DExplore - Differential Gene Expression Analysis

DExplore- User's guide

1. Introduction

DExplore is an online and easy-to-use tool for detecting differentially expressed (DE) genes using data from the international public repository NCBI GEO. DExplore can also be used for performing functional enrichment analysis using a built-in version of the well-established web tool WebGestalt[1] (www.webgestalt.org).

The user may also upload data that have not been submitted to the GEO yet. Since the user's data are only temporarily saved on the server and are automatically deleted as soon as the user exits the platform or their session expires, there is no danger of the data being copied or used from anyone other than the user.

DExplore is built using R programming language and Bioconductor and can be used by researchers with no specialized programming skills required.

The user interface is quite simple and the application runs exclusively online, so the user does not have to download nor store raw data in their computer.

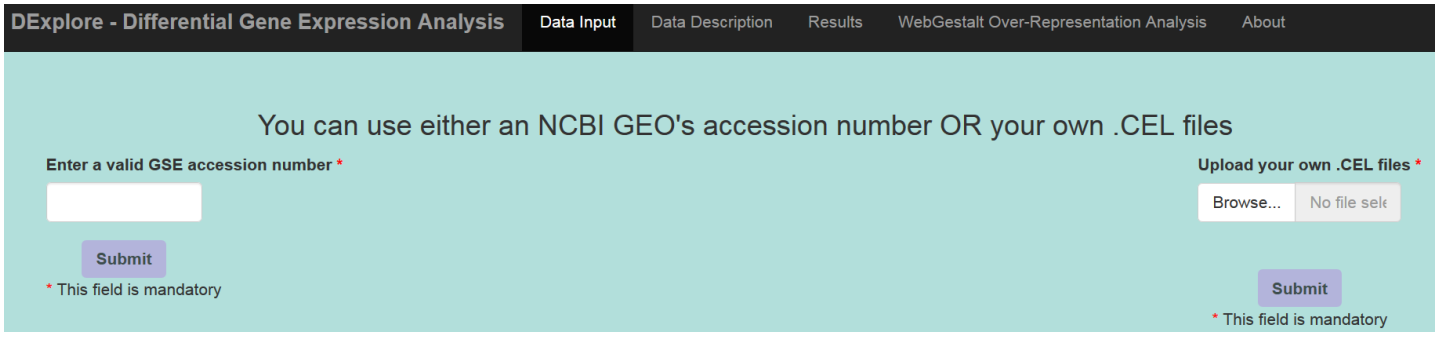
In addition, both the source code and the docker image can be accessed and downloaded using the links on the homepage.

Currently, DExplore can only be used to analyze single-channel mRNA microarray experiments for Affymetrix platforms but will be expanded in the near future to be used for other commercially available platforms such as Illumina and Agilent.

2. How to use

DExplore is comprised of four tab panels; Data Input, Data Description, Results and WebGestalt Over-Representation Analysis.

Do not forget to refresh DExplore between analyses.

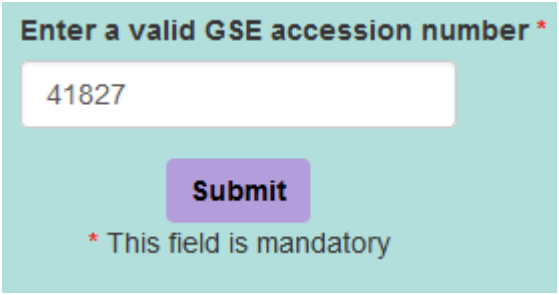


Analyses differ depending on the type of data the user wants to use.

a. Using DExplore with NCBI GEO's microarray data

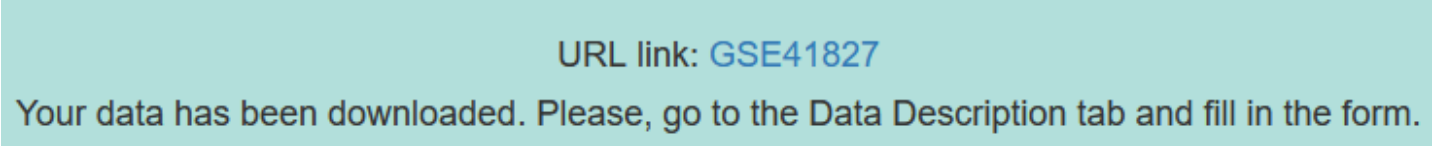
The first thing the user has to do is to enter a valid GSE accession number from the NCBI GEO Database (https://www.ncbi.nlm.nih.gov/geo/) and press the button "Submit" in the "Data Input" tab.

Important notice: enter only the number without writing "GSE" (e.g. for Series "GSE41827", enter "41827").



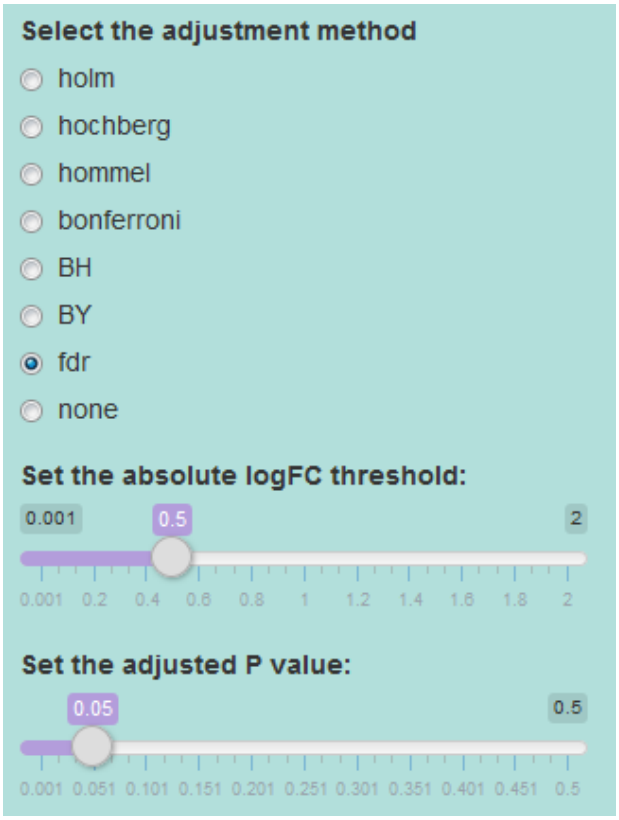
If the user enters an invalid GSE accession number, DExplore shows an error message and the user has to refresh the webpage and enter a valid accession number.

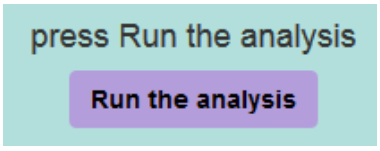
After a few minutes (time depends on the submitted data size), DExplore provides a hyperlink to the GEO's corresponding page.



In the Data Description tab, the user has to choose the parameters for the analysis, i.e. the platform to be used (this is important for experiments that use microarray chips from more than one platforms), the criterion for the comparison that will be carried out, which samples are to be treated as control, to which samples the controls are to be compared. After choosing each of the aforementioned parameters, the user should press the submit button.

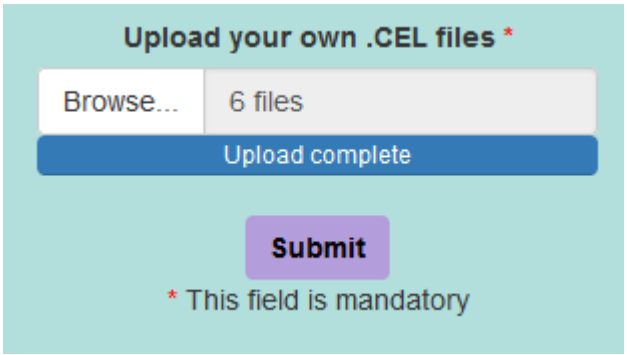
Subsequently, some statistical parameters necessary for the analysis should be selected; i.e. the method to be used for adjusting the p-value for multiple comparisons, the absolute $\log_2$ Fold Change threshold and the adjusted p-value threshold (see Appendix below). The user may either choose to use the default parameters, which are False Discovery Rate – "fdr" for multiple comparisons adjustment, 0.5 as the absolute $\log_2$FC threshold and 0.05 as the adjusted p-value threshold for the analysis, or either change those values as they see fit. Then, the user has to press the "Run the analysis" button and wait a few minutes.



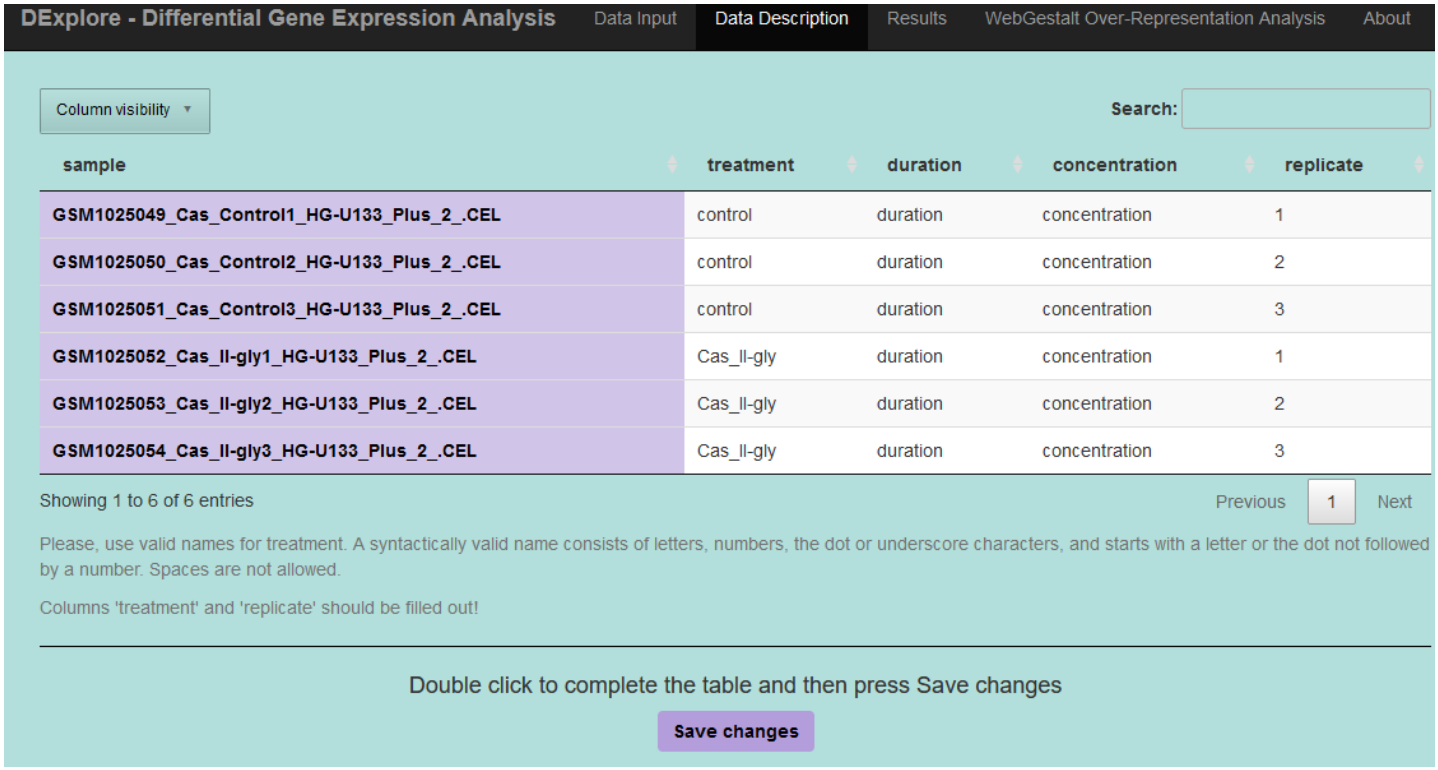


b. Using DExplore with your own .CEL files

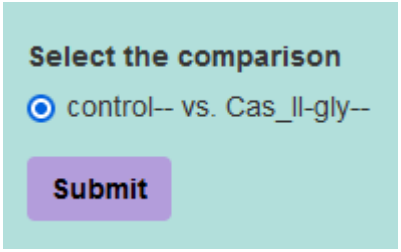
If the user wishes to use data that haven't been submitted to the NCBI GEO, the right side of the tab panel can be used to browse their computer and select the files for the analysis. The user should use raw data stored in .CEL format (one file per experimental condition and replicate). After selecting the files and pressing the button "Submit", DExplore provides a table showing information about the experimental design in the "Data Description" tab.



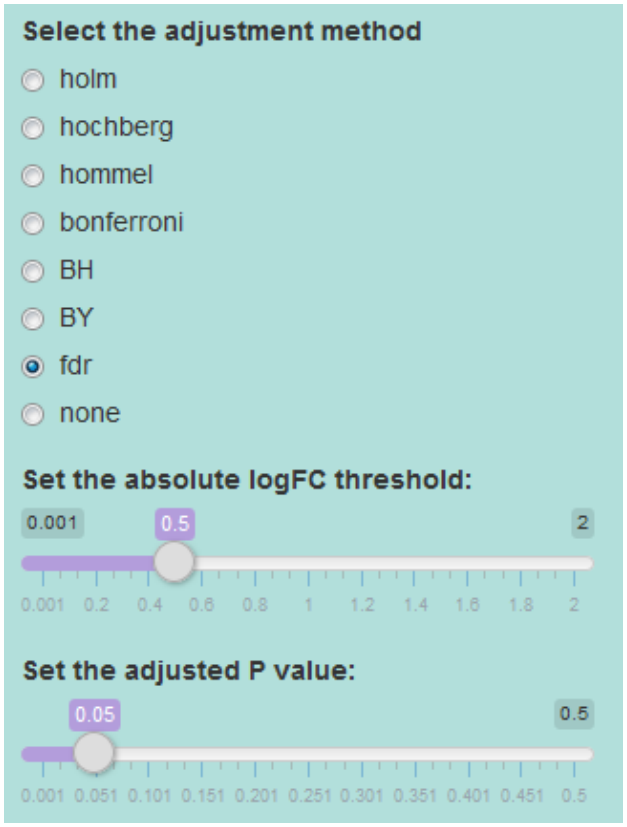
The user has to fill in the table providing information about the experimental design, e.g. which samples are control, which is the treatment for treated samples (type of treatment, duration and concentration for chemical substances or dose for radiation). It is not necessary to fill "duration" and "concentration" columns if they do not affect the experimental design. However, the columns "treatment" and "replicate" must be filled-out. After finishing completing the table, the user should press the "Save changes" button.

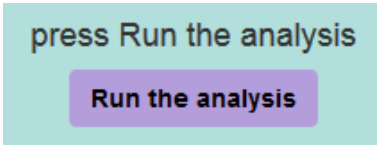


Once the changes have been saved, the user must choose which of the possible comparisons will be carried out, e.g. in cases of two different treatment methods, A and B, and the control samples: (1) compare A to control, (2) compare B to control, (3) compare A to B.



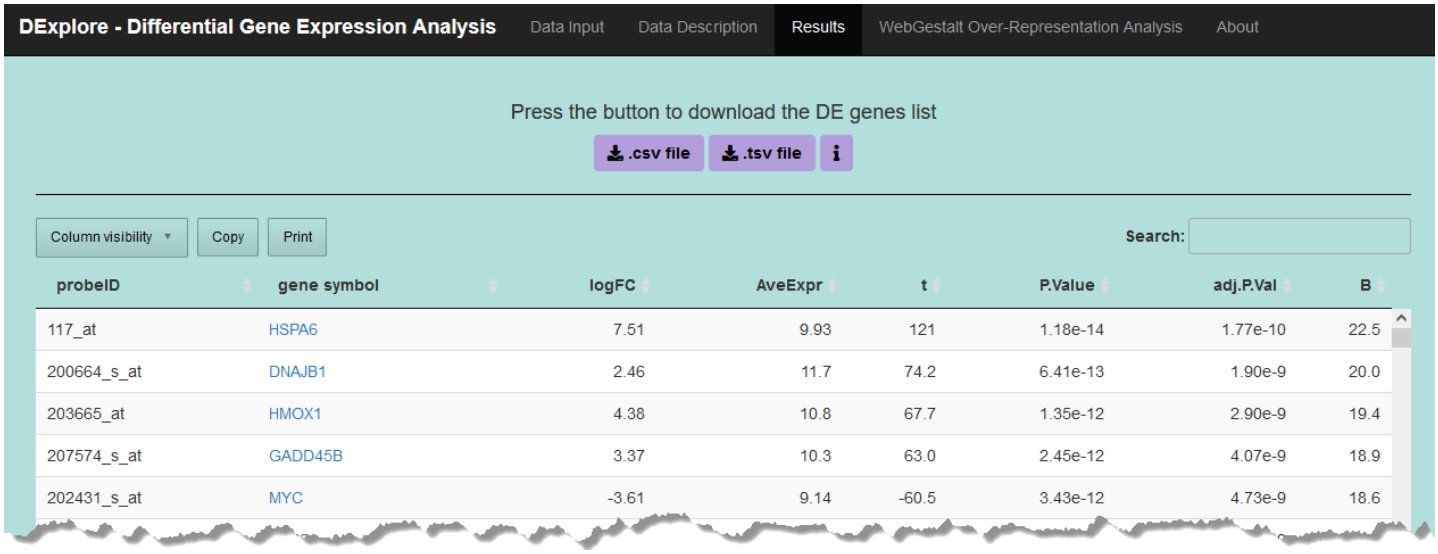
Subsequently, some statistical parameters necessary for the analysis should be selected; i.e. the method to be used for adjusting the p-value for multiple comparisons, the absolute $\log_2$ Fold Change threshold and the adjusted p-value threshold (see Appendix below). The user may either choose to use the default parameters, which are False Discovery Rate – "fdr" for multiple comparisons adjustment, 0.5 as the absolute $\log_2$FC threshold and 0.05 as the adjusted p-value threshold for the analysis, or either change those values as they see fit. Then, the user has to press the "Run the analysis" button and wait a few minutes.





3. Results

When the analysis is done, in the "Results" tab the user can see a list of the differentially expressed genes (both the probe ID used by Affymetrix and the gene symbol are provided) and some statistical values for each of them. In the gene symbol column, there is also a hyperlink to NCBI's Gene database, in case the user wants to explore the genes on the list.

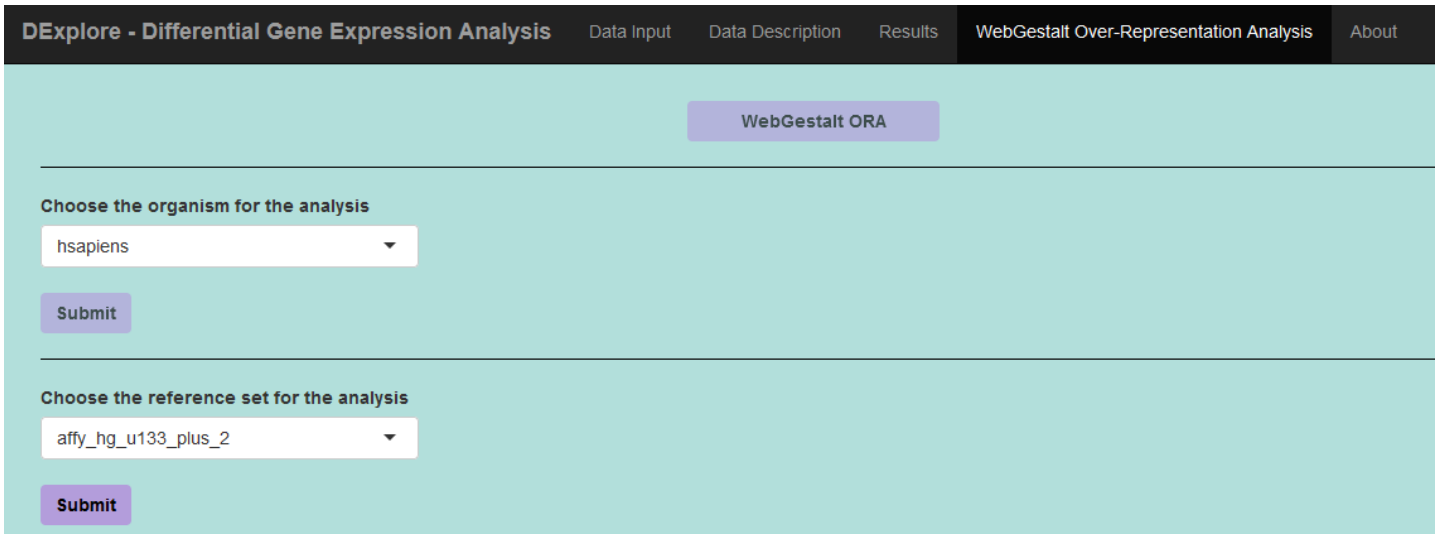


probeID	gene symbol	logFC	AveExpr	t	P.Value	adj.P.Val	B
117_at	HSPA6	7.51	9.93	121	1.18e-14	1.77e-10	22.5
200664_s_at	DNAJB1	2.46	11.7	74.2	6.41e-13	1.90e-9	20.0
203665_at	HMOX1	4.38	10.8	67.7	1.35e-12	2.90e-9	19.4
207574_s_at	GADD45B	3.37	10.3	63.0	2.45e-12	4.07e-9	18.9
202431_s_at	MYC	-3.61	9.14	-60.5	3.43e-12	4.73e-9	18.6

There is also the possibility to print the list (in order to save as a .pdf file), copy it to clipboard and then paste it in another file, or download the list as a .csv or a .tsv file in order to use it for further analyses.

4. WebGestalt Over-Representation Analysis

After having detected the differentially expressed genes under two experimental conditions, it is a common practice to proceed to a functional enrichment analysis. DExplore enables the user to perform functional enrichment analysis using the well-established web tool WebGestalt (www.webgestalt.org). By pressing the WebGestalt ORA button, the user will be prompted to choose one of the organisms supported from WebGestalt platform and one of the reference sets to which the list of differentially expressed genes will be compared.



After a few minutes, DExplore renders the results of Over-Representation Analysis, that can be both explored through a browser and downloaded to the user's computer for future use.

For more information regarding the WebGestalt web tool as well as the methods used for over-representation functional enrichment analysis, please visit www.webgestalt.org.

5. Appendix

p-value and Multiple Comparisons

A p-value provides information about whether a statistical hypothesis test is significant or not and it also provides some indication on "how significant" the result is: the smaller the p-value the stronger the evidence against the null hypothesis. Most importantly, it does this without committing to a particular level of significance as traditional hypothesis tests and confidence intervals do[2].

In statistics, the multiple comparisons, multiplicity or multiple testing problem occurs when one considers a set of statistical inferences simultaneously or infers a subset of parameters selected based on the values observed. The more inferences are made the more likely erroneous inferences are to occur. Several statistical techniques have been developed to prevent this from happening, allowing significance levels for single and multiple comparisons to be directly compared. These techniques generally require a stricter significance threshold for individual comparisons, so as to compensate for the number of inferences being made.

Adjustment for Multiple Comparisons

A typical microarray study generates a gene expression matrix with tens of thousands of rows—probe sets representing genes. Assume we have 10,000 genes and we are performing 10,000 univariate tests. If the significance level $\alpha = 0.05$ is used, then for each of these tests we allow a 5% chance of making a Type I error (=the rejection of a true null hypothesis, aka a "false positive" finding). This means that we expect 5% of the 10,000 genes to be deemed significant (significantly differentially expressed) by chance alone—amounting to 500 false positives. To control the overall probability of a Type

I error, we have to apply a correction for multiple testing. Whether we are repeatedly performing a t test, an ANOVA F test, or any other (univariate or multivariate) test resulting in a p-value, we have to adjust the individual raw p-values for multiplicity in order to control the overall posterior false positive rate.

When performing multiple tests, rather than considering the significance level of individual tests, we should use a procedure that controls one of the Type I error rates defined for testing multiple null hypotheses. Among the commonly used Type I error rates are the family-wise error rate (FWER) and the false discovery rate (FDR).

Family-wise error rate (FWER)

The family-wise error rate is defined as the probability of at least one Type I error (i.e., at least one false positive) over all tests. This probability for a single test is equal to the significance level α of the test. However, if we perform M independent tests, this probability is equal to $1 - (1- \alpha)^M$, which for a high M is close to 1.

False discovery rate (FDR)

The false discovery rate is the expected proportion of false positives among the rejected null hypotheses (i.e., among all genes reported as differentially expressed). When all null hypotheses are true (i.e., none of the tested genes is differentially expressed), FDR is equal to FWER, but otherwise it is smaller.

Generally, procedures controlling the FWER are more conservative than those controlling FDR[3]. The best known among the FWER-controlling procedures are: the classical single-step Bonferroni adjustment, the single-step Sidak procedure, and the step-down Holm procedure. The most popular among the FDR-controlling procedures is the step-up Benjamini and Hochberg procedure. The single-step procedures apply the same multiplicity adjustment to each individual α or raw p-value, whereas adjustments made by the stepwise approaches depend on the rank of the gene among all tested genes and on the outcomes of the tests for other genes.[4]

Adjustment Methods provided by DExplore

DExplore allows you to select an adjustment method for your analysis among the Bonferroni correction ("bonferroni"), the correction introduced by Holm (1979)[5] ("holm"), by Hochberg (1988)[6] ("hochberg"), by Hommel (1988)[7] ("hommel"), by Benjamini & Hochberg (1995)[8] ("BH" or its alias "fdr"), and by Benjamini & Yekutieli (2001)[9] ("BY"). A pass-through option ("none") is also included.

The first four methods are designed to give strong control of the family-wise error rate. There seems to be no reason to use the unmodified Bonferroni correction, since it is overriden by Holm's method, which is also valid under arbitrary assumptions.

Hochberg's and Hommel's methods are valid when the hypothesis tests are independent or when they are non-negatively associated.[10,11] Hommel's method is more powerful than Hochberg's, but the difference is usually small and the Hochberg p-values are faster to compute.

The "BH" (aka "fdr") and "BY" method of Benjamini, Hochberg, and Yekutieli control the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. The false discovery rate is a less stringent condition than the family-wise error rate, and thus these methods are more powerful than the rest.

For a more detailed review of adjustment methods commonly used, see Shaffer, J. P. Multiple Hypothesis Testing. *Annu. Rev. Psychol.* (1995). doi:10.1146/annurev.ps.46.020195.003021.

6. References

1. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019 : gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, 199–205 (2019).
2. Wright, S. P. Adjusted P-Values for Simultaneous Inference. *Biometrics* (1992). doi:10.2307/2532694
3. Dudoit, S. & Laan, M. J. van der. *Multiple Testing Procedures with Applications to Genomics*. *Springer* (2009). doi:10.1007/978-0-387-98135-2
4. Dziuda, D. M. *Data Mining for Genomics and Proteomics*. *Analysis of Gene and Protein Expression Data* (John Wiley & Sons, Inc., 2010). doi:10.1002/9780470593417
5. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Stat.* (1979). doi:10.2307/4615733
6. Hochberg, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* (1988). doi:10.1093/biomet/75.4.800
7. Hommel, G. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika* (1988). doi:10.1093/biomet/75.2.383
8. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B …* (1995). doi:10.2307/2346101
9. Benjamini, Y. & Yekutieli, D. The control of the false dicovery rate in multiple testing under depency. *Ann. Stat.* (2001).
10. Sarkar, S. K. Some probability inequalities for ordered MTP2 random variables: A proof of the Simes conjecture. *Ann. Stat.* (1998). doi:10.1214/aos/1028144846
11. Sarkar, S. K., Chang, C. K. & Chang, C. K. The simes method for multiple hypothesis testing with positively dependent test statistics. *J. Am. Stat. Assoc.* (1997). doi:10.1080/01621459.1997.10473682