False discovery rate estimation on breast cancer diagnosis

Authors: Lúcia Moreira
Nirbhaya Shaji
Rafael Belokurows

Date: 10/12/2019

Summary

This work deals with the estimation of the false discovery rate (FDR) for minimizing the type I error on breast cancer diagnosis using a multiple test procedure. Different algorithms exist for FDR estimation; we have considered 5 different algorithms: one of them estimates that parameter based on the local FDR (*twilight*), 3 of them provide simultaneously an estimate for FDR based on both tail-area and local FDR estimation (*qvalue*, *fdrtool* and *SAGx)* while *multtest* package only provides estimates for the tail-area FDR.

The breast cancer diagnosis dataset considers 30 variables based on breast tissue images relevant for classifying as malignant or benign tumors from those tissue samples. Our exploratory data analysis started by studying the distribution of the variables in our dataset. Histogram based analysis showed that most of the variables are not normally distributed although symmetry was considered in the difference of the medians of the two classes. According to this, we have considered a non-parametric test for extracting the p-values to be used as inputs in all the algorithms for FDR estimation herein considered. Despite of that, in some FDR algorithms we make the comparison between the parametric testing and non-parametric testing taking also into account the pedagogical perspective present on this work.

Our results showed that a multiple testing analysis maybe be not very relevant for this dataset because of the low number of available variables and because most the variables show extremely small p-values, as obtained from the 30 single tests, but most importantly also because the non-relevant variables have quite high p-values. So, a quite clear relevant /non-relevant variable threshold separation is obtained.

Overall, all the algorithms seem to agree with each other on the choice of the non-relevant variables in this dataset aiming a breast cancer diagnosis, but more detailed conclusions and comparisons can be found in the relevant sections of our report. This work was inspired by the paper: BMC Bioinformatics 2008, 9:303 [1].

Introduction

Multiple testing refers to any instance that involves the simultaneous testing of more than one hypothesis[2]. If decisions about the individual hypotheses are based on the unadjusted marginal p-values, then there is typically a large probability that some of the true null hypotheses will be rejected.

Take the case of *m = 100* hypotheses being tested at the same time, all of them being true, with a significance level α, i.e P(making an error) = *0.05*, one expects five true hypotheses to be rejected.

In general if we perform *m* hypothesis tests, then

$$P(\text{Not making an error}) = 1 - \alpha,$$

$$P(\text{Not making an error in m tests}) = (1 - \alpha)^{m}$$

$$P(\text{Making at least one error in m tests}) = 1 - (1 - \alpha)^{m}$$

So, if all tests are mutually independent, then the probability that at least one true null hypothesis will be rejected rises to 1− (1 - 0.05)^100= 0.994.

Fig. 1 shows, for $\alpha = 0.05$, how the probability for at least 1 false positive grows with m, the number of hypotheses.

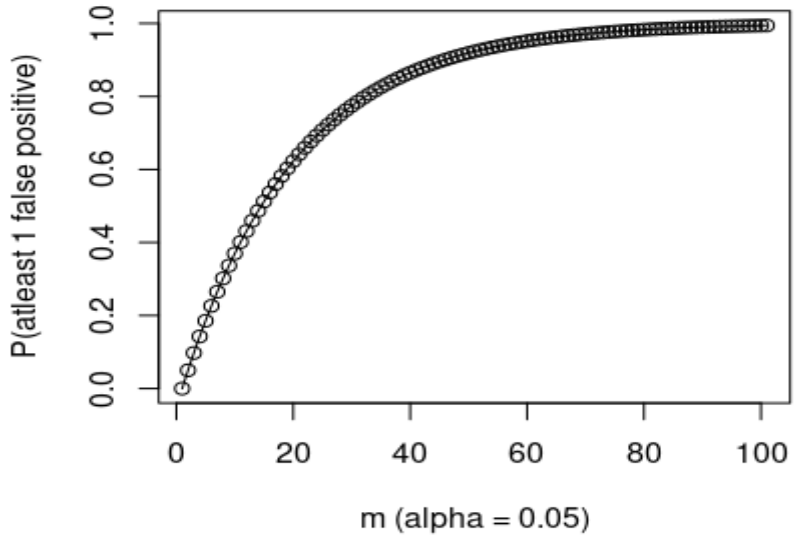

Fig 1.- Probability of at least one false positive

For the problem of simultaneously testing a null hypothesis H_s against the alternative hypothesis H'_s, for $s = 1,...,m$, a **multiple testing procedure (MTP)** is a rule which makes

some decision about each H_s. The term **false discovery** refers to the rejection of a true null hypothesis based on a MTP.

Accounting for the multiplicity (the potential inflation of the type I error rate as a result of multiple testing) of individual tests can be achieved by controlling an appropriate error rate. **Familywise Error Rate (FWE)** is the traditional or classical expression for the probability of one or more false discoveries:

$$FWE = P\{\text{reject at least one hypothesis } H_s : s \in I(P)\},$$

Where I(P) denote the set of true null hypotheses, that is, $s \in I(P)$ if and only if H_s is true.

Control of the FWE means that, for a given significance level α, we need:

$$FWE \leq \alpha$$

for any probability distribution.

Control of the FWE allows one to be $1-\alpha$ confident that there are no false discoveries among the rejected hypotheses. The strictness of FWE does not allow even a single true hypothesis to be rejected. However, in the case where m is very large, the corresponding MTP might result in lower ability to reject false null hypotheses. Which leads to the need of less strict versions of FWE.

Let *F denote the number of false rejections* and let *R denote the total number of rejections*. The **false discovery proportion (FDP)** is defined as:

$$FDP = (F/R) \mid \{R > 0\}$$

Instead of FWE, we may consider the probability of the FDP exceeding a small, pre-specified proportion:

$$P\{FDP > \gamma\}, \text{ for some } \gamma \in [0,1)$$

The choice of $\gamma = 0$ simplifies to the traditional FWE. Another alternative to the FWE is the **false discovery rate (FDR)**, defined to be the expected value of the FDP:

$$FDRP = EP[FDP]$$

This project looks into different False discovery rate (FDR) control algorithms while working with the Breast Cancer Wisconsin (Diagnostic) Data Set [8] taking the null hypothesis as *there is no difference in the medians/means between the groups identified as benign and malignant* and the alternative is that there is a difference.

Maintaining control over false positives is a significant challenge in analyzing data which are large in nature. FWE is appropriate when you want to guard against any false positives.

However, in many cases we can live with a certain number of false positives e.g: genomics. In these cases, the more relevant quantity to control is the *proportion of false positives* among the set of rejected hypotheses [5].

FDR is the expected proportion of Type 1 errors. FDR adjusts the p values in a way that limits the number of false positives that are reported as “significant”. “Adjusting p-values for the number of hypothesis tests performed” means to control the Type I error rate.

The false discovery rate (FDR) as introduced by Benjamini and Hochberg [3] is defined as the expected proportion of variables falsely called relevant for diagnosis among all variables indeed relevant (i.e. the expected value of the proportion of false positives among the rejected hypotheses). The shortcoming of this classical FDR is that it does not refer to a single variable but a collection or list of variables [6,7].

Efron et al. [4] introduced the local FDR, an analogous measure of uncertainty referring to single data. It is defined as the probability that a variable is truly not relevant for diagnosis given an observed p-value.

Error rates based on the proportion of false positives (e.g., FDR) are especially more appealing for large-scale testing problems because those error rates do not increase exponentially with the number of tested hypotheses. Such large-scale testing problems can be encountered in genomics. On the other hand, error rates based on the number of false positives (e.g., gFWER) could be wiser for smaller-scale testing problems [5].

When relevant, this work uses the same naming conventions established on Strimmer’s paper [1], which are: “fdr” denotes the ***local*** *false discovery rate*, “Fdr” denotes the ***tail area****-based false discovery rate*, and “FDR” is a generic term encompassing both variants. “FNDR”, on the other hand, is the abbreviation for the false non-discovery rate (1- FDR).

Dataset

Data Set Information

Dataset comprised of 569 observations of attributes of digitized images of tumor masses that were found in real biopsies [8]. The study and data collection were made by the University of Wisconsin with a procedure called “Fine Needle Aspiration”, which provides a minimally invasive way to examine a small amount of tissue from the tumor. First, the tissue is extracted, then the physician draws an outline of the tumor and lastly a Machine Learning algorithm adjusts the outline to the exact shape of the nuclei. This allows for precise and automated analysis of nuclear size, shape and texture [9].

Data Attributes

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32) There are 10 attributes and three detailed pieces of information about each one: The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Other important characteristics of the dataset:

- All feature values are recorded with four significant digits.
- There are no missing attribute values (NAs).

Exploratory Analysis

The first step before analyzing and performing tests with the chosen dataset was to perform a quick exploratory analysis to get a better sense of the way the data was collected and organized, and ultimately to determine whether the data were normally distributed, which is one of the main factors to be taken into account when deciding the type of test(s) to be run.

The software used was R, a programming language and free software environment for statistical computing and graphics. Its basic package contains several of the functions needed for statistical analysis and there are several more packages which contains specific functions, including some of those needed for the activities developed in this paper.

The target variable of the study is a categorical variable, with possible values: “B” and “M”, indicating whether the breast tumor was diagnosed as Benign or Malignant, respectively. As shown in Figure 2, there was an uneven distribution in those values, with 357 tumors labeled as Benign and 212 as Malignant.

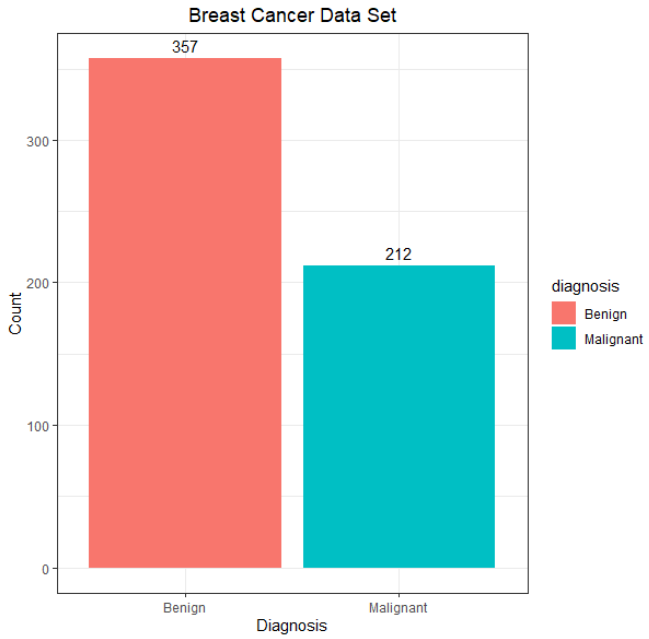

Fig. 2 Number of cases of Benign and Malignant breast tumors as found in the dataset used in this study

Since the dataset is constituted almost solely of numerical variables, a second analysis that was made was a plot of the density function of each variable, which as shown in Figure 3, indicates that none of the variables appear to be normally distributed or even symmetrical. All of the variables appear to have its values concentrated on the left side of the plot, i.e. they present positive skewness.

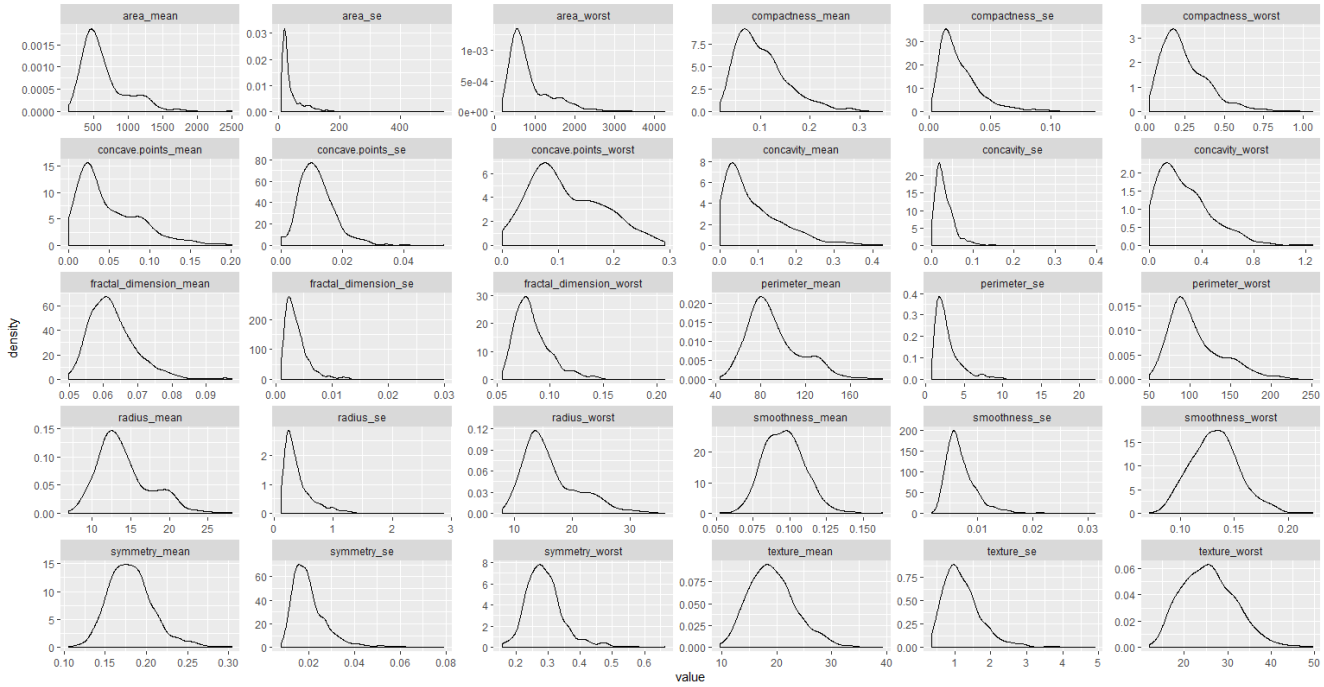

Fig. 3 Density of all variables. None of them show signs of normality neither of symmetry.

Similar conclusions can be reached by looking at the following grid with the boxplot of all variables. Figure 4 shows that none of the variables show clear signs of symmetry, with a few displaying some definite outliers as well. However, we will assume that the differences in the medians are symmetrical.

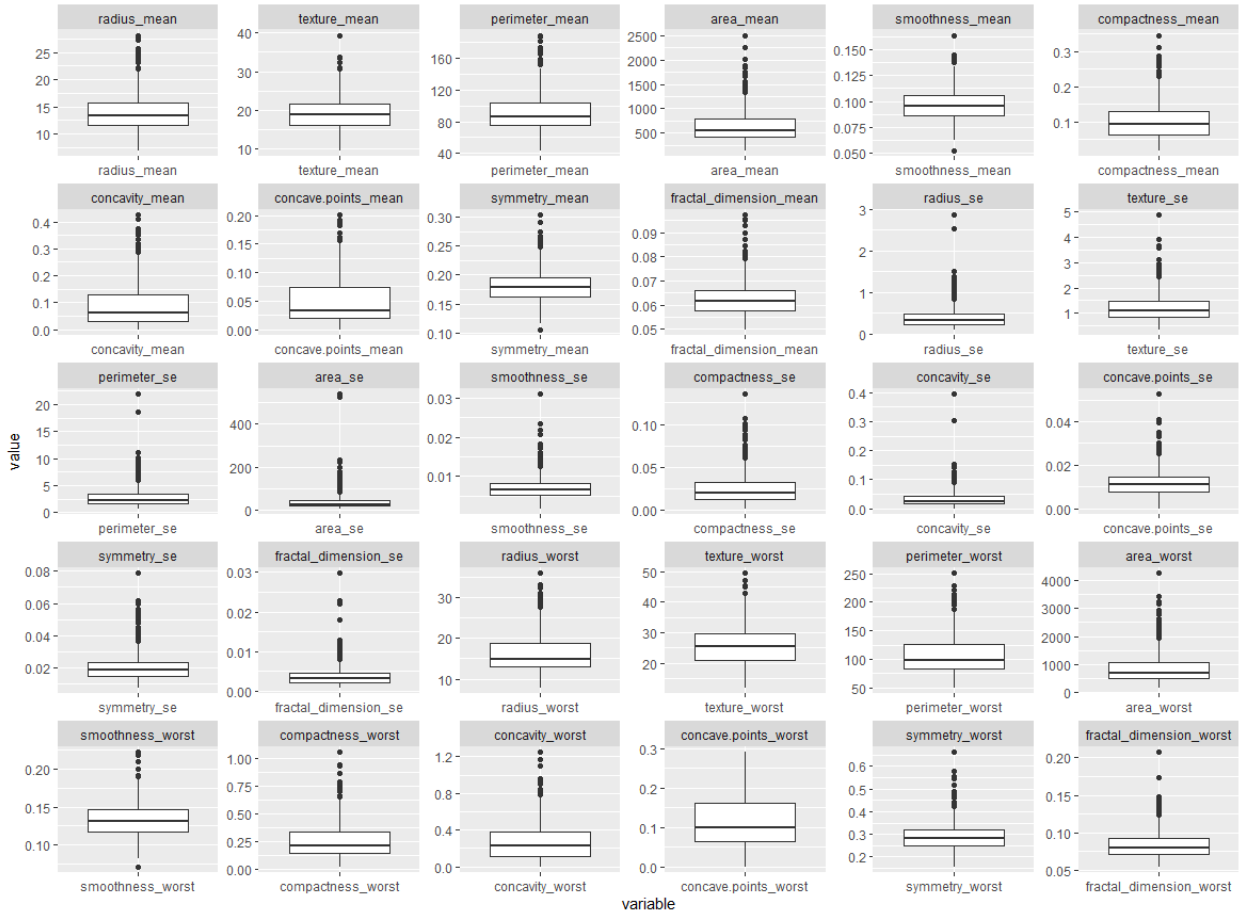

Fig. 4 - Boxplot chart of all variables showing little to no signs of symmetry within each variable.

Comparison of Algorithms to control FDR

- Multitest package - Tail area-based false discovery rate
- Twilight package - Local false discovery rate
- fdrtool package - Simultaneous local and tail area-based FDR estimation
- qvalue - Tail-based FDR
- SAGx - Local and tail-based

Multtest package - Tail area-based false discovery rate

Multtest package uses p-values as inputs in the Benjamini-Hochberg (BH) algorithm for estimating the "classic" tail area-based Fdr [5]. The BH rule used was the most popular and

simpler conservative estimator of Fdr. Several methods for controlling the chosen Type I error rate are available in multtest [5]. Procedures are provided to control Type I error rates defined as tail probabilities for arbitrary functions of the numbers of false positives and rejected hypotheses. We have used the FDR-controlling augmentation procedure and compared it with the non-augmented Benjamini-Hochberg (BH) algorithm and the Benjamini & Yekutieli. The augmentation procedure uses a FWER-controlling MTP that is augmented to control the Type I error rate, such as the gFWER and TPPFP. Following, two FDR controlling procedures can then be derived from the TPPFP controlling augmentation multiple testing procedure, one conservative and other restrictive. Using the function fwer2fdr(), which take FWER adjusted p-values as input, it returns augmentation adjusted p-values for control of the FDR.

Accordingly, given a multivariate dataset and user-supplied choices for the test statistics and the Type I error rate and its target level, the main user-level function of the package returns adjusted p-values directly via the type one argument of the main function MTP. Following, the fwer2fdr() function uses these adjusted p-values as inputs that are then augmented adjusted p-values. The non-parametric robust test was chosen assuming equal variance between each of the two samples so a Wilcoxon rank sum or Mann-Whitney test was performed for non-paired samples. The type I error rate was controlled by the step-down common-quantile (minP) procedure [5]. One hundred bootstrap samples were also considered. The target nominal type I error rate chosen is 0.05.

Figure 4 shows that the augmented conservative and restrictive methods as well as the BY algorithm all reject the null hypothesis in 27 variables while the BY method rejects 26 variables. The Benjamini-Hochberg (BH) algorithm estimates the "classic" tail area-based Fdr. Multiple testing in our dataset shows that the median difference between malignant and benign sample breast tissue tests are significantly different for a 0.05 level of confidence in 26 (for BH) or 27 of the 30 tests (variables), indicating that only 4 (or 3) tests are not relevant for breast cancer diagnosis.

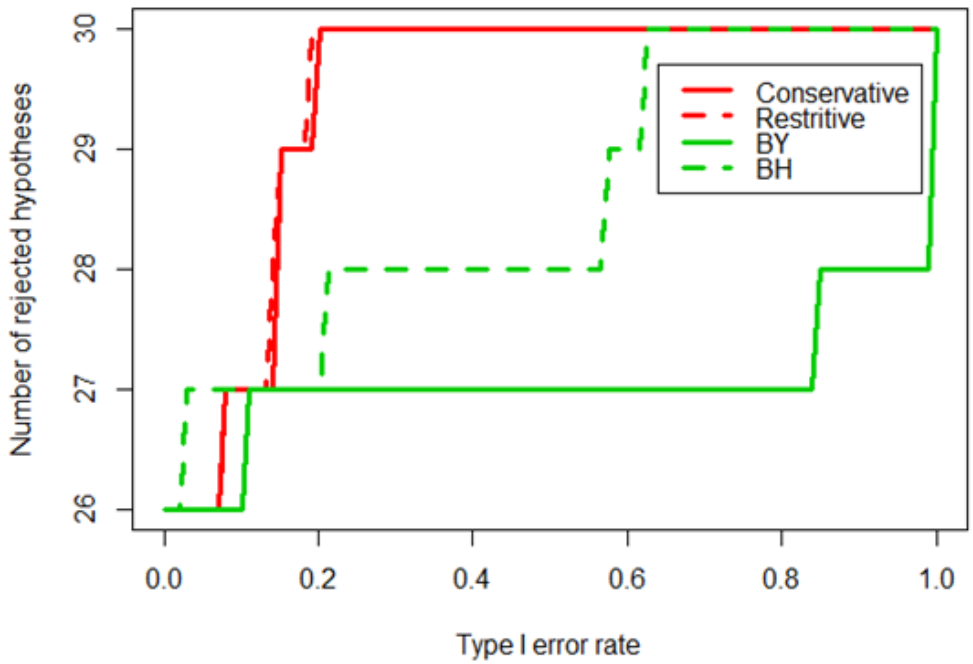

Fig. 4 - Number of rejected hypotheses as a function of the Type I error rate for the augmented both restrictive and conservative multtest based algorithms and comparison with BH and BY algorithms.

Looking into more detail on the corrected p-values for each method and comparing them with the raw p-values, we can see that the raw p-values reject 27 null hypotheses out of the 30 simple tests and accept the alternative hypothesis in the 3 variables indexes numbers 10, 12 and 15 (Fig. 5, left plot, grey points) for a 0.05 confidence. But in this case and for a 95 % confidence, the probability of having a type I error rate would be 1-0.95^30=0.786. So performing, multiple testing allows maintain the type I error at the 5 % level.

Fig, 5 (middle) shows that both the augmented restrictive and conservative multiple testing methods agree with each other on the variable indexes to be not rejected (index numbers 10, 12,15,19) (Fig. 5, middle plot). In its turn, the BH method (right plot, blue points) rejects 27 variables, while the BY method rejects 26 variables by not rejecting also the indexes 10,12,15,19. Those indexes of the non-relevant variables for cancer diagnosis are: index 10 - mean of the *fractal* dimension, index 12 - standard deviation of texture, index: 15 - standard deviation of smoothness and index 19 - standard deviation of the symmetry.

So, all these 4 methods mostly agree on the choice of the more relevant variables for breast cancer diagnosis, despite the simple BH method includes one extra variable considered for rejection once is lower than the confidence level. Also it should be pointed out that the raw p-values are in agreement with the tail based FDR methods and gives exactly the same result as the BH method (right plot, blue points). So, when the number of variables/tests are

very small sized and the raw p-values are infinitely small maybe a multiple testing may not be that such a relevant procedure.

Below follows the multiple testing considering the local FDR controlling type I error.

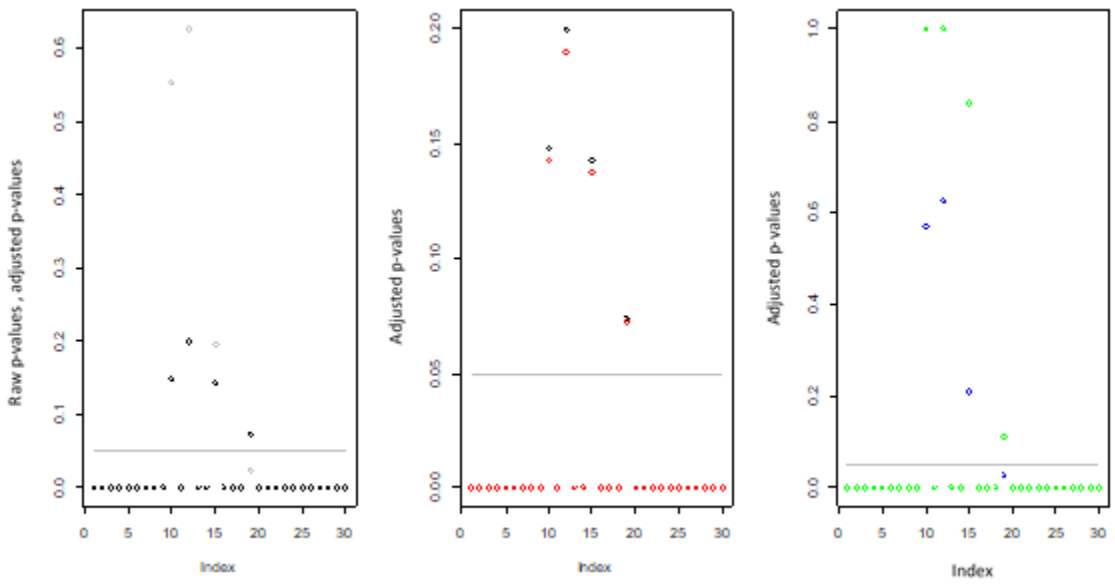

Fig. 5- Raw p-values (non-parametric, grey), conservative (black) and restrictive (red) augmented based adjusted values and BY (green) and BH (blue) adjusted p-values for each of the variable indexes. Flat line is the 0.05 threshold.

Twilight - Local false discovery rate

Twilight algorithm is a penalized stochastic/heuristic downhill search algorithm to estimate the local FDR that performs a successive exclusion procedure based on the Kolmogorov-Smirnov goodness-of-fit as truncation point [6,7]. In a nutshell, the algorithm works as follows: starting with a set of observed p-values, it successively removes p-values until the set of remaining p-values follows a uniform distribution. This set represents the variables/tests that are not significant for diagnosis in our dataset. Given its uniform, the percentage of p-values in the uniform part and the observed overall p-values density are used to estimate the local FDR. The procedure, however, relies on the assumptions that variables/tests are independent from each other and p-values follow a uniform distribution for the non-relevant variables [6,7]. For estimating the local FDR, the main function twilight used a set of p-values as input, computed from a two-sample t test (parametric, non-paired, a non-parametric test was also considered but the twilight function returned an error). The local FDR estimator's mean is assessed on the average of 3 bootstrap samples of the input p-values for a 95 % confidence interval.

Fig. 6 shows that 5 variables are not significant (indexes nr. 10, 12, 15,19, and 20) when using this method and the null hypothesis is rejected for the other 25 variables. Comparing with the tail based methods one can see that local FDR methods includes more variables as not significant for breast cancer diagnosis, however most of these same variables are selected in the local based FDR and tail area based FDR.

Differences between the tail area based algorithms may also be explained in terms of the different estimators for the FDR by itself. Moreover, in the tail area based method a non-parametric approach was considered. Moreover, because only very few variables are not relevant in this dataset and the number of tests/variables is quite small, this algorithm has shown difficulties to achieve a uniform distribution from the p-values of only 5 variables/tests (see Fig. 6). So this algorithm may present a bigger error in estimating the local FDR, because only 3 bootstrap samples could be used. Moreover, using a parametric two sample t test for raw p-values determination may introduce larger errors once the variables do not follow normal distributions.

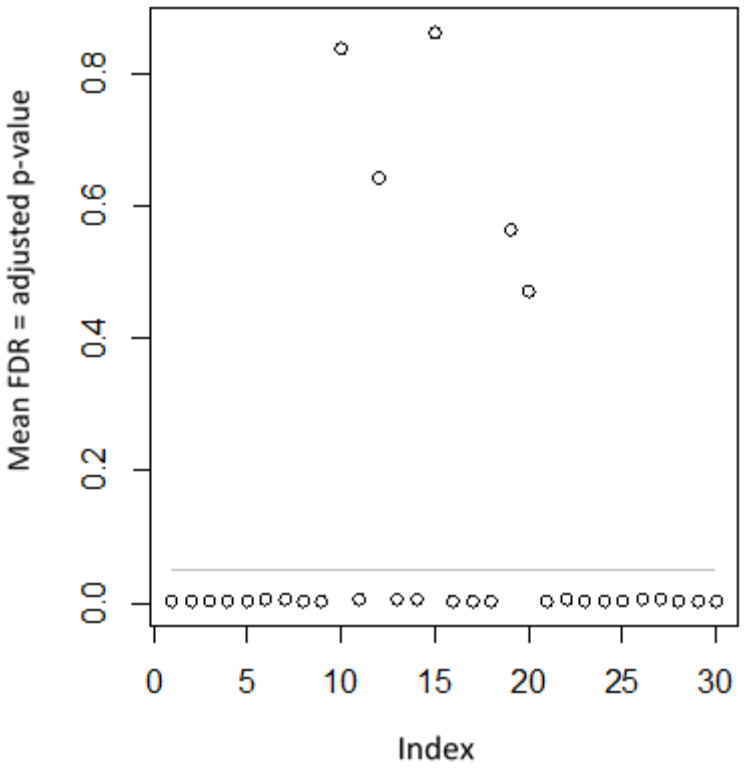

Fig. 6- Mean local FDR from 3 bootstrap samples for each of the variable indexes from the *twilight* analysis (parametric based single testing). Flat line is the 0.05 threshold.

fdrtool package - Simultaneous local and tail area-based FDR estimation

This package is a unifying algorithm that estimates simultaneously both the tail area-based Fdr as well as the density-based fdr (=q-value resp. local false discovery rate) [1]. The approach is semi-parametric and is based on a modified Grenander density estimator, empirical null model, with selection of the truncation point by FNDR. Moreover, it remains (largely) compatible with the well-established "locfdr" and "qvalue" algorithms [1].

As stated by Strimmer [1], the use of the Grenander density estimator provides two main benefits: it explicitly incorporates monotonicity constraints (to preserve ordering of original test statistics) and provides simultaneous estimates of both the Probability Density Function and Cumulative Density Function (necessary to allow computation of both tail-area and local-based FDR).

Figures 7 and 8 show the non-parametric obtained p-values and parametric t-test based p-values output from the fdrtool function. For the non-parametric two sample test (Fig. 7), both tail-based FDR (left) and local FDR (right) give similar results on the variables that are not rejected (indexes nr. 10, 12,15 and 19). So both methods reject 26 variables and overall is majorly in accordance with the other tail based FDR and local FDR algorithms.

Regarding the parametric t-test results, Fig. 8 shows that the tail based area FDR fails to reject the null hypothesis for variable indexes 10, 12, 15, 19 and 20, i.e. rejects 26 variables, while the local FDR rejects 24 variables and fails to reject the null hypothesis for indexes (10, 12, 15, 17, 19 and 20). This is the method that has rejected the higher number of variables under the null hypothesis.

However, the confidence level to compare the results is not clear because looking at the output from this function, the tail-area based FDR (q-values) were converted to very small values so the 0.05 significance level may not be straightforwardly applied here. So one maybe should look instead for the overall q-values and visually make a decision. This is what was performed in this work: visually inspect the most different q-values from the others. Also, looking at the local FDR estimation (Fig. 8, right) and applying for e.g. the commonly used 0.05 confidence level this would imply all variables to be rejected!. So once again the threshold was decided based on visual inspection. These topics should be further discussed with colleagues during the project presentation.

Non-parametric single testing p-values as input

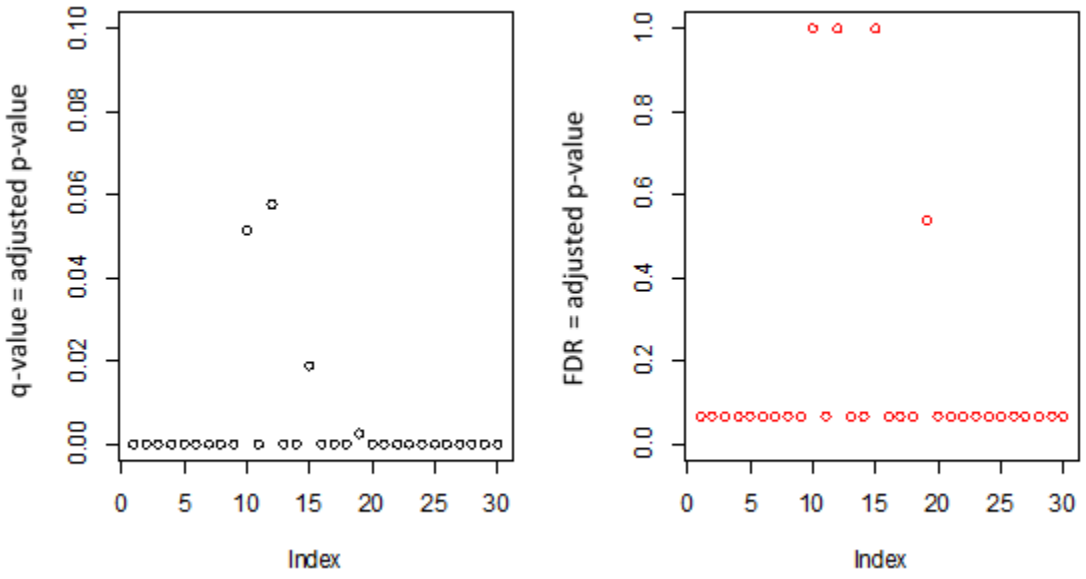

Fig. 7- Tail area-based FDR (left) and local based FDR (right) for each of the variable indexes from the *fdrtool* analysis (non-parametric based single testing).

Parametric single testing p-values as input (t-test)

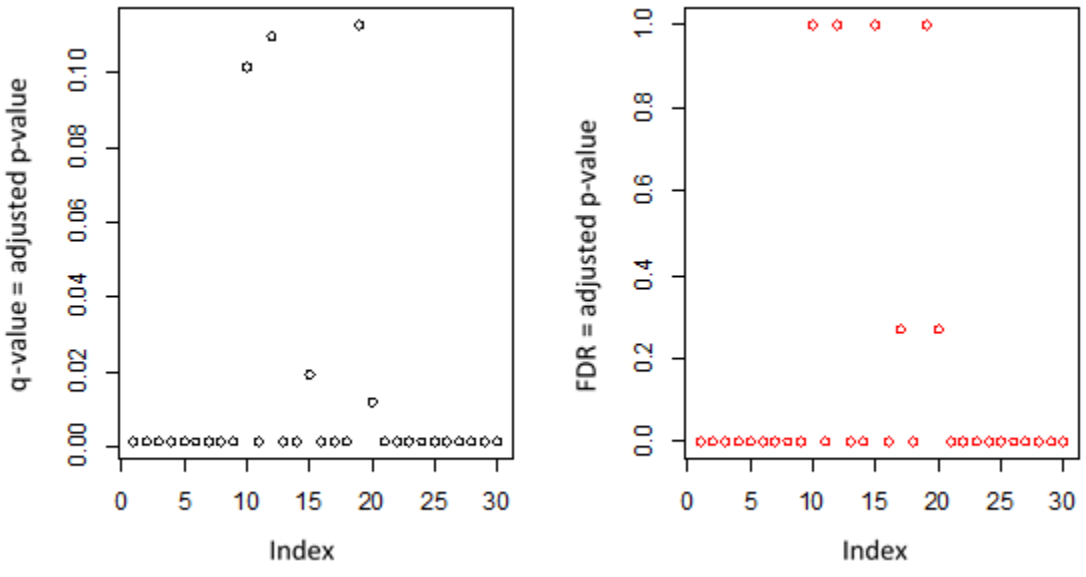

Fig. 8- Tail area -based FDR (left) and local based FDR (right) for each of the variable indexes from the *fdrtool* analysis (parametric based single testing).

Qvalue

The qvalue package performs false discovery rate (FDR) estimation from a collection of p-values or from a collection of test-statistics with corresponding empirical null statistics. This package produces estimates of three key quantities: q-values, the proportion of true null hypotheses (denoted by π_0), and local false discovery rates.

The Q-value, as denoted by Storey [10] is the FDR based measure of significance that can be calculated simultaneously for multiple hypothesis tests. Since the FDR does not necessarily increases with an increasing significance threshold, the q-value is determined as the minimum FDR at which the test is called significant.

Based on multiple tests performed on the 30 numerical predictor variables of our dataset, the p-values for significance of those tests were extracted and then their respective q-values were calculated.

One important caveat is that the p-values obtained from the multiple tests had to be rounded to 4 digits because the package apparently doesn't handle well really small p-values as were obtained on our tests.

Table 1- Comparison of the no. of variables significant under a given significance level.

statistic/ significance level	**<0.001**	**<0.01**	**<0.05**	**<0.10**	**<1**
p-value	25	25	**26**	26	30
q-value	25	25	**26**	26	30
local FDR	25	25	25	25	25

As it can be seen on Table 1 and Figure 9, using the q-value, just as with the p-value, right above the 0.1 level, 4 variables are no longer considered significant, that is, they have a significance value over 0.1.

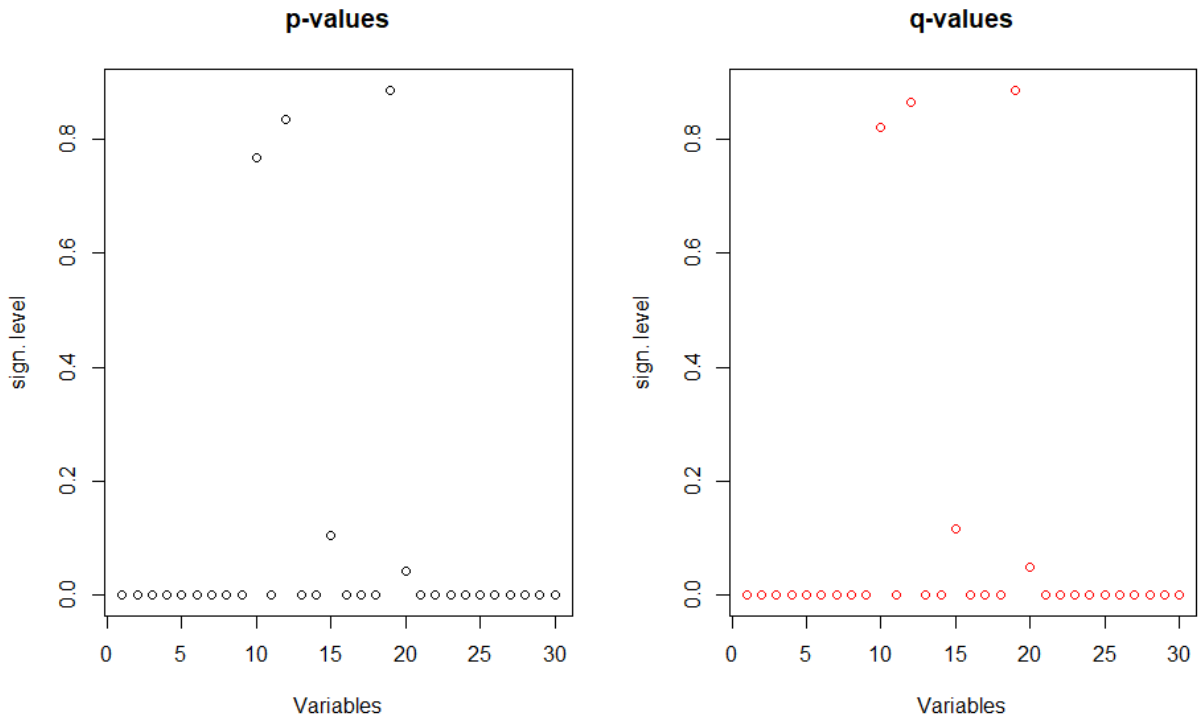

Fig. 9 - Comparison of the p-values obtained from the multiple tests and the calculated q-values. The significance of four of the variables is above the 0.05 level for both p and q.

This package also allows to estimate the lambda (pi0) parameter, which is an estimate of the overall proportion of true null hypotheses, but with the dataset object of this study it wasn't possible to obtain this estimate, due to the inability of the function to deal with p-values so small.

Moreover, the local FDR was computed, and as is represented in Table 2, we can see that 5 out of the 30 variables are determined as not significant under any significance level. Those variables were: mean of the fractal dimension, standard deviation of texture, standard deviation of smoothness, standard deviation of the symmetry and standard deviation of the fractal dimension.

The q-value calculated didn't differ greatly from the p-value obtained on the actual tests, probably because most of the p-values were very low, below 0.0001 (they were rounded to 4 digits to allow the use of the qvalue package), but the variable fractal_dimension_mean, which had a p-value and q-value very close to the significance level of 0.05, was deemed insignificant with the FDR measure. So, the exclusion of the 4 variables in Table 2 (bold) are pretty in agreement with all the above methods herein presented (which can be observed by number of the index/variables in the plots herein presented.

Table 2. P-values obtained from the tests; Q-values and FDR calculated with package *qvalue*.

variable	p	q	fdr
radius_mean	0.0001	0.0001	0.0001
texture_mean	0.0001	0.0001	0.0001
perimeter_mean	0.0001	0.0001	0.0001
area_mean	0.0001	0.0001	0.0001
smoothness_mean	0.0001	0.0001	0.0001
compactness_mean	0.0001	0.0001	0.0001
concavity_mean	0.0001	0.0001	0.0001
concave.points_mean	0.0001	0.0001	0.0001
symmetry_mean	0.0001	0.0001	0.0001
radius_se	0.0001	0.0001	0.0001
perimeter_se	0.0001	0.0001	0.0001
area_se	0.0001	0.0001	0.0001
compactness_se	0.0001	0.0001	0.0001
concavity_se	0.0001	0.0001	0.0001
concave.points_se	0.0001	0.0001	0.0001
radius_worst	0.0001	0.0001	0.0001
texture_worst	0.0001	0.0001	0.0001
perimeter_worst	0.0001	0.0001	0.0001
area_worst	0.0001	0.0001	0.0001
smoothness_worst	0.0001	0.0001	0.0001
compactness_worst	0.0001	0.0001	0.0001
concavity_worst	0.0001	0.0001	0.0001
concave.points_worst	0.0001	0.0001	0.0001
symmetry_worst	0.0001	0.0001	0.0001
fractal_dimension_worst	0.0001	0.0001	0.0001
fractal_dimension_se	0.0422	0.0487	**1.0000**
smoothness_se	**0.1053**	**0.1170**	**1.0000**
fractal_dimension_mean	**0.7667**	**0.8215**	**1.0000**
texture_se	**0.8354**	**0.8642**	**1.0000**
symmetry_se	**0.8871**	**0.8871**	**1.0000**

SAGx

SAGx is a package used in Bioinformatics which uses both local and tail area-based FDR estimation with the use of the functions discussed by Broberg, which combine the local FDR, Poisson Regression and Isotonic Regression [11].

Using the function fava.fdr available under this package, we are able to calculate the FDR values for each p-value obtained from our multiple tests. As can be seen in Figures 10 and Table 3, the computed local FDR is greater than the tail-based FDR (labeled simply as FDR).

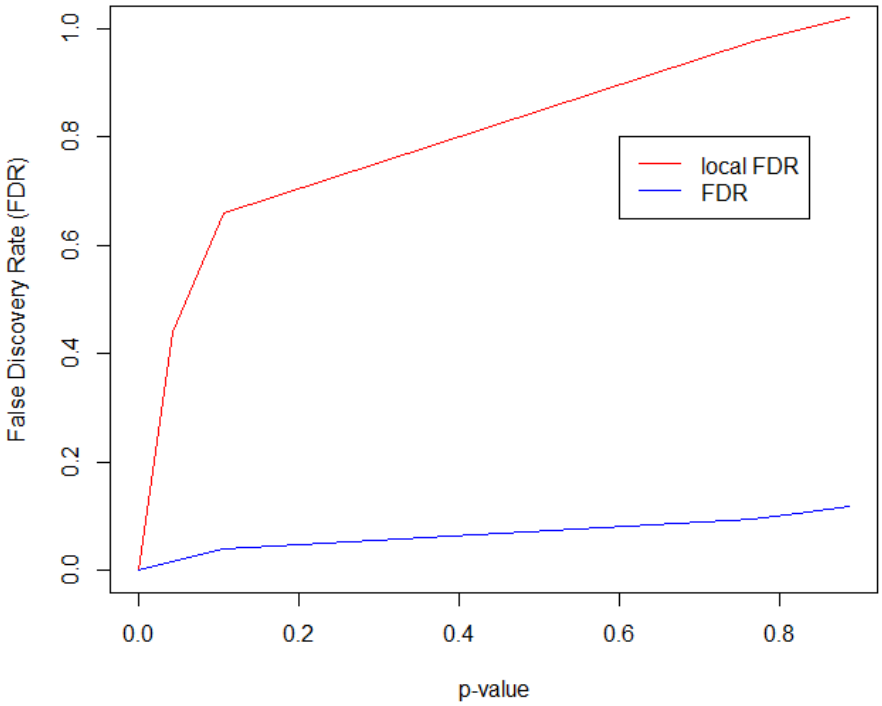

Fig. 10 - Scatter plot of the local false discovery rate and the false discovery rate as estimated by function pava.fdr

As it can be seen in Table 3, with concern to the local FDR, 5 variables were deemed not significant under the previously established level of significance of 0.05, while as for the tail-based FDR, only 3 variables were. Not surprisingly, the 5 variables with higher FDR values were the same between the SAGx and Qvalue packages and mostly agreeing with the FDR values from the other packages.

Table 03. Comparison between our p-value x localfdr x fdr according to the calculations performed with the SAGx package.

variables	p	local fdr	tail fdr
radius_mean	0.0001	0.0001	0.0001
texture_mean	0.0001	0.0001	0.0001
perimeter_mean	0.0001	0.0001	0.0001
area_mean	0.0001	0.0001	0.0001
smoothness_mean	0.0001	0.0001	0.0001
compactness_mean	0.0001	0.0001	0.0001
concavity_mean	0.0001	0.0001	0.0001
concave.points_mean	0.0001	0.0001	0.0001
symmetry_mean	0.0001	0.0001	0.0001
radius_se	0.0001	0.0001	0.0001
perimeter_se	0.0001	0.0001	0.0001
area_se	0.0001	0.0001	0.0001
compactness_se	0.0001	0.0001	0.0001
concavity_se	0.0001	0.0001	0.0001
concave.points_se	0.0001	0.0001	0.0001
radius_worst	0.0001	0.0001	0.0001
texture_worst	0.0001	0.0001	0.0001
perimeter_worst	0.0001	0.0001	0.0001
area_worst	0.0001	0.0001	0.0001
smoothness_worst	0.0001	0.0001	0.0001
compactness_worst	0.0001	0.0001	0.0001
concavity_worst	0.0001	0.0001	0.0001
concave.points_worst	0.0001	0.0001	0.0001
symmetry_worst	0.0001	0.0001	0.0001

fractal_dimension_worst	0.0001	0.0001	0.0001
fractal_dimension_se	0.0422	**0.4406**	0.0170
smoothness_se	**0.1053**	**0.6604**	0.0408
fractal_dimension_mean	**0.7667**	**1.0000**	**0.0751**
texture_se	**0.8354**	**1.0000**	**0.1070**
symmetry_se	**0.8871**	**1.0000**	**0.1367**

Conclusions

This report was inspired on the work presented on paper: BMC Bioinformatics 2008, 9:303 [1], which compared different tail-area and local FDR based algorithms in 2008, when trying to develop a unified tail-area and local based single algorithm. Some of those algorithms are presently updated and more powerful and actually most of them now indeed perform simultaneous tail-area and local based FDR estimations.

We have used 5 different algorithms for FDR estimation in a breast cancer diagnosis setting: one of the algorithms estimates that parameter based on the local FDR (*twilight*), 3 of them provide simultaneously an estimate for FDR based on both tail-area and local FDR estimation (*qvalue*, *fdrtool* and *SAGx)* while the *multtest* package only provides estimates for the tail-area FDR. Multiple test analysis showed that ca. 4-5 variables are not relevant for the breast cancer diagnosis according to the data set herein considered and all the package agree on the selection of the non-relevant variables. However, such a conclusion is not very different from the single-test p-value analysis because the relevant variables show very small raw p-values either in both the parametric and non-parametric testing approaches and also because the number of tests/variables is very reduced in this dataset. Moreover, the non-relevant variables show quite high raw p-values. However, this work was very relevant from a pedagogical point of view.

References

1. A unified approach to false discovery rate estimation. Strimmer K. BMC Bioinformatics. 2008. doi: 10.1186/1471-2105-9-303.
2. Multiple Testing. Romano et. al. Palgrave Macmillan (ed.) 2010 DOI 10.1057/978-1-349-95121-5_2914-1
3. Yoav Benjamini and Yosef Hochberg, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1 (1995), pp. 289-300

4. Bradley Efron, Robert Tibshirani, John D Storey, Virginia Tusher, “Empirical Bayes Analysis of a Microarray Experiment”, Journal of the American Statistical Association , Volume 96, 2001 - Issue 456, Pages 1151-1160.
5. Katherine S. Pollard, Sandrine Dudoit, and Mark J. van der Laan, “Multiple Testing Procedures: R multtest Package and Applications to Genomics”, University of California, Berkeley, Berkeley Division of Biostatistics Working Paper Series, Year 2004, Paper 164.
6. Stefanie Scheid and Rainer Spang, “twilight; a Bioconductor package for estimating the local false discovery rate”, BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 12 2005, pages 2921–2922.
7. Stefanie Scheid and Rainer Spang, “A Stochastic Downhill Search Algorithm for Estimating the Local False Discovery Rate”, IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 1, NO. 3, 2004.
8. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
9. W Nick Street, William H Wolberg and O L Mangasarian, “Nuclear Feature Extraction For Breast Tumor Diagnosis”, ISTSPIE - International Symposium on Electronic Imaging Science and Technology Volume 1905 Pages 861-870, San Jose, California, 1993.
10. J. D. Storey. A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B, 64:479–498, 2002.
11. P Broberg. Statistical methods for ranking differentially expressed genes. Genome Biology, 4:R41, 2003. doi: http://dx.doi.org/10.1186/gb-2003-4-6-r41. URL http: //genomebiology.com/2003/4/6/R41