The problem of multiple testing and its solutions for genom-wide studies

How to cite: Gyorffy B, Gyorffy A, Tulassay Z:. The problem of multiple testing and its solutions for genom-wide studies. **Orv Hetil**, *2005;146(12):559-563*

ABSTRACT

The problem of multiple testing and its solutions for genome-wide studies. Even if there is no real change, the traditional p = 0.05 can cause 5% of the investigated tests being reported significant. Multiple testing corrections have been developed to solve this problem. Here the authors describe the one-step (Bonferroni), multi-step (step-down and step-up) and graphical methods. However, sometimes a correction for multiple testing creates more problems, than it solves: the universal null hypothesis is of little interest, the exact number of investigations to be adjusted for can not determined and the probability of type II error increases. For these reasons the authors suggest not to perform multiple testing corrections routinely. The calculation of the false discovery rate is a new method for genome-wide studies. Here the p value is substituted by the q value, which also shows the level of significance. The q value belonging to a measurement is the proportion of false positive measurements when we accept it as significant. The authors propose using the q value instead of the p value in genome-wide studies.

Free keywords: multiple testing, Bonferroni-correction, one-step, multi-step, false discovery rate, q-value

List of abbreviations:
FDR: false discovery rate
FWER: family familywise error rate
SNP: Single Nucleotide Polymorphism
PFP: proportion of false positives

INTRODUCTION

A common feature in all of the 'omics studies is the inspection of a large number of simultaneous measurements in a small number of samples. Then, one must decide whether the findings are truly causative correlations or just the byproducts of multiple hypothesis testing. This is particularly important in transcriptomic, proteomic and genomic tests such as DNA and protein microarrays, deep sequencing, SNP-arrays, TaqMan gene expression assays which have gained widespread use and are now part of everyday practice in research. The majority of these methods are based on techniques having a high intrinsic variability between individual measurements. These are not only related to the biological unpredictability, but also to the efficiency of various techniques, like hybridization, which can change on a broad scale. There is no mathematical formula available to correct these intrinsic unevenness, therefore the application of a multiple testing correction method is the preferable way to avoid over-inflating the results in any of these studies.

When measuring 20 different parameters in a patient simultaneously and setting the significance threshold at the traditionally acknowledged 0.05, on average one parameter will be reported to be different as compared to the healthy reference population even without any biological significance. Not taking into account the possible impact of multiple simultaneous testing can greatly increase the probability of false positive findings. In 'omics studies we test many hypotheses simultaneously, thus the overall consequences of this 5% are drastically magnified. For example, when investigating 10 thousand genes and setting p to 0.05, 500 genes will be found as significant even without any real correlation. Clearly, appropriate multiple testing corrections are needed for the correct interpretation of the results.

The first multiple testing correction tests were set up in line with the statistical hypotheses of Neyman and Pearson in the 1920s [1]. Back then, the main goal was to predict the number of defects in industrial production. How can we estimate the overall number of defect light bulbs when we check 20 out of 1000?

In today's genomic research the correction protects us from making over-optimistic assumptions after finding few minimally significant parameters when randomly investigating a large set of parameters. Similar scenario arises when the same test is performed on different subgroups, for example in the investigation of the role of gene polymorphisms in several diseases. Furthermore, in studies randomly searching for correlations without any pre-defined hypothesis the use of multiple testing correction is of utmost importance.

A DECISION TREE FOR METHOD SELECTION

In general, the aim of using a statistical test can be "exploration" or "validation". As for "validation" the expected end-results are limited, these are not subject of multiple testing corrections in most cases. In "experimental" tests sometimes the need for a multiple testing correction can be worked out by combining the data: should a gene deliver a significant difference in both male and female patients, the combination of the two groups should also result in a significant correlation. While such a permutation immediately eliminates the need for multiple hypothesis testing, the significance can be smaller is some groups.

How to select the best suited method? Since all these methods are based on the p-values, the actual decision is always on a scale between accepting more false positives or more false negatives. How can we organize simple decision nodes for the selection of the best technique? For this, we have set up three nodes: (1.) Do you have a pre-defined set of candidate features? (2.) Do you perform fishing for results? (3.) Do you have large noise? A statistician might argue that this rule of thumb selection do not correspond to given statistical hypotheses. However, our aim was not to give an ultimate commandment but to set up the nodes according to our actual research experience as representing the most common questions arising in real-life studies.

Ad 1: In a genomic study, the measurement of a large number of genes is usually more practical than measuring a set of candidate genes. For example one selects a set of genes related to cancer as those having binding sites for a given transcription factor. In this case, the researcher hypothesizes the role of the transcription factor in the pathogenesis of the disease. Performing a genome-wide microarray for the measurement of mRNA expression changes is much cheaper than setting up a custom microarray. However, in this case the researcher does not have to correct for the total number of genes on the array. The actual question in these scenarios is "how many of my significant genes are not significant". Therefore, in such cases one should define the rejection area - the chance of making a false discovery among the "significant" tests. This can be made by computing the false discovery rate (FDR) and the q-value.

Ad 2: In many studies we do not have any pre-existing hypothesis and we randomly explore the possibility of correlations between the measured parameters and the investigated condition. This process is sometimes referred to as "fishing for results". However, during the analysis of a SNP array having 100.000 SNPs the random correlation is highly probable. Clearly, this scenario will result in the highest proportion of false positive findings. Therefore,

the most stringent multiple testing control must be applied, and here we recommend a one-step multiple correction technique.

Ad 3: When choosing the best suited method for a particular experiment it is important to take into consideration that stepwise procedures offer a more dynamic approach to control for false positives, than single-step procedures. Therefore stepwise procedures are more suitable for those who are not fishing for candidate variables. Studies remaining after the first two nodes require an intermediate balance between false positives and false negatives. Here, the selection can be based on the intrinsic noise of the used technology. Machineries having high noise are for example DNA microarrays, while low-noise technologies are RT-PCR assays (techniques with higher reproducibility will generally have lower noise). For these a step-up, for the others a step-down technique can be recommended.

The above decision tree is summarized in **Figure 1**. Using this tree one can rank the available multiple testing methods for the selection of the most appropriate one for a given study. In the next chapter we will briefly summarize traditional and newer techniques of multiple testing as well as the more subtle approaches like the computation of the false discovery rate and the q-value. These descriptions are aimed to allow the understanding of the purpose of the test and thus we have not elaborated on all details for each algorithm - for a more detailed description of a particular methods please check the full references (see **Table 1**). Related software tools will be listed in the proceeding chapter.

SYNOPSIS OF THE MOST COMMON TECHNIQUES

Generally, in any hypothesis test we compare the null hypothesis to our "alternative" hypothesis. Type I error means rejecting the null hypothesis when it is true and type II error means accepting the null hypothesis, while the alternative hypothesis is true. When designating the rejection level by α *(alpha)*, we reject every part having a type I error smaller than α.

Order p-values in ascending order: i=1, ... n, and let be H_i the null hypothesis belonging to p_i. A common feature of the various methods is rejecting the null hypothesis for the smallest p-values. The difference between them is in the proportion of rejected hypotheses.

1. One-step methods

The p-values are compared to a factor computed using α and n, and for p values smaller than α the null hypothesis is rejected. A common characteristic of the one-step methods is the independence between individual hypotheses and the simultaneously measured other hypotheses. A typical one-step method is the Bonferroni-correction, in which the p-value is multiplied with the number of measurements [2]:

$$p'_i = np_i \leq \alpha.$$

Consider following example: our aim is p <0.05 and five measurements are executed - here the threshold (for p') will be 0.01. The significance will of course remain 0.05, and not the calculated 0.01. The Bonferroni method can be used in a reverse order: not the p'-value is generated, but the original p-value is multiplied by the number of measurements (in our case p=0.05 for five measurements equals with p' = 0.25 - this is only possible until p' tops at 1).

Another well-known one-step method is the Sidak correction [3, 4], in which the null hypothesis is rejected where:

$$p'_i = 1 - (1 - p_i)^n \leq \alpha$$

One-step methods for 500-600 or more tests can increase terribly the type II error, making these approach not suitable for many of the genomic and proteomic analyses, but for those who want to establish strong biomarkers.

As a general reference for one-step and multi-step methods, their principles are depicted in **Figure 2**.

2. Multi-step methods

The test results are arranged incrementally. One can start at the smallest and to the largest p-value - this is called step-down method (Holm [5],Westfall and Young [6]), or start at the largest and go the smallest p value - these are step-up methods (Hochberg[7], Simes [8]). A one-step method can be advanced very simply to a step-down method: consider the smallest p-value adjusted for n multiplicity. In case we can reject the associated hypothesis, this p-value can be removed from the pool of the p-values. Then, the next lowest (but higher than the previous one) p-value is adjusted (for n-1). This must be carried on as long a hypothesis is not rejected. This method can be built on both Bonferroni-correction, and the Sidak-method.

In performing a step-up method, one must start at the largest p-value, and move towards smaller values. After the first significant p-value is reached, all null hypotheses with a smaller p-value are rejected. Generally, step-up methods deliver the highest number of significant p values, while the Bonferroni-correction delivers the smallest number. Step-down methods can be used as a balance between trimming the p values and retaining as many significant as possible.

3. Controlling the false discovery rate

The Bonferroni, step-down, and step-up techniques are controlling family familywise error rate (FWER), the probability of making one or more false discoveries, or type I errors among all the hypotheses. In 'omics studies, one can compute the false discovery rate (FDR) proposed by Benjamini and Hochberg [9] which gives better control of the number of rejected hypotheses than FWER methods. FDR is defined as the expected (E) proportion of erroneously rejected hypotheses (V) among all rejected hypotheses (R):

$$FDR = E\left[V / R \mid R>0\right] \times Prob\,(R>0)$$

Thus, Benjamini and Hochberg defined FDR as the proportion of incorrectly rejected hypotheses relative to the total number of rejected hypotheses.

What is the difference between FWER and FDR? In FWER a fixed error rate is used to define the rejected area, in FDR a fixed rejected area is used to define the error rate. For example, a p-value of 0.05 means that an average of 5% of the true null hypotheses are accepted as significant ones, while 5% FDR means that in 5% of the variables accepted as significant the null hypotheses is actually true.

Taking into account of the possibility of R = 0, the positive FDR can be computed using a modification suggested by Storey:

$$pFDR = E\left[V / R \mid R>0\right]$$

Positive FDR is the probability that the findings are false [10].

It is important to discuss the possible relationship between variables. Theoretically, three possibilities arise: 1. no connection between the variables, 2. a loose relationship between the variables, 3 a general relationship exist between the variables. For the first and the second cases the pFDR is very accurate [11], but it cannot be used in the third case. In a general 'omics study, the most likely form of dependence between variables is loose - the

main reason behind this is the existence of genetic regulatory pathways, and the requirement for the coordinated regulation of multiple genes to achieve a phenotypic change. An additional important factor is that technology issues, like cross hybridization generated by similar genes also increase the dependence between the variables.

When using pFDR we assume the independence of the p-values from each other. Therefore, we must mention the method of Fernando et al. for multiple dependent tests [12], which do not depend on the correlation between the tests and the number of performed tests. Their "proportion of false positives" (PFP) is calculated in a manner similar to pFDR, where V and R values are estimated separately:

$$PFP = E(V) / E(R)$$

The PFP and FDR are often mistakenly equated, but their difference is actually very important: the PFP controls the proportion of accumulated false positives in the performed measurements, while pFDR controls the expected proportion of false positives for each experiment.

4. Q-value

The q value is very similar to the p value in pFDR and it shows the level of significance in genomic studies. The q value of a variable shows the proportion of false positive measurements in case the variable is accepted as significant. The q values increase parallel to the p values.

The details of the calculation of the q-value, and the estimation of π_0 were described by Storey [10, 11]. Briefly, *t* is a threshold between 0 and 1 below of which we accept all *p* values significant. Denote *m* the number of observed p-values above the t threshold (p_1,p_2,...p_m). π_0 is the estimated rate of the true null hypotheses ($\pi_0=m_0/m$), and S(t) is the number of rejected hypotheses depending on t. Then the FDR is:

$$FDR(t) = (\pi_0 * m * t) / S(t),$$

An equivalent definition of the q-value is the minimum FDR that can be attained when calling that feature significant:

$$q(p_i) = \min FDR(t)$$

One may use the *q* values as an exploratory guide as of which features to investigate further, or one may also take all features with significant *q* values to manage FDR. Most importantly, a systematic use of *q* values in genomic tests could yield a balance between false positives and true positives and give a standard measure of significance that can be universally interpreted.

In summary, the computation of the FDR is advanced over conventional tests (one must, however, define the rejected proportion in advance). The positive FDR (the chance of a false discovery among all discoveries) is an updated version of the FDR (the chance of making a false discovery among all tests). In these, the publication of the q-value (the minimal pFDR threshold above which the results are rejected) adds additional information. When having a pre-compiled set of candidate features and still searching for a liberal criteria compared to the one-step method corrected p-value, the publication of the q-value can save us from drowning in false negative results.

A survival guide of how to compute the FDR and the q value using R is presented on **Figure 3**. A comparison of the results of the various techniques is depicted on **Table 1**. A supporting homepage for the computation of the methods described is available at www.kmplot.com/multipletesting.

SOFTWARE PACKAGES

Many of the available multiple testing correction methods are incorporated in actual statistical packages dealing with microarrays or with general statistical issues (see **Table 2**). Unfortunately, no package is capable to compute all available algorithms at once, thus at the time of the study design one must also pick the appropriate software.

Computation of the basic algorithms can be easily performed in any spreadsheet-software. For those aiming for an in-depth analysis of given algorithm, we can suggest the use of the R statistical environment (http://www.r-project.org). Here, several Bioconductor libraries (http://www.bioconductor.org), like the fdrtool [13], the multtest by Pollard et al. and the twilight [14] can deliver computations of FDR and pFDR. R-based microarray-focused multiple-testing correction libraries are fdrame [15] and OCplus [16].

The q-value can be computed easily with the qvalue package also running in the R statistical environment (http://genomine.org/qvalue). This program takes a list of *p* values and computes their *q* values and estimates π_0.

The popular Significance Analysis of Microarrays implements the computation of the FDR and the q value (http://www-stat.stanford.edu/~tibs/SAM/). However, Tusher et al. implemented a slightly different approach, where an upper and a lower cutoff values are defined (t_1 and t_2 - these are not symmetric) based on the quantile-quantile plot [17]. Then, the FDR is computed for both sets of smaller and larger genes and the total sum of the nonsense genes is divided by the number of significant genes to get the FDR.

LIMITATIONS OF MULTIPLE TESTING

Although the application of a multiple testing correction provides stronger evidence for the parameter under evaluation, we must mention some of its limitations.

A common hindering of these tests is the increasing loss of power parallel to the reduction of significance. In other words the probability of rejecting a real correlation will inevitably increase and important findings might be suddenly insignificant [18]. An extreme example would be the omission of the treatment of a patient suffering acute myocardial infarction, when the creatine-kinase level loses its significance after correction for hundreds of other laboratory parameters tested.

Another principal flaw of the methods is the dependence on the overall number of performed tests. A gene might have an elevated expression in a clinical study - as long as other genes have not been measured, which will shift the level of significance. Taking the most extreme approach, researchers should correct to the absolute number of all tests performed in their career - which, of course, would be a futile endeavor.

While many methods are used after careful consideration, Bonferroni correction is the most common method, not even cited in most cases. However, especially Bonferroni correction pinpoints a fundamental setback of multiple testing [19]. Bonferroni correction assumes a general null-hypothesis, when nothing is significant (all null-hypotheses are true simultaneously). This is, nonetheless an irrelevant assumption in most of the cases, since our goal is to find exactly the differences. When we compute 20 clinical tests, and we find significant difference in one of the tests after Bonferroni correction, then the two groups can be considered as dissimilar. Therefore, we can discard the general null-hypothesis. Thus, the Bonferroni correction gives an exact answer to a question never asked [20].

In summary, the blind compulsory application of a multiple testing correction would instigate a "salami tactics" (researchers would publish only one p-value at a time to retain

significance). Meta-analyses would completely disappear, since re-computation would shrink all remaining significance. Therefore, a clear understanding of the potential and limitations of the these techniques is necessary for actual selection of the best suited correction method.

SUMMARY

As research in medicine is getting truly multidisciplinal involving mathematics, statistics and cutting edge molecular genetics, researchers cannot neglect statistical issues in a study as a result of lack of proficiency. Multiple testing corrections cannot be used to correct technological errors, but will help in the elucidation of the results. Our aim was to give a general reference summary of the most popular multiple testing correction methods which can be used in clinical research.

Multiple testing corrections were developed for repeated measures occurring in industrial production. Using these tests, it is possible to estimate the proportion of malfunctioning products sold or the proportion of withdrawn functional products. Similarly, scientists are acquainted with the fact of rejecting a proportion of true null hypotheses and true alternative hypotheses during their lifetime. The number of these is, in point of fact, completely independent of the number of performed tests. As the shortcomings of multiple hypotheses testing can bring more frustration than advantage, the habitual application of such a test is not recommended in clinical research. In these, the results should be evaluated using conventional tests and the issues of multiple hypothesis testing should be evaluated in the discussion.

In contrast, in true genomic studies dealing with large datasets the use of multiple hypothesis testing cannot be neglected. We recommend to use simple methods as these are already capable to reduce the large number of potential candidates to be able to select the most robust variables qualified for subsequent in-depth analysis.

ACKNOWLEDGEMENTS

The study was supported by the OTKA PD83154 grant and by the Alexander von Humboldt Stiftung.

FIGURES AND TABLES.

Figure 1. A quick guide for selecting the most appropriate multiple testing correction method.

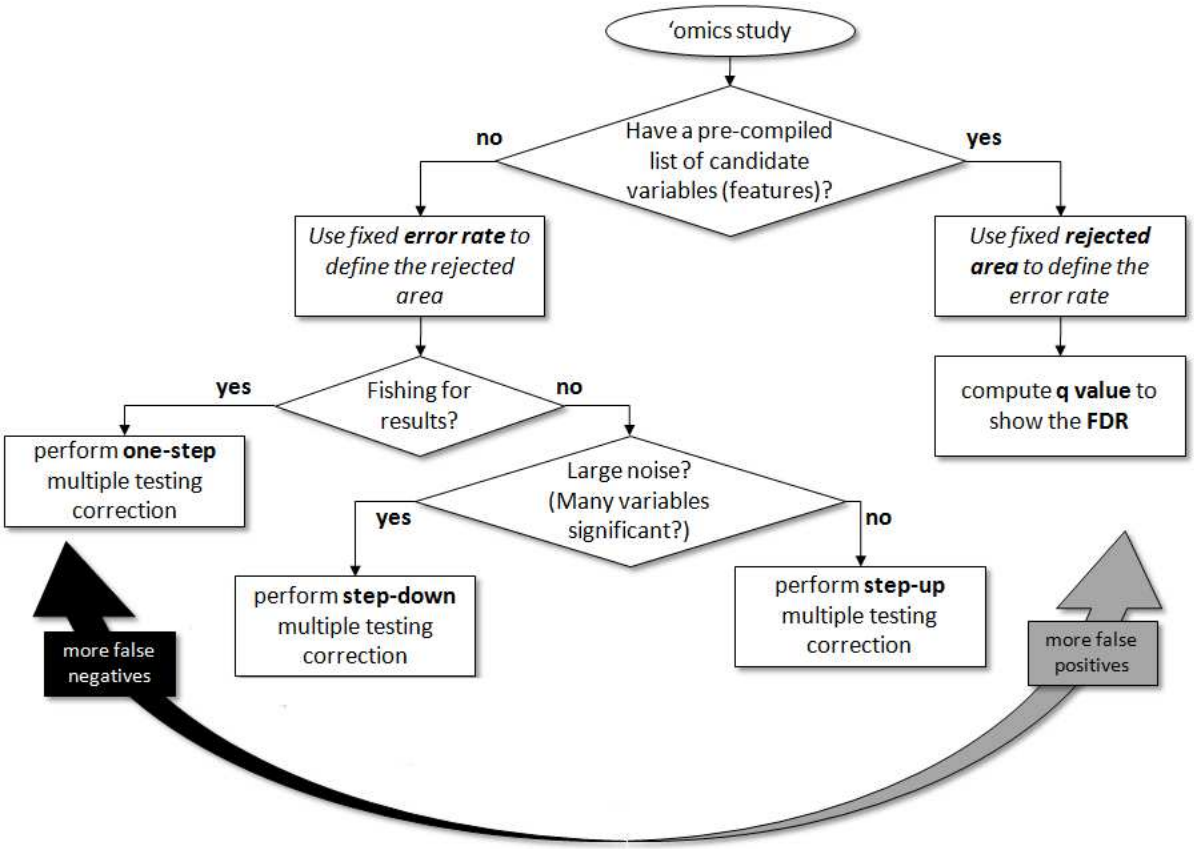

Figure 2. Overview of the three most common multiple hypothesis testing correction methods. Step-up methods deliver the highest number of significant p values, while the Bonferroni-correction delivers the smallest number. Step-down methods can be used as a balance between trimming the p values and retaining as many significant as possible.

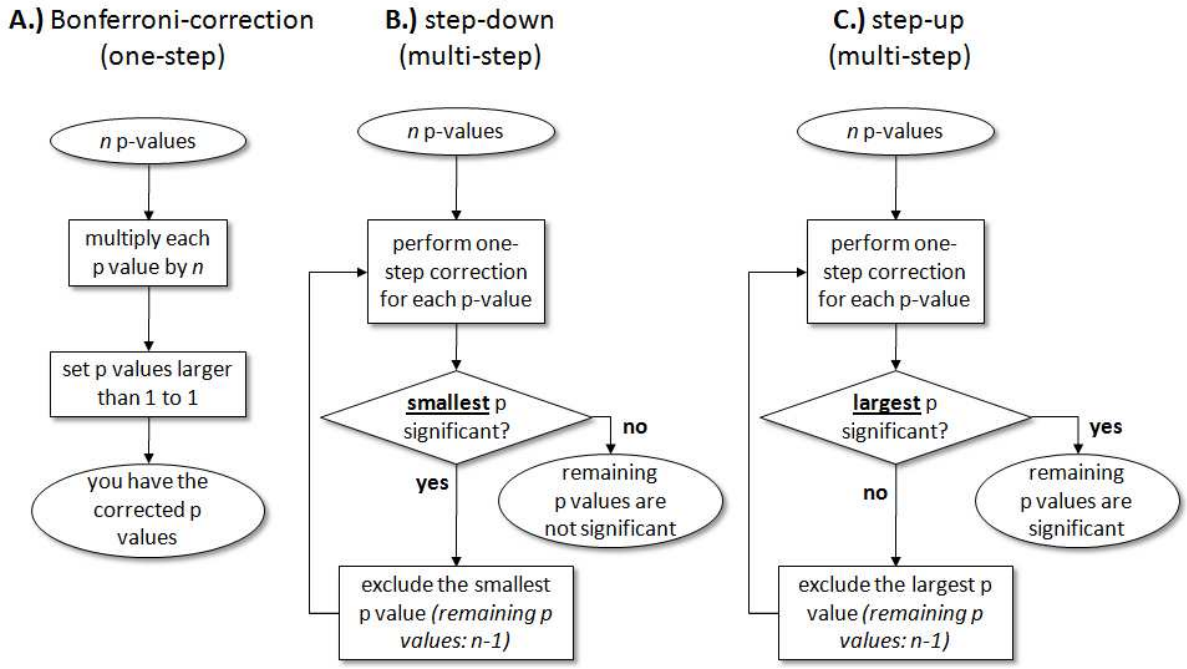

Figure 3. A step-by-step guide on easily computing the false discovery rate (FDR) and the q value in the R statistical environment for a list of p values. A cutoff at 5% FDR means that in 5% of the variables accepted as significant the null hypotheses is actually true. The q value of a variable shows the proportion of false positive measurements in case the variable is accepted as significant. The brainwaver library computes the FDR as described by Benjamini and Yekutieli (2001).

1. **Install** latest version of **R** for Windows

http://cran.r-project.org/bin/windows/base/

start R

2. copy the contents of following box into the **R console:**

FDR	q value

`#installation of necessary package:`
`install.packages("brainwaver")`
`#loading package:`
`library(brainwaver)`
`#combining p values into "pvalues":`
`pvalues<-c(3.92E-05,3.08E-04,4.87E-04,2.59E-03,8.06E-03,1.04E-02,1.26E-02,1.40E-02,1.53E-02,1.62E-02,4.28E-02,6.47E-02,1.62E-01,2.45E-01,2.69E-01,2.93E-01,6.72E-01,8.54E-01,9.30E-01,9.43E-01)`
`#computing FDR at 0.05:`
`FDR<-compute.FDR(pvalues,0.05)`
`#displaying the computed cut-off:`
`FDR`
	`#installation of necessary package:`
	`install.packages("qvalue")`
	`#loading package:`
	`library(qvalue)`
	`#combining p values into "pvalues":`
	`pvalues<-c(3.92E-05,3.08E-04,4.87E-04,2.59E-03,8.06E-03,1.04E-02,1.26E-02,1.40E-02,1.53E-02,1.62E-02,4.28E-02,6.47E-02,1.62E-01,2.45E-01,2.69E-01,2.93E-01,6.72E-01,8.54E-01,9.30E-01,9.43E-01)`
	`#computing q values:`
	`qobj<-qvalue(pvalues)`
	`#displaying q values for each p value:`
	`qobj$qvalues`

For more information see: *http://cran.r-project.org/web/packages/brainwaver/brainwaver.pdf*	***For more information see:*** *http://cran.r-project.org/web/packages/qvalue/qvalue.pdf*

Table 1. A comparison of the results of applying different multiple testing correction methods on 20 p values. Bonferroni, step-down and step-up values were computed in Excel as of Figure 2. False discovery rate (FDR) and the q values were computed in R using the script of Figure 3. The significant values (at 0.05) are marked by grey shading.

list of p values	Bonferroni correction	Step-down	Step-up	FDR below 5% at	q-value
0.00004	**0.001**	**0.001**	**0.00004**	**0.00004**	**0.001**
0.0003	**0.006**	**0.006**	**0.001**	**0.0003**	**0.003**
0.0005	**0.010**	**0.009**	**0.001**	**0.0005**	**0.003**
0.0027	0.052	**0.044**	**0.010**	**0.0027**	**0.012**
0.008	0.161	0.129	**0.040**	0.008	**0.029**
0.010	0.208	0.156	0.062	0.010	**0.029**
0.013	0.252	0.176	0.088	0.013	**0.029**
0.014	0.280	0.182	0.112	0.014	**0.029**
0.015	0.306	0.184	0.138	0.015	**0.029**
0.016	0.324	0.178	0.162	0.016	**0.029**
0.043	0.856	0.428	0.471	0.043	0.070
0.065	1.000	0.582	0.776	0.065	0.097
0.162	1.000	1.000	1.000	0.162	0.224
0.245	1.000	1.000	1.000	0.245	0.315
0.269	1.000	1.000	1.000	0.269	0.322
0.293	1.000	1.000	1.000	0.293	0.329
0.672	1.000	1.000	1.000	0.672	0.711
0.854	1.000	1.000	1.000	0.854	0.848
0.930	1.000	1.000	1.000	0.930	0.848
0.943	1.000	0.943	1.000	0.943	0.848

Table 2. The most widely used multiple-testing correction methods and the implementation of these in statistical analysis software packages.

Method	**Cited**	**References**	Microarray-focused									General statistic					
			Genespring	SAM	BxArray	SNPAnalyzer	Lipsia	Stata	Acuity 4.0	Winsteps	ArrayNorm	R-stats	TM4	GeneMaths XT	Gene Pattern	HBDStat!	Acuity
One-step																	
Bonferroni	910	[2]	X		X			X	X	X	X	X	X	X	X	X	X
Sidak	374	[4]			X			X	X							X	X
Step-down																	
Holm	3295	[5]	X		X			X	X			X	X				X
Westfall-Young	1002	[6]	X				X						X				
Step-up																	
Hochberg	1336	[7]			X				X			X					X
Benjamini-Hochberg	6647	[9]	X			X			X	X		X	X	X	X	X	X
Benjamini-Yekutieli	744, 123	[21, 22]			X							X					
Other																	
Tusher (SAM)	4365, 504	[17, 23]		X									X				
Storey (q-value)	1736, 892	[10, 11]					X										

References

1. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 1928;20 A:175-240.

2. Bland JM, Altman DG. Multiple significance tests: The Bonferroni method. British Medical Journal 1995;310:170.

3. Sidak Z. On probabilities of rectangles in multivariate student distributions: Their dependence on correlations. Ann Math Stat 1971;42:169-75.

4. Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 1967;62:626-33.

5. Holm S. A simple sequential rejective multiple test procedure. Scandinavian Journal of Statistics, 1979;6:65-70.

6. Westfall PH, Young SS. Resampling-Based Multiple Testing 1993.

7. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 1988;75:800-2.

8. Simes JR. An improved bonferroni procedure for multiple test of significance. Biometrika 1986;73:751-4.

9. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc 1995;57:289-300.

10. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society. Series B: Statistical Methodology 2002;64:479-98.

11. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America 2003;100:9440-5.

12. Fernando RL, Nettleton D, Southey BR, Dekkers JCM, Rothschild MF, Soller M. Controlling the Proportion of False Positives in Multiple Dependent Tests. Genetics 2004;166:611-9.

13. Strimmer K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. Bioinformatics 2008;24:1461-2.

14. Scheid S, Spang R. twilight; a Bioconductor package for estimating the local false discovery rate. Bioinformatics 2005;21:2921-2.

15. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003;19:368-75.

16. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics 2005;21:3017-24.

17. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America 2001;98:5116-21.

18. Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: Should we lock the crazy aunt in the attic? British Medical Journal 2001;322:989-91.

19. Perneger TV. What's wrong with Bonferroni adjustments. British Medical Journal 1998;316:1236-8.

20. Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. American Journal of Epidemiology 1995;142:904-8.

21. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 2001;29:1165-88.

22. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference 1999;82:171-96.

23. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association 2001;96:1151-60.