

Multiple Testing Corrections

Contents at a glance

I. What are Multiple Testing Corrections?	2
A. Introduction	2
B. Importance of Multiple testing corrections	2
II. Where to find Multiple Testing Corrections?	3
III. Technical details	4
A. Bonferroni correction	5
B. Bonferroni Step-down (Holm) correction	5
C. Westfall and Young Permutation	6
D. Benjamini and Hochberg False Discovery Rate	6
IV. Interpretation	7
V. Recommendations.....	7
VI. Frequently asked questions.....	8
VII. References	9

I. What are Multiple Testing Corrections?

A. Introduction

Multiple testing corrections adjust p-values derived from multiple statistical tests to correct for occurrence of false positives. In microarray data analysis, false positives are genes that are found to be statistically different between conditions, but are not in reality.

B. Importance of Multiple testing corrections

A typical microarray experiment measures several thousand genes simultaneously across different conditions. 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).

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. Wouldn't you say that you have a much higher chance to sample the red marble? Indeed, there is a 64% chance in the latter case (see Table below).

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.

Example: 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.

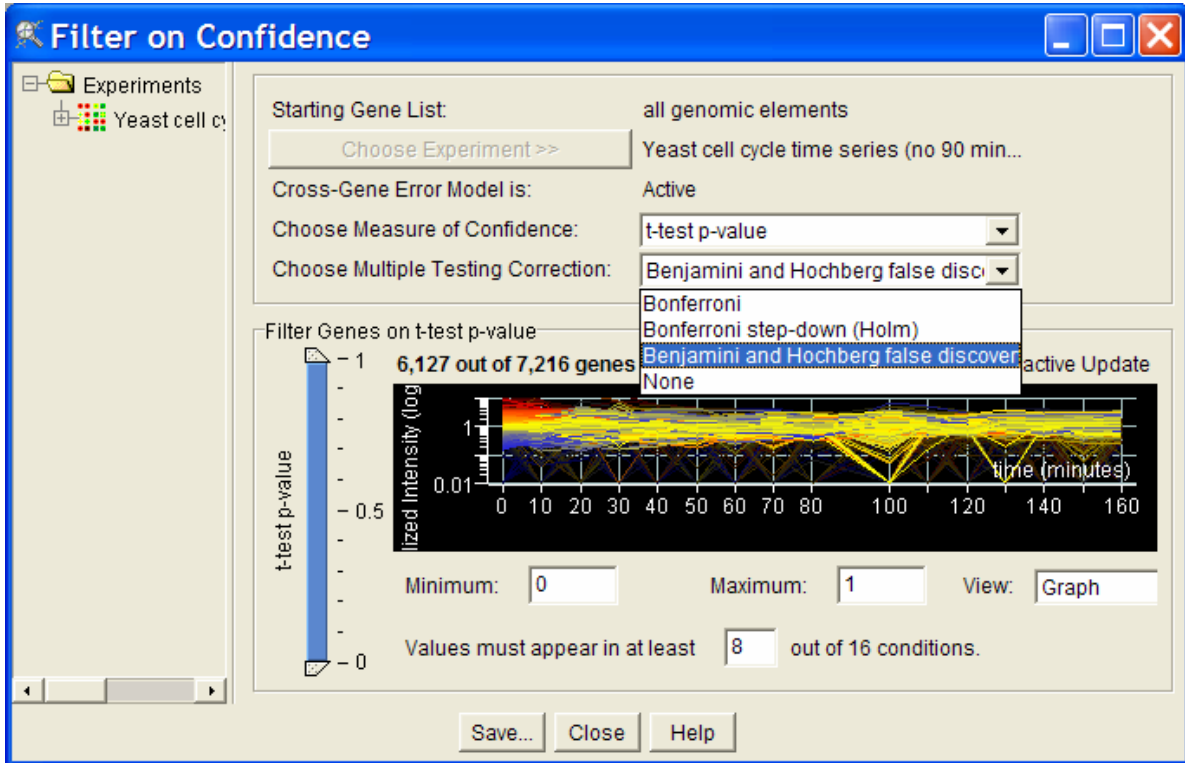
Number of genes tested (N)	False positives incidence	Probability of calling 1 or more false positives by chance ($100(1-0.95^N)$)
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

This is why it is important to correct the p-value of each gene when performing a statistical test on a group or genes. Multiple testing correction adjusts the individual 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*.

II. Where to find Multiple Testing Corrections?

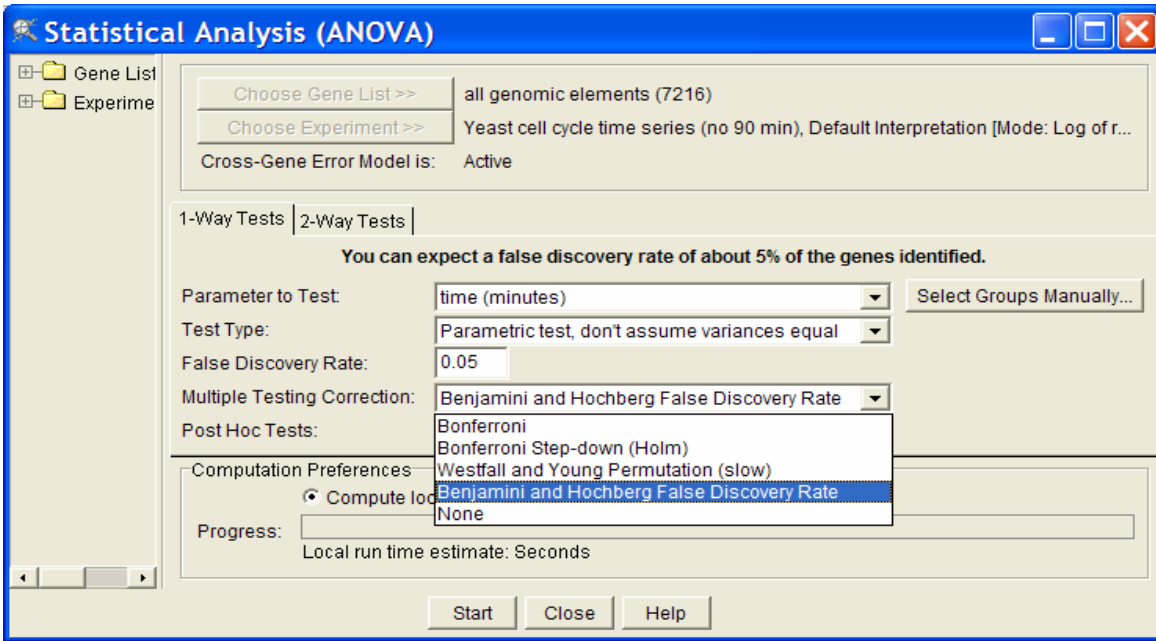
The Multiple testing corrections can be found in two types of filters involving different statistical tests:

- 1) Filter on Confidence, under the **Filtering** menu:



The multiple testing corrections are available for a filter on the t-test p-value only. The t-test p-value filter uses the p-value calculated for each gene based on a one-sample t-test. The multiple testing corrections will be applied on these p-values.

- 2) The Statistical Analysis tool, found under the **Tools** menu. They are available for 1-way and 2-way tests.



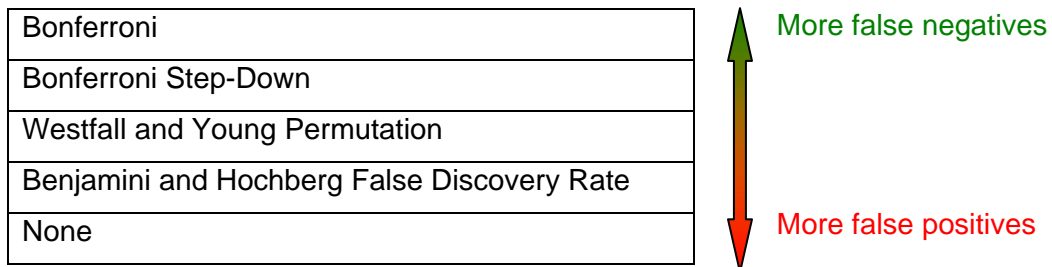
The multiple testing corrections are available for all test types. Correction will be applied on the p-value calculated based on a two-sample t-test or ANOVA. Depending on what multiple testing correction is selected, the p-value cutoff wording changes, referring to the error rate type (e.g., False Discovery Rate with the Benjamini and Hochberg False Discovery Rate multiple testing correction). See section IV for details on interpretation.

III. Technical details

GeneSpring offers four types of multiple testing corrections:

- A. Bonferroni
- B. Bonferroni Step-down (Holm)
- C. Westfall and Young Permutation
- D. Benjamini and Hochberg False Discovery Rate

The methods are listed in order of their stringency, with the Bonferroni being the most stringent, and the Benjamini and Hochberg FDR being the 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.



Each method is described in more details below. The descriptions are only approximate for clarity purposes. In the examples, an error rate of 0.05 and a gene list of 1000 genes are assumed.

A. 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:

$$\text{Corrected P-value} = \text{p-value} * n < 0.05$$

As a consequence, if testing 1000 genes at a time, the highest accepted individual 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.

B. Bonferroni Step-down (Holm) correction

This correction is very similar to the Bonferroni, but a little less stringent:

- 1) The p-value of each gene is ranked from the smallest to the largest.
- 2) The first p-value is multiplied by the number of genes present in the gene list: if the end value is less than 0.05, the gene is significant:

$$\text{Corrected P-value} = \text{p-value} * n < 0.05$$

- 3) The second p-value is multiplied by the number of genes less 1:

$$\text{Corrected P-value} = \text{p-value} * n-1 < 0.05$$

- 4) The third p-value is multiplied by the number of genes less 2:

$$\text{Corrected P-value} = \text{p-value} * n-2 < 0.05$$

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

Because it is a little less corrective as the p-value increases, this correction is less conservative. However the Family-wise error rate is very similar to the Bonferroni correction (see table in section IV).

C. 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:

- 1) P-values are calculated for each gene based on the original data set and ranked.
- 2) The permutation method creates a pseudo-data set by dividing the data into artificial treatment and control groups.
- 3) P-values for all genes are computed on the pseudo-data set.
- 4) The successive minima of the new p-values are retained and compared to the original ones.
- 5) 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.

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.

D. Benjamini and Hochberg False Discovery Rate

This correction is the least stringent of all 4 options, and therefore tolerates more false positives. There will be also less false negative genes. Here is how it works:

- 1) The p-values of each gene are ranked from the smallest to the largest.
- 2) The largest p-value remains as it is.
- 3) The second largest p-value is multiplied by the total number of genes in gene list divided by its rank. If less than 0.05, it is significant.
Corrected p-value = $p\text{-value} * (n/n-1) < 0.05$, if so, gene is significant.
- 4) The third p-value is multiplied as in step 3:
Corrected p-value = $p\text{-value} * (n/n-2) < 0.05$, if so, gene is significant.

And so on.

Example:

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

As you can see from the example above, 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.

IV. Interpretation

The interpretation of the number of false-positive genes after multiple testing correction can be summarized in the table below.

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).

Family-wise error rates allow very few occurrences of false positives.

False discovery error rate allows a percentage of called genes to be false positives.

V. Recommendations

1. 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.

2. The Bonferroni correction is the most stringent test of all, but offers the most conservative approach to control for false positives.
3. The Westfall and Young Permutation is the only correction accounting for genes coregulation. However, it is very slow and is also very conservative.
4. 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. Example:

If there are 10,000 genes in the gene list, assuming a p-value cutoff of 0.05, the individual p-value for a gene must be at least $0.05/10,000=0.000005$ to pass a statistical test after a Bonferroni Multiple Testing Correction.

If there are 4000 genes in the gene list, the maximal individual p-value must be only 0.0000125.

VI. Frequently asked questions

Q. Why do I get more genes with a smaller gene list than with all genes list when I perform a one-way ANOVA using a Multiple Testing Correction?

A. As multiple testing corrections depend on how many genes are tested, the larger the gene list, the more stringent the correction will be. For instance, the Bonferroni correction will multiply the p-values of each gene by the number of tests performed. The more tests (or the more genes, since there is one test per gene), the smaller the p-value must be to pass the restriction.

Q. Why should I use a Multiple Testing Correction? If I select one, no genes pass the restriction.

A. Even though no genes pass the statistical restriction, it is important to keep in mind that genes that pass a restriction without multiple testing correction might all be false positives, thus not significant at all. If you have 10,000 genes in your genome, and perform a statistical analysis, a p-value cutoff of 0.05 allows a 5% chance of error. That means that 500 genes out of 10,000 could be found to be significant by chance alone.

Q. What should I do if no genes pass the statistical test when I apply the multiple testing correction?

A. To improve your statistical results, try one or more of the following suggestions:

- 1) Increase the p-value cutoff or error rate.
- 2) Increase the number of replicates in your experiment.
- 3) Select a smaller list of genes to use with your analysis. The smaller the list, the less stringent the multiple testing correction will be (see first question).

- 4) Select a less stringent or no multiple testing correction. If you choose to apply no multiple testing correction, rank the genes by their p-values to inspect them manually. Genes with the smallest p-values will be the most reliable.

VII. References

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society B*, 57, 289 -300.

Dudoit, S. et al. (2000). "Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments". Technical Report # 578.

Ge Y. et al. (2003). "Resampling-based Multiple Testing for Microarray Data Analysis". Technical Report # 633.

Holm, S. (1979). "A Simple Sequentially Rejective Bonferroni Test Procedure," *Scandinavian Journal of Statistics*, 6, 65 -70.

Westfall P.H., Young S.S. (1993) *Resampling-based Multiple Testing*. Wiley, New York.