# FALSE DISCOVERY RATE CONTROL IN VOLCANO PLOT FOR OMICS DATA

Célia Vidal & Vivien Dupont

Data Science in Health & Biostatistics, 3rd year, Ensai

## False discovery rate

By construction of hypothesis testing there is always a chance to wrongly reject the null hypothesis. For a single test, this is known as the type I error. When working with Omics data, the datasets are usually huge. That means as many chances to be wrong when hypothesis testing. That why we introduce the False Discovery Rate (FDR) as the expectation of the false discovery proportion (FDP):

$$FDR = \mathbb{E}\left(\frac{\#\, false\ positives}{\#\, Hypothesis}\right) = \mathbb{E}(\text{FDP})$$

## Benjamini-Hochberg method

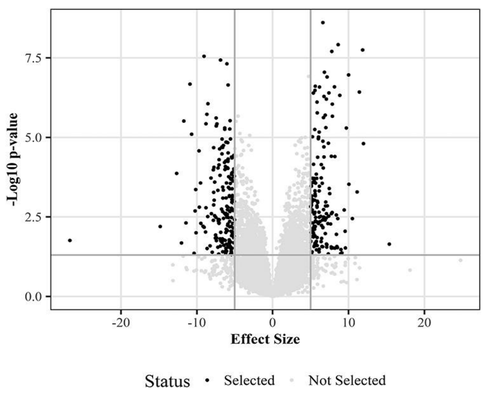
The Benjamini-Hochberg (BH) procedure for controlling the FDR is one of the most widely used procedures for p-value adjustment. It controls the proportion of false positives at a level $\alpha$ (usually 0.05). It reads as follow :

- Sort p-values : $p_{(1)} \leq \cdots \leq p_{(m)}$
- Define $\hat{I} = \max\{k | p_{(k)} \leq \alpha \frac{k}{m}\}$
- Reject all $i$ such that $p_i \leq p_{\hat{I}}$

## Volcano plot

When to many discoveries remain after application of the Benjamini-Hochberg procedure, the double filtering procedure can be applied, which keeps only the p-values that have a large effect size. This double filtering procedure is represented by a Volcano plot.

FIGURE 1: Volcano plot




The Volcano plot is a scatter plot of -log10 of p-values among significant p-values after Benjamini-Hochberg correction (y-axis) versus effect size (x-axis). The most interesting discoveries are on the right and left corners.

## Limitations & Objective

Nothing guarantee that a FDR-controlled subset is FDR-controlled itself. Intuitively, in the case of volcano plots, we can see that the denominator of the FDP will be reduced because we discard features with lesser effect size but the numerator (i.e. the number of false positives) may not decrease at the same rate. Situations where it doesn't work includes:

- Given the distribution of an estimate : $\widehat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma_i^2)$, we can see that $\widehat{\beta}_i$ will have a large effect size not only when its true effect $\beta_i$ is large but also when its variance $\sigma_i^2$ is large. When false positives features have a larger variance than true positives features, the FDP will inflate.
- There is generally more false positives than true features in a study, so even when the variances distributions are the same the biggest effect feature have a large probability to be a false positives.

Two approaches are proposed to overcome this weakness, the Focused Benjamini-Hochberg and the Closed testing.

## Focused Benjamini-Hochberg

The focused BH method is a variant of the classical BH procedure that guarantees FDR-control over a subset of discoveries by applying a filter to the corrected p-values. The procedure works as follow:

**Input:** p-values $\boldsymbol{p} = (p_1, \ldots, p_m)$, filter $\mathfrak{F}$

1 **for** $r \in \{0, p_1, \ldots, p_m\}$ **do**

2 | Compute $\widehat{FDP}\,(r) = \frac{m \times r}{\#\mathfrak{F}(\{j : p_j \leq r\}, \boldsymbol{p})}$;

3 **end for**

4 Compute $r^* = \max\{r \in \{0, p_1, \ldots, p_m\} : \widehat{FDP}(r) \leq \alpha\}$;

5 Compute the classical BH rejection set $\mathcal{R}* = \{j : p_j \leq r^*\}$;

6 Compute $\mathcal{U}^* = \mathfrak{F}(\mathcal{R}*, \boldsymbol{p})$;

**Result** : Filtered rejection set $\mathcal{U}^*$

In the case of Volcano plots, we shall select $\mathfrak{F}$ such that it selects the $k$ largest $|\hat{\beta}_i|$ or all the $|\hat{\beta}_i|$ above a certain threshold $\tau$.

## Closed testing

Instead of controlling directly the FDR, we can use a closed testing approach to estimate the FDP with a certain probability :

$$\mathbb{P}(FDP(D) \geq \overline{FDP}(D) \text{ for all subset } D) \geq 1 - \alpha$$

Closed testing procedure is valid aver all subsets of D. That insures us that the procedure works even with double filtration like volcano plots. The procedure works as follow:

**Input:** p-values $\boldsymbol{p} = (p_1, \ldots, p_m)$, $D = \{g : |\hat{\beta}_g| \geq \tau \,\&\, p_g \leq p\}$ the volcano plot subset

1 Compute $h_\alpha = \max\{\mathrm{i} \in \{0, \ldots, \mathrm{m}\} : \mathrm{i}p_{(m-i+j)} > j\alpha, \text{for j} = 1, \ldots, \mathrm{i}\}$;

2 Compute $\bar{t}(D) = \max_{1 \leq u \leq \#D} 1 - u + \{g \in D : h_\alpha p_g \leq u\alpha\}$;

3 Compute $\text{median}\overline{FDP}(D) = 1 - \frac{\bar{t}(D)}{\#D}$ with $\alpha = 0{,}5$;

4 adjust $\tau$ and $p$ such that $\text{median}\overline{FDP}(D) \leq 0{,}1$;

**Result:** $D$, the FDP bounded volcano plot set

## Results

In the real case study, the database contains 49 individuals and 58037 genes. Authors use a resampling method with 6 individuals for the test sample and 37 individuals for the validation sample. If among the 37 individuals there are significant genes for BH (adjusted P-value for BH <0.05) then they are considered differentially expressed.

- If differentially expressed gene in the validation and test sample → true positive (TP);
- If gene differentially expressed in the validation sample and not differentially expressed in the test sample (or vice versa) → false positive (FP);
- If gene not differentially expressed in validation and test sample → true negative (TN).

17% of the genes selected by the Volcano plot in the test sample were not significant in the validation. Filtering the results may lead to FDP inflation and that a large effect size does not necessarily imply that the result is a true positive.

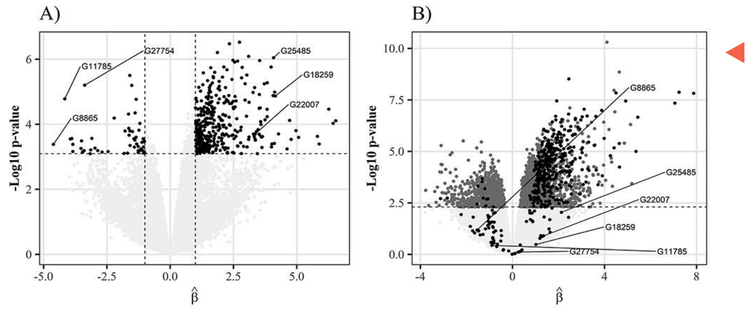



FIGURE 2: Volcano plot of the real case study with the Benjamini-Hochberg method

When we use the Closed testing method, we obtain an estimate medianFDP equals to 12%, that is better than the previous Benjamini-Hochberg method (17%). It is also closer to the expected value because the medianFDP is conservative and it never underestimate the FDP value. Morover, the confidence interval of the medianFDP includes the true value of FDP because it goes to 0% à 76%.

| | Selected | True Discovery ( 95% CI ) | False Discovery ( 95% CI ) |
|---|---|---|---|
| Number | 420 | 371 ( 100 , 420 ) | 49 ( 0 , 320 ) |
| Proportion | - | 0.88 ( 0.24 , 1 ) | 0.12 ( 0 , 0.76 ) |

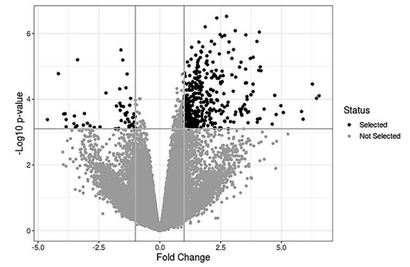



FIGURE 3: Volcano plot of the real case study with the Closed testing method

## Conclusion

FDR inflation occurs when FDR control procedures such as BH are combined with the Volcano Plots double filter procedure. FDR control on a set of discoveries does not imply FDR control on subsets of these discoveries. FDR inflation is high when the variance of differentially expressed genes is lower than the variance of non-differentially expressed genes. FDR inflation is less important when there are high correlations or a low proportion of null genes. Closed testing and focused Benjamini-Hochberg are two alternatives to the Benjamini-Hochberg procedure that allow double screening while controlling the FDR.

## Bibliography

1. Ebrahimpoor (2021) Inflated false discovery rate due to volcano plots: problem and solutions
2. Katsevich (2020) Filtering the rejection set while preserving false discovery rate control