Briefings in Bioinformatics, 22(5), 2021, 1–12

https://doi.org/10.1093/bib/bbab053
Method Review

Inflated false discovery rate due to volcano plots: problem and solutions

Mitra Ebrahimpoor and Jelle J. Goeman

Corresponding author: Mitra Ebrahimpoor, Medical statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands; E-mail: m.ebrahimpoor@lumc.nl

Abstract

Motivation: Volcano plots are used to select the most interesting discoveries when too many discoveries remain after application of Benjamini–Hochberg's procedure (BH). The volcano plot suggests a double filtering procedure that selects features with both small adjusted *P*-value and large estimated effect size. Despite its popularity, this type of selection overlooks the fact that BH does not guarantee error control over filtered subsets of discoveries. Therefore the selected subset of features may include an inflated number of false discoveries. **Results:** In this paper, we illustrate the substantially inflated type I error rate of volcano plot selection with simulation experiments and RNA-seq data. In particular, we show that the feature with the largest estimated effect is a very likely false positive result. Next, we investigate two alternative approaches for multiple testing with double filtering that do not inflate the false discovery rate. Our procedure is implemented in an interactive web application and is publicly available.

Introduction

Advances in DNA sequencing technology allow simultaneous measurement for thousands of features in one study. Generally, the goal is to compare the expression level of mRNAs/genes between two conditions (e.g. healthy vs. diseased) and identify differentially expressed features. In such high-dimensional settings, the chance of committing type I errors increases as well as the expected number of such false discoveries. This is a well-known issue and multiple testing procedures are adopted to avoid it. Popular multiple testing procedures control the false discovery rate (FDR), which is the expected proportion of type I errors among the discoveries. The most widely used such method in genomics is the Benjamini–Hochberg (BH) procedure [2]. BH provides a set of discoveries that maintains a prespecified FDR level, typically 0.05. FDR-based methods are generally more powerful than their competitors, making them a favorable choice for exploratory research.

Once the features are tested and multiple testing corrections are applied, the resulting list of potentially interesting discoveries can be rather long, so that researchers want to reduce it. It is not always desirable to prioritize the features merely based on the *P*-values, so researchers additionally use the estimated effect size (or fold-change) for *post hoc* filtering of the discoveries. This is known as a 'double filtering' procedure [22], where the results are first filtered based on significance and then according to the magnitude of change. The most famous double filtering tool in genomics is the volcano plot [6], that is widely used to visualize the results of genomic experiments. The volcano plot is a scatter-plot of the statistical significance ($-\log_{10}$ *p*-values on Y-axis) against the magnitude of effect (estimated

Mitra Ebrahimpoor is currently working as a PhD student at the Medical Statistics section, Department of Biomedical Data Sciences, Leiden University Medical Center. Her research focuses on statistical methods for high-dimensional data especially genomics.
Jelle J. Goeman is currently working as a professor at the Medical Statistics section, Department of Biomedical Data Sciences, Leiden University Medical Center. His research focuses on methods for large multiple hypothesis testing problems with applications in gene expression and neuroscience.
Submitted: 24 July 2020; **Received (in revised form):** 1 February 2021

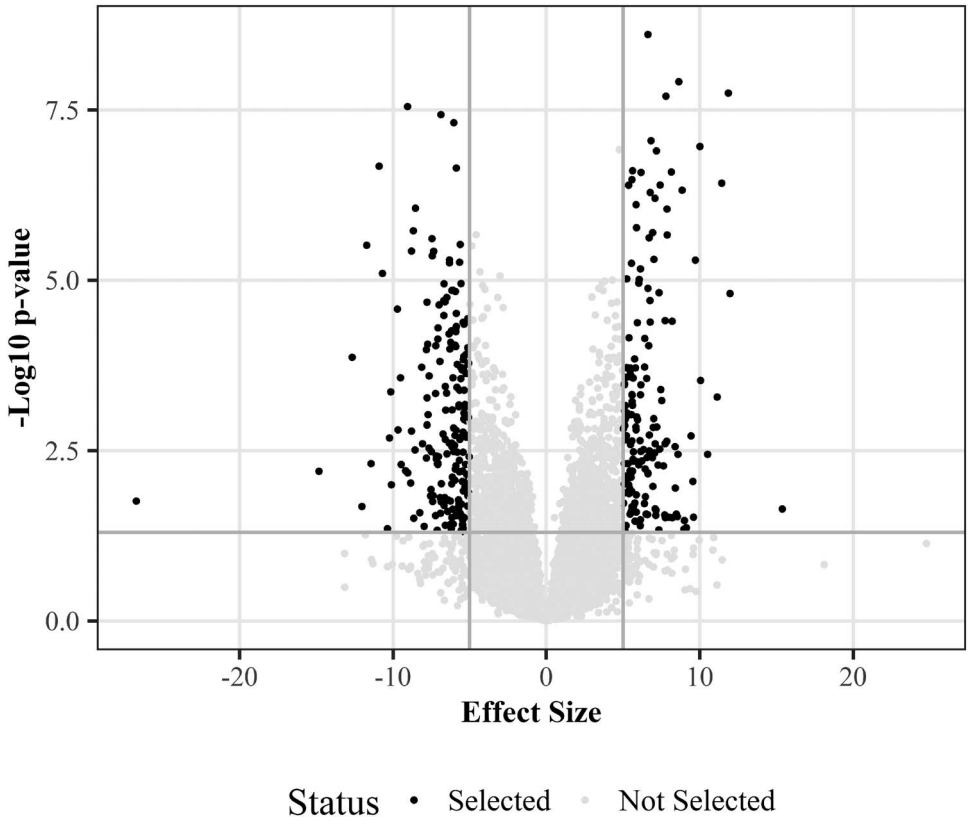

Figure 1. Example volcano plot. Points on top-right and top-left corners are considered the most promising findings.

effect size or fold change on X-axis). As shown in Figure 1, the most strongly up/down-regulated features lie towards the tails of X-axis and the most statistically significant features are towards the top. Discoveries on the upper-left and upper-right corners are considered most promising.

The intuition behind volcano plots is simple: it aims to select features that are not only significant but also carry the largest effect size. Nevertheless, the reliability of findings, especially in exploratory studies depends on the adequate control of type I error. Many researchers believe that the BH method will provide this necessary error control even when double filtering is adopted. Unfortunately, this is not true. Although BH controls the upper-bound for FDR over *all* rejected features, it provides no such guarantees over any subset of the features [1, 8, 12]. Consequently, BH is not guaranteed to control FDR after double filtering. As depicted by the simulations in this paper, the inflation of the type I error may be severe under certain conditions. In this paper we explain and quantify the problem of inflated type I error when double filtering with volcano plots. Moreover, we suggest alternative multiple testing procedures that do preserve the error control after volcano plot-type filtering. The rest of this paper is organized in two main sections. In Section 2, we explain the volcano plot problem and present examples of type I error inflation using RNA-seq data and simulation experiments. In Section 3, we introduce solutions for tackling this problem: controlling FDR using focused BH or controlling either the median or upper-bound of the false discovery proportion (FDP) using closed testing with Simes local tests. We briefly review the application of each method for volcano plot-type selection of findings and then compare them using simulation studies. Finally, we introduce Active Volcano Plot, a shiny app designed to create the volcano plots with valid type I error control.

Volcano plot problem

In this section, we describe and explain the potential problem of inflated type I error with volcano plots. We illustrate the severity of the problem using both simulations and real RNA-seq data.

Notation and problem setting

The most commonly adopted analysis pipeline in genomics includes normalizing expression values, fitting a model to test the effect of interest for each feature and applying multiple testing corrections. After multiple testing, if the number of significant features, i.e. discoveries, is large, a volcano plot may be adopted to illustrate the results and select the most relevant discoveries.

In the next two sections, we will review the analysis pipeline leading to the volcano plot problem. To keep the notations simple, we define the problem based on a simple linear regression model and with BH for multiple testing, but the problem arises also with more complicated designs or models, and when using different FDR-controlling methods.

Testing individual features

Let n denote the number of subjects in the study and m the number of features under investigation. Let, also, y_{gi} and x_i correspond to the gene expression and phenotypic data of gth

feature of the ith subject, respectively.

$$Y_g = \beta_{0g} + \beta_{1g}x + \epsilon_g, \tag{1}$$

where β_{0g} is the intercept, x is an $n \times 1$ vector denoting the phenotype variable and β_{1g} is the corresponding regression coefficient for gth feature for effect of interest. Also, $\epsilon_g \sim N(0, \sigma_g^2)$ is the residual error. Here the interest is differential expression analysis, so the null hypothesis for feature g is

$$H_{0g} : \beta_{1g} = 0. \tag{2}$$

This hypothesis may be tested based on a classical t-test, where the test statistic is

$$t_g = \frac{\hat{\beta}_{1g}}{\sqrt{\hat{\sigma}_g^2/n}}, \tag{3}$$

where $\hat{\beta}_{1g}$ and $\hat{\sigma}_g^2$ are the maximum likelihood estimates. In the following, we will simply write β_g for β_{1g}.

In many genomic studies, n is very small-even as small as 2 or 3, which leads to low power due to lack of degrees of freedom for the classical t-statistic. A variation of the test, called the moderated t-test, overcomes this by estimating σ_g^2 using empirical Bayes method [21], which shrinks large estimated $\hat{\sigma}_g^2$ downward and small $\hat{\sigma}_g^2$ upward. For more details on the moderated t, refer to the supplementary material.

The FDR inflation mechanism

Once the P-values are calculated for each feature, they need to be corrected for multiple testing. The most popular procedure for this is the BH procedure, which controls FDR. Control of FDR at level α means that the expected proportion of false discoveries among the discoveries is at most α: most of the discoveries are correct.

For some studies, BH finds 'too many' significant features and the researcher may want to reduce the number of findings. An intuitive way to do this is to discard findings with small estimated effect size, since larger effect sizes tend to be biologically more interesting. This is what is done using the volcano plot: among the BH-significant findings, the researcher considers only those findings that have an estimated effect size larger than some threshold, that is usually chosen after seeing the data. This double filtering (on P-value and effect size) is illustrated in Figure 1. First, the features are ordered by their BH-adjusted P-values, and non-BH-significant ones are discarded. Next, the features are ordered according to the absolute estimated effect size ($|\hat{\beta}_g|$). Then either k features with the largest effect size estimates or all features with a $|\hat{\beta}_g|$ larger than a certain threshold are selected. This way, researchers aim to target the most biologically relevant discoveries.

A problem with this procedure is that it ignores the fact that the guarantee of FDR-control was only over the full BH-corrected set, and not necessarily over the filtered subset. In general, a subset of FDR-controlled discoveries is not FDR-controlled [1, 4, 8, 12]. To understand why, note that FDR is the expected False Discovery Proportion (FDP), which is the ratio of the number of false discoveries over the total number of discoveries. Filtering reduces the denominator of the FDP, and may reduce the numerator as well, but not necessarily at the same rate. To retain FDR control with double filtering, the selection step should either retain the ratio of false to true discoveries, or discard false discoveries at a higher rate. Only if the double selection procedure can guarantee this, then we can guarantee FDR control over the double filtered results.

Examples of double filtering that are generally valid include sub-selection by the P-values again, retaining small P-values only. A valid selection would also be to select on the true effect size (β_g), selecting larger effect sizes. It might seem intuitive that selection on large estimated effect size would also be valid, but this is not true. To see this, look at the distribution of $\hat{\beta}_g \sim N(\beta_g, c_j\sigma_g^2)$. Large estimated effect sizes $|\hat{\beta}_g|$ occur not only when $|\beta_g|$ is large, but also when σ_g is large, since the standard error of $\hat{\beta}_g$ is large when σ_g is large. Selecting for large $|\hat{\beta}_g|$ is therefore not only selecting for large $|\beta_g|$, but also for large σ_g. The former selection will always decrease FDP, but the latter may increase it.

The obvious situation in which selecting for large σ_g would increase FDP is when null features tend to have larger variance than features under the alternative. This may or may not be the case in practice, but it becomes clear that an assumption for valid FDR control after double filtering is that such a negative association between $|\beta_g|$ and σ_g does not exist. We will see this clearly in the simulation results. If such a negative association cannot be excluded, double filtering may result in lack of FDR control.

However, even if σ_g varies between features, but is not associated with $|\beta_g|$, we may see lack of FDR control as a result of double filtering. Large $|\hat{\beta}_g|$ may be caused by either large σ_g or large $|\beta_g|$. If there is large heterogeneity in the values of σ_g, then features with large $|\hat{\beta}_g|$ are predominantly features with large σ_g. Even when null and non-null features have the same distribution of σ_g, a feature with a large σ_g is more likely a null feature, simply because there are usually more null than non-null features. Therefore, within the doubly selected subset of features, we may see an association between σ_g and $|\beta_g|$ due to a collider effect, even if none existed before filtering. Double filtering can therefore result in lack of FDR control even if there is no association between σ_g and $|\beta_g|$.

These phenomena we have described for the regular t-test. They are even more severe for the moderated t-test. While the moderated t-test does not influence the estimation of β_g, it does change the calculation of the P-value. Features with large σ_g tend to have large $\hat{\sigma}_g$, which is shrunk downward by the moderated t-test, resulting in a larger t-test statistic and a too small P-value, smaller than the regular t-test. Similarly, features with small σ_g tend to have a too large p-value. Under the prior assumptions of the moderated t-test, the mixture of too small and too large P-values normally evens out to give valid inference, as explained in the supplementary material. However, when selecting for large $|\hat{\beta}_g|$, which implies selecting for large σ_g as explained above, the moderated t-test therefore also inadvertently selects for P-values that are biased downward. This makes lack of FDR control after double filtering an even bigger problem with the moderated tests than with regular tests, and we will see this in the simulation results.

Problem Illustration 1: Simulation Study

In this section, we will illustrate the potential problem with volcano plots using simulations. We generate data based on a linear model and analyze them according to the routine volcano plot pipeline described earlier. Then the observed proportion of false discoveries among the volcano plot selected features is calculated under different simulation scenarios.

Simulation set-up

To mimic a usual application of volcano plots, we generated observations for $m = 20\,000$ features from n subjects in two groups. A set of h features were randomly selected to be truly differentially expressed (DE). The null hypothesis was true for the other $m - h$ features.

We generating true variances $\sigma_1^2, \ldots, \sigma_m^2$ based on a scaled inverse chi-squared distribution with degrees of freedom $d_0 = 4$ and scale parameter $\sigma_0^2 = 4$. Let r_g be the rank of σ_g^2, with $r_g = 1$ for the largest value of σ_g^2 and $r_g = m$ for the smallest. We randomly selected the h differentially expressed features based on a multinomial distribution with weights

$$w_g = \left(\frac{r_g}{m+1}\right)^{\lambda/2}\left(\frac{m+1-r_g}{m+1}\right)^{-\lambda/2}$$

and without replacement. The weights are defined in such a way that $\lambda > 0$ indicates an over-representation of large values of σ_g^2 among non-null features, $\lambda < 0$ indicates an over-representation of small values of σ_g^2 among non-null features, and finally $\lambda = 0$ indicates equal distribution of σ_g^2 values among true and false discoveries. Theoretically, we expect inflation of type I error to be more prominent with larger λ, as explained in the previous section. The regression coefficients β_g were set to zero for false discoveries and $\beta_g \sim N(0, \gamma^2\sigma_0^2)$ for truly DE features, where γ is the effect-size. The log-expression values were generated independently according to the model (1). The P-values were calculated for both the classical t-test and the moderated-t statistic.

Once the p-values were corrected for multiple comparisons using the BH procedure, features with an adjusted P-value < 0.05 were labelled significant. Following the volcano plot selection, these statistically significant features were sorted by their estimated $|\hat{\beta}_g|$ and those with the largest effect size were selected. For a selected feature, g, if $\beta_g \neq 0$, it was marked as true discovery and otherwise it was considered a false discovery.

We repeated the simulation experiment 1000 times for every combination of the simulation parameters $h/m = (0.01, 0.05, 0.1, 0.15, 0.2, 0.4, 0.8)$, $\gamma = (1, 1.5, 2)$, $\lambda = (-2, -1, 0, 1, 2)$ and three different sample sizes $n = (6, 12, 24)$. These values cover a realistic range of parameters in a common genomic experiment. For each repetition, the empirical FDP was calculated as the proportion of null features among features selected by the volcano plot, or 0 if no features were selected. FDR was calculated as the mean FDP.

Simulation results

For every combination of parameter values in the simulation experiment, we calculated the FDR for a double filtering procedure that selects the top $k = 1, \ldots, 100$ among the BH-significant results, i.e. the k BH-significant features with highest ranked values of $|\hat{\beta}_g|$. In case there were fewer than 100 features with BH-adjusted P-value < 0.05, all significant features were selected. Figure 2 gives the FDR (averaged over repetitions) for one typical scenario. A frequency plot of original values are presented as supplementary material (Figures S1-S3).

Generally, when $\lambda > 0$ we see inflated FDR for all values of k, although the severity of the problem clearly decreases with k. Naturally, as k increases and more features are selected the FDR converges towards its theoretical level of $\alpha\pi_0$, where $\pi_0 = 1 - h/m$ is the proportion of truly null features.

To investigate the way the FDR inflation depends on the parameters of the simulation we focus on the highest ranked ('top') feature, which we defined as the feature with the largest $|\hat{\beta}_g|$ among the significant ones, i.e. setting $k = 1$ in the previous figure. Figure 3 shows how often a volcano plot-selected top-feature is a false positive.

As expected from the theory in Section 2.2, FDR inflation is greater when null genes tend to have large variance ($\lambda > 0$), and mostly absent when the reverse holds ($\lambda < 0$). Inflation is still present when $\lambda = 0$. Other relationships we see, is that FDR inflation tends to increase as the proportion of active features increases, as well as when the sample size increases. We explain this by noting that as these parameters increase, there are more and more BH-rejected features to select from. With very low sample size and very low effect size, the power is too low to detect many features as significant, hence few discoveries are made, both true and false, and there is little selection even if k is small. Large effect size γ seems to diminish the problem, perhaps because with very large effect size high $|\hat{\beta}_g|$ will be more often due to large $|\beta_g|$ than due to large σ_g.

We have only presented the results for the classical t-test. A similar pattern was observed for the moderated-t, though with an even higher FDR inflation, (as expected, see Section 2.2). These plots are presented in supplementary material. Also in the Supplementary Material is an small investigation into the double filtering problem under dependence between features; we see there that the FDR inflation becomes smaller in some scenarios if genes are correlated, but the problem does not disappear.

Problem Illustration 2: RNA-seq Study

Here we analyze a real RNA-seq example to portray the volcano plot-problem in a realistic setting. In real data we cannot tell the true from the false discoveries, so it is not straightforward to show that FDR is inflated when volcano plots are used. As a solution, we adopted a resampling procedure to approximate the unknown FDP among features with large $|\hat{\beta}_g|$ estimates. In contrast to the simulation above, this analysis does not impose effect size, variability in σ_g or the value of λ, but takes it from real data. Our point here is to show that FDR inflation due to volcano plots may also happen in real data settings. Moreover, the real data also have a realistic correlation structure. We created several random analysis sets and evaluated the results based on the results of corresponding validation sets.

RNA-seq data analysis

The data-set includes RNA sequencing on blood samples collected from healthy children and those evaluated with diarrheal disease where the pathogens present are known. The goal of the study was to determine host response gene signatures specific to Salmonella, Shigella and rotavirus that distinguish them from healthy controls. A detailed description of the data collection procedure and primary results are available in [7]. Here we focus on differential expression between the group with Shigella infection (SH) and healthy controls (HC). RNA-seq counts were obtained through the recount project with accession number *SRP059039* [9]. The data included raw counts of 58037 features from 37 SH and 12 HC samples. The corresponding count data were normalized and low expressed features were removed using *edgeR* package [19].

To evaluate the reproducibility of the volcano plot results, we performed a resampling procedure. At each repetition, 6 subjects were randomly selected from each phenotype group as the analysis set, the rest were considered as the validation set of 37 samples. This validation set was used in each repetition

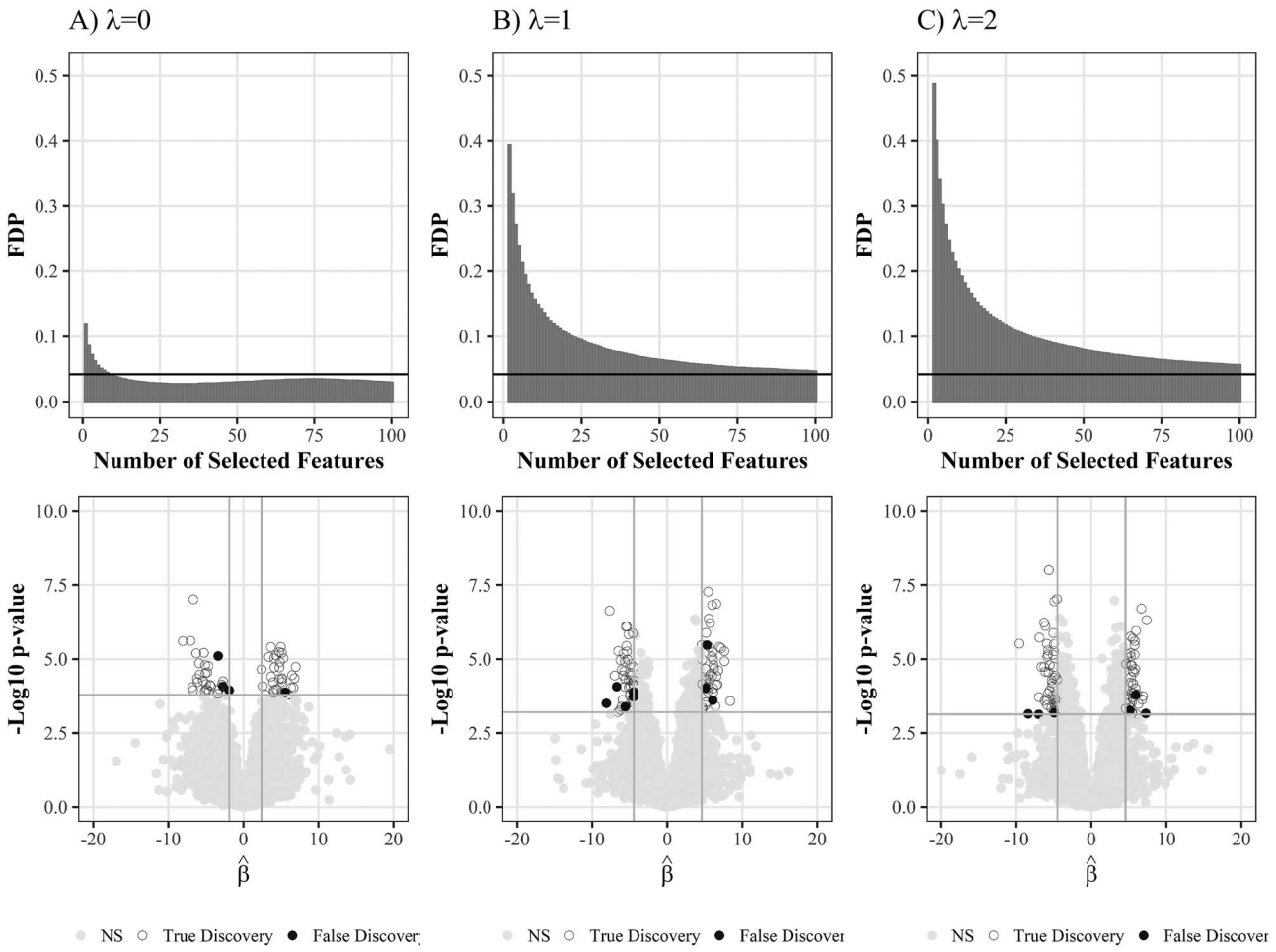

Figure 2. Simulation result. Observed FDR by selecting top $k = 1, 2, \ldots, 100$ features according to the volcano plot. The data were generated using the procedure explained in Section 2.3 with $\gamma = 1$, $h/m = 0.15$, $n = 12$ and **(A)** $\lambda = 0$, (B) $\lambda = 1$, (C) $\lambda = 2$. The bars represent the mean FDR over simulations and the horizontal lines indicate the theoretical FDR-level of BH ($\pi_0 \alpha = 0.043$). The volcano plots are single realizations of the 1000 repetitions that are averaged for bar plots at $k = 100$. The dark points are BH-selected false discoveries, the white points are BH-selected true discoveries; the gray points are discarded in volcano plot selection when $k = 100$.

to determine a ground truth of differential expression: all BH-significant genes (BH-adjusted *P*-value < 0.05) in the validation set were considered truly DE. To ensure that putative null genes were truly null, we permuted the gene expressions of null genes over the subjects. The same permutation was used for all null genes to preserve their correlation structure. Based on this ground truth, false discovery proportions could be determined for a double filtering procedure on the analysis set. We used the same double filtering as in the simulation above: to select top genes, significant genes (BH-adjusted *P*-value < 0.05) were sorted by their $|\hat{\beta}_g|$ value and $k = 1, 2, \ldots, 100$ genes with largest values were selected. Discoveries were classified as true or false according to the ground truth from the corresponding validation set. All analyses (in analysis and validation set) were performed using *voom* to derive the raw *P*-value and the estimate $\hat{\beta}_g$ per gene [18]. The procedure was repeated 1000 times and FDP results were averaged to derive the FDR estimates.

RNA-seq data results

As above, the bar plot presented in Figure 4 portrays the estimated FDR by selecting 1 up to 100 features with the largest $|\hat{\beta}_g|$ values. The average proportion of discoveries was 0.004 and 0.071 for analysis and validation sets, respectively. On average, the top feature ($k = 1$) in the analysis set is not significant in the validation set 16% of the time. To take a closer look, we plotted the volcano plot of both the analysis and validation sets for one replication in Figure 5. For this specific repetition, the validation set included 3357 (10%) significant genes with BH-adjusted *P*-value < 0.05. In the corresponding analysis set, there were 546 (2%) significant features after BH adjustment, and further selection with $|\hat{\beta}_g| > 1$ resulted in 420 top genes. The estimated FDR for top genes is much higher than expected nominal level, which is $(1 - 0.1) \times 0.05 = 0.045$ for this example. Namely, 17% of genes selected by volcano plot in the analysis set (black dots in Figure 5 A) were not significant in the validation set (black dots in Figure 5 B). Once again, these results confirm that filtering the findings may lead to an inflation of FDP and that large effect size does not necessarily imply that the finding is a true positive. In fact, our results suggest that a high $|\hat{\beta}_g|$ may be an indication of a false positive result.

Volcano plots with Type I error control

Despite the potential type I error inflation, volcano plots do have desirable properties. It is easy to create a volcano plot for almost any study design and it illustrates two attributes at the

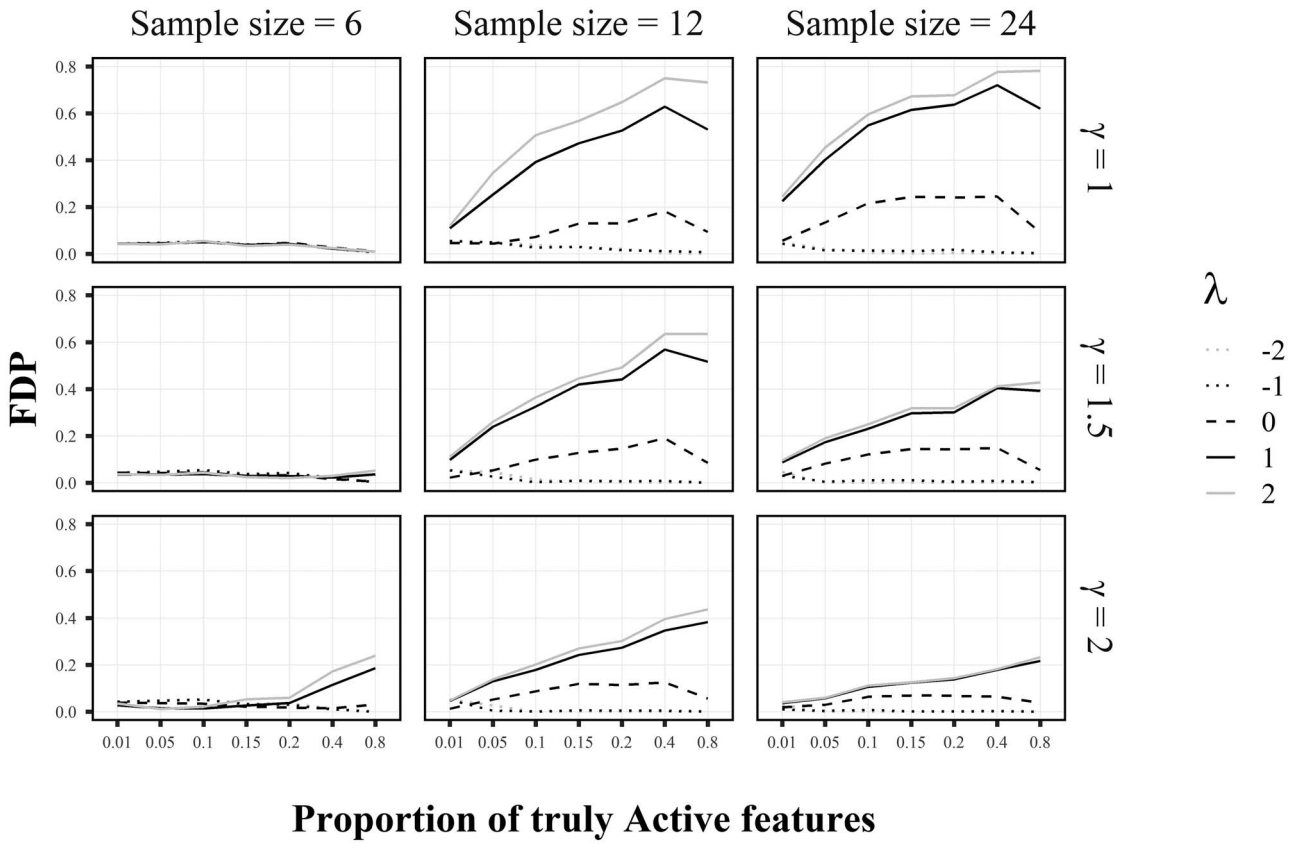

Figure 3. Simulation result. How often is the most highly ranked feature selected by volcano plot a type I error? This value can be as high as 0.77 for small sample size (= 24) and small effect size (= 1). Note that the type I error rate is largest when the variance of null features is higher ($\lambda > 0$) but also happens when there is no such association ($\lambda = 0$). It tends to get smaller as the effect of non-null features gets stronger.

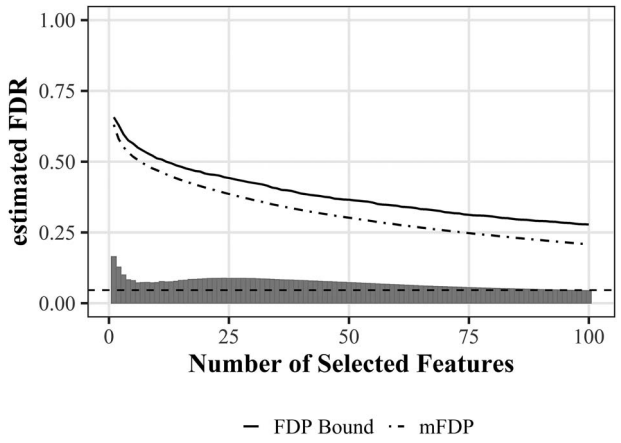

Figure 4. RNA-seq data analysis. FDR when selecting top $1, 2, \ldots, 100$ features with BH-adjusted P-value< 0.05.Bars represent the mean FDP over 1000 simulations and the dashed horizontal line is the average expected FDR of regular BH at $\pi_0\alpha = 0.046$. The curves are the average values of $\overline{FDP}$ (—) and mFDP(−. − .) as defined in Section 3.1.

same time: evidence for the presence of an effect and its estimated magnitude. Here, we suggest two alternative approaches to create a volcano plot, which will maintain the type I error control over the selected discoveries: controlling the FDP based on closed testing and controlling FDR based on focused BH (fBH). First, we briefly introduce each of the methods, and review their application in the volcano plot setting. Then we investigate the close relationship between the methods and compare their performance using simulation studies.

FDP control

A recent multiple comparisons procedure to limit the number of false discoveries is to control the False Discovery Proportion. [11] introduced a closed testing procedure [16] that controls FDP over all possible subsets of the collection of hypotheses. For any chosen set of discoveries D, the method provides an estimate along with a confidence bound for the FDP of any such subset. These confidence bounds bound the true FDP value with probability at least $1 - \alpha$. As the bounds are simultaneous over all 2^m possible sets D, *post hoc* choice or modification of D will not change the confidence level of the simultaneous bound. This feature makes FDP estimation an attractive alternative to FDR-correction when the volcano plots are used, since it promises FDP control, regardless of the way the discoveries have been selected. [13] introduced a shortcut to estimate the bounds in linear time when Simes local tests [20] are adopted. Here we focus on application of this procedure for volcano plots.

Let $M = \{1, \ldots, m\}$ be the index set of the features. Let $T_{DE} \subseteq M$ denote the unknown index set of the truly active features. Then the proportion of false discoveries in a set $D \subseteq M$ is given by

$$\mathrm{FDP}(D) = 1 - \frac{t(D)}{|D|},$$

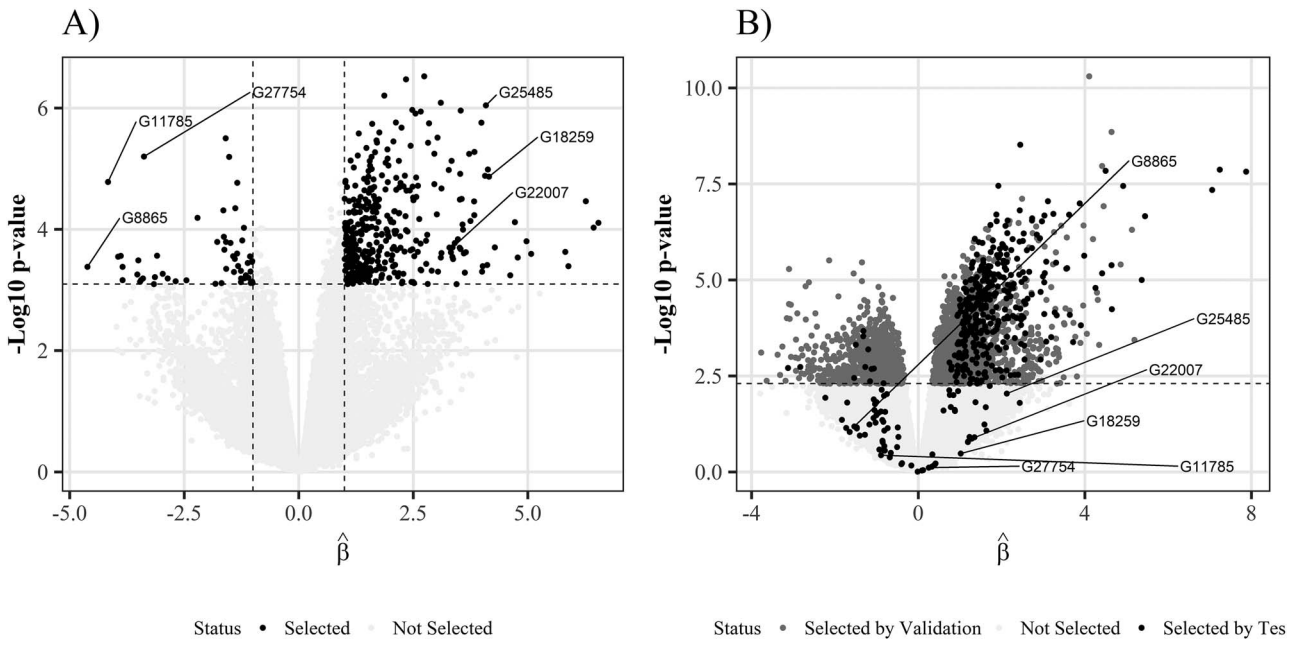

Figure 5. RNA-seq data analysis. **A)** Volcano plot of a single sub-sampled analysis set, black points are the genes that are both significant (small BH-adjusted *P*-value) and highly regulated (large $\hat{\beta}$). **B)** Scatter plot of the corresponding validation set where the significant genes are denoted by dark gray. Black points represent the points selected by volcano plot in the analysis set. Some of the black dots are labelled to illustrate how the results change for the same feature in the validation set.

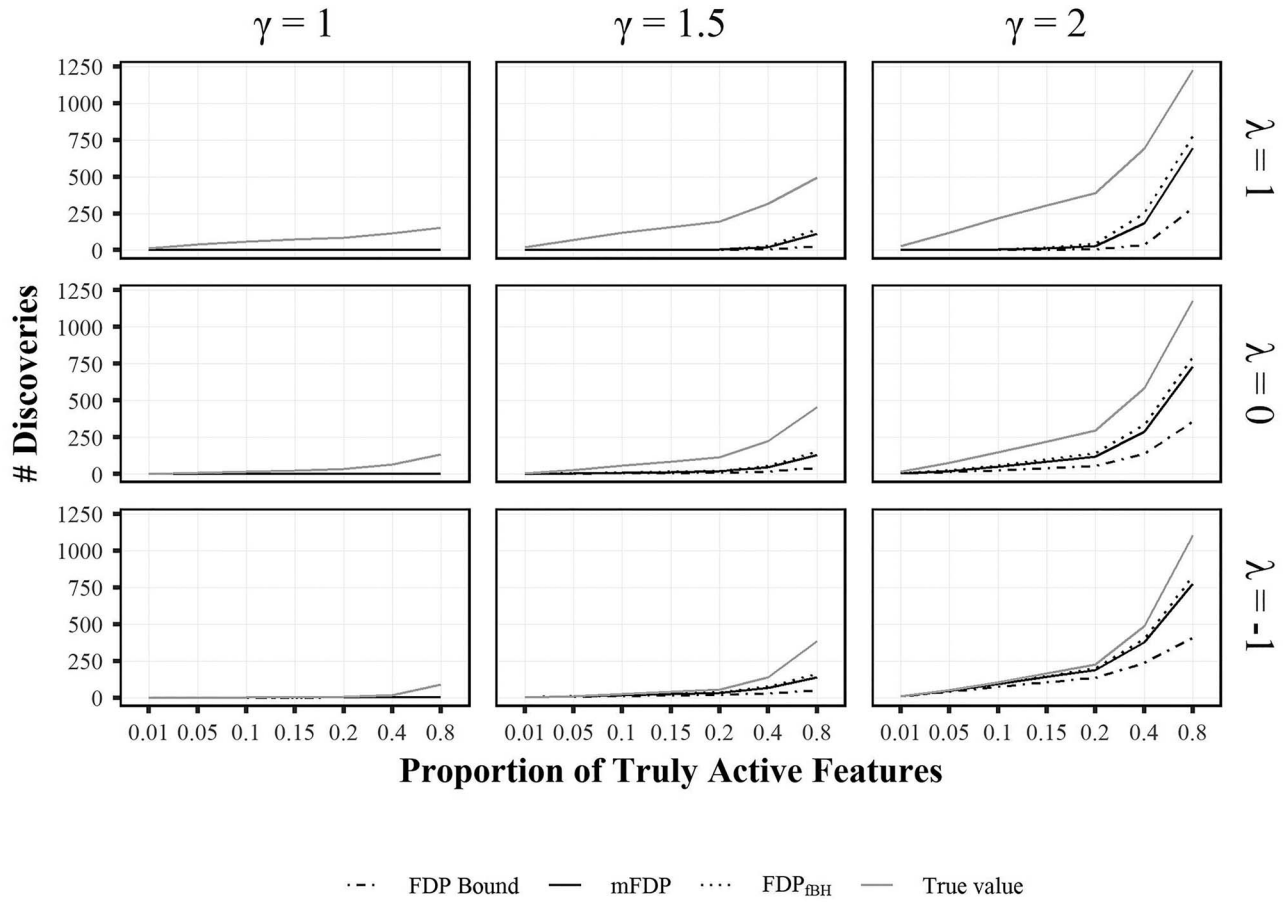

Figure 6. Simulation study 1. Number of discoveries made by controlling the FDP bound $\overline{FDP}$ (-.-.), mFDP (—) and FDP_{fBH} (.....) is compared to the actual number of discoveries (-.-.). Discoveries are selected based on a predefined τ and the *P*-value threshold is chosen by each method to restrict *FDP* at level $\alpha = 0.05$. There are six samples in each group.

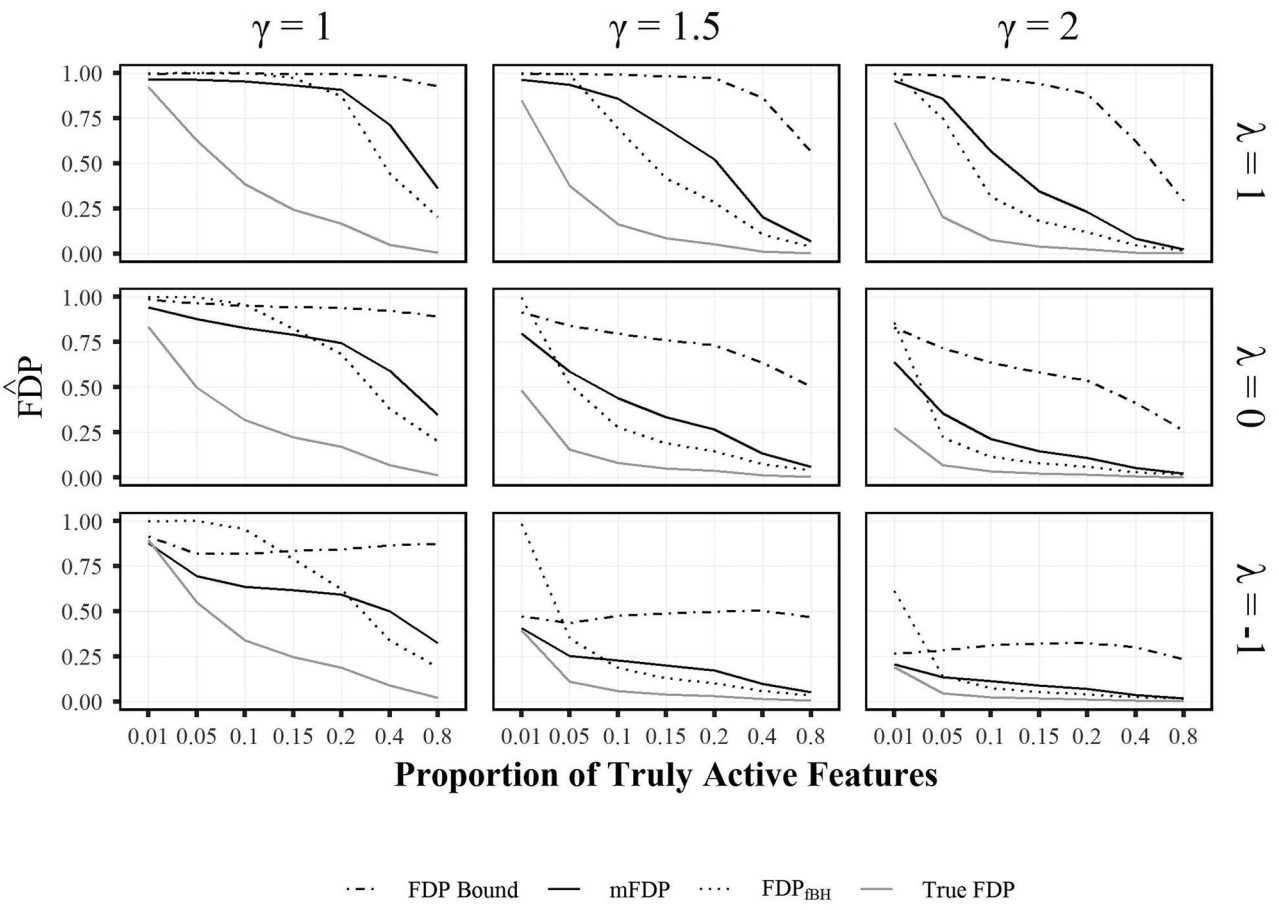

Figure 7. Simulation study 2. FDP is estimated based on the FDP bound $\overline{\text{FDP}}$ (-.-.), mFDP (—) and FDP_{fBH} (.....) for the set of discoveries selected based on thresholds $\tau = 3$ and $P = 0.001$. The actual value of FDP is also shown (——). There are six samples in each group.

where $t(D) = |T_{D\mathcal{E}} \cap D|$, and $|\cdot|$ denotes the size of a set.

As shown by [13], closed testing can be adopted to build simultaneous confidence bounds for FDP such that

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

The bound $\overline{\text{FDP}}(D)$ is given by

$$\text{FDP}(D) = 1 - \frac{\bar{t}(D)}{|D|},$$

where $\bar{t}(D)$ is the $(1-\alpha)$ lower confidence bound for $t(D)$, given by

$$\bar{t}(D) = \max_{1 \leq u \leq |D|} 1 - u + \{g \in D : h_\alpha p_g \leq u\alpha\}, \tag{1}$$

where

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

The 'midpoint' of the confidence interval, i.e. the lower confidence bound at $\alpha = 0.5$, can be used as a point estimate for the proportion of false discoveries in each set. We will denote this median point estimate by mFDP(D), found by choosing $\alpha = 0.5$. Note that mFDP(D) is also calculated based on a simultaneous confidence bounds and hence it is robust against choosing set D interactively on the basis of the data. This procedure is implemented in the R-package *hommel* [10]. Given the raw p-vales for all features, users can calculate the mFDP estimate and the corresponding confidence bound for any subset of features.

In the case of volcano plots, we can write $D = \{g : |\hat{\beta}_g| \geq \tau \ \& \ p_g \leq p\}$ to denote the set of features selected by volcano plot. Note that τ can be adjusted so that D is the set of top k features. Both mFDP and $\overline{FDP}$ can be estimated for various choices of D for various values of τ and P. In case a certain FDP value is of interest the researcher is free to adjust τ and P thresholds such that either the bound or the median FDP of the selected discoveries are below that level, without compromising the validity of these bounds. For instance, the user can freely tune the thresholds to select such that the final set of discoveries has mFDP less than or equal to 0.1. The set D of discoveries may depend on the data, and the way the set depends on the data does not have to be predefined.

Focused BH

Focused BH [15] is a variant of the classical BH procedure that guarantees FDR-control over a subset of discoveries selected in

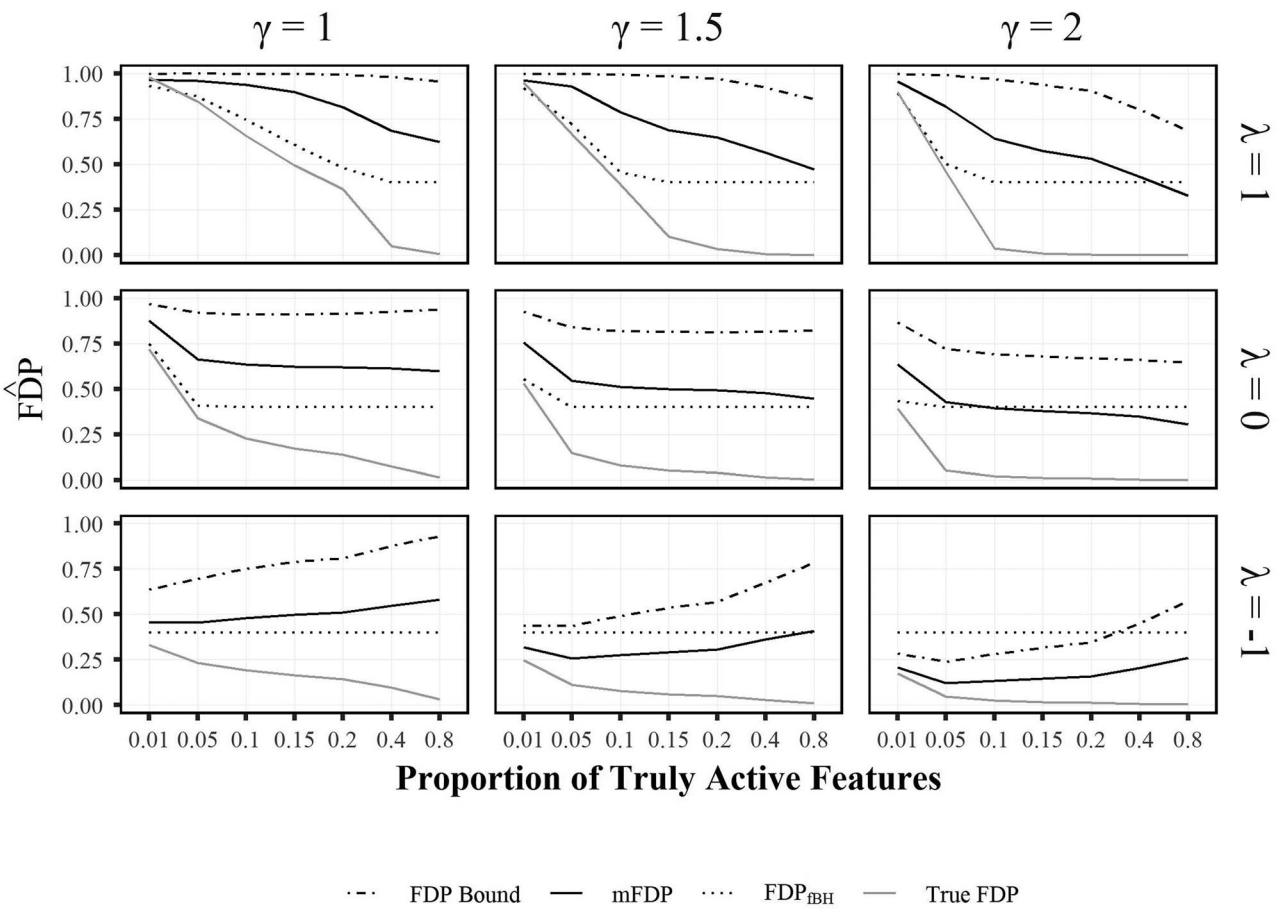

Figure 8. Simulation study 3. FDP is estimated using the FDP bound $\overline{\text{FDP}}$ (-.-.), mFDP (—) and FDP_{fBH} (.....) for the selected discoveries. 50 discoveries with the largest $\hat{\beta}$ values were selected among features with *P*-value ≤ 0.001. The actual value of FDP is also shown (-.-.). There are 6 samples in each group.

a predefined way. We will denote this subset by the index set F. The FDP_{fBH} estimate can be presented as a function of $r \in [0, 1]$,

$$\text{FDP}_{\text{fBH}}(r) = \frac{m \cdot r}{|\{g \in F : p_g \leq r\}|}. \tag{2}$$

Let r^* be the maximum value of r for which $\text{FDP}_{\text{fBH}}(r) \leq \alpha$. It has been shown that the set $D = \{g \in F : p_g \leq r^*\}$ is the set of discoveries with FDR-control over the subset F, under various conditions on the filter.

In the case of volcano plots, two type of filters are relevant. The first is a filter that selects only the features with estimated $\hat{\beta}_g$ above a pre-chosen threshold τ. The second is a filter that selects the k features with the largest estimated $|\hat{\beta}_g|$, where k is chosen *a priori*. Although these filters are of the general type explored in [15], neither one fulfills the conditions for which FDR control has been proven. We will investigate the performance of the method using these filters empirically.

Similarities between the methods

These two alternative methods essentially adopt two very different strategies to select the set D of top features but interestingly, they are still related. If B_γ was the set of discoveries given by fBH at level γ, then it can be shown that

$$\alpha\, \overline{\text{FDP}}_\alpha(B_\gamma) \leq \gamma\, \overline{\text{FDP}}_\alpha(M). \tag{3}$$

Now by replacing γ with $\alpha\gamma$, we have $P(\overline{\text{FDP}}(B_{\alpha\gamma}) > \gamma) \leq \alpha$. So by increasing the fBH error rate by a factor α, we can control the 95% upper quantile of FDP at level α. Furthermore, we can simplify the inequality in terms of mFDP. Given that $0 \leq \overline{\text{FDP}}(M) \leq 1$, the following holds,

$$0 \leq \text{mFDP}(B_\gamma) \leq 2\gamma\, \overline{\text{FDP}}_{1/2}(M) \leq 2\gamma. \tag{4}$$

Indicating that the estimated FDP is at most 2γ. Given that in genomics $\overline{\text{FDP}}(M)$ is rarely close to 1, the actual value can even be smaller. This loss of power is very small given the added flexibility that the mFDP is simultaneously estimated for all $2^m - 1$ subsets and not only the B_γ subset. A formal proof of this inequality is provided in the supplementary material.

We note that while the corollary (4), viewed simply as a property of fBH, could also have been proven directly using Markov's inequality, the underlying result (3) is a much more fundamental result on the similarity between the two methods [compare 13,discussion of Lemma 5].

Simulation study: comparing solutions

Direct comparison of mFDP and fBH results is not straightforward. We have designed three simulation studies, where the resulting set of discoveries or its properties are comparable across the approaches. For all simulations, the data are generated based on the linear model described in Section 2.3, but the procedure to select top discoveries D is different.

Simulation Study 1

The aim of this study was to compare the number of of discoveries made by the methods while each is controlling their respective error rate. Values of τ and α were pre-specified and discoveries were selected by choosing the P-value threshold P for features with $|\hat{\beta}_g| \geq \tau$, such that $\overline{\text{FDP}}$, mFDP or FDP_{fBH} was controlled at level α. Results for $\tau = 3$, $\alpha = 0.05$ and $n = 12$ are presented in Figure-6, plots for other scenarios are provided as supplementary material.

The three approaches have very similar results when the actual number of discoveries is low (small sample size, small γ, few truly active features). As expected, all methods underestimate the number of discoveries when $\lambda > 0$, and controlling $\overline{\text{FDP}}$ is always more conservative than controlling mFDP or fBH. For $\lambda < 0$ or large n ($n = 24$), the estimation bias for all methods including $\overline{\text{FDP}}$ is close to zero. In general, mFDP and fBH are similar but for $\lambda > 0$, the number of discoveries made by fBH is slightly higher compared to mFDP.

Simulation Study 2

The aim of this study was to compare the proportion of truly active features estimated by each approach. We defined the set D by fixing the τ and P thresholds, then calculated $\overline{\text{FDP}}(D)$, mFDP(D) and $\text{FDP}_{\text{fBH}}(D)$. Here the goal was not controlling the error rates at a certain level but rather comparing the FDP estimate of the same set of discoveries based on different approaches. Simulations were repeated for different values of τ and P.

Figure-7 illustrates the estimates for $\tau = 3$ and $P = 0.001$ and a sample size of 12; other conditions are presented as supplementary.

For relatively large effect sizes, fBH and mFDP are close to each other and both conservative when $\lambda \geq 0$. The FDP bound is very conservative except for either a large sample size, large γ or $\lambda < 0$. Although mFDP overestimates the true FDP in most cases, it starts to be less conservative with a more stringent P threshold (as can be seen in plots in supplementary e.g. Figure S17 and S14). It may even underestimate the FDP, when h/m is very small and both P and τ are small. This is not the case with $\overline{FDP}$. fBH slightly overestimates the FDP but is less biased than mFDP. Nevertheless, fBH has an inter-related association with P and τ thresholds. Similar to mFDP, the estimate gets closer to truth as the P threshold decreases. More specifically, when the P threshold is very low, fBH severely overestimates the FDP and is more conservative than $\overline{\text{FDP}}$ in some cases. On the other hand, if τ becomes very large, fBH starts to overestimate again, likely due to the fact that only a few features are selected. This is explained by the formula presented in (2); $\text{FDP}_{\text{fBH}}(D)$ increases as the size of filter decreases relative to mt.

Simulation Study 3

Here again, the aim of the study was to compare the proportion of truly active features estimated by different approaches. Threshold α was predefined but instead of fixing τ, we chose k features with the largest $\hat{\beta}$ values, making the value of τ data-dependent. Then FDP(D) was estimated as before based on the three approaches. The simulation was repeated for different values of k and P.

Figure 8 illustrates the estimates for $k = 50$, $P = 0.001$ and $n = 12$, here again other cases are presented as supplementary. The general scheme is similar to simulation study 2. As the signal increases (larger n or γ and larger h/m) both $\overline{\text{FDP}}$ and mFDP become less conservative. This is not the case for fBH. By definition, for a very small P threshold (r), the numerator of FDP_{fBH} is very small, hence $\text{FDP}_{\text{fBH}}(D)$ will underestimate the true FDP as relatively more features are selected (larger denominator).

RNA-seq data analysis

As an illustration, we will briefly discuss application of mFDP on the RNA-seq example introduced in the Section 2.4. For each analysis set, we calculated both mFDP and $\overline{\text{FDP}}$ for $k = 1, \ldots, 100$ genes with largest $|\hat{\beta}|$. The average over repetitions are plotted in Figure 4 for both methods. Due to the inflation of type I error, it is clear that this data set is not a case of $\lambda < 0$. As expected both mFDP and its corresponding bound are conservative, however they never underestimate the FDP value. Finally mFDP is more conservative for small values of k, which was also expected based on the simulation results. As explained earlier this type of double filtering is data-dependent so we did not apply fBH.

Interactive web application: active volcano plot

We have developed an interactive web application, Active Volcano Plot, through which the researcher can create a Volcano Plot and get the mFDP and FDP bound estimates for the selected discoveries. For this, we used the R package Shiny [5], which enables interactive plotting via a graphic user interface directly from R code.

The user can freely evaluate various sets of discoveries D by adjusting τ and P thresholds. The mFDP and $\overline{FDP}$ values will update instantly as the thresholds change, hence the word *Active*. Active Volcano Plot app is available at https://mebpr.shinyapps.io/activevp and the codes are accessible via https://github.com/mitra-ep/ActiveVolcanoPlot. An interactive interface requires a method that remains valid when the user adaptively chooses the set D of discoveries. For this reason only the mFDP and $\overline{\text{FDP}}$ methods are implemented, and not focused BH.

Upon submission of the current manuscript, we came across a closely related shiny app developed by Blanchard and others [3, 17]. This app can be accessed via https://pneuvial.github.io/sanssouci/.

As an example, we have used the selected analysis set in Figure 5 to create a volcano plot using this tool. Remember that the BH procedure at level 0.045 was applied, however, the classic volcano plot selection resulted in an FDP of 0.17. The screenshot in Figure 9 shows the same volcano plot where the median estimate mFDP is 0.12, much closer to the expected value and $\overline{\text{FDP}}$ is (0,0.76), which includes the true value of FDP.

Discussion

We have shown, using simulation experiments and empirical analysis of an RNA-seq data, that the features with largest estimated effect size, selected by a Volcano Plot can have an inflated type I error rate. It is important to stress that the problem occurs when FDR-controlling procedures such as BH are combined with double filtering procedure of volcano plots. FDR is a powerful

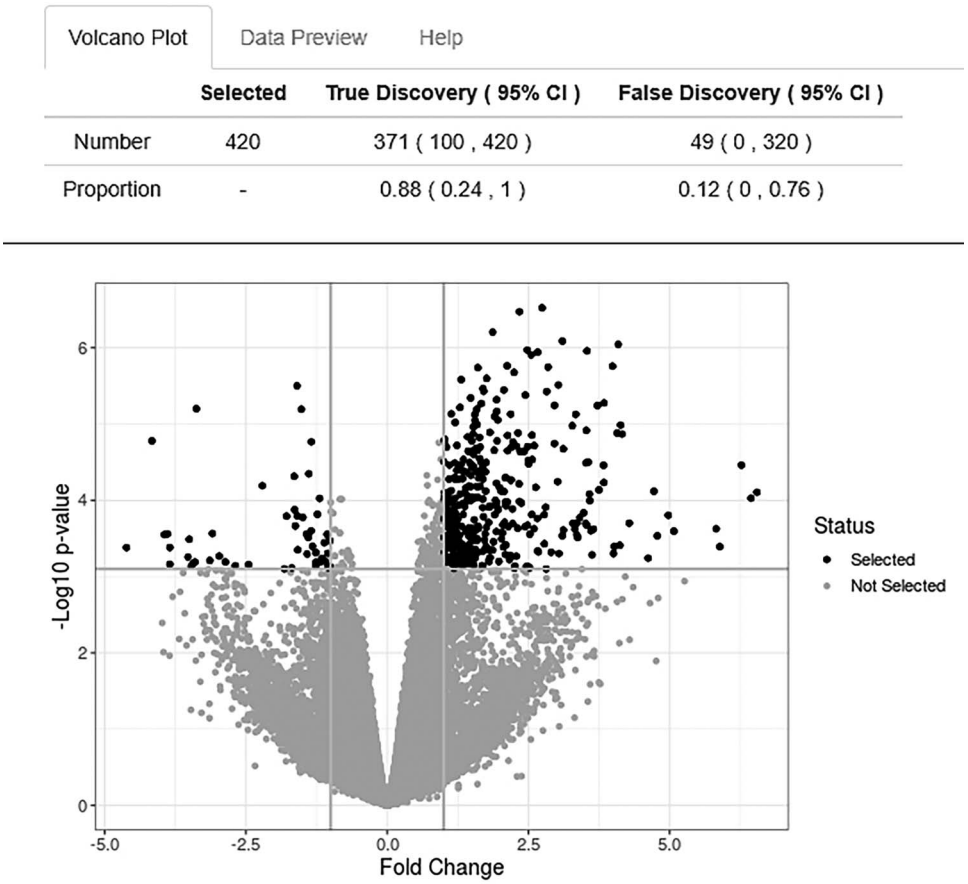

Figure 9. Shiny App: Active Volcano Plot. A classic volcano plot is made by selecting $P = 0.0008$ (3.1 in log-scale) and $\tau = 1$. The table above the plot presents the mFDP estimate along with the corresponding bounds for the selected features.

error rate in genomics studies but it is not suitable for combining with double filtering, since it does not have the sub-setting property: FDR control on a set of discoveries does not imply FDR control on subsets of those discoveries.

FDR inflation due to volcano plots may happen in many settings, but is especially pronounced when the variance of the truly differentially expressed features is less than the variance of the features that are not differentially expressed. There is no a priori biological reason why null features should have higher or lower variances than non-null features. We believe that either may happen, depending on the underlying biology, making double filtering methods hazardous in practice. We have shown that the conditions for FDR inflation may occur in real data settings, but we do not know how often, and we recommend that this is investigated. We also note that FDR inflation due to double filtering is aggravated when moderated *t*-tests are used instead of regular statistical tests. On the other hand FDR inflation is less severe if circumstances simultaneously exist in the data that deflate FDR, such as strong correlations or a low proportion of null features.

Volcano plots and double filtering should be seen as exploratory methods that change the collection of discoveries *post hoc*. Tailored methods are needed that are able to deal with such selective inference. FDP control [13] and focused BH (fBH) [15] are two alternatives to the BH procedure, which can control type I error even for volcano-plot-selected features, and in general for filtered discoveries. We have compared these methods extensively. While fBH may lead to more discoveries under certain conditions, it is limited to one *a priori*-chosen filter. On the other hand, by controlling FDP bound ($\overline{\text{FDP}}$) or median FDP (mFDP) the researcher can freely adjust and update the filter, even after seeing the data. Such interactive use of the method is implemented for volcano plots in the Active Volcano Plot app. There is more research to be done developing and comparing methods in this area. For example, permutation-based versions of fBH [15] and FDP-bound control [14] are available, may result in improvement of power, but are not studied here.

Key Points

- Classic volcano plots use double-filtering based on FDR-adjusted *P*-values and the estimated effect size, which can lead to serious inflation of the false discovery rate among selected features.
- Features with the highest fold-change are relatively more likely to be a false positive discovery.
- Closed testing with Simes local tests and focused-BH methods allow double filtering while preserving the false discovery rate.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Data Availability Statement

Codes used for simulations and data analysis mentioned in this manuscript are publicly available via GitHub (https://github.com/mitra-ep/ActiveVolcanoPlot) and are archived on Zenodo (https://zenodo.org/record/4459929).

Funding

This research was funded by Netherlands Organization for Scientific Research (NWO) Vidi grant number 639.072.412.

References

1. Barber RF, Candés EJ. Controlling the false discovery rate via knockoffs. *Ann Stat* 2015; **43**(5): 2055–85.
2. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 1995; **57**(1): 289–300.
3. Blanchard G, Neuvial P, Roquain E. Post-hoc confidence bounds on false positives using reference families. *Ann Stat* 2020; **48**(3): 1281–303.
4. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci* 2010; **107**(21): 9546–51.
5. Chang W, Cheng J, Allaire JJ, *et al*. *shiny: Web Application Framework for R*, 2019, R Package Version 1.3.2.
6. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol* 2003; **4**(4): 210.
7. DeBerg HA, Zaidi MB, Altman MC, *et al*. Shared and organism-specific host responses to childhood diarrheal diseases revealed by whole blood transcript profiling. *PLoS One* 2018; **13**(1):e0192082.
8. Finner H, Roters M. On the false discovery rate and expected type i errors. *Biom J* 2001; **43**(8): 985.
9. Frazee AC, Langmead B, Leek JT. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics* 2011; **12**(1): 449.
10. Goeman J, Meijer R, Krebs T. *hommel: Methods for Closed Testing with Simes Inequality, in Particular Hommel's Method*, 2019a, R Package Version 1.5.
11. Goeman JJ, Solari A. Multiple testing for exploratory research. *Stat Sci* 2011; **26**(4): 584–97.
12. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Stat Med* 2014; **33**(11): 1946–78.
13. Goeman JJ, Meijer RJ, Krebs TJP, *et al*. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika* 2019b.
14. Hemerik J, Solari A, Goeman JJ. Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika* 2019; **106**(3): 635–49.
15. Katsevich E, Sabatti C, Bogomolov M. Filtering the rejection set while preserving false discovery rate control. 2018; arXiv:1809.01792v3.
16. Marcus RUTH, Peritz ERIC, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**(3): 655–60.
17. Neuvial P, Sadacca B, Blanchard G, *et al*. *sansSouci: Post Hoc Multiple Testing Inference*, 2020, R package version 0.9.4.
18. Ritchie ME, Phipson B, Wu D, *et al*. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; **43**(7): e47–7.
19. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009; **26**(1): 139–40.
20. Simes RJ. An improved bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**(3): 751–4.
21. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004; **3**(1): 1–25.
22. Zhang S, Cao J. A close examination of double filtering with fold change and t test in microarray analysis. *BMC Bioinformatics* 2009; **10**(1): 402.