

Finding Differentially Expressed Genes Statistics in GeneSpring 7

吳漢銘

2005年9月20日



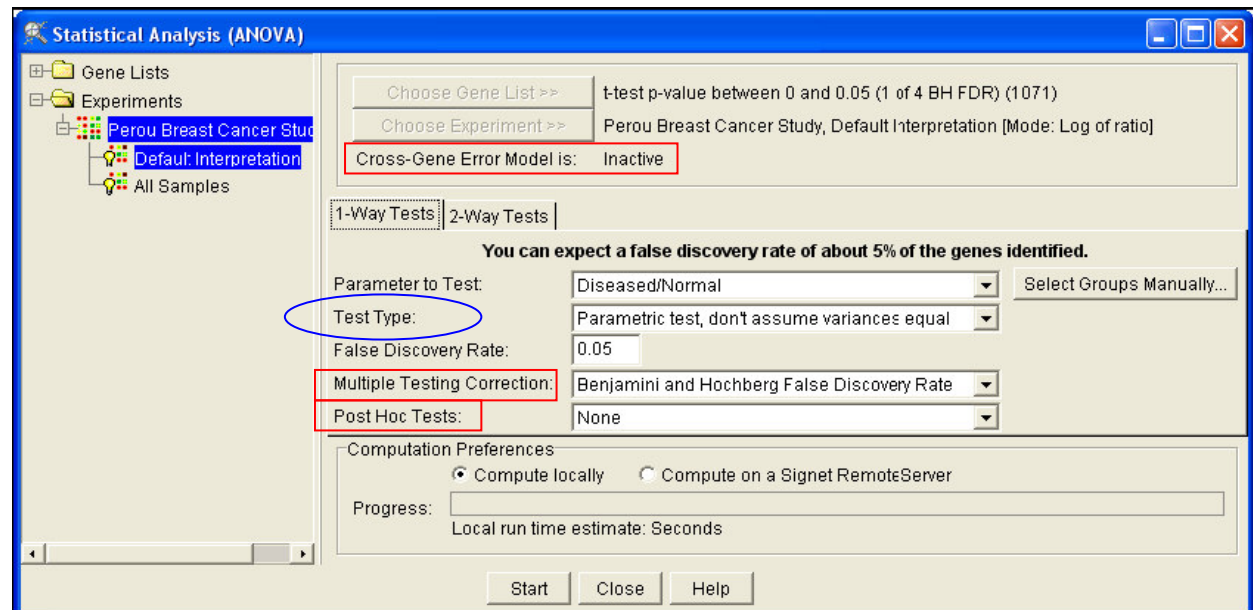
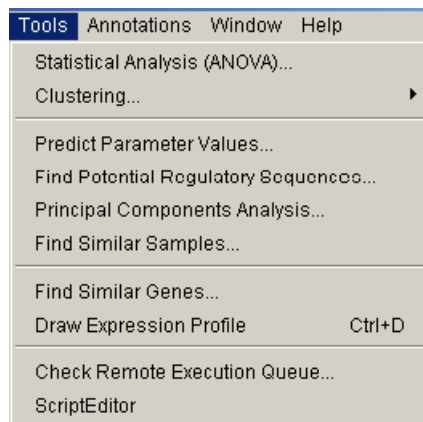
中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica

hmwu@stat.sinica.edu.tw
<http://www.sinica.edu.tw/~hmwu>

- One Sample and Two-sample t-test
- 1-way ANOVA (analysis of variance)
 - ◆ Cross-Gene Error Model (CGEM)
 - ◆ Multiple Testing Corrections (MTC)
 - ◆ Post Hoc Tests
- Interpreting ANOVA Results

Chapter 13 GeneSpring Manual 7.2

source: GeneSpring Manual 7.2



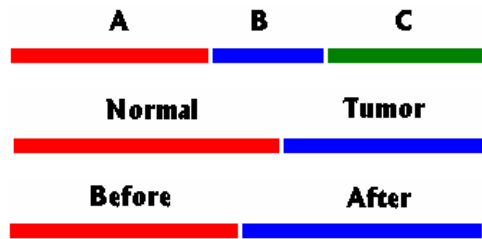
Analysis Guides

<http://www.chem.agilent.com/scripts/generic.asp?lpage=34708&indcol=Y&prodcol=Y>

GeneSpring Tutorials

<http://www.chem.agilent.com/Scripts/Generic.ASP?lPage=34743&indcol=Y&prodcol=Y>

ANOVA in GeneSpring 7.2



→ 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
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

p-values

0.067
0.052
0.013 *
0.016 *
0.112
0.017 *
0.059
0.063
0.516
-0.009 *
0.068
0.030 *
0.002 *
0.423
0.084
0.048
0.018 *
0.538
0.053
0.074
0.764
0.423
0.723

Statistical Analysis: ANOVA

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

Source: http://www.chem.agilent.com/cag/bsp/SiG/Downloads/pdf/one_way_anova.pdf

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

- Multiple Testing Corrections (MTC)
- Post Hoc Tests

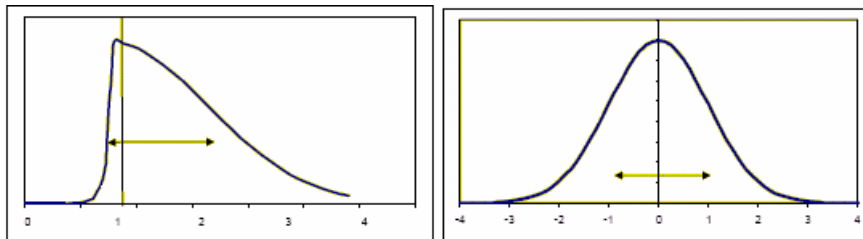
Parametric vs. Non-Parametric Test

4 / 43

有母數/無母數檢定

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.

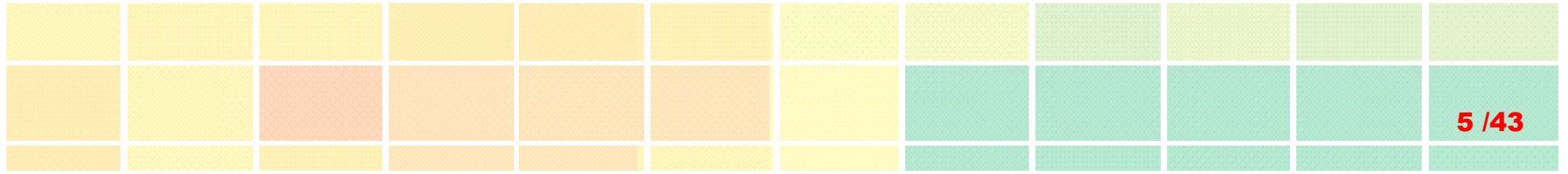


- Data for a gene in all samples for a condition (ratio)
- More accurate results if data fit a normal distribution
(Use log mode interpretation)

Non-Parametric Tests

A non-parametric test is used in place of its parametric counterpart 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



Hypothesis Testing

Hypothesis Testing

6 / 43

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.



Terminology in Hypothesis Testing

7 / 43

- 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)
 - ◆ $H_1: \mu < 2.5$.
 - ◆ $H_1: \mu \neq 2.5$.
- 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.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 (\dots = \mu_n)$

The p -values

8 / 43

- 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 H_0 | 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 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)

If A Result is Statistically Significant

9 / 43

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.

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

$$H_0 : \mu = \mu_0$$

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

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.

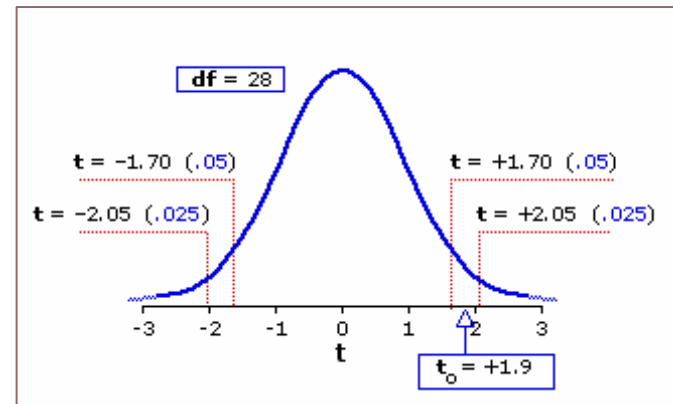
n : number of observations in the sample.

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

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).

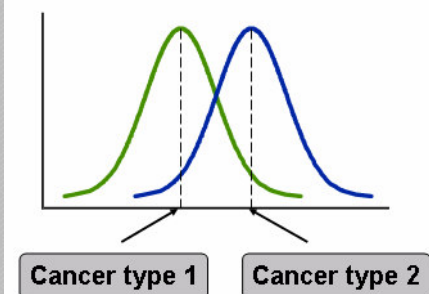
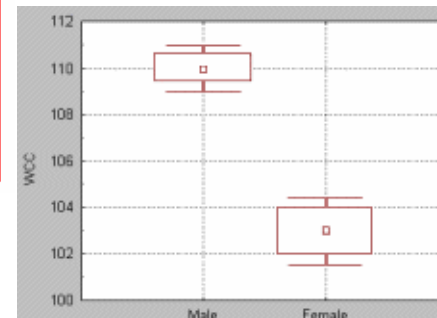
Test Statistic:

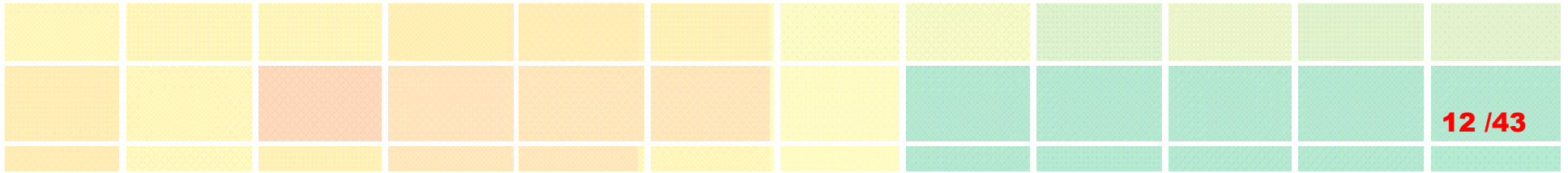
$$t_0 = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

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

for heterogeneous variances:
adjusted df

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





One-Way Analysis of Variance (ANOVA)

Analysis Guides: One-way ANOVA

http://www.chem.agilent.com/cag/bsp/SiG/Downloads/pdf/one_way_anova.pdf

GeneSpring Tutorials: Two-Way ANOVA View

http://www.chem.agilent.com/cag/bsp/SiG/Downloads/Tutorial/2_way_anova.viewlet/2_way_anova_viewlet_swf.html

GeneSpring Tutorials: One-Way ANOVA & Post-hoc Tests

http://www.chem.agilent.com/cag/bsp/SiG/Downloads/Tutorial/1_way_anova_and_post_hoc.viewlet/1_way_anova_and_post_hoc_viewlet_swf.html

One-Way ANOVA

13 / 43

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.

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)

One-Way ANOVA (conti.)

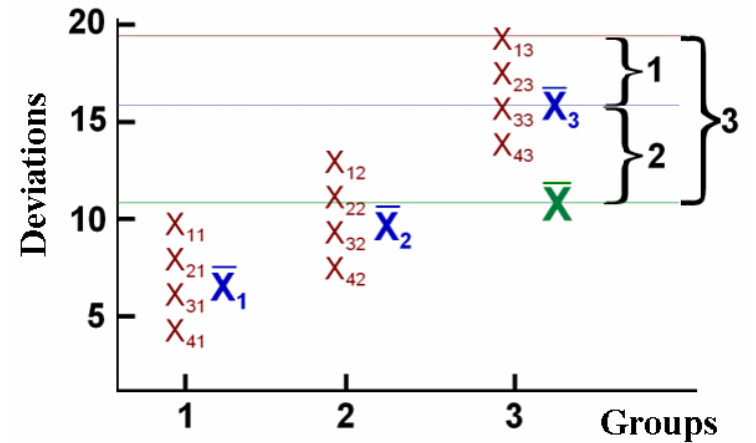
Groups

1	2	...	j	...	k
X_{11}	X_{12}	...	X_{1j}	...	X_{1k}
X_{21}	X_{22}	...	X_{2j}	...	X_{2k}
			...		
X_{i1}	X_{i2}	...	X_{ij}	...	X_{ik}
\vdots			\vdots		X_{nk}
X_{n1}	X_{n2}	...	X_{nj}	...	

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

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

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



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

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

$$X_{ij} = \mu_j + \epsilon_{ij} \quad \begin{matrix} i = 1, \dots, n_j \\ j = 1, \dots, k \end{matrix}$$

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

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

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

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

16 / 43

Welch's F Test

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

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

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

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

$$i = 1, \dots, n_j$$

$$j = 1, \dots, k$$

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

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

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

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

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

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

Wilcoxon Rank-Sum Test

(Mann-Whitney U 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 not less than the critical value 77, so we reject H_0 .

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

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 n_i should be **at least 5** for the approximation to be valid.

$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
 $H_1 : \mu_i \neq \mu_j$ for at least one set of i and j

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

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

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

Groups						Rank Data					
1	2	...	j	...	k	1	2	...	j	...	k
X_{11}	X_{12}	...	X_{1j}	...	X_{1k}	R_{11}	R_{12}	...	R_{1j}	...	R_{1k}
X_{21}	X_{22}	...	X_{2j}	...	X_{2k}	R_{21}	R_{22}	...	R_{2j}	...	R_{2k}
				
X_{i1}	X_{i2}	...	X_{ij}	...	X_{ik}	R_{i1}	R_{i2}	...	R_{ij}	...	R_{ik}
⋮			⋮		$X_{n_k k}$	⋮			⋮		$R_{n_k k}$
$X_{n_1 1}$	$X_{n_2 2}$...	$X_{n_i j}$...		$R_{n_1 1}$	$R_{n_2 2}$...	$R_{n_i j}$...	

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

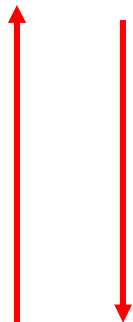
$$x = G(\alpha) = G(F(x))$$

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

Recommendations

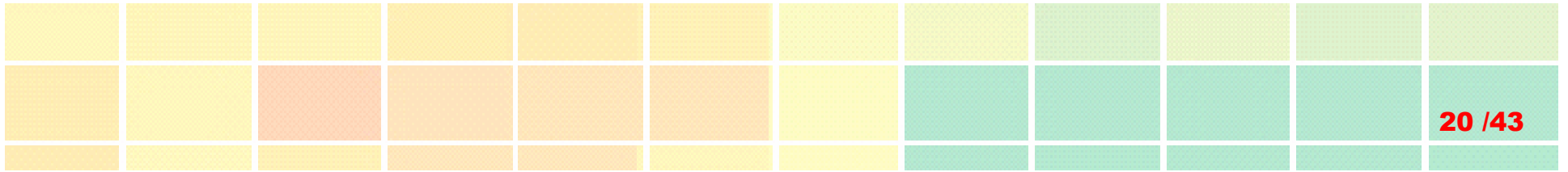
- Student's t-test/ANOVA (variances assumed equal) should be used
 - ◆ if very few replicates are available, or
 - ◆ if some groups being analyzed do not have replicates.
 - ◆ At least one condition has replicates.
- **Default:** the Welch test (variances not assumed equal) should be used
 - ◆ for most cases.
 - ◆ if variance in the population is unknown.
- Error Model should be used
 - ◆ only when there are **no replicates**.
 - ◆ The parametric test, use all available error estimate, is similar to Welch test but has better variance estimates for small sample sizes.
- Non-parametric test makes the least assumptions about your data but should be used
 - ◆ only when there are more than 5 replicates per group.

More Power



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

Fewer Assumptions



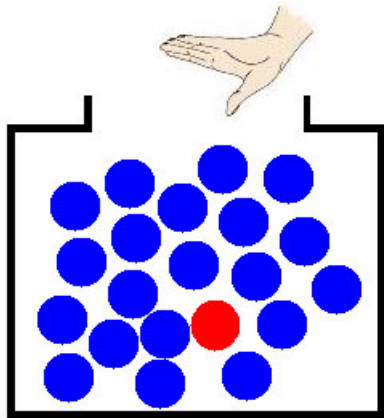
Multiple Testing

Analysis Guides: Multiple Testing Corrections

<http://www.chem.agilent.com/cag/bsp/SiG/Downloads/pdf/mtc.pdf>

Multiplicity of Testing

21 / 43



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

$$P(X \geq 1)$$

$$= 1 - P(X = 0)$$

$$= 1 - 0.95^n$$

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%

Assigning Significance and Multiplicity of Testing

22 / 43

- There is a serious consequence of performing statistical tests on many genes in parallel, which is known as **multiplicity of p-values**.

Example: Take a large supply of reference sample, label it with **Cy3** and **Cy5**.

- Since every sample hybridized to the arrays is the same reference sample, we know that 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

Question: Is this gene truly differentially expressed, or could it be a false positive result?

Multiple Testing Corrections (MTC)

23 / 43

- When testing for potential differential expression across those conditions, each gene is considered independently from one another.
- In other words, a t-test or ANOVA is performed on each gene separately.
- The incidence of *false positives* (or genes falsely called differentially expressed when they are not) is proportional to the number of tests performed and the critical significance level (p-value cutoff).

When a two-sample t-test is performed on a gene, the probability by which the gene's expression level will be considered significantly different between two groups of samples is expressed by the **p-value**.

- ◆ The p-value is the probability that a gene's expression level are different between the two groups due to chance.
- ◆ A p-value of 0.05 signifies a 5% probability that the gene's mean expression value in one condition is different than the mean in the other condition by chance alone.
- ◆ If 10,000 genes are tested, 5% or 500 genes might be called significant by chance alone.

Multiple Testing

- 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

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

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

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

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

Power = Reject the false null hypothesis

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

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

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

Multiple Testing Corrections (MTC)

25 / 43

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 default	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).



- 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

26 / 43

- 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 = $p\text{-value} * n < 0.05$.
 - ◆ 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)

27 / 43

- 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 = $p\text{-value} * (n - i + 1) < 0.05$
- 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:

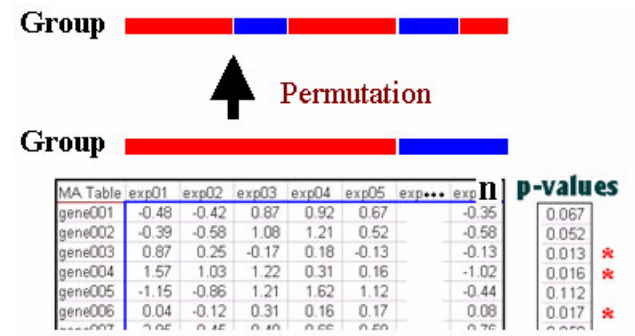
Let $n=1000$, error rate=0.05

Gene name	p-value before correction	Rank	Correction	Is gene significant after correction?
A	0.00002	1	$0.00002 * 1000 = 0.02$	$0.02 < 0.05 \rightarrow$ Yes
B	0.00004	2	$0.00004 * 999 = 0.039$	$0.039 < 0.05 \rightarrow$ Yes
C	0.00009	3	$0.00009 * 998 = 0.0898$	$0.0898 > 0.05 \rightarrow$ 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**.



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

(4) Benjamini and Hochberg FDR

29 / 43

- 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.

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

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	$0.1 > 0.05 \rightarrow$ No
B	0.06	999	$1000/999 * 0.06 = 0.06006$	$0.06006 > 0.05 \rightarrow$ No
C	0.04	998...	$1000/998 * 0.04 = 0.04008$	$0.04008 < 0.05 \rightarrow$ Yes

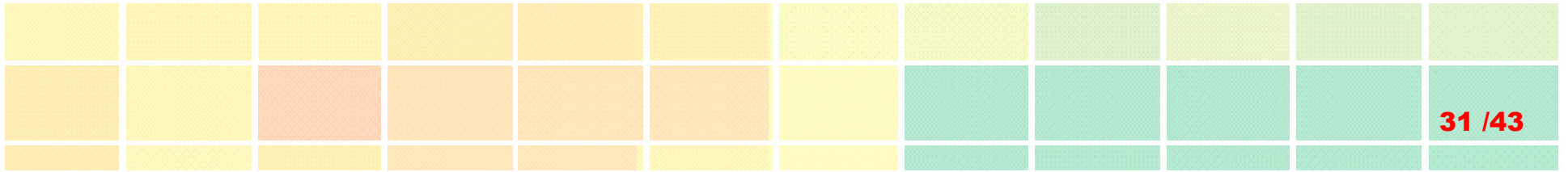
Recommendations

30 / 43

- The default multiple testing correction 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 in the Filter on Confidence or the Statistical Analysis tool. (GeneSpring)

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



Cross-Gene Error Model

Analysis Guides: Cross-Gene Error Model

http://www.chem.agilent.com/cag/bsp/SiG/Downloads/pdf/error_model.pdf

GeneSpring Tutorials: Using the Cross-Gene Error Model

http://www.chem.agilent.com/cag/bsp/SiG/Downloads/Tutorial/cross_gene_error_model.viewlet/cross_gene_error_model_viewlet_swf.html

Cross-Gene Error Model

32 / 43

- The standard deviation of replicates estimates the precision of the expression intensity for a gene.
- However, this is a very imprecise measurement if few replicates are available and not possible if no replicates are available.

Why use the Error Model

- The parametric tests require error information from the normalized data for assessing significance.
- The Cross-Gene Error Model provides a more accurate estimate for the precision of a gene by combining *measurement variation* and *between-sample variation* information.

If no replicates are available

- ◆ Calculate SD and SE
- ◆ Display error bars
- ◆ T-test p-value
- ◆ Color by significance
- ◆ Filter for reliable genes using global error variances
- ◆ Statistical analyses

If replicates are available

- ◆ More precision for all the calculations above.

Cross-Gene Error Model

33 / 43

Requirement: data is normalized such that 1.0 is point of reference.

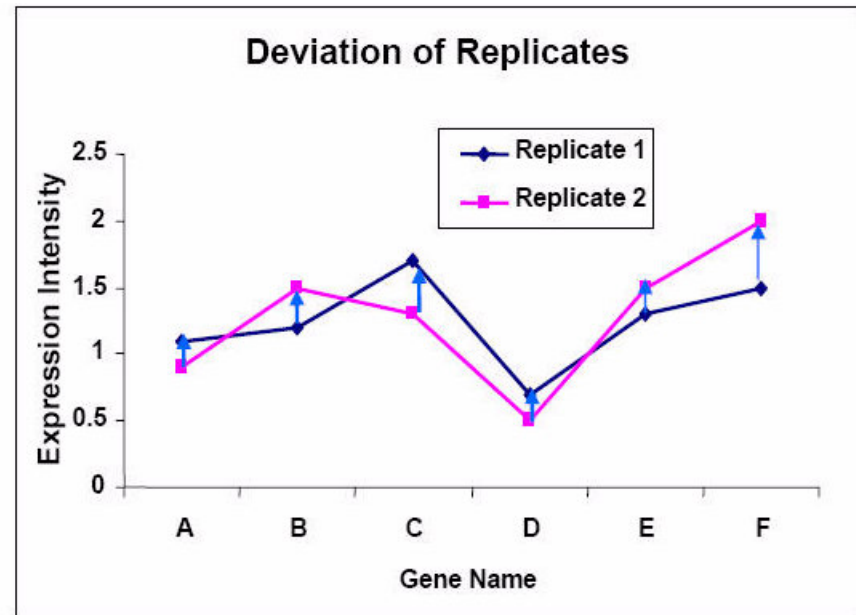
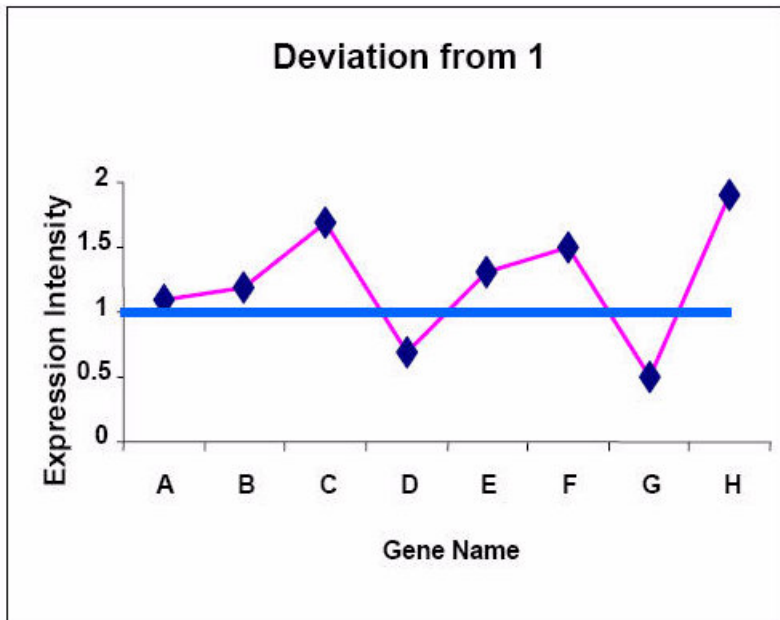
(1) Deviation from 1.0 Model

- Assumes a general-purpose array was used, where most genes have little biological variability.
- Assumes most genes are not changing cross experimental conditions
- Assumes non-changing genes whose measurement occur in the same range have similar error
- Coefficients for the two component error model equation are fitted for each chip.
- Changes in expression level are due to measurement error for most genes

(2) Replicates Model

- Assumes variability between replicates is similar for all genes with similar measurement level.
- Assumes most genes are not changing cross treatment or experimental conditions
- Assumes the non-changing gene whose measurements occur in the same range have similar standard error.
- Variability b/w replicates is similar for all genes with similar measurement levels
- Coefficients for the two component error model equation are fitted for each set of replicates chips.

Deviations



Control strength = $C(\text{Per chip}) * C(\text{Per Gene})$

Normalized value = Raw Signal / Control Strength

How Does the CGEM Work?

35 / 43

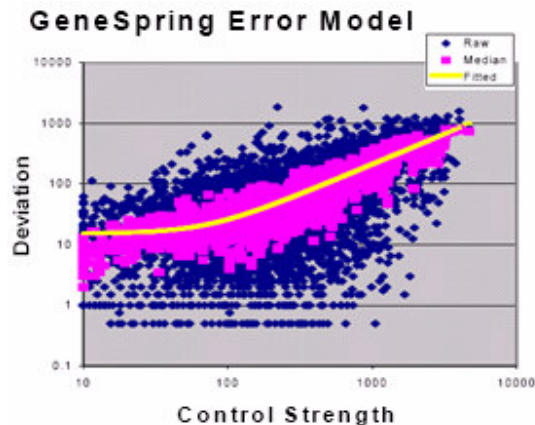
- Deviation for each gene is computed and these values are then ordered by increasing control strength. (blue)
- Medians of the deviation and control strength are calculated for each set of 11 points. (pink)
- The deviation on squared control strength is computed on the reduced data set and a separate curve is fitted for each sample or for each replicate group.

$$S(\text{raw})^2 = a^2 + b^2 C^2 \longrightarrow \frac{\text{raw}}{\text{control}} \longrightarrow S(\text{norm})^2 = a^2 / C^2 + b^2$$

a = fixed (absolute) error (base), $\pm 0.05 \mu\text{g}$

b = proportional error (proportion), $\pm 5\%$

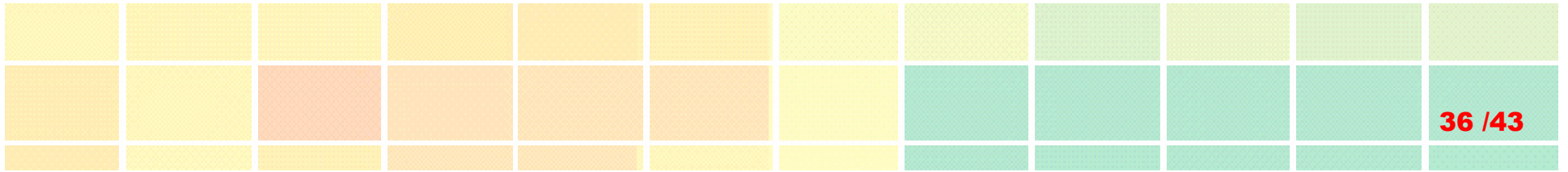
C = control value (control strength)



In two-color array, the control value is the intensity from the control channel.

In one-color array, the control value is a synthetic value that is the product of the normalization steps.

Stated another way, the control value is the factor that your raw value was divided by to obtain the normalized value. For example, if a per-chip and a per-gene normalization were applied to your data set, then the control value is:
Control = C(per-chip) * C(per-gene)



Post Hoc Tests

GeneSpring Tutorials: One-Way ANOVA & Post-hoc Tests

http://www.chem.agilent.com/cag/bsp/SiG/Downloads/Tutorial/1_way_anova_and_post_hoc.viewlet/1_way_anova_and_post_hoc_viewlet.swf.html

Post Hoc Tests

Applicable when comparing more than 2 groups.

One-way ANOVA model

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 (\dots = \mu_n)$$

If H_0 is 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

38 / 43

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

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

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

- 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

39 / 43

honestly significant difference (HSD)

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

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.

- $(k \text{ vs } 1, k \text{ vs } 2, \dots, k \text{ vs } k-1) (k-1 \text{ vs } 1, k-1 \text{ vs } 2, \dots, k-1 \text{ vs } k-2) \dots (\dots 2 \text{ vs } 1)$

$r = k, df, \alpha \rightarrow$ 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.

Interpreting ANOVA Results

40 / 43

1-way ANOVA without Post Hoc Test

New Gene List (592 genes)

Name: 1-Way ANOVA

Folder:

Notes: P: 24 HOUR COMPOUND STUDY SET Default Interpretation - Genes from reliable genes with statistically significant differences when grouped by 'Drug Agent'; parametric test, variances not assumed equal (Welch ANOVA). p-value cutoff 0.05, multiple testing correction: Benjamini and Hochberg False Discovery Rate. This restriction tested 1,651 genes. About 5.0% of the identified genes would be expected to pass the restriction by chance.

Gene Lists

- Simplified Gene Ontology
 - Cyp11 b2 [AA924224] Rattus norvegicus cytochrome P-450 11-beta hyc
 - Lnk [AI138146] Rattus norvegicus Lnk1 mRNA, complete cds
 - Gpx1 [AA964788] Rat mRNA for glutathione peroxidase
 - [AA899219] EST, Highly similar to tubulin T beta15 [R.norvegi
 - Neo1 [AA997838] Rattus norvegicus neogenin mRNA, partial cds

Similar lists: Show as List Show as Navigator

P value	List Name
0.0	1-Way ANOVA
0.007232821	Genes differentiating Agent
0.008014404	Cytochrome

NOTE:

When overlap occurs,

Eg.

$$\begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \\ \hline & & & \\ \hline & & & \end{array}$$

- The only conclusion is $\mu_1 \neq \mu_3 = \mu_4$
- No conclusion about X_2

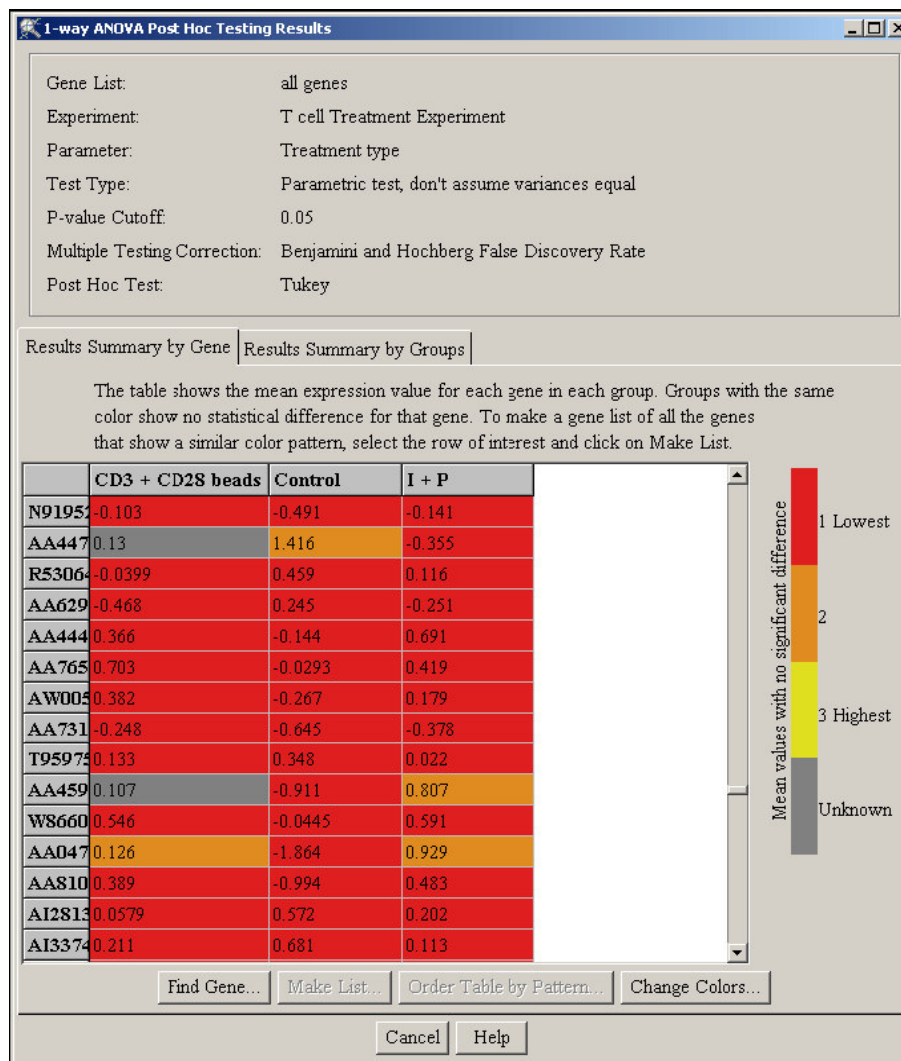
■ Why do I get zero genes passing the restriction when I perform statistical analysis?

- ◆ Analysis criteria might be too stringent (low p-value cut-off and conservative multiple testing correction)
- ◆ Not enough replicates in each group resulting in insufficient power to detect real differences between groups under study
- ◆ Biologically, there may not be differential gene expression.

Interpreting ANOVA Results

41 / 43

1-way ANOVA with Post Hoc Test, Summary by Gene Tab



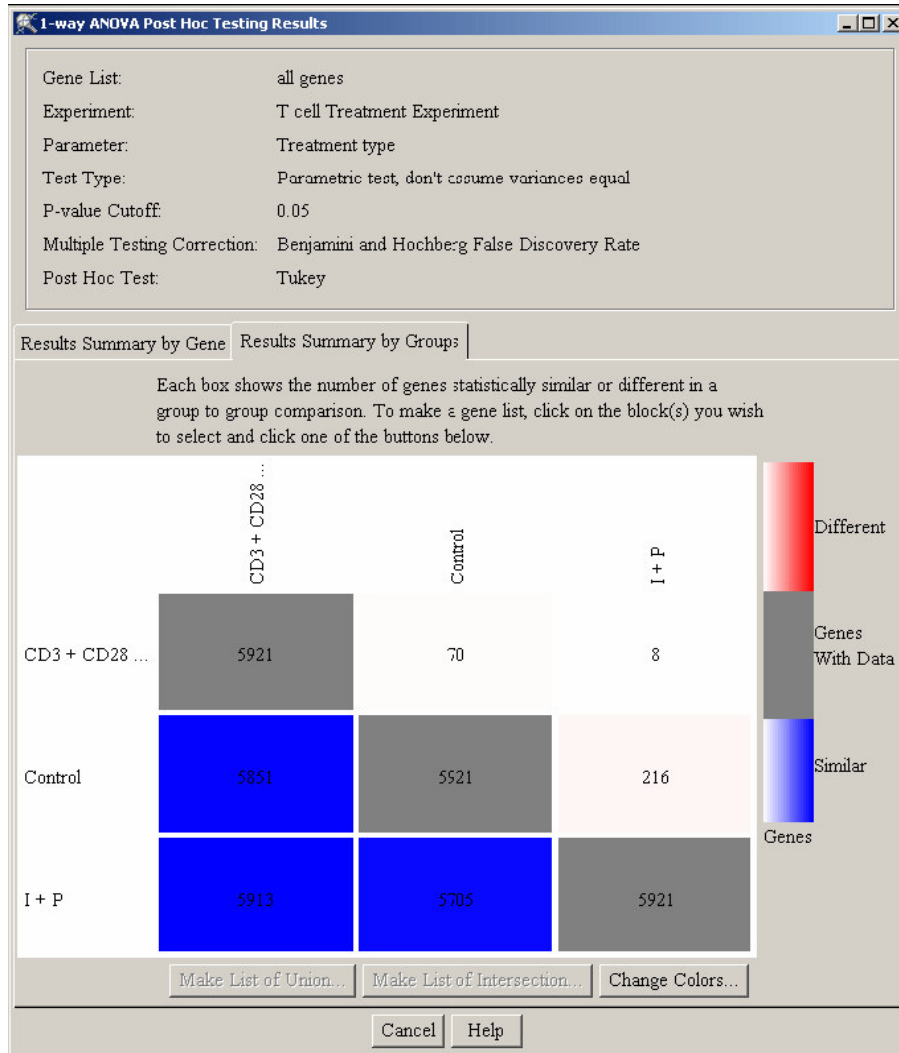
- The Results Summary by Gene tab displays the mean expression level by group for each significant gene.
- Figure lists all the genes considered differentially expressed by statistical criteria.
- For each gene, the coloring indicates which groups differ significantly from the others.
- Groups of the same color show no significant difference for that gene.
- Groups of different colors differ significantly from each other.
- Groups with the highest color differential have the most significant difference.
- A group colored grey is considered to be unknown because the significance of its mean difference cannot be determined with confidence from the test used.

Source: GeneSpring Manual 7.2

Interpreting ANOVA Results

42 / 43

1-way ANOVA with Post Hoc Test, Summary by Groups Tab



Source: GeneSpring Manual 7.2

- The Results Summary by Groups tab displays a matrix with rows and columns indexed by parameter values.
- Each cell corresponds to a combination of groups.
- The numbers in the lower half of the matrix represent the number of genes that differ significantly between the groups.
- The numbers in the upper half are the genes which show no significant difference.
- Figure indicates the total number of genes that are statistically differentially expressed between the groups being compared in the matrix.
- Greater color saturation indicates greater difference (or similarity).
- Total number of genes analyzed is shown in the box colored grey.

To Be Continued...

43 / 43



hmwu@stat.sinica.edu.tw

<http://www.sinica.edu.tw/~hmwu/Talks/index.htm>



309


中央研究院

Clustering and Characterizing Data

Statistics in GeneSpring 7

吳漢銘

2005年9月21日

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

hmwu@stat.sinica.edu.tw
<http://www.sinica.edu.tw/~hmwu>