$$

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$X_{ij} = \mu_j + \epsilon_{ij} \qquad i = 1, \cdots, n_j \quad j = 1, \cdots, k$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}[(X_{ij}-\bar{X}_j)+(\bar{X}_j-\bar{X})]^2$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}_j-\bar{X})^2$$

ANOVA Table

Source	SS	df	MS	F	p
Between	SS_B	$p-1$	MS_B	MS_B/MS_W	< 0.05
Within	SS_W	$N-p$	MS_W		
Total	SS_T	$N-1$			

$$SS_{Total} = SS_{Within} + SS_{Between}$$

$$F = \frac{MS_{Between}}{MS_{Within}}$$

Reject H_0, if $F_{obs} > F_{\{\alpha,k-1,N-k\}}$

Welch ANOVA

Welch's F Test

- Use when the sample sizes are unequal.
- Use when the sample sizes are equal but small.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$X_{ij} = \mu_j + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma_j^2)$$

$$i = 1, \cdots, n_j$$

$$j = 1, \cdots, k$$

$$\bar{X}' = \frac{\sum_{j=1}^{k} w_j \bar{X}_j}{\sum_{j=1}^{k} w_j}$$

$$F' = \frac{\frac{\sum_{j=1}^{k} w_j (\bar{X}_j - \bar{X}')^2}{k-1}}{1 + \frac{2(k-2)}{k^2-1} \sum_{j=1}^{k} (\frac{1}{n_j - 1})(1 - \frac{w_j}{\sum_{j=1}^{k} w_j})^2}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2}{n_j - 1}$$

$$df' = \frac{k^2 - 1}{3 \sum_{j=1}^{k} (\frac{1}{n_j - 1})(1 - \frac{w_j}{\sum_{j=1}^{k} w_j})^2}$$

$$w_j = \frac{n_j}{s_j^2}$$

$$\text{Reject } H_0, \text{ if } F'_{obs} > F_{\{\alpha, k-1, df'\}}$$

Other t-Statistics

B-statistic

Lonnstedt and Speed, Statistica Sinica 2002: parametric empirical Bayes approach.

- B-statistic is an estimate of the posterior log-odds that each gene is DE.
- B-statistic is equivalent for the purpose of ranking genes to the penalized t-statistic $t = \frac{\bar{M}}{\sqrt{(a+s^2)/n}}$, where a is estimated from the mean and standard deviation of the sample variances s^2.

$$M_{gj}|\mu_g, \sigma_g \sim N(\mu_g, \sigma_g^2)$$

$$B_g = \log \frac{P(\mu_g \neq 0|M_{gj})}{P(\mu_g = 0|M_{gj})}$$

Penalized t-statistic

Tusher et al (2001, PNAS, SAM)

Efron et al (2001, JASA)

$$t = \frac{\bar{M}}{(a+s)/\sqrt{n}}$$

Lonnstedt, I. and Speed, T.P. Replicated microarray data. *Statistica Sinica* , 12: 31-46, 2002

General Penalized t-statistic

(Lonnstedt et al 2001)

$$t = \frac{b}{s^* \times SE}$$

multiple regression model

Penalized two-sample t-statistic

$$t = \frac{\bar{M}_A - \bar{M}_B}{s^* \times \sqrt{1/n_A + 1/n_B}}, \quad \text{where } s^* = \sqrt{a + s^2}$$

Robust General Penalized t-statistic

Non-parametric Statistics

- Do not assume that the data is normally distributed.
- There are two good reasons to use non-parametric statistic.
  - *Microarray data is noisy:*
    - there are many sources of variability in a microarray experiment and outliers are frequent.
    - The distribution of intensities of many genes may not be normal.
    - Non-parametric methods are robust to outliers and noisy data.
  - *Microarray data analysis is high throughput:*
    - When analysising the many thousands of genes on a microarray, we would need to check the normality of every gene in order to ensure that t-test is appropriate.
    - Those genes with outliers or which were not normally distributed would then need a different analysis.
    - It makes more sense to apply a test that is distribution free and thus can be applied to all genes in a single pass.

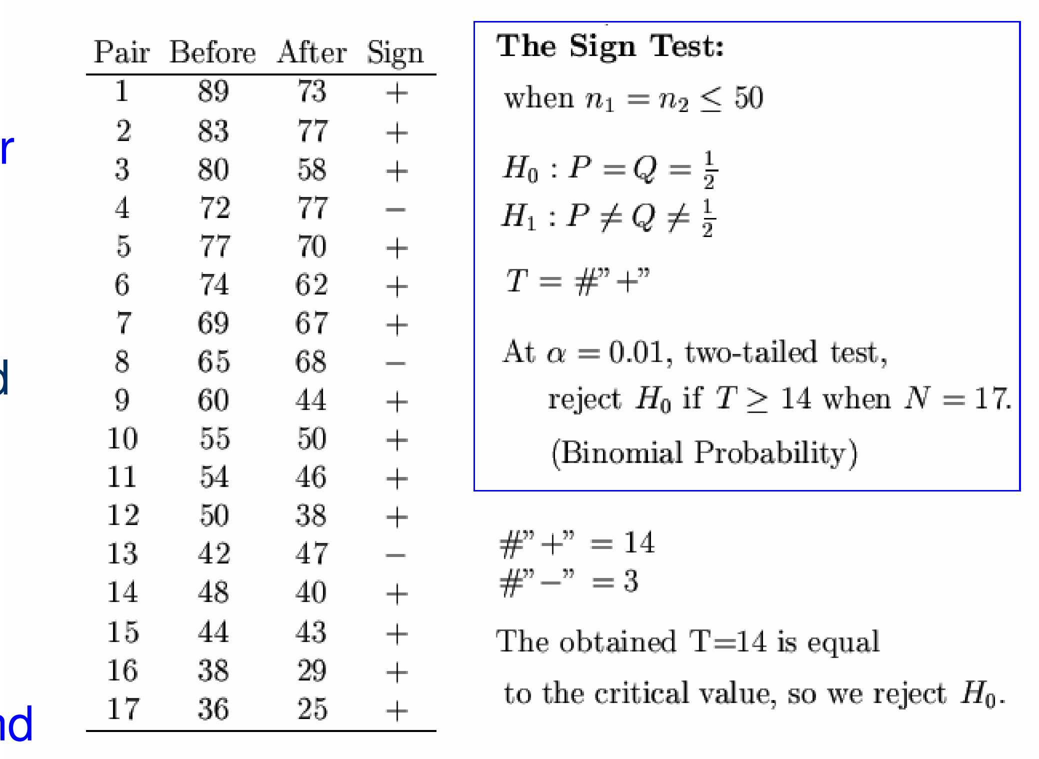
Sign Test

- Given *n* pairs of data, the sign test tests the hypothesis that the median of the differences in the pairs is zero.

- The test statistic is the number of positive differences.

- If the null hypothesis is true, then the numbers of positive and negative differences should be approximately the same.

- In fact, the number of positive differences will have a Binomial distribution with parameters *n* and *p*.



Pair	Before	After	Sign
1	89	73	+
2	83	77	+
3	80	58	+
4	72	77	−
5	77	70	+
6	74	62	+
7	69	67	+
8	65	68	−
9	60	44	+
10	55	50	+
11	54	46	+
12	50	38	+
13	42	47	−
14	48	40	+
15	44	43	+
16	38	29	+
17	36	25	+

The Sign Test:

when $n_1 = n_2 \leq 50$

$H_0 : P = Q = \frac{1}{2}$
$H_1 : P \neq Q \neq \frac{1}{2}$

$T = \#"+"$

At $\alpha = 0.01$, two-tailed test,
reject H_0 if $T \geq 14$ when $N = 17$.
(Binomial Probability)

$\#"+" = 14$
$\#"-" = 3$

The obtained T=14 is equal to the critical value, so we reject H_0.

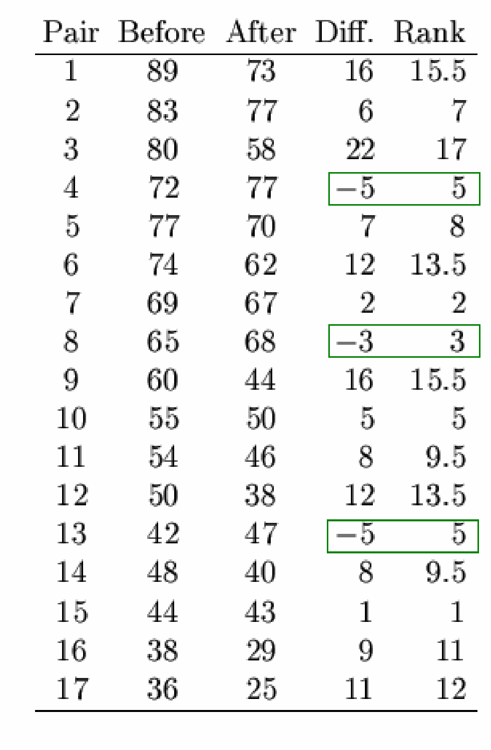
Wilcoxon Signed-Rank Test (paired)

- Null hypothesis: the population median from which both samples were drawn is the same.

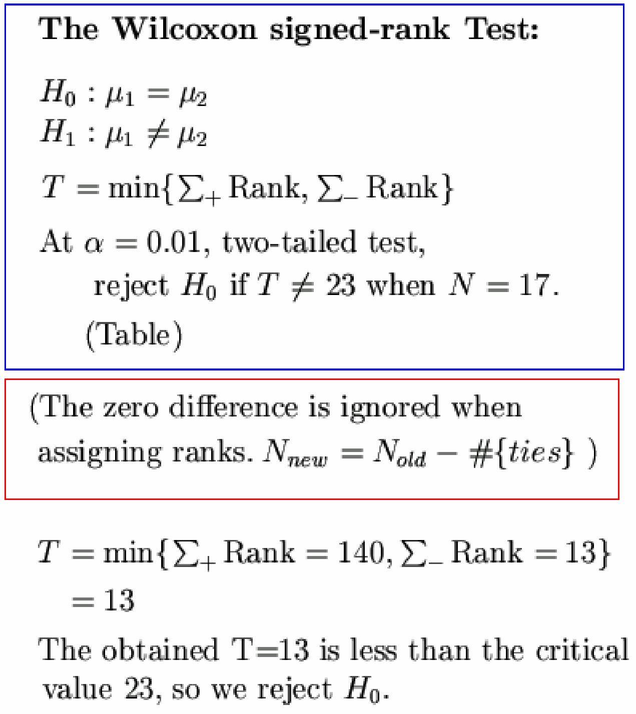
- The sum of the ranks for the "positive" (up-regulated) values is calculated and compared against a precomputed table to a p-value.

  - Sorting the absolute values of the differences from smallest to largest.
  - Assigning ranks to the absolute values.
  - Find the sum of the ranks of the positive differences.

- If the null hypothesis is true, the sum of the ranks of the positive differences should be about the same as the sum of the ranks of the negative differences.



Pair	Before	After	Diff.	Rank
1	89	73	16	15.5
2	83	77	6	7
3	80	58	22	17
4	72	77	-5	5
5	77	70	7	8
6	74	62	12	13.5
7	69	67	2	2
8	65	68	-3	3
9	60	44	16	15.5
10	55	50	5	5
11	54	46	8	9.5
12	50	38	12	13.5
13	42	47	-5	5
14	48	40	8	9.5
15	44	43	1	1
16	38	29	9	11
17	36	25	11	12



The Wilcoxon signed-rank Test:

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

$T = \min\{\sum_+ \text{Rank}, \sum_- \text{Rank}\}$

At $\alpha = 0.01$, two-tailed test,
reject H_0 if $T \neq 23$ when $N = 17$.
(Table)

(The zero difference is ignored when assigning ranks. $N_{new} = N_{old} - \#\{ties\}$)

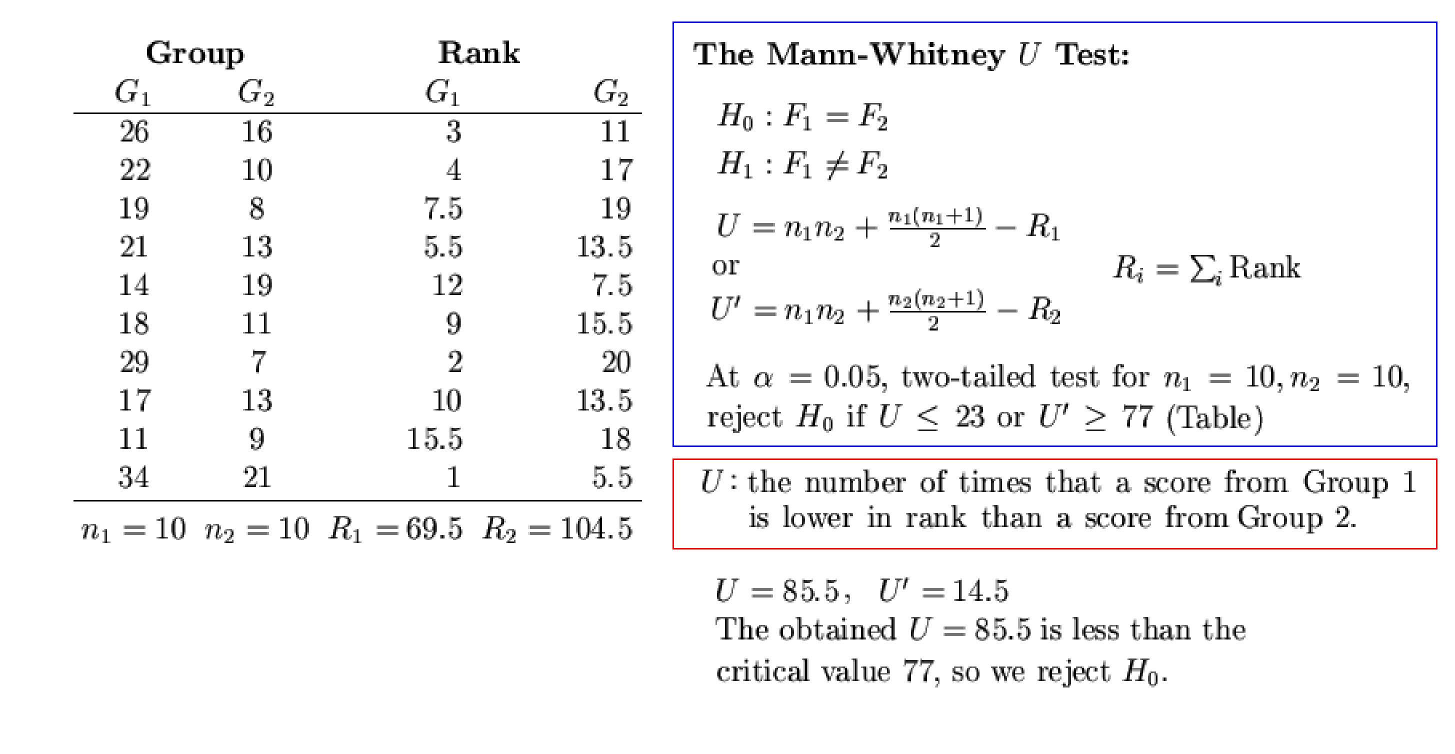
$T = \min\{\sum_+ \text{Rank} = 140, \sum_- \text{Rank} = 13\}$
$= 13$

The obtained T=13 is less than the critical value 23, so we reject H_0.

Mann-Whitney Test

(Wilcoxon Rank-Sum Test, unpaired)

- The data from the two groups are combined and given ranks. (1 for the smallest, 2 for the second smallest,...)
- The ranks for the larger group are summed and that number is compared against a precomputed table to a p-value.



Group		Rank	
G_1	G_2	G_1	G_2
26	16	3	11
22	10	4	17
19	8	7.5	19
21	13	5.5	13.5
14	19	12	7.5
18	11	9	15.5
29	7	2	20
17	13	10	13.5
11	9	15.5	18
34	21	1	5.5
$n_1 = 10$	$n_2 = 10$	$R_1 = 69.5$	$R_2 = 104.5$

The Mann-Whitney U Test:

$H_0 : F_1 = F_2$

$H_1 : F_1 \neq F_2$

$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

or

$U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$

$R_i = \sum_i \text{Rank}$

At $\alpha = 0.05$, two-tailed test for $n_1 = 10, n_2 = 10$, reject H_0 if $U \leq 23$ or $U' \geq 77$ (Table)

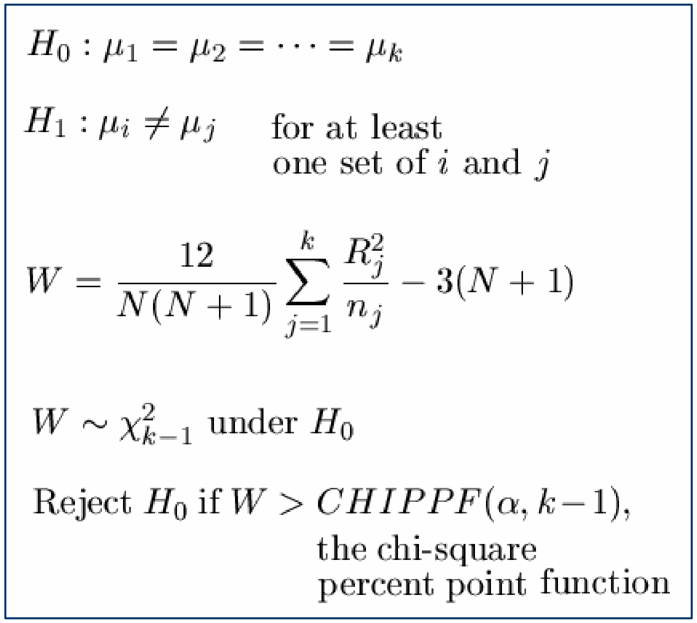
U: the number of times that a score from Group 1 is lower in rank than a score from Group 2.

$U = 85.5, \quad U' = 14.5$

The obtained $U = 85.5$ is less than the critical value 77, so we reject H_0.

Kruskal-Wallis Test

- The Kruskal Wallis test can be applied in the one factor ANOVA case. It is a non-parametric test for the situation where the ANOVA normality assumptions may not apply.
- Each of the *ni* should be at least 5 for the approximation to be valid.



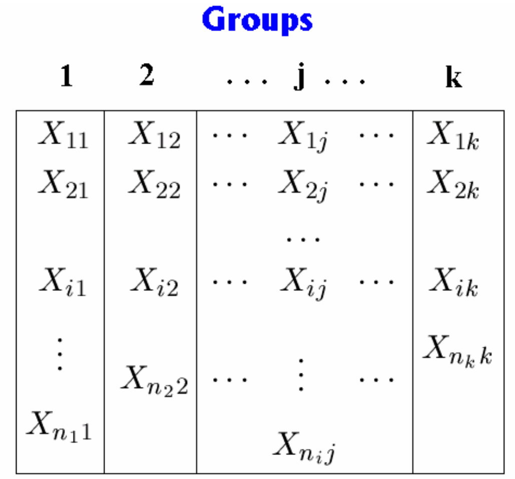
$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

$H_1 : \mu_i \neq \mu_j$ for at least one set of i and j

$$W = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$

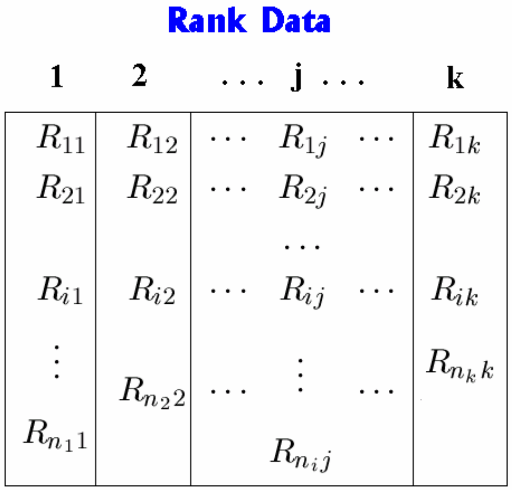
$W \sim \chi^2_{k-1}$ under H_0

Reject H_0 if $W > CHIPPF(\alpha, k-1)$, the chi-square percent point function



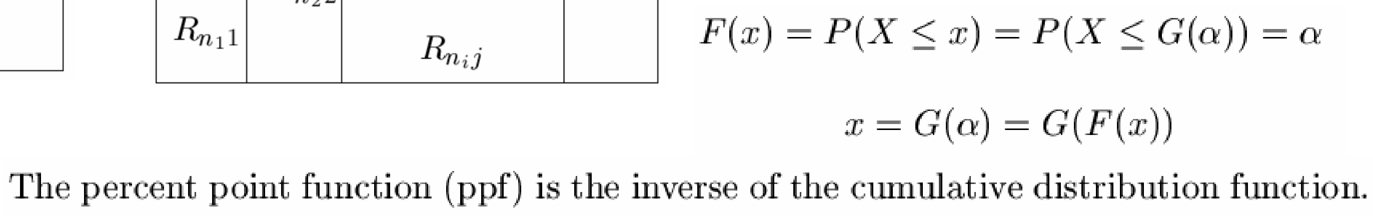
Groups

1	2	… j …	k
X_{11}	X_{12}	$\cdots$ X_{1j} $\cdots$	X_{1k}
X_{21}	X_{22}	$\cdots$ X_{2j} $\cdots$	X_{2k}
		$\cdots$	
X_{i1}	X_{i2}	$\cdots$ X_{ij} $\cdots$	X_{ik}
$\vdots$			$X_{n_k k}$
	$X_{n_2 2}$	$\cdots$ $\vdots$ $\cdots$	
$X_{n_1 1}$		$X_{n_i j}$	



Rank Data

1	2	… j …	k
R_{11}	R_{12}	$\cdots$ R_{1j} $\cdots$	R_{1k}
R_{21}	R_{22}	$\cdots$ R_{2j} $\cdots$	R_{2k}
		$\cdots$	
R_{i1}	R_{i2}	$\cdots$ R_{ij} $\cdots$	R_{ik}
$\vdots$			$R_{n_k k}$
	$R_{n_2 2}$	$\cdots$ $\vdots$ $\cdots$	
$R_{n_1 1}$		$R_{n_i j}$	



$$F(x) = P(X \leq x) = P(X \leq G(\alpha)) = \alpha$$

$$x = G(\alpha) = G(F(x))$$

The percent point function (ppf) is the inverse of the cumulative distribution function.

Parametric vs. Non-Parametric Test

Parametric Tests

- Assume that the data follows a certain distribution (normal distribution).
- Assuming equal variances and Unequal variances.
- More powerful.
- Not appropriate for data with outliers.

Non-Parametric Tests

When certain assumptions about the underlying population are questionable (e.g. normality).

- Does not assume normal distribution
- No variance assumption
- Ranks the order of raw/normalized data across conditions for analyses
- Not affected by interpretation mode (GeneSpring)
- Decrease effects of outliers (Robust)
- Not recommended if there is less than 5 replicates per group
- Needs a high number of replicates
- Less powerful

t-test	Non-parametric
Easy	Easy
Powerful	Robust
Widely Implemented	widely implemented
Not appropriate for data with outliers	Less powerful

Note:

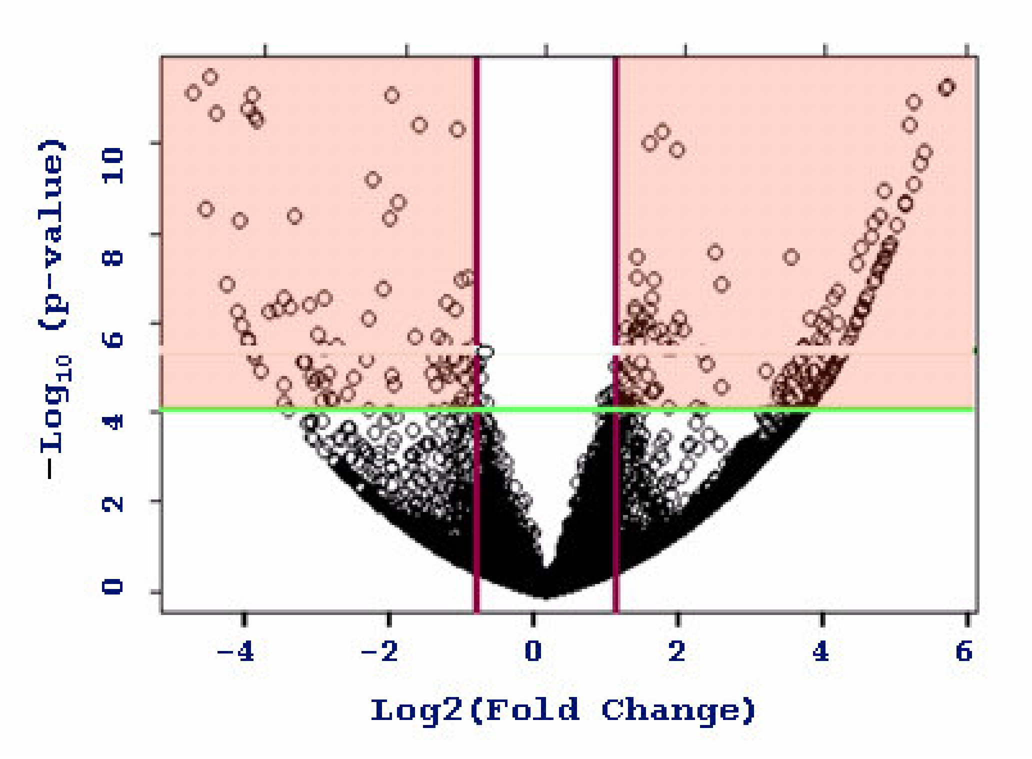
Because of the loss of power, classical non-parametric statistics have not become popular for use with microarray data, and instead bootstrap methods trend to be preferred.

Bootstrap Analysis
Robust
Powerful
Requires use of specialist packages or programming.

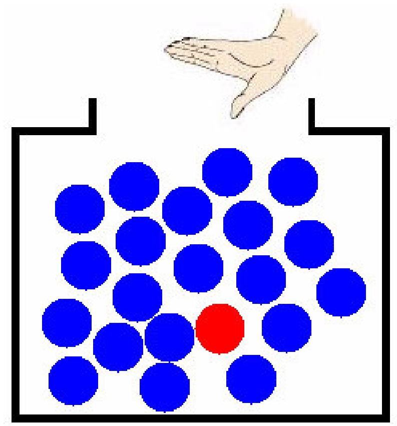
Volcano Plot

The Y variate is typically a probability (in which case a -log10 transform is used) or less commonly a p-value.

The X variate is usually a measure of differential expression such as a log-ratio.



Multiple Testing



Imagine a box with 20 marbles: 19 are blue and 1 is red.

What are the odds of randomly sampling the red marble by chance?

It is 1 out of 20.

Now let's say that you get to sample a single marble (and put it back into the box) 20 times.

Have a much higher chance to sample the red marble.

This is exactly what happens when testing several thousand genes at the same time:

Imagine that the red marble is a false positive gene: the chance that false positives are going to be sampled is higher the more genes you apply a statistical test on.

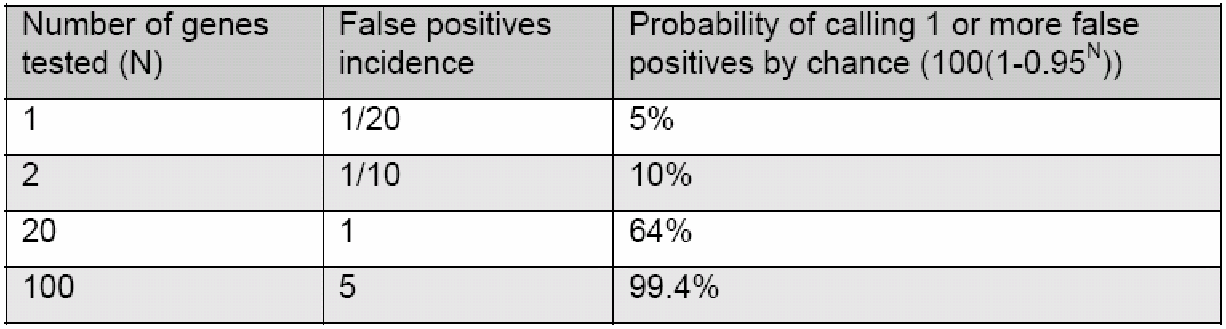
X: false positive gene

$P(X>=1)$

$= 1-P(X=0)$

$= 1- 0.95^n$

Multiplicity of Testing



Number of genes tested (N)	False positives incidence	Probability of calling 1 or more false positives by chance ($100(1-0.95^N)$)
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

Multiplicity of Testing

- There is a serious consequence of performing statistical tests on many genes in parallel, which is known as multiplicity of p-values.
- Take a large supply of reference sample, label it with Cy3 and Cy5: no genes are differentially expressed: all measured differences in expression are experimental error.
  - By the very definition of a p-value, each gene would have a 1% chance of having a p-value of less than 0.01, and thus be significant at the 1% level.
  - Because there are 10000 genes on this imaginary microarray, we would expect to find 100 significant genes at this level.
  - Similarly, we would expect to find 10 genes with a p-value less than 0.001, and 1 gene with p-value less than 0.0001
  - The p-value is the probability that a gene's expression level are different between the two groups due to chance.

Question:

1. How do we know that the genes that appear to be differentially expressed are truly differentially expressed and are not just artifact introduced because we are analyzing a large number of genes?
2. Is this gene truly differentially expressed, or could it be a false positive results?

Types of Error Control

- Multiple testing correction adjusts the p-value for each gene to keep the overall error rate (or false positive rate) to less than or equal to the user-specified p-value cutoff or error rate individual.

Multiple Testing

	# Reject H_0	# not Reject H_0	
# true H_{0j}	V	U	m_0
# true H_{1j}	S	T	m_1
	R	m - R	m

V : false positives = Type I errors

T : false negatives = Type II errors

Type One Errors Rates

$$\text{PCER} = \frac{E[\text{V}]}{\text{m}}$$

$$\text{PFER} = E[\text{V}]$$

$$\text{FWER} = p(\text{V} \geq 1)$$

$$\text{FDR} = E[\frac{\text{V}}{\text{R}}] \quad \text{if } \text{R} > 0$$

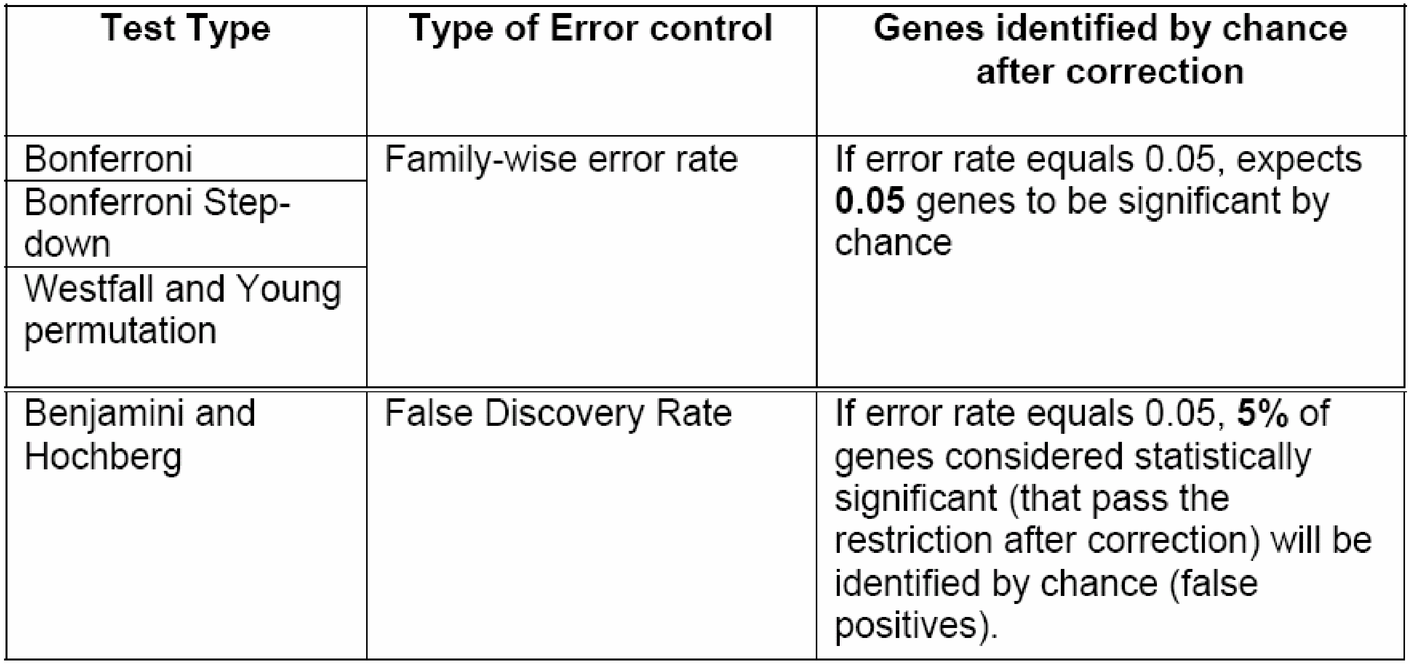
Power = Reject the false null hypothesis

$$\text{Any-pair Power} = p(\text{S} \geq 1)$$

$$\text{Per-pair Power} = \frac{E[\text{S}]}{\text{m}_1}$$

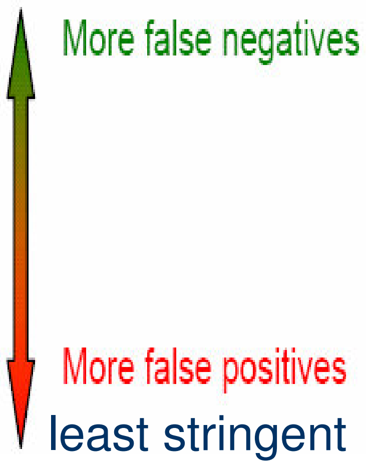
$$\text{All-pair Power} = p(\text{S} = \text{m}_1)$$

Multiple Testing Corrections



Test Type	Type of Error control	Genes identified by chance after correction
Bonferroni	Family-wise error rate	If error rate equals 0.05, expects **0.05** genes to be significant by chance
Bonferroni Step-down		
Westfall and Young permutation		
Benjamini and Hochberg	False Discovery Rate	If error rate equals 0.05, **5%** of genes considered statistically significant (that pass the restriction after correction) will be identified by chance (false positives).

most stringent



least stringent

- The more stringent a multiple testing correction, the less false positive genes are allowed.
- The trade-off of a stringent multiple testing correction is that the rate of *false negatives* (genes that are called non-significant when they are) is very high.
- FWER is the overall probability of false positive in all tests.
  - Very conservative
  - False positives not tolerated
- False discovery error rate allows a percentage of called genes to be false positives.

(1) Bonferroni

- The p-value of each gene is multiplied by the number of genes in the gene list.

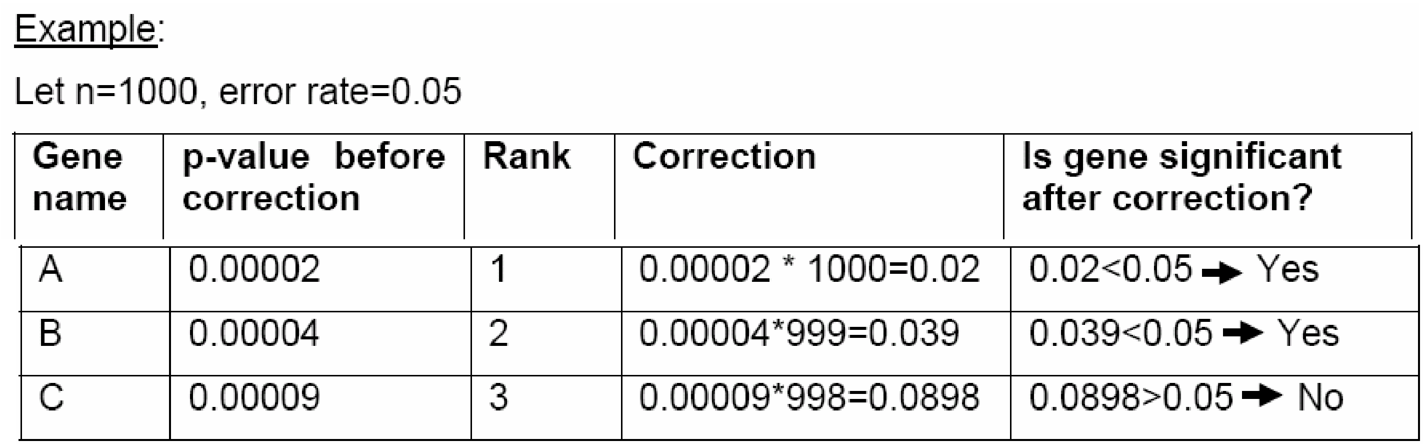
- If the corrected p-value is still below the error rate, the gene will be significant:
  - Corrected p-value= p-value * n <0.05.
  - If testing 1000 genes at a time, the highest accepted individual un-corrected p-value is 0.00005, making the correction very stringent.

- With a Family-wise error rate of 0.05 (i.e., the probability of at least one error in the family), the expected number of false positives will be 0.05.

(2) Bonferroni Step Down (Holm)

- This correction is very similar to the Bonferroni, but a little less stringent.
- The p-value of each gene is ranked from the smallest to the largest.
  - The *i*th p-value is multiplied by the number of genes present in the gene list

    Corrected P-value= p-value * (n – i + 1) < 0.05
- if the end value is less than 0.05, the gene is significant.
- It follows that sequence until no gene is found to be significant.



Example:

Let n=1000, error rate=0.05

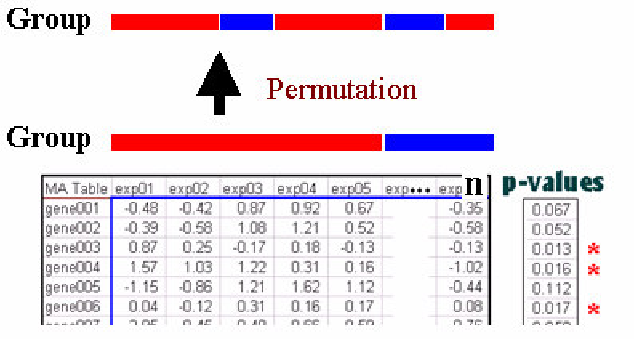
Gene name	p-value before correction	Rank	Correction	Is gene significant after correction?
A	0.00002	1	0.00002 * 1000=0.02	0.02<0.05 → Yes
B	0.00004	2	0.00004*999=0.039	0.039<0.05 → Yes
C	0.00009	3	0.00009*998=0.0898	0.0898>0.05 → No

(3) Westfall and Young Permutation

- Both Bonferroni and Holm methods are called single-step procedures, where each p-value is corrected independently.
- The Westfall and Young permutation method takes advantage of the *dependence structure* between genes, by permuting all the genes at the same time.

- The Westfall and Young permutation follows a step-down procedure similar to the Holm method, combined with a bootstrapping method to compute the p-value distribution.
- Because of the permutations, the method is very slow.
- The Westfall and Young permutation method has a similar Family-wise error rate as the Bonferroni and Holm corrections.

- P-values are calculated for each gene based on the original data set and ranked.
- The permutation method creates a pseudo-data set by dividing the data into artificial treatment and control groups.
- P-values for all genes are computed on the pseudo-data set.
- The successive minima of the new p-values are retained and compared to the original ones.
- This process is repeated a large number of times, and the proportion of resampled data sets where the minimum pseudo-p-value is less than the original p-value is the adjusted p-value.



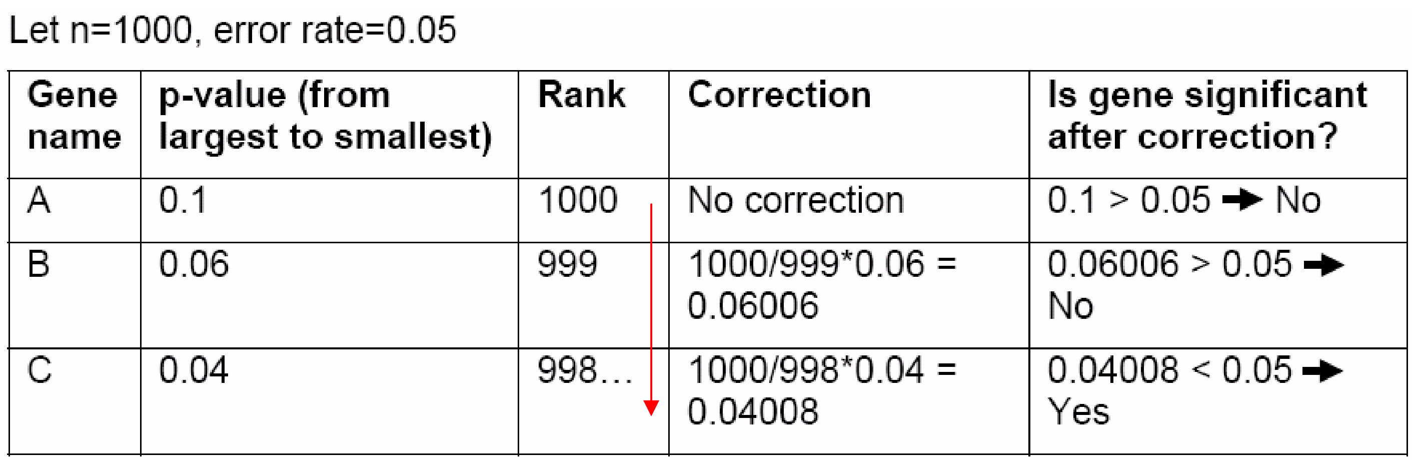
MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp n	p-values	
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35	0.067	
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58	0.052	
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13	0.013	*
gene004	1.57	1.03	1.22	0.31	0.16		-1.02	0.016	*
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44	0.112	
gene006	0.04	-0.12	0.31	0.16	0.17		0.08	0.017	*

$$\text{corrected p-value:} \approx \frac{\#\{\, p^* < 0.05 \,\}}{n\,!}$$

(4) Benjamini and Hochberg FDR

- This correction is the least stringent of all 4 options, and therefore tolerates more false positives.
- There will be also less false negative genes.
- The correction becomes more stringent as the p-value decreases, similarly as the Bonferroni Step-down correction.
- This method provides a good alternative to Family-wise error rate methods.
- The error rate is a proportion of the number of called genes.
- FDR: Overall proportion of false positives relative to the total number of genes declared significant.

$$\text{Corrected P-value} = \text{p-value} * (n / R_i) < 0.05$$



Let n=1000, error rate=0.05

Gene name	p-value (from largest to smallest)	Rank	Correction	Is gene significant after correction?
A	0.1	1000	No correction	0.1 > 0.05 → No
B	0.06	999	1000/999*0.06 = 0.06006	0.06006 > 0.05 → No
C	0.04	998…	1000/998*0.04 = 0.04008	0.04008 < 0.05 → Yes

Recommendations

- The default multiple testing correction in GeneSpring is the Benjamini and Hochberg False Discovery Rate.
- It is the least stringent of all corrections and provides a good balance between discovery of statistically significant genes and limitation of false positive occurrences.
- The Bonferroni correction is the most stringent test of all, but offers the most conservative approach to control for false positives.
- The Westfall and Young Permutation is the only correction accounting for genes coregulation. However, it is very slow and is also very conservative.
- As multiple testing corrections depend on the number of tests performed, or number of genes tested, it is recommended to select a prefiltered gene list.

If There Are No Results with MTC

- increase p-cutoff value
- increase number of replicates
- use less stringent or no MTC
- add cross-validation experiments

Post Hoc Tests

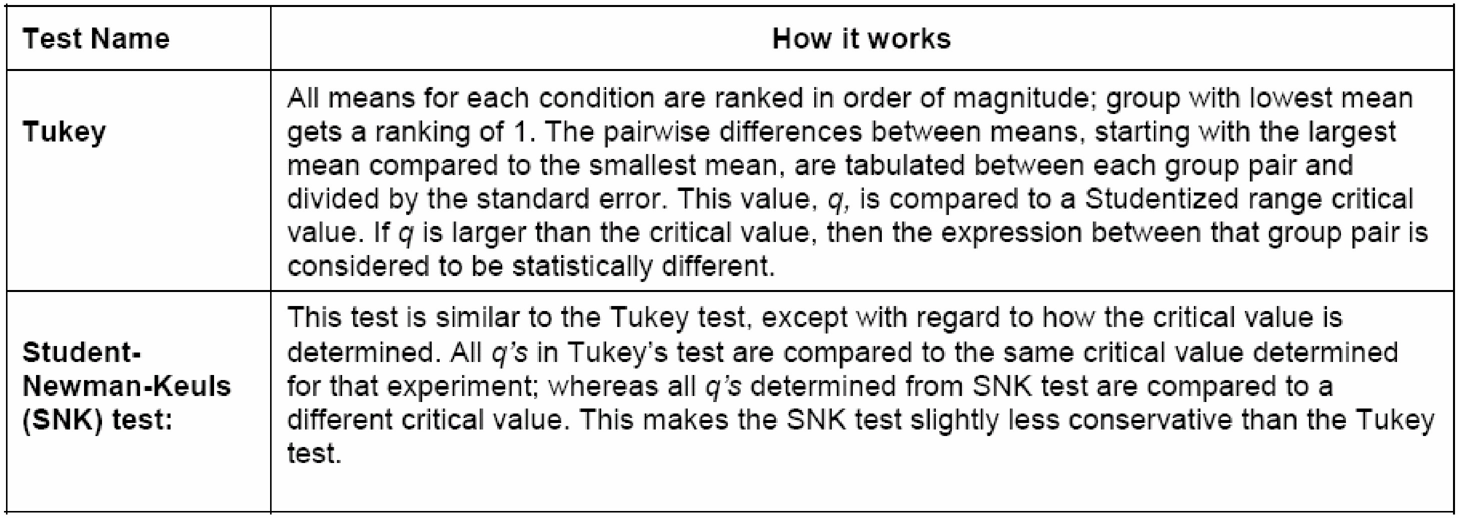
Applicable when comparing more than 2 groups.

One-way ANOVA model

$$\textbf{H0: } \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \ (\ldots = \mu_n)$$

If H0is rejected for a **gene**, there is still no information about where differences are observed.

How does one determine which specific differences are significant?



Test Name	How it works
Tukey	All means for each condition are ranked in order of magnitude; group with lowest mean gets a ranking of 1. The pairwise differences between means, starting with the largest mean compared to the smallest mean, are tabulated between each group pair and divided by the standard error. This value, *q*, is compared to a Studentized range critical value. If *q* is larger than the critical value, then the expression between that group pair is considered to be statistically different.
Student-Newman-Keuls (SNK) test:	This test is similar to the Tukey test, except with regard to how the critical value is determined. All *q's* in Tukey's test are compared to the same critical value determined for that experiment; whereas all *q's* determined from SNK test are compared to a different critical value. This makes the SNK test slightly less conservative than the Tukey test.

Student-Newman-Keuls (SNK) Test

assuming
equal sample sizes and
homogeneity of variance

Group	A	B	C	D
Mean	2	3	7	8

alpha = 0.01
$n = 5$
$df = 16$

$$\sqrt{\frac{MSE}{n}} = \sqrt{\frac{.5}{5}} = 0.316$$

$$q = \frac{\overline{X}_i - \overline{X}_j}{\sqrt{\frac{MSE}{n}}}$$

- Parametric and non-parametric
- Unequal sample sizes
- Variance assumption

"r" is the number of means spanned by a given comparison.

r, df, alpha → studentized range statistic q

1. $r = 4$, $q_{.01} = 5.19$

A vs D: $q = \frac{8-2}{0.316} = 18.99$, $p < 0.01$

2. $r = 3$, $q_{.01} = 4.79$

a. A vs C: $q = \frac{7-2}{0.316} = 15.82$, $p < 0.01$

b. B vs D: $q = \frac{8-3}{0.316} = 15.82$, $p < 0.01$

3. $r = 2$, $q_{.01} = 4.13$

a. A vs B: $q = \frac{3-2}{.316} = 3.16$, $p > 0.01$

b. B vs C: $q = \frac{7-3}{.316} = 12.66$, $p < 0.01$

c. C vs D: $q = \frac{8-7}{.316} = 3.16$, $p > 0.01$

Tukey's HSD Test

Honestly Significant Difference (HSD)

$$HSD = q\sqrt{\frac{MS_{Within}}{n}} \qquad \frac{M_1 - M_2}{\sqrt{MS_w\left(\frac{1}{n}\right)}}$$

Tukey's HSD Post-hoc test is applied in exactly the same way that the Student-Newman-Keuls is, with the exception that r is set at k for all comparisons.

- (k vs 1, k vs 2,.., k vs k-1) (k-1 vs 1, k-1 vs 2,…, k-1 vs k-2) …(…2 vs 1)

$$r = k\ ,\ df, \text{ alpha} \rightarrow \text{ studentized range statistic } q$$

- All alpha's in Tukey's test are compared to the same critical value.
- All alpha's in SKN test are compared to a different critical value.
- This test is more conservative (less powerful) than the Student-Newman-Keuls.

Bootstrap Analysis

- Bootstrap Sample/data
  - with replacement: different individuals in the bootstrap data could have the same value from the real data.
  - without replacement: each of the real values is only used once in the bootstrap data.

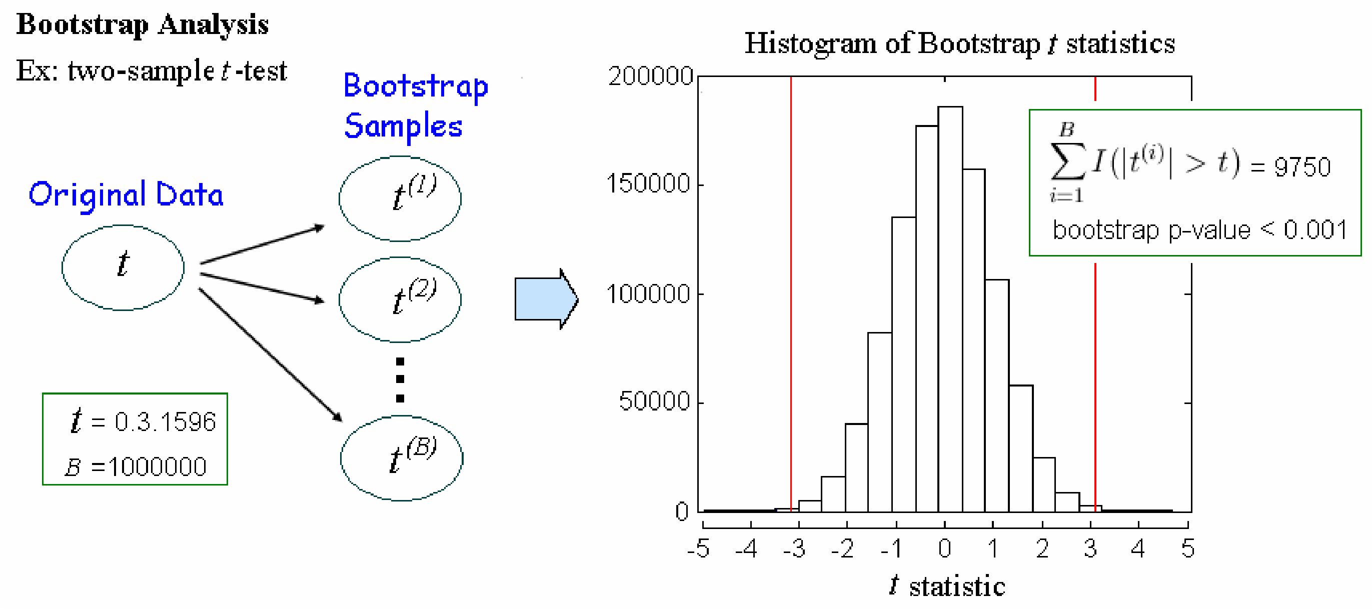
- Bootstrap analysis are more appropriate for microarray analysis than either t-test or classical non-parametric tests.
  - don't require that the data are normally distributed.
  - robust to noise and experimental artifacts.

- Under the null hypothesis; there is no difference in gene expression between the two groups. If that were the case, then any of the measurements in the data could have been observed in any of the individuals.
  - ex: any of the AML patients could have had any of the 38 measurements associated with both the AML and ALL patients.

- The bootstrap works by constructing a large number of random data sets by resampling from the original data, in which each individual is randomly allocated one of the measurements from the data, which could be from either of the groups.

Bootstrap Analysis (conti.)

- The bootstrap data sets look like the real data, in that they have similar values, but are biologically nonsense because the values have been randomized.
- Aim: the aim of the test is to compare some property of the real data with a distribution of the same property in random data sets.



Steps of Bootstrap Analysis

1. We generate an empirical distribution using the t-statistics calculated from the randomized bootstrap data.

2. The t-statistic from the real data is compared with the distribution of t-statistics from the bootstrap data.

3. We calculate an empirical p-value by computing the proportion of bootstrap statistics that have a more extreme value than the t-statistic from the real data.
   - if the real t-statistic is in the belly of the distribution, then it is indistinguishable from t-statistics generated from randomized data.
   - if the statistic from the real data is towards the edge of the bootstrap distribution, then it is unlikely that the experimental result can have arisen by chance, and we would conclude that the gene is significantly differentially expressed.

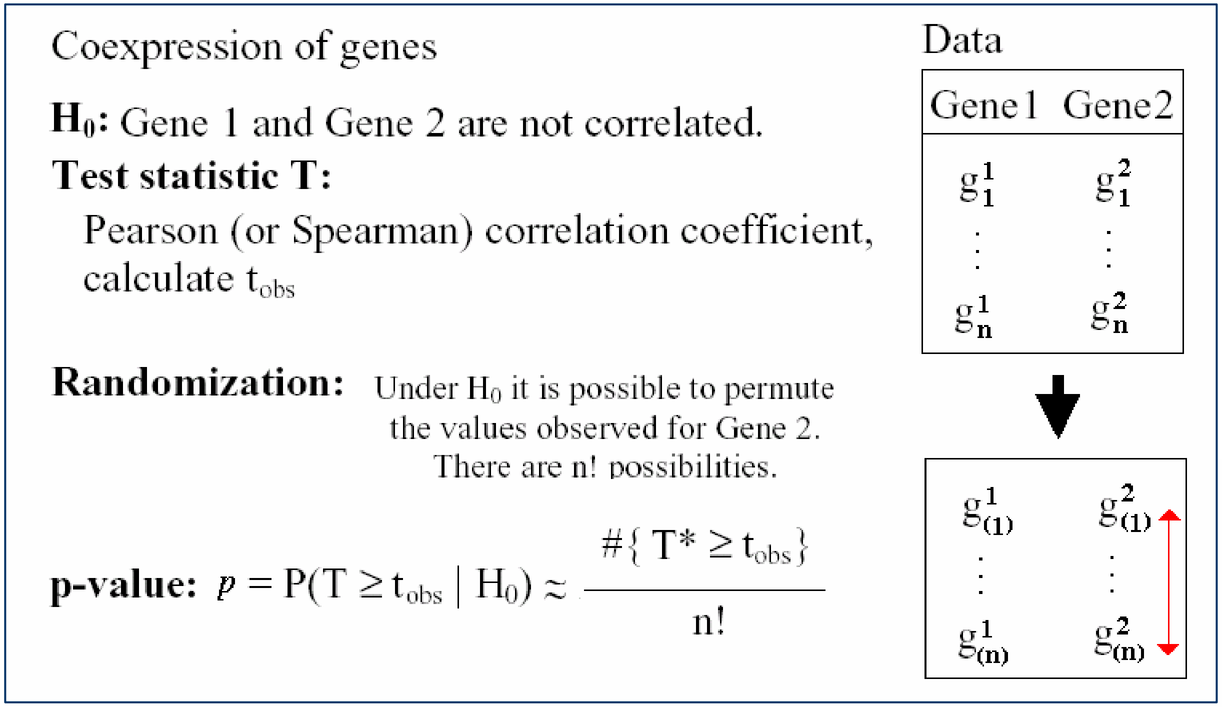
Permutation Test

- The permutation test is a test where the null-hypothesis allows to reduce the inference to a randomization problem.
- The process of randomizations makes it possible to ascribe a probability distribution to the difference in the outcome possible under H**0**.
- The outcome data are analyzed many times (once for each acceptable assignment that could have been possible under H**0**) and then compared with the observed result, without dependence on additional distributional or model-based assumptions.
- Perform a permutation test (general):
  1. Analyze the problem, choice of null-hypothesis
  2. Choice of test statistic **T**
  3. Calculate the value of the test statistic for the observed data: **tobs**
  4. Apply the randomization principle and look at all possible permutations, this gives the distribution of the test statistic **T** under H0.
  5. Calculation of p-value:

$$p = \mathrm{P}(\mathrm{T} \geq \mathrm{t}_{\mathrm{obs}} \mid \mathrm{H}_0) \approx \frac{\#\{\mathrm{t}^* \geq \mathrm{t}_{\mathrm{obs}}\}}{\#\ \mathrm{permutations}}$$

Ref: Mansmann, U. (2002), Practical microarray analysis: resampling and the Bootstraap.Heidelberg.

Permutation Test (conti.)



Coexpression of genes

H$_0$: Gene 1 and Gene 2 are not correlated.

Test statistic T:

Pearson (or Spearman) correlation coefficient, calculate t_{obs}

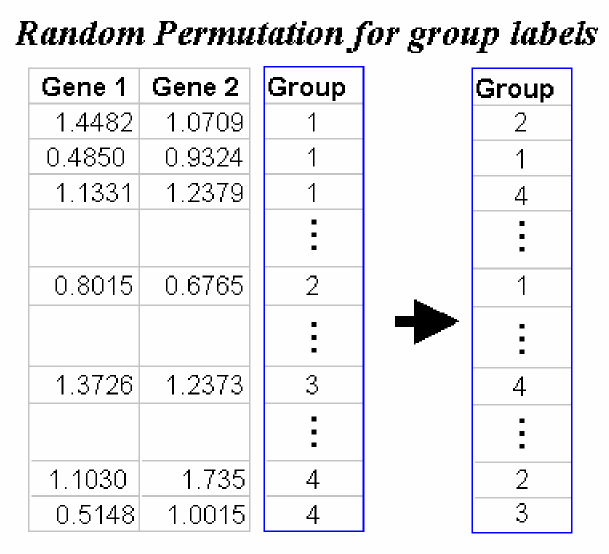
Randomization: Under H$_0$ it is possible to permute the values observed for Gene 2. There are n! possibilities.

p-value: $p = P(T \ge t_{obs} \mid H_0) \approx \dfrac{\#\{T^* \ge t_{obs}\}}{n!}$

Data

Gene1	Gene2
g^1_1	g^2_1
⋮	⋮
g^1_n	g^2_n

$g^1_{(1)}$	$g^2_{(1)}$
⋮	⋮
$g^1_{(n)}$	$g^2_{(n)}$



Random Permutation for group labels

Gene 1	Gene 2	Group		Group
1.4482	1.0709	1		2
0.4850	0.9324	1		1
1.1331	1.2379	1		4
		⋮		⋮
0.8015	0.6765	2	→	1
		⋮		⋮
1.3726	1.2373	3		4
		⋮		⋮
1.1030	1.735	4		2
0.5148	1.0015	4		3

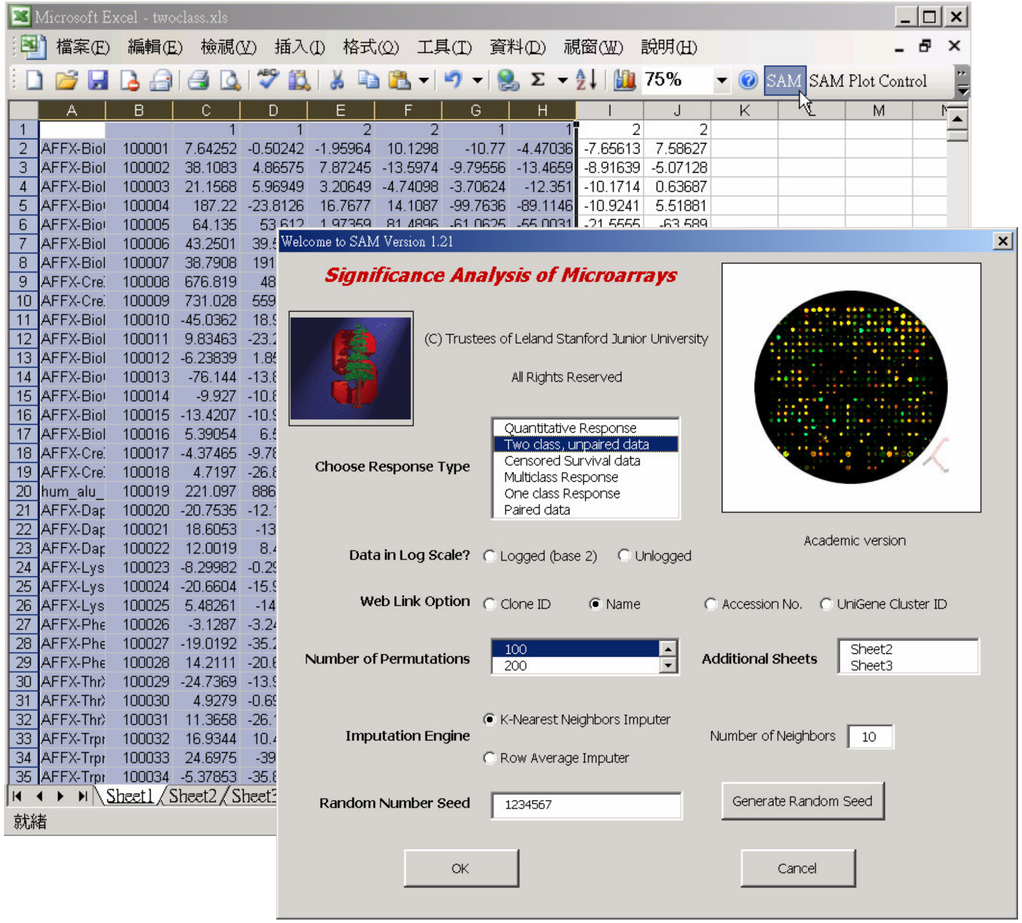
The permutation test allows determining the statistical significance of the score for every gene.

SAM: Significance Analysis of Microarrays

SAM assigns a score to each gene in a microarray experiment based upon its change in gene expression relative to the standard deviation of repeated measurements.

- *SAM plot*: the number of observed genes versus the expected number. This visualizes the outlier genes that are most dramatically regulated.
- ***False discovery rate***: is the percent of genes that are expected to be identified by chance.
- *q-value*: the lowest false discovery rate at which a gene is described as significantly regulated.

SAM does not do any normalization!


Microsoft Excel - twoclass.xls
檔案(F) 編輯(E) 檢視(V) 插入(I) 格式(O) 工具(T) 資料(D) 視窗(W) 說明(H)
75%
SAM
SAM Plot Control
Sheet1
Sheet2
Sheet3
就緒

	A	B	C	D	E	F	G	H	I	J
1			1	1	2	2	1	1	2	2
2	AFFX-Biol	100001	7.64252	-0.50242	-1.95964	10.1298	-10.77	-4.47036	-7.65613	7.58627
3	AFFX-Biol	100002	38.1083	4.86575	7.87245	-13.5974	-9.79556	-13.4659	-8.91639	-5.07128
4	AFFX-Biol	100003	21.1568	5.96949	3.20649	-4.74098	-3.70624	-12.351	-10.1714	0.63687
5	AFFX-Bio	100004	187.22	-23.8126	16.7677	14.1087	-99.7636	-89.1146	-10.9241	5.51881
6	AFFX-Bio	100005	64.135	53.612	1.97359	81.4896	-61.0625	-55.0031	-21.5555	-63.589
7	AFFX-Biol	100006	43.2501	39.5						
8	AFFX-Biol	100007	38.7908	191						
9	AFFX-Cre	100008	676.819	48						
10	AFFX-Cre	100009	731.028	559						
11	AFFX-Biol	100010	-45.0362	18.9						
12	AFFX-Biol	100011	9.83463	-23.2						
13	AFFX-Biol	100012	-6.23839	1.85						
14	AFFX-Bio	100013	-76.144	-13.8						
15	AFFX-Bio	100014	-9.927	-10.8						
16	AFFX-Biol	100015	-13.4207	-10.9						
17	AFFX-Biol	100016	5.39054	6.5						
18	AFFX-Cre	100017	-4.37465	-9.78						
19	AFFX-Cre	100018	4.7197	-26.8						
20	hum_alu_	100019	221.097	886						
21	AFFX-Dap	100020	-20.7535	-12.1						
22	AFFX-Dap	100021	18.6053	-13						
23	AFFX-Dap	100022	12.0019	8.4						
24	AFFX-Lys	100023	-8.29982	-0.29						
25	AFFX-Lys	100024	-20.6604	-15.9						
26	AFFX-Lys	100025	5.48261	-14						
27	AFFX-Phe	100026	-3.1287	-3.24						
28	AFFX-Phe	100027	-19.0192	-35.2						
29	AFFX-Phe	100028	14.2111	-20.6						
30	AFFX-Thr	100029	-24.7369	-13.9						
31	AFFX-Thr	100030	4.9279	-0.69						
32	AFFX-Thr	100031	11.3658	-26.1						
33	AFFX-Trpr	100032	16.9344	10.4						
34	AFFX-Trpr	100033	24.6975	-39						
35	AFFX-Trpr	100034	-5.37853	-35.8						

Welcome to SAM Version 1.21

Significance Analysis of Microarrays

(C) Trustees of Leland Stanford Junior University

All Rights Reserved

Academic version

Choose Response Type: Quantitative Response / **Two class, unpaired data** / Censored Survival data / Multiclass Response / One class Response / Paired data

Data in Log Scale? ○ Logged (base 2) ○ Unlogged

Web Link Option: ○ Clone ID ● Name ○ Accession No. ○ UniGene Cluster ID

Number of Permutations: **100** / 200

Additional Sheets: Sheet2 / Sheet3

Imputation Engine: ● K-Nearest Neighbors Imputer ○ Row Average Imputer

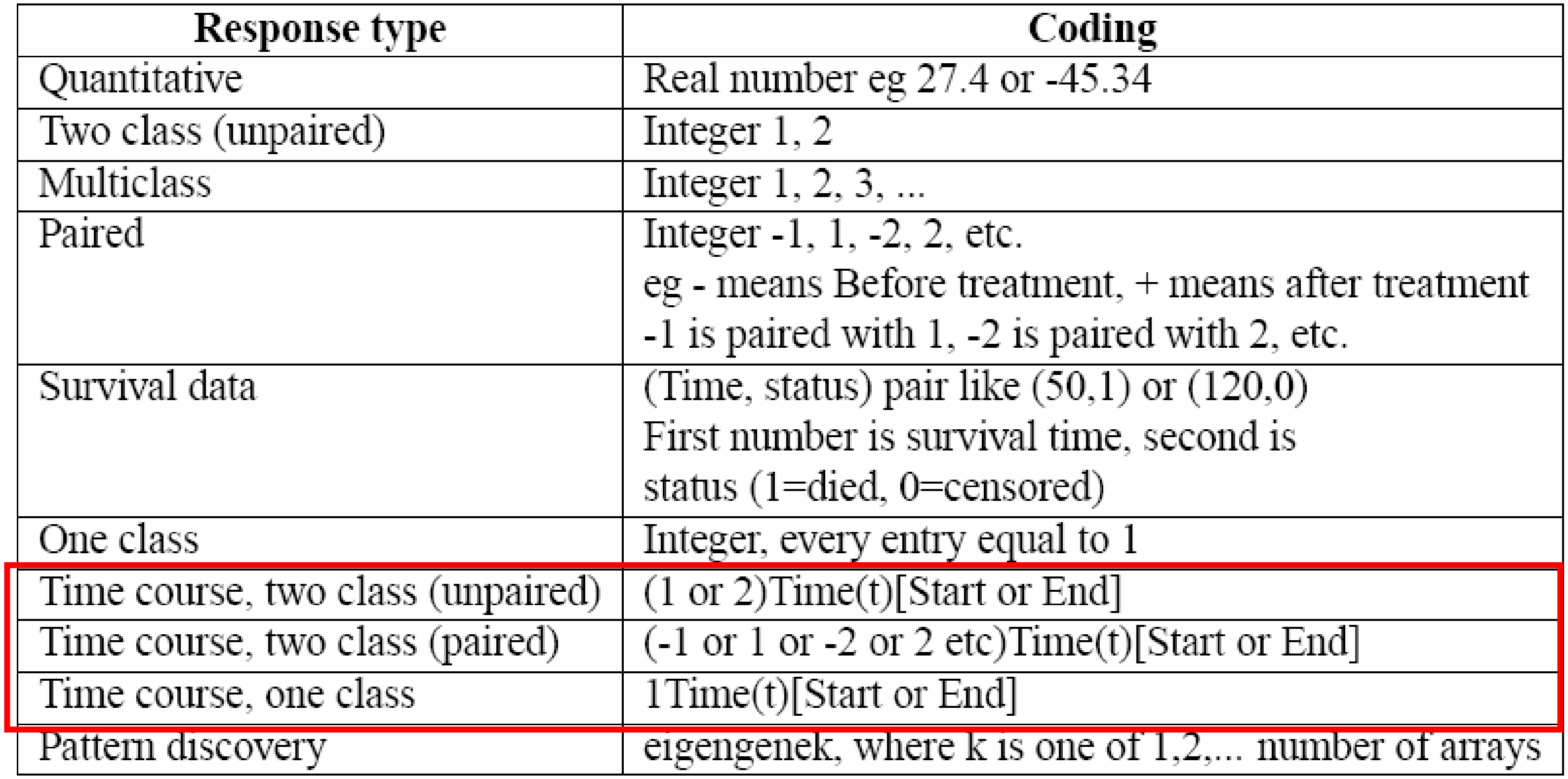
Number of Neighbors: 10

Random Number Seed: 1234567

Tusher VG, Tibshirani R, Chu G.(2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 98(9):5116-21.

SAM: Significance Analysis of Microarrays
http://www-stat.stanford.edu/~tibs/SAM/

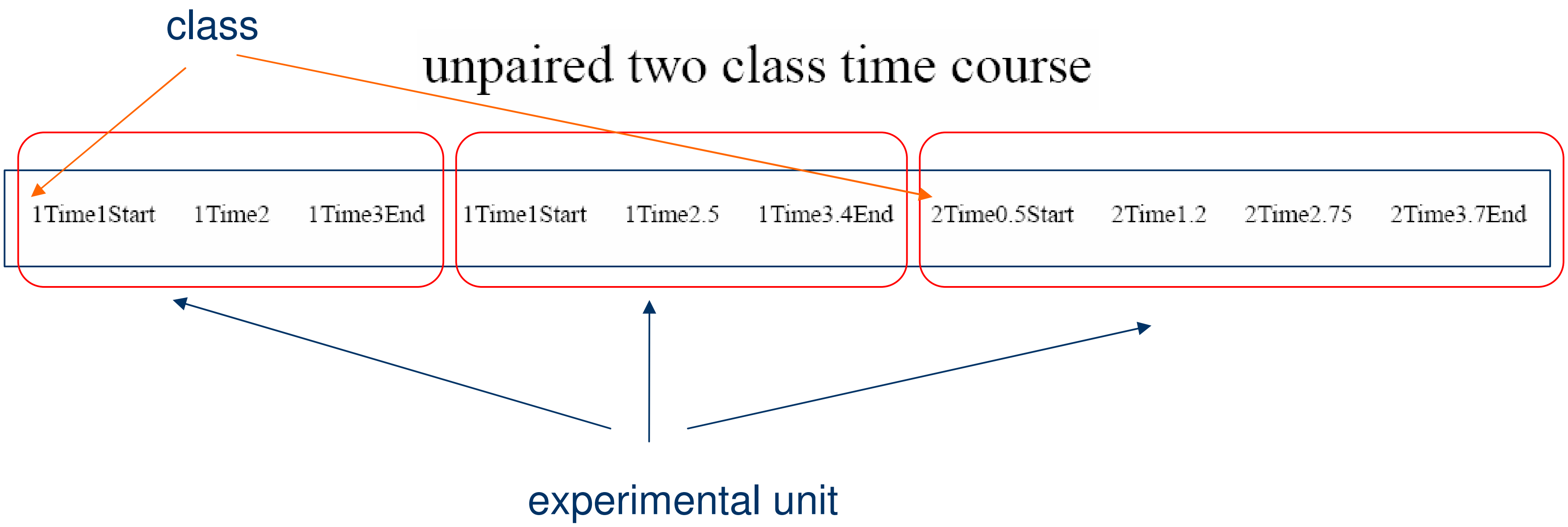
SAM: Response Type



Response type	Coding
Quantitative	Real number eg 27.4 or -45.34
Two class (unpaired)	Integer 1, 2
Multiclass	Integer 1, 2, 3, ...
Paired	Integer -1, 1, -2, 2, etc. eg - means Before treatment, + means after treatment -1 is paired with 1, -2 is paired with 2, etc.
Survival data	(Time, status) pair like (50,1) or (120,0) First number is survival time, second is status (1=died, 0=censored)
One class	Integer, every entry equal to 1
Time course, two class (unpaired)	(1 or 2)Time(t)[Start or End]
Time course, two class (paired)	(-1 or 1 or -2 or 2 etc)Time(t)[Start or End]
Time course, one class	1Time(t)[Start or End]
Pattern discovery	eigengenek, where k is one of 1,2,... number of arrays

SAM Users guide and technical document

SAM: Time Series


class

unpaired two class time course

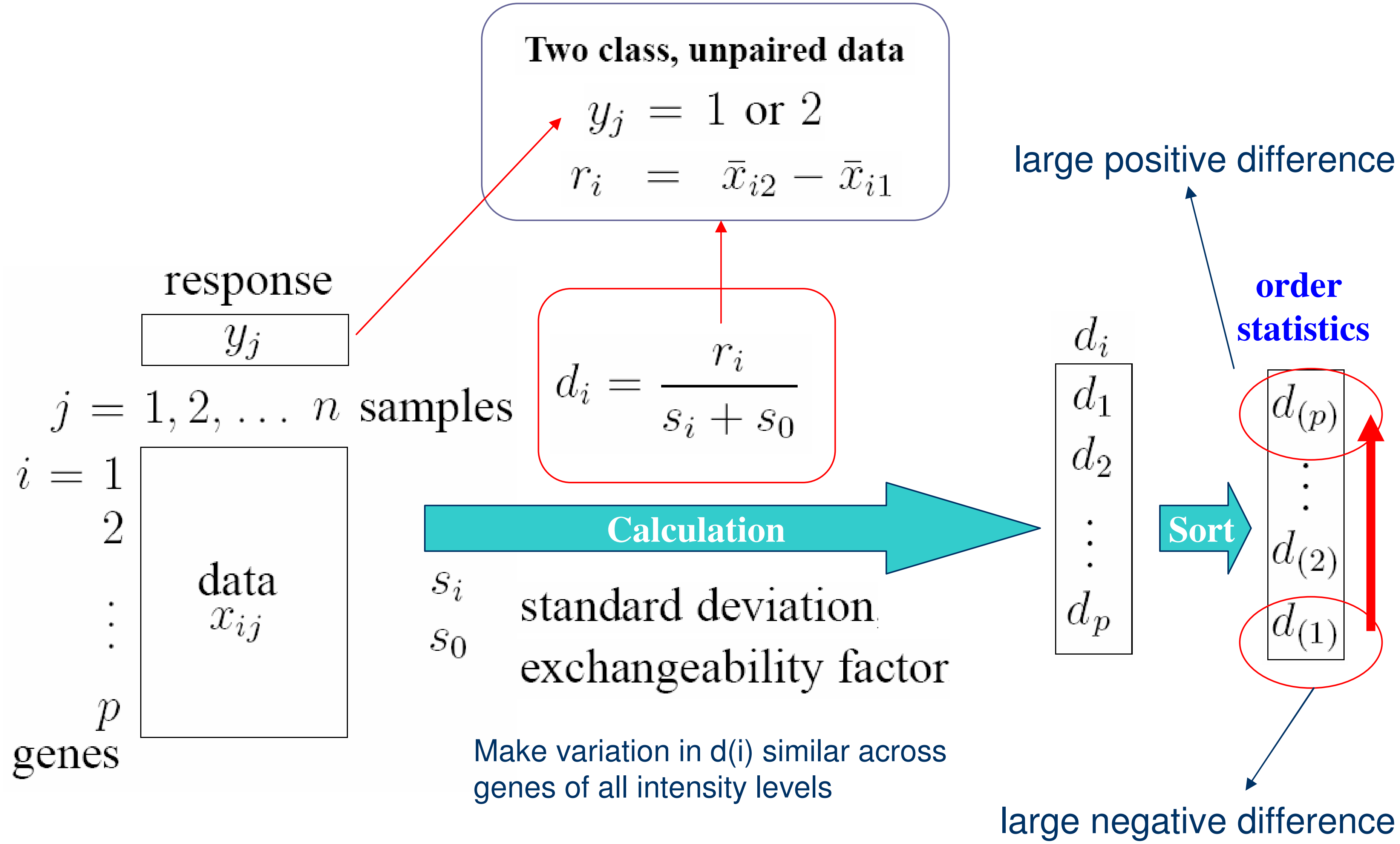
1Time1Start 1Time2 1Time3End 1Time1Start 1Time2.5 1Time3.4End 2Time0.5Start 2Time1.2 2Time2.75 2Time3.7End

experimental unit

- Paired data time courses: class label is -1, or 1 or -2 or 2.
- One class time courses: class label is a 1.

NOTE: SAM summarizes each time course by a ***slope*** or a ***signed area***, and then treats the summarized data in the same way as it treats two class, one class, or a two-class paired design.

SAM: Significance Analysis of Microarrays



Two class, unpaired data

$$y_j = 1 \text{ or } 2$$

$$r_i = \bar{x}_{i2} - \bar{x}_{i1}$$

response

y_j

$j = 1, 2, \ldots\ n$ samples

$i = 1$
2
$\vdots$
p

genes

data

x_{ij}

$$d_i = \frac{r_i}{s_i + s_0}$$

Calculation

s_i standard deviation

s_0 exchangeability factor

Make variation in d(i) similar across genes of all intensity levels

d_i

d_1
d_2
$\vdots$
d_p

Sort

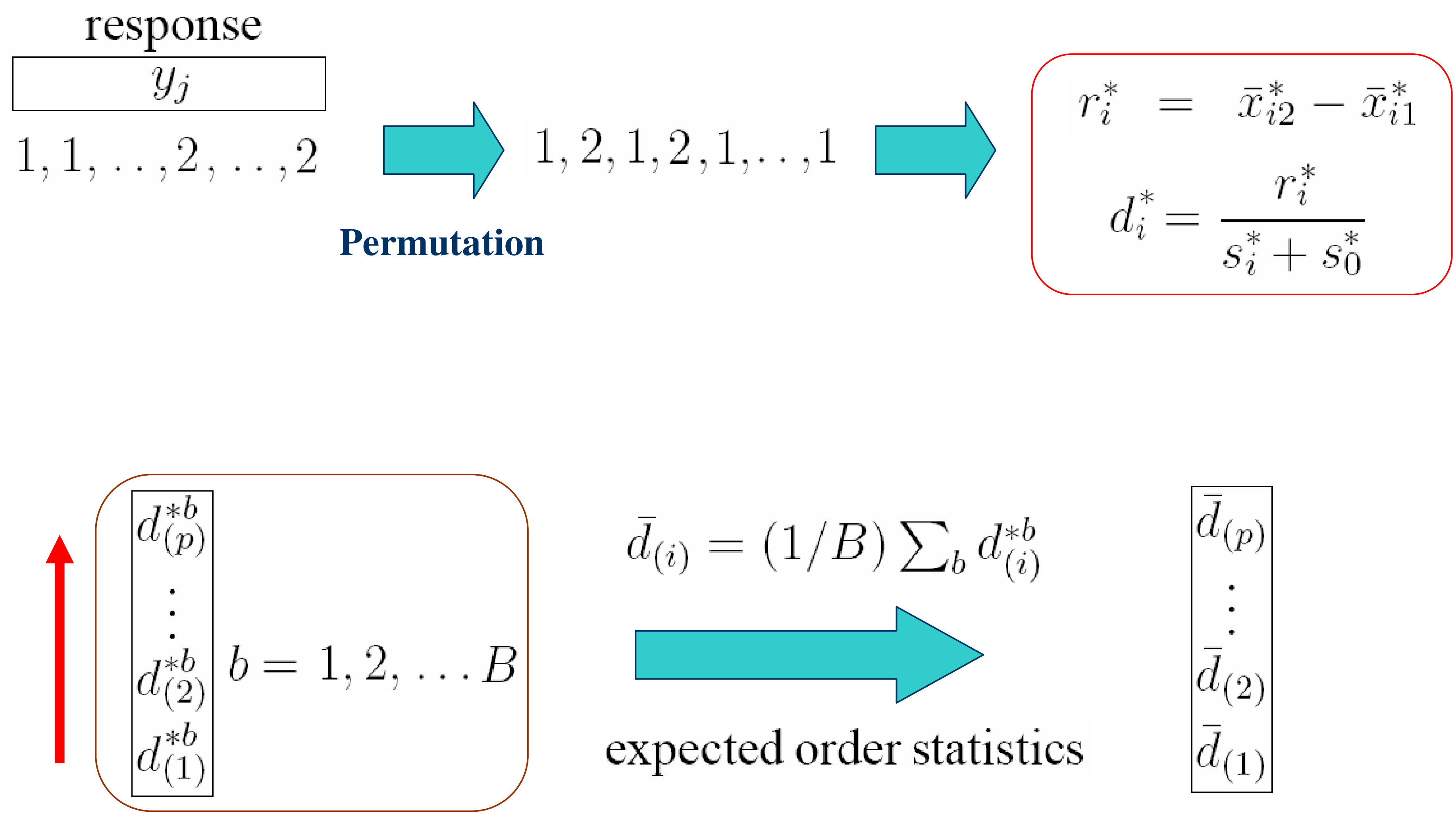
order statistics

$d_{(p)}$
$\vdots$
$d_{(2)}$
$d_{(1)}$

large positive difference

large negative difference

SAM: Expected Test Statistics



response

y_j

$1, 1, \dots, 2, \dots, 2$

Permutation

$1, 2, 1, 2, 1, \dots, 1$

$$r_i^* = \bar{x}_{i2}^* - \bar{x}_{i1}^*$$

$$d_i^* = \frac{r_i^*}{s_i^* + s_0^*}$$

$$\begin{array}{|c|}\hline d_{(p)}^{*b} \\ \vdots \\ d_{(2)}^{*b} \\ d_{(1)}^{*b} \\ \hline\end{array} \quad b = 1, 2, \dots B$$

$$\bar{d}_{(i)} = (1/B) \textstyle\sum_b d_{(i)}^{*b}$$

expected order statistics

$$\begin{array}{|c|}\hline \bar{d}_{(p)} \\ \vdots \\ \bar{d}_{(2)} \\ \bar{d}_{(1)} \\ \hline\end{array}$$

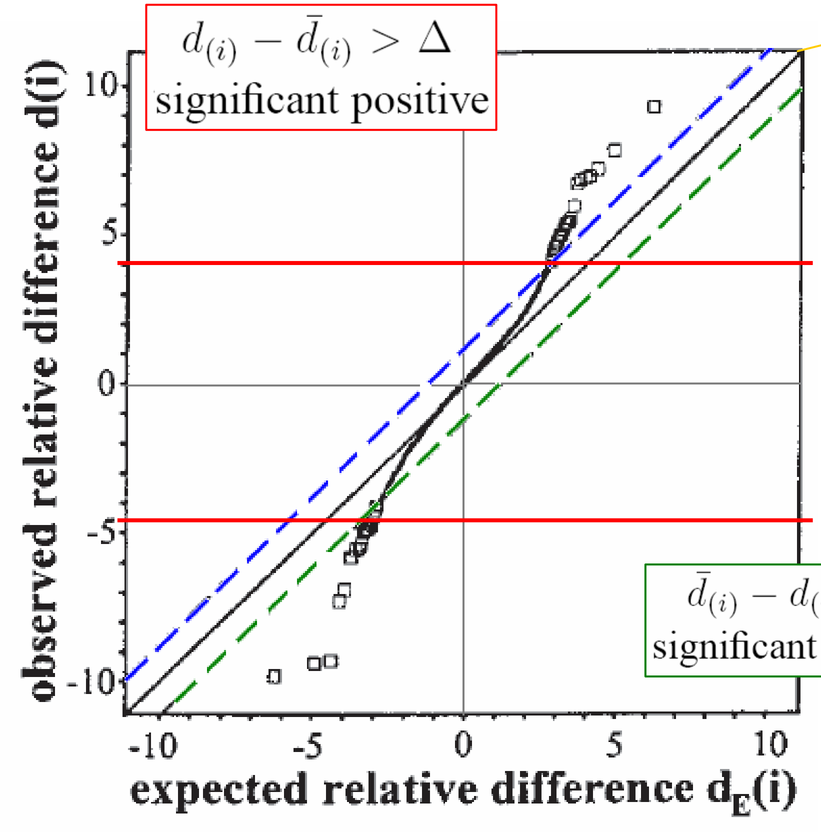
SAM Plot

$$\begin{array}{c} d_{(p)} \\ \vdots \\ d_{(2)} \\ d_{(1)} \end{array}$$

vs

$$\begin{array}{c} \bar{d}_{(p)} \\ \vdots \\ \bar{d}_{(2)} \\ \bar{d}_{(1)} \end{array}$$

Points for genes with evidence of induction



$d_{(i)} - \bar{d}_{(i)} > \Delta$
significant positive

$\bar{d}_{(i)} - d_{(i)} > \Delta$
significant negative

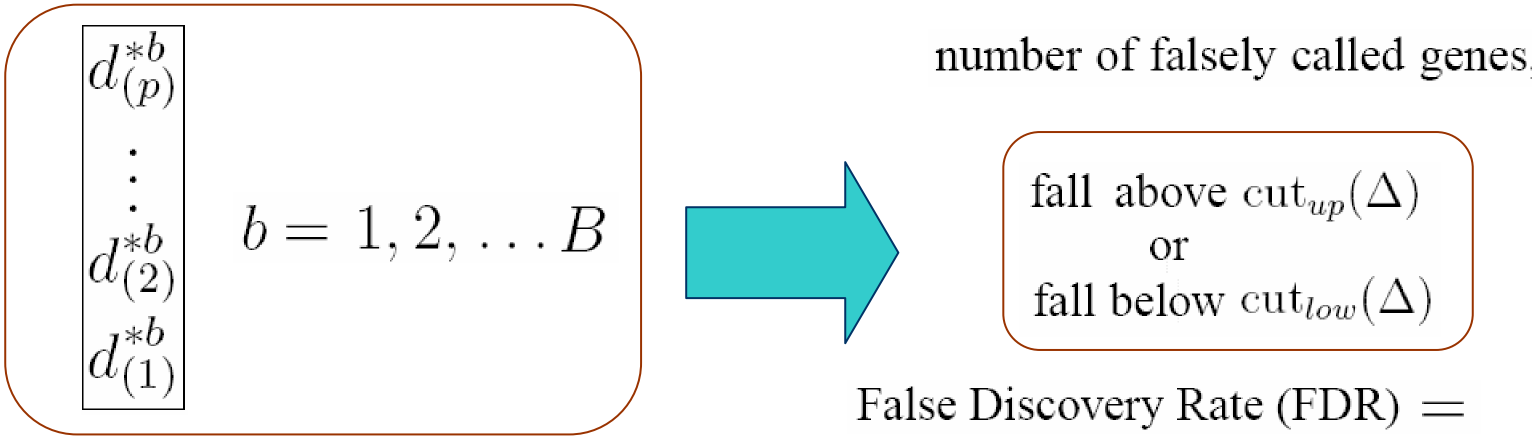
$d_{(i)} = \bar{d}_{(i)}$

upper cut-point $\text{cut}_{up}(\Delta)$

lower cut-point $\text{cut}_{low}(\Delta)$

Points for genes with evidence of repression

Estimating FDR for a Selected Δ



$$\boxed{\begin{matrix} d^{*b}_{(p)} \\ \vdots \\ d^{*b}_{(2)} \\ d^{*b}_{(1)} \end{matrix}} \quad b = 1, 2, \ldots B$$

number of falsely called genes

fall above $\text{cut}_{up}(\Delta)$
or
fall below $\text{cut}_{low}(\Delta)$

$$\text{False Discovery Rate (FDR)} = \frac{\text{[median (or 90th percentile) of the number of falsely called genes]}}{\text{[the number of genes called significant]}}$$

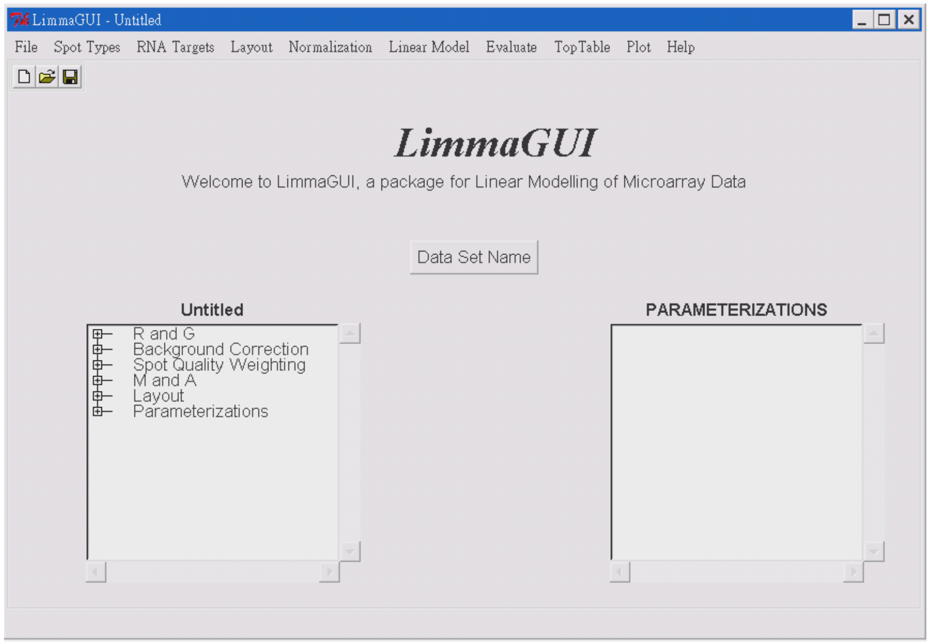
in the original data.

The q-value of a gene is the false discovery rate for the gene list that includes that gene and all genes that are more significant. It is computed by finding the smallest value of $\hat{\Delta}$ for which the gene is called significant, and then is the FDR corresponding to $\hat{\Delta}$.

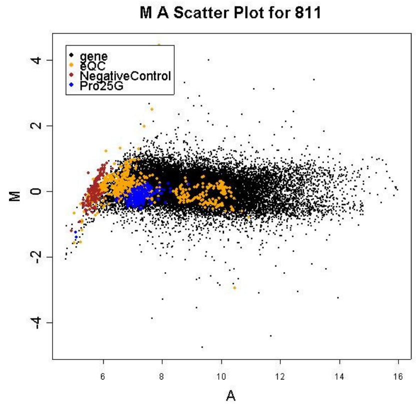
John D. Storey (2002) A direct approach to false discovery rates, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (3), 479–498.

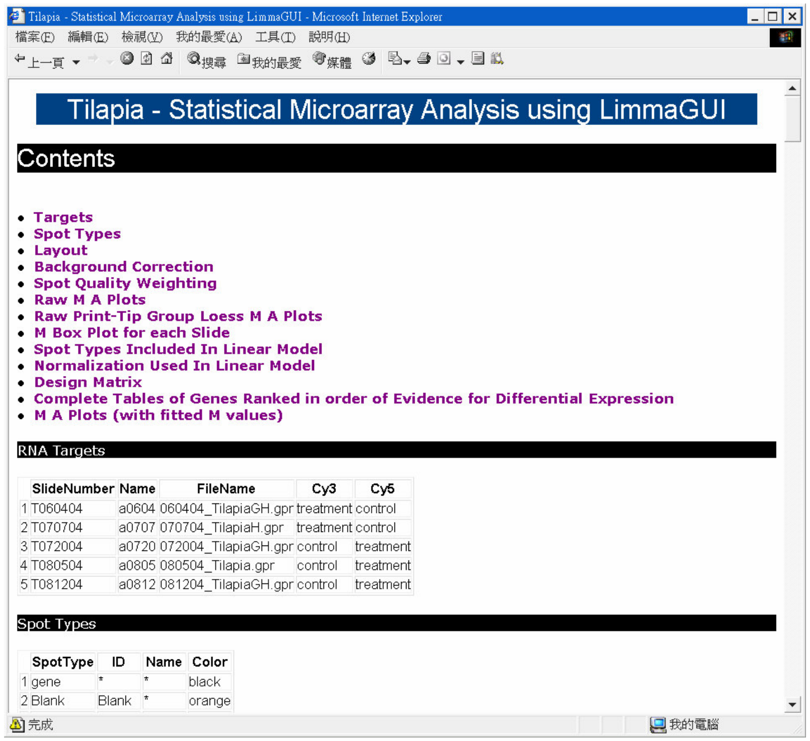
- The **q-value** gives the scientist a hypothesis testing error measure for each observed statistic with respect to pFDR.
- The **p-value** accomplishes the same goal with respect to the **type I error**, and the **adjusted p-value** with respect to **FWER**.

Software: Limma, LimmaGUI, affylmGUI



Welcome to LimmaGUI, a package for Linear Modelling of Microarray Data




Tilapia - Statistical Microarray Analysis using LimmaGUI - Microsoft Internet Explorer
檔案(F)
編輯(E)
檢視(V)
我的最愛(A)
工具(T)
說明(H)
上一頁
搜尋
我的最愛
媒體
完成
我的電腦

Tilapia - Statistical Microarray Analysis using LimmaGUI

Contents

- **Targets**
- **Spot Types**
- **Layout**
- **Background Correction**
- **Spot Quality Weighting**
- **Raw M A Plots**
- **Raw Print-Tip Group Loess M A Plots**
- **M Box Plot for each Slide**
- **Spot Types Included In Linear Model**
- **Normalization Used In Linear Model**
- **Design Matrix**
- **Complete Tables of Genes Ranked in order of Evidence for Differential Expression**
- **M A Plots (with fitted M values)**

RNA Targets

	SlideNumber	Name	FileName	Cy3	Cy5
1	T060404	a0604	060404_TilapiaGH.gpr	treatment	control
2	T070704	a0707	070704_TilapiaH.gpr	treatment	control
3	T072004	a0720	072004_TilapiaGH.gpr	control	treatment
4	T080504	a0805	080504_Tilapia.gpr	control	treatment
5	T081204	a0812	081204_TilapiaGH.gpr	control	treatment

Spot Types

	SpotType	ID	Name	Color
1	gene	*	*	black
2	Blank	Blank	*	orange

Limma: Linear Models for Microarray Data

http://bioinf.wehi.edu.au/limma/

LimmaGUI: a menu driven interface of Limma

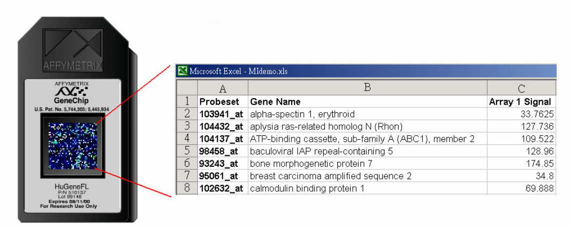
http://bioinf.wehi.edu.au/limmaGUI

- Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, Chapter 23. (To be published in 2005)
- Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3, No. 1, Article 3.

Reference

Reference for Finding Differential Expressed Genes

http://idv.sinica.edu.tw/hmwu/SMDA/DifferentialExpressedGene/index.htm



	A	B	C
1	**Probeset**	**Gene Name**	**Array 1 Signal**
2	**103941_at**	alpha-spectin 1, erythroid	33.7625
3	**104432_at**	aplysia ras-related homolog N (Rhon)	127.736
4	**104137_at**	ATP-binding cassette, sub-family A (ABC1), member 2	109.522
5	**98458_at**	baculoviral IAP repeal-containing 5	128.96
6	**93243_at**	bone morphogenetic protein 7	174.85
7	**95061_at**	breast carcinoma amplified sequence 2	34.8
8	**102632_at**	calmodulin binding protein 1	69.888

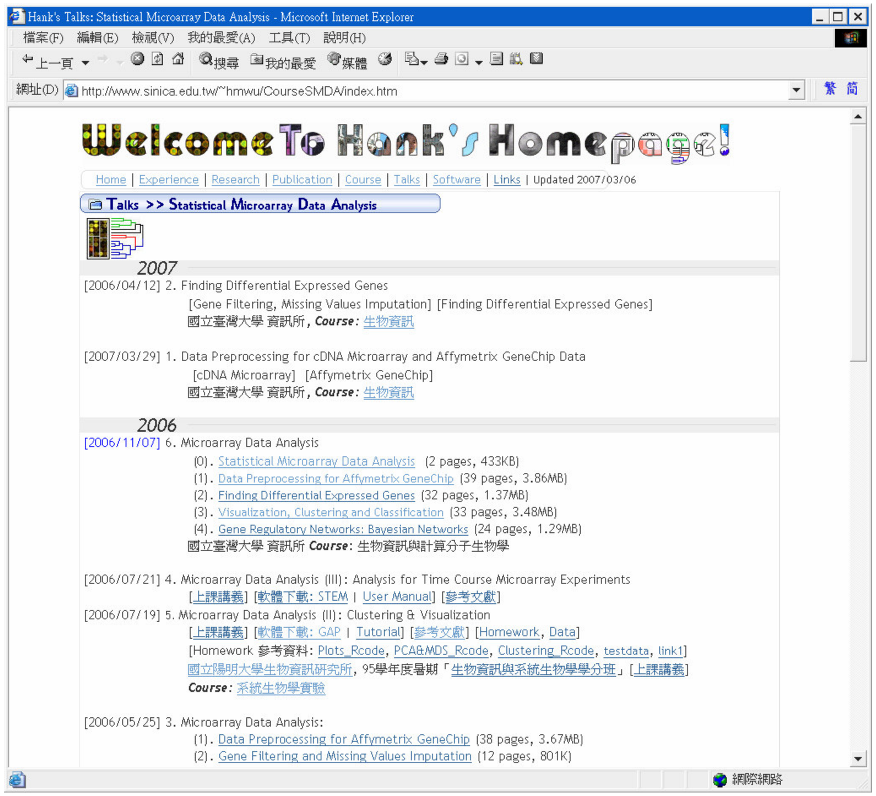
吳漢銘

E-mail: hmwu@stat.sinica.edu.tw

http://idv.sinica.edu.tw/hmwu

中央研究院 統計科學研究所

Institute of Statistical Science, Academia Sinica


Hank's Talks: Statistical Microarray Data Analysis - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
媒體
網址(D)
繁 簡
網際網路

http://www.sinica.edu.tw/~hmwu/CourseSMDA/index.htm

Welcome To Hank's Homepage!

Home | Experience | Research | Publication | Course | Talks | Software | Links | Updated 2007/03/06

Talks >> Statistical Microarray Data Analysis

2007

[2006/04/12] 2. Finding Differential Expressed Genes
[Gene Filtering, Missing Values Imputation] [Finding Differential Expressed Genes]
國立臺灣大學 資訊所, ***Course***: 生物資訊

[2007/03/29] 1. Data Preprocessing for cDNA Microarray and Affymetrix GeneChip Data
[cDNA Microarray] [Affymetrix GeneChip]
國立臺灣大學 資訊所, ***Course***: 生物資訊

2006

[2006/11/07] 6. Microarray Data Analysis
(0). Statistical Microarray Data Analysis (2 pages, 433KB)
(1). Data Preprocessing for Affymetrix GeneChip (39 pages, 3.86MB)
(2). Finding Differential Expressed Genes (32 pages, 1.37MB)
(3). Visualization, Clustering and Classification (33 pages, 3.48MB)
(4). Gene Regulatory Networks: Bayesian Networks (24 pages, 1.29MB)
國立臺灣大學 資訊所 ***Course***: 生物資訊與計算分子生物學

[2006/07/21] 4. Microarray Data Analysis (III): Analysis for Time Course Microarray Experiments
[上課講義] [軟體下載: STEM | User Manual] [參考文獻]
[2006/07/19] 5. Microarray Data Analysis (II): Clustering & Visualization
[上課講義] [軟體下載: GAP | Tutorial] [參考文獻] [Homework, Data]
[Homework 參考資料: Plots_Rcode, PCA&MDS_Rcode, Clustering_Rcode, testdata, link1]
國立陽明大學生物資訊研究所, 95學年度暑期「生物資訊與系統生物學學分班」[上課講義]
Course: 系統生物學實驗

[2006/05/25] 3. Microarray Data Analysis:
(1). Data Preprocessing for Affymetrix GeneChip (38 pages, 3.67MB)
(2). Gene Filtering and Missing Values Imputation (12 pages, 801K)