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

2006/11/07

吳漢銘

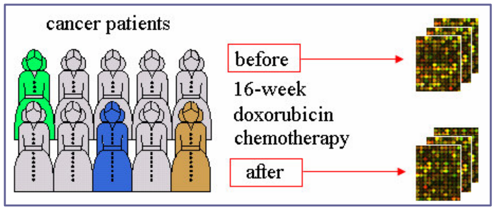
hmwu@stat.sinica.edu.tw

http://www.sinica.edu.tw/~hmwu

中央研究院 統計科學研究所

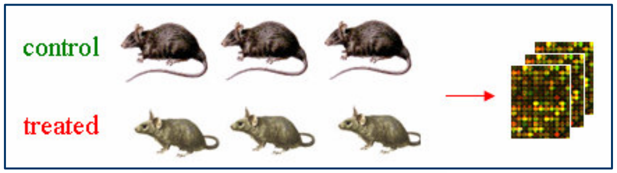
Institute of Statistical Science, Academia Sinica

Outlines



Dependent samples

- Introduction
- Hypothesis Testing



Independent samples

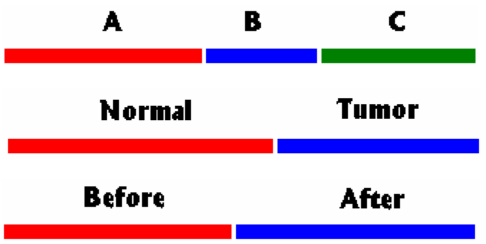
Comparison	Two Groups		More than two Groups
Hypothesis Testing	Paired data	Unpaired data	Complex data
Parametric (variance equal)	One sample t-test	Two-sample t-test	One-Way Analysis of Variance (ANOVA)
Parametric (variance not equal)	Welch t-test		Welch ANOVA
Non-Parametric (無母數檢定)	Wilcoxon Signed-Rank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

- Multiple Testing Corrections
- Software: R: limma Package

Reference for Finding Differential Expressed Genes

http://www.sinica.edu.tw/~hmwu/Talks/DE2006index.htm

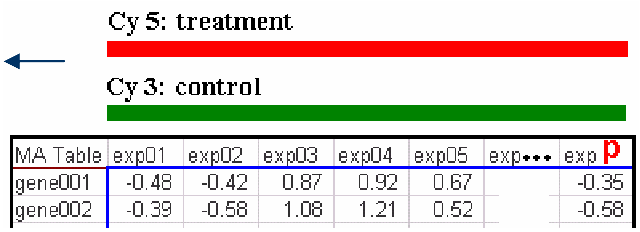
Finding Differentially Expressed Genes



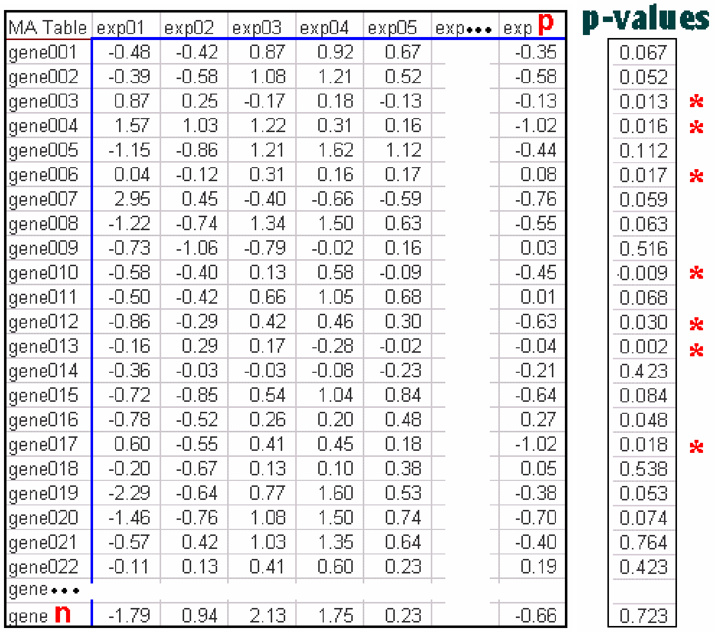
→ More than two samples

→ Two-sample (independent) ←

→ Paired-sample (dependent)

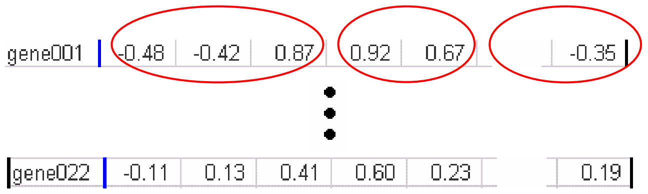

Cy 5: treatment
Cy 3: control

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

- 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.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D, (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.)

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. 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

- There are four types of small round blue cell tumors of childhood: neuroblastoma (NB), non-Hodgkin lymphoma (NHL), rhabdomyosarcoma (RMS) and Ewing tumours (EWS). Sixty-three samples from these tumours, 12, 8, 20 and 23 in each of the groups, respectively, have been hybridised 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, s2=200. => subtract bg => s1=200, s2=100 ==> 2-fold.
  (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.

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: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ $(\ldots = \mu_n)$

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.

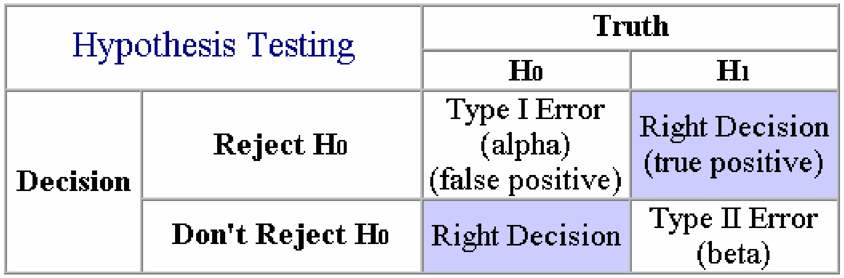
Decision Rule

- 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)

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$_1$: μ≠35 alternative hypothesis (μ>35; μ<35)
one-sided

α signifcant level: =0.05

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

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

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

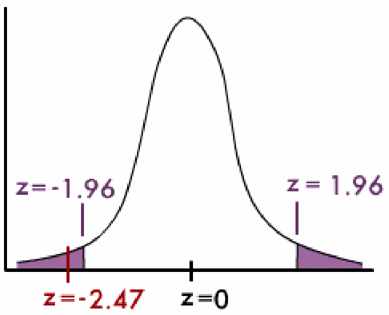
Sample Data: =33.6
test statistic: z=-2.47

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

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

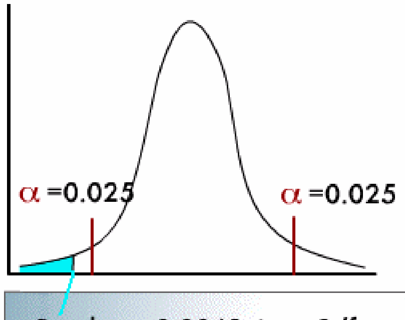
$$\text{p-value} = P_{H_0}(|Z| > z_0),\ z_0 = \frac{\bar{X}-m}{\sigma/\sqrt{n}}$$

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.

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**

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.

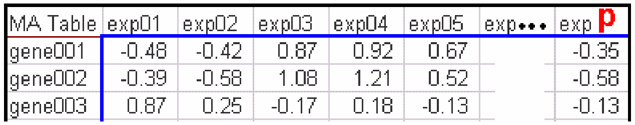
S: sample standard deviation.

n: number of observations in the sample.

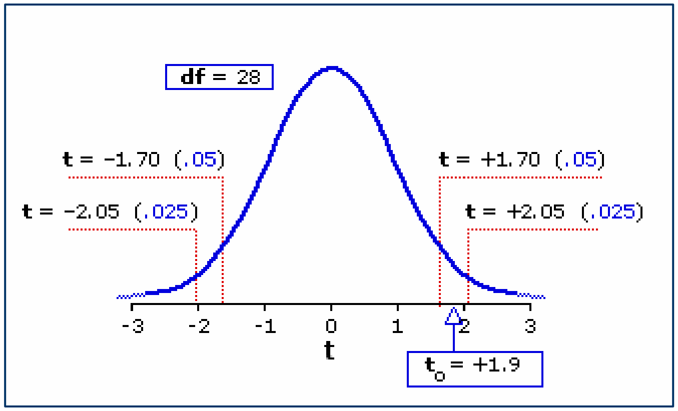
- Reject H_0 if $|t_0| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ:
  $\bar{X} - t_{\alpha/2} S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2} S/\sqrt{n}$
- p-value $= P_{H_0}(|\mathbf{T}| > t_0)$, $\mathbf{T} \sim t_{n-1}$.

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: $\bar{d} - t_{\alpha/2} S/\sqrt{n} \leq \mu_d < \bar{d} + t_{\alpha/2} S/\sqrt{n}$
- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_d)$, $\mathbf{T} \sim t_{n-1}$.

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).

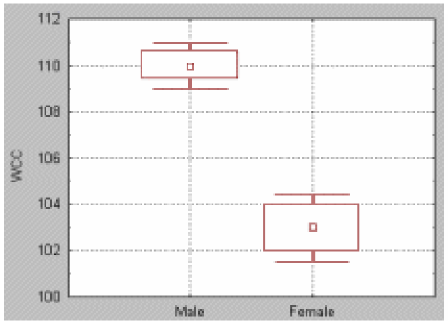
Test Statistic:

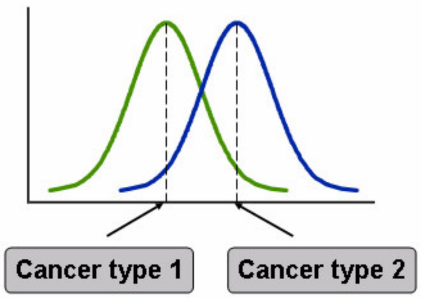
$$t_0 = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

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

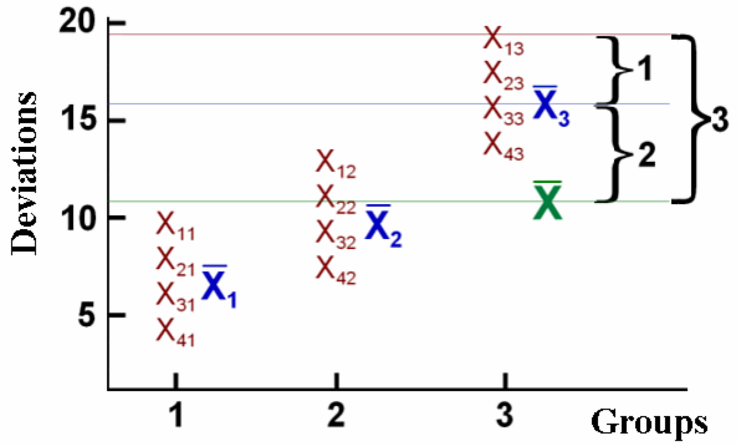
Groups

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

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

Reference for Finding Differential Expressed Genes
http://www.sinica.edu.tw/~hmwu/Talks/DE2006index.htm

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.

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.

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.

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)

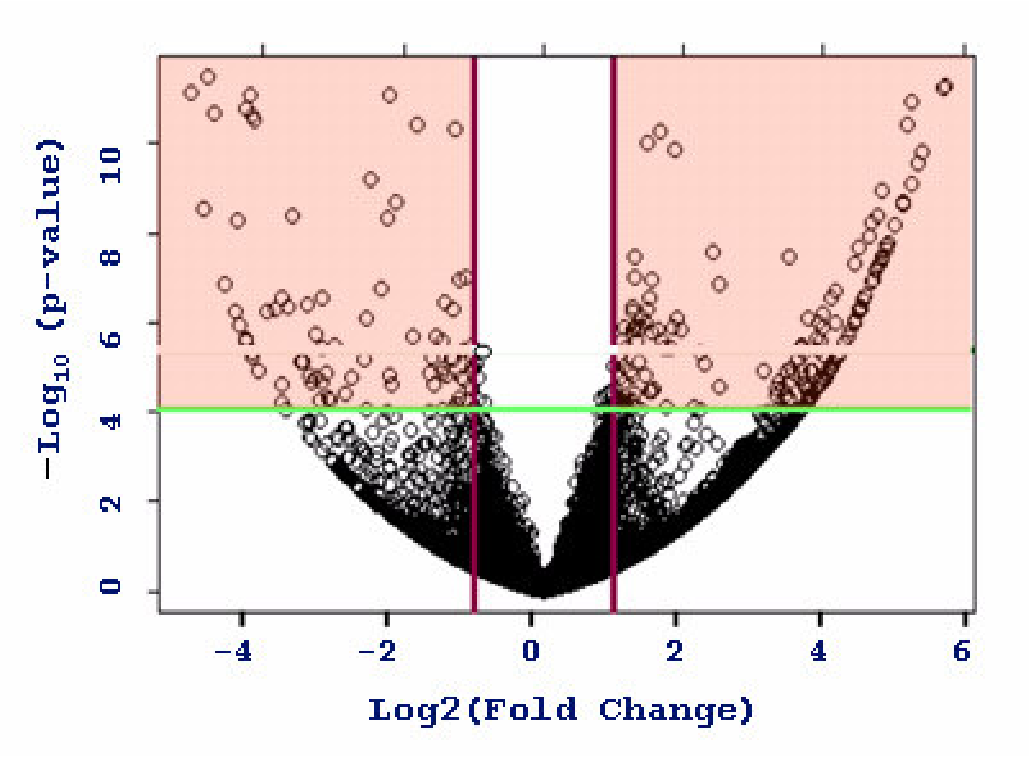
- Decrease effects of outliers (Robust)
- Not recommended if there is less than 5 replicates per group
- Needs a high number of replicates
- Less powerful

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

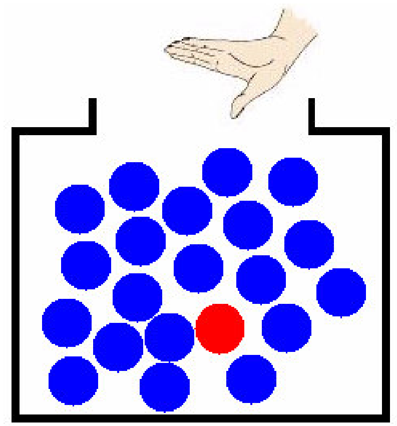
Volcano Plot

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:

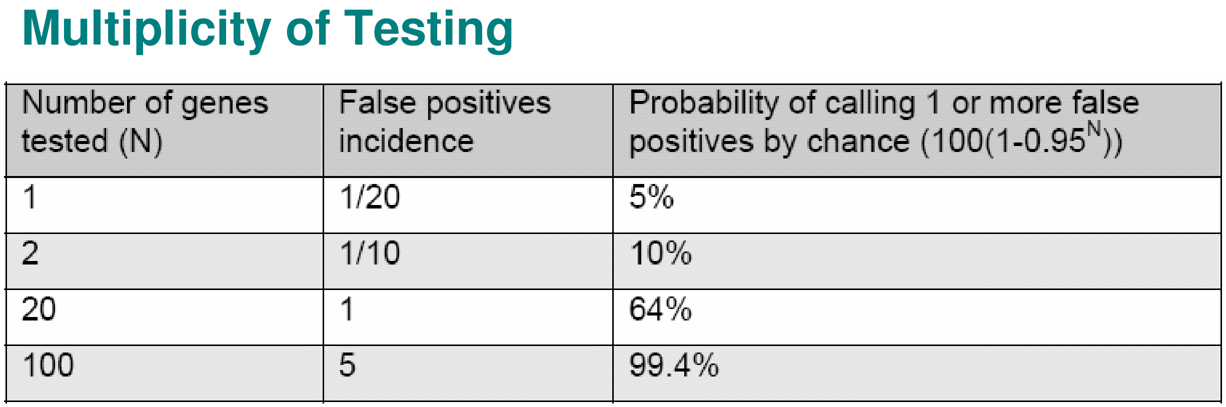
Imagine that the red marble is a false positive gene: the chance that false positives are going to be sampled is higher the more genes you apply a statistical test on.

X: false positive gene

$P(X>=1)$

$= 1-P(X=0)$

$= 1- 0.95\text{^}n$



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

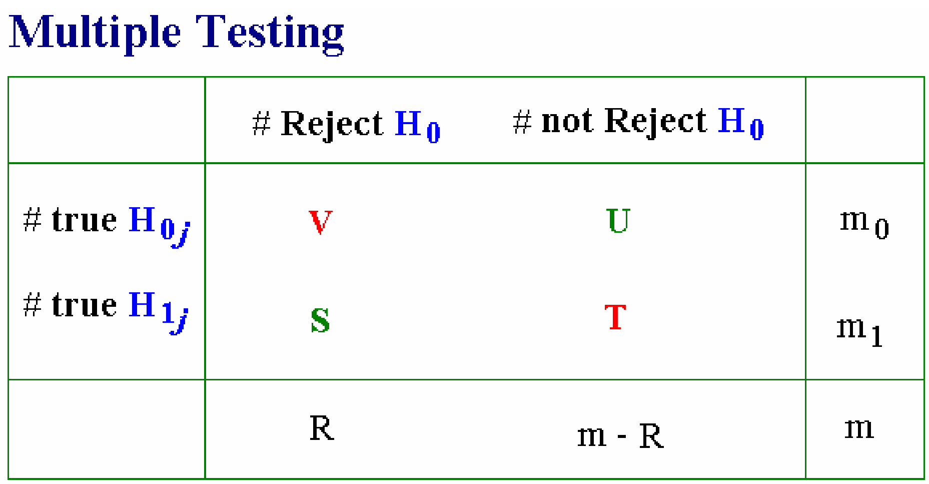
- There is a serious consequence of performing statistical tests on many genes in parallel, which is known as multiplicity of p-values.
- Take a large supply of reference sample, label it with Cy3 and Cy5: no genes are differentially expressed: all measured differences in expression are experimental 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.

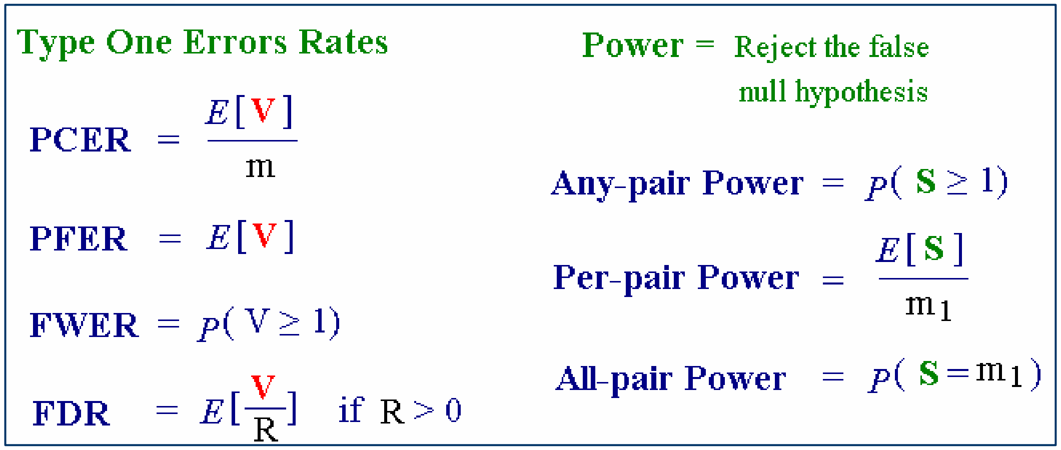


Multiple Testing

	# 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

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

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

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

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

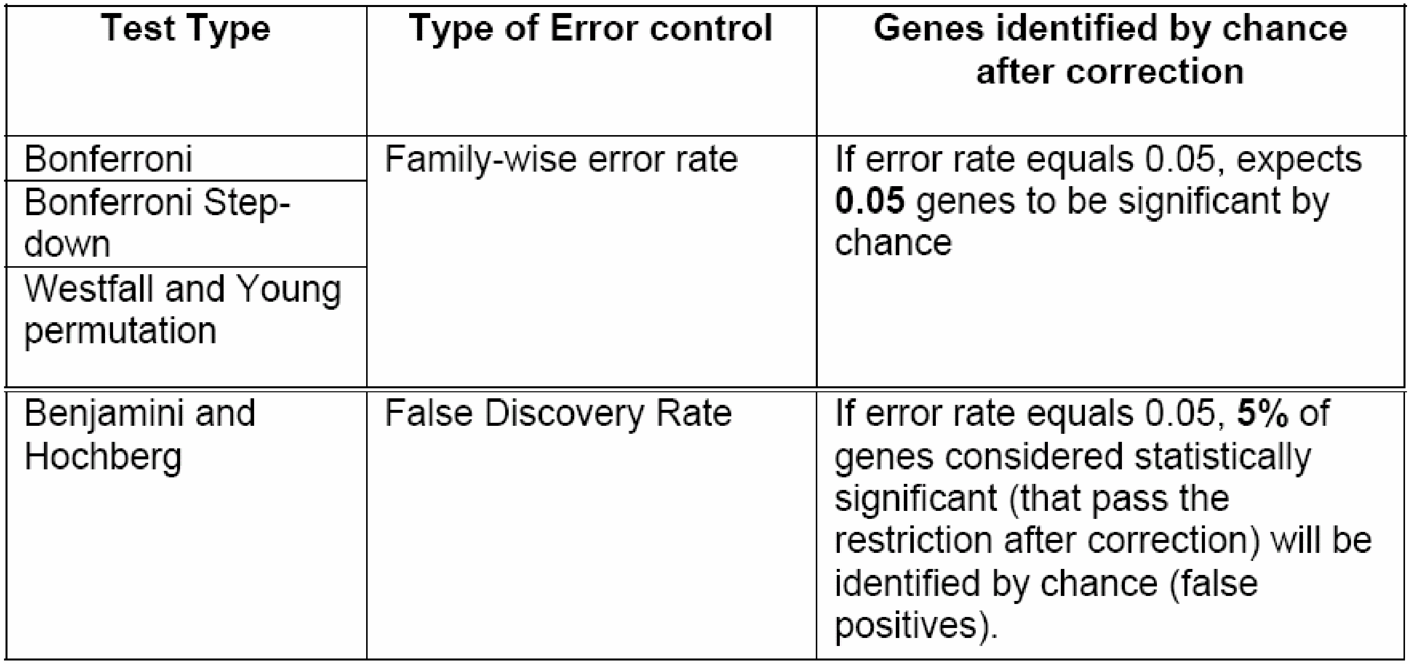
Power = Reject the false null hypothesis

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

$$\text{Per-pair Power} = \frac{E[S]}{m_1}$$

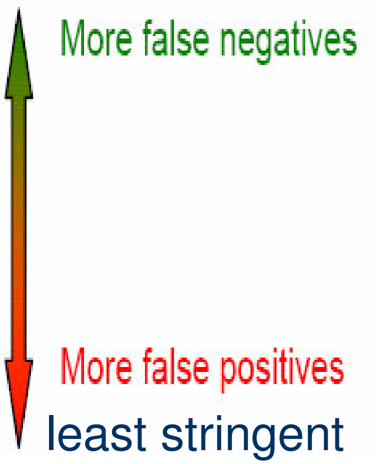
$$\text{All-pair Power} = p(S = m_1)$$

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.

Bonferroni Correction

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

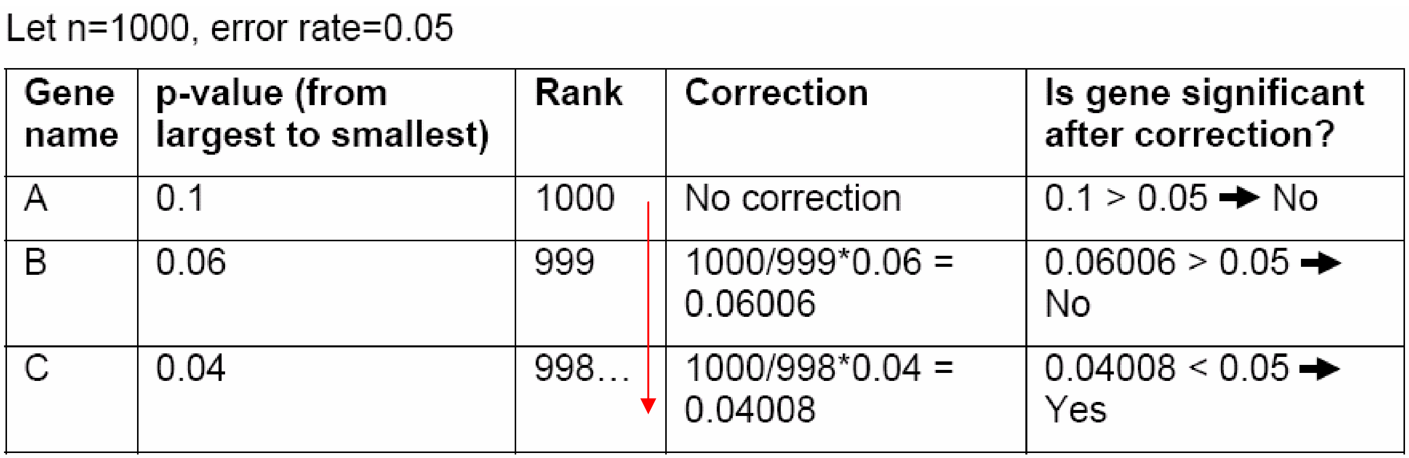
- If the corrected p-value is still below the error rate, the gene will be significant:
  - Corrected p-value= p-value * n <0.05.
  - If testing 1000 genes at a time, the highest accepted individual un-corrected p-value is 0.00005, making the correction very stringent.

- 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.

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.

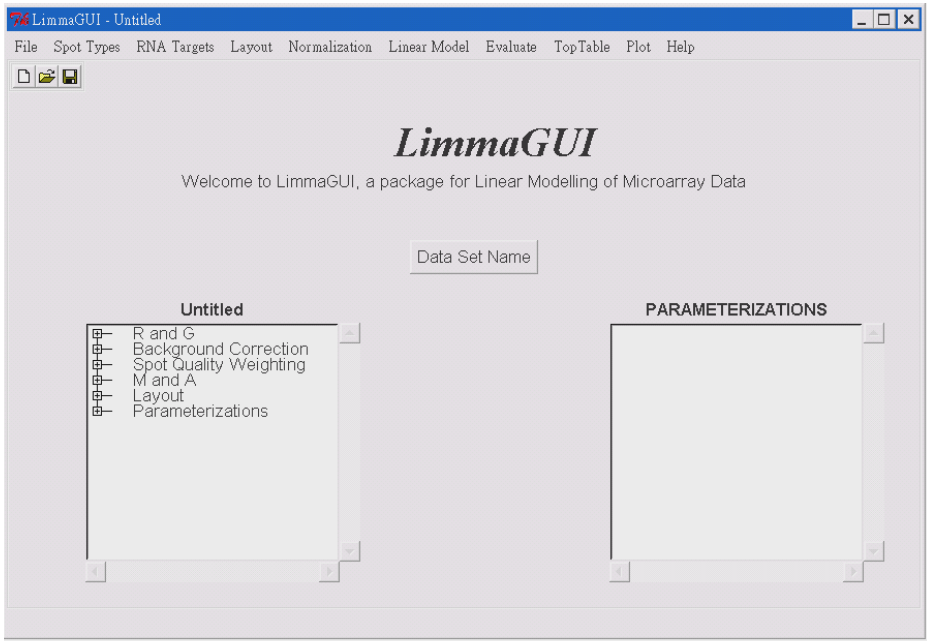
- The Westfall and Young Permutation is the only correction accounting for genes coregulation. However, it is very slow and is also very conservative.

- 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

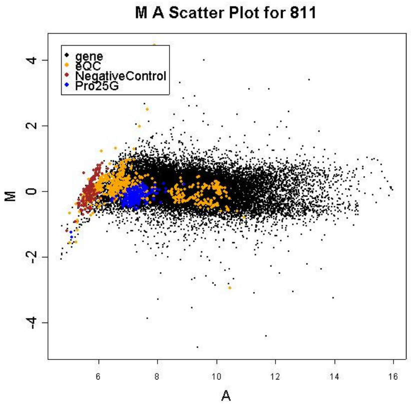
- increase p-cutoff value
- increase number of replicates
- use less stringent or no MTC
- add cross-validation experiments

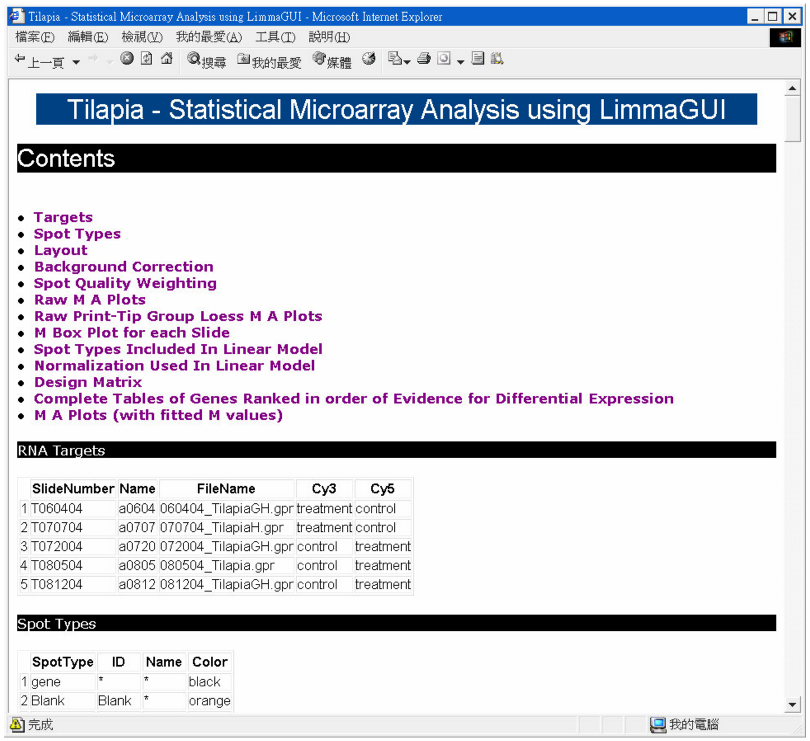
Software: Limma, LimmaGUI, affylmGUI



LimmaGUI

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.