Post hoc inference for differential gene expression studies

Application of post hoc inference methods for differential expression studies in genomics

Pierre Neuvial

2021-04-20

Note: this document is adapted from the vignette "Post hoc inference for differential gene expression studies" available from the R package SansSouci (https://pneuvial.github.io/sanssouci/).

Introduction

We demonstrate how the `sansSouci` (https://github.com/pneuvial/sanssouci) package may be used to obtain post hoc confidence bounds on false positives in the case of differential gene expression analysis. After showing the output of a classical differential analysis based on False Discovery Rate control we illustrate the application of basic post-hoc bounds derived from probabilistic inequalities. Then, we introduce more powerful post hoc methods (introduced by Blanchard, Neuvial, and Roquain (2020)) that yield tighter bounds by adapting to unknown dependence by randomization. Finally we demonstrate the use of these methods on two applications of post hoc methods:

- confidence curves (envelopes) for the true or false positives
- statistical inference on volcano plots

The methods described in this vignette are described in detail in the book chapter Blanchard, Neuvial, and Roquain (2019) and in the paper Blanchard, Neuvial, and Roquain (2020). A shiny application for volcano plots is also available from https://shiny-iidea-sanssouci.apps.math.cnrs.fr/ (https://shiny-iidea-sanssouci.apps.math.cnrs.fr/).

```
library("plotly")
library("ggplot2")
library("sansSouci")
```

Set the seed of the random number generator for numerical reproducibility of the results:

```
set.seed(20210419)
```

Motivation: a differential gene expression study

We focus on differential gene expression studies in cancerology. These studies aim at identifying genes whose mean expression level differs significantly between two (or more) populations, based on a sample of gene expression measurements from individuals from these populations. Specifically, we consider a data set studied in Bourgon, Gentleman, and Huber (2010).

```
data(expr_ALL, package = "sansSouci.data")
dat <- expr_ALL
rm(expr_ALL)
```

This data set consists of gene expression measurements for $n = 79$ patients with B-cell acute lymphoblastic leukemia (ALL) Chiaretti et al. (2005). These patients are classified into two subgoups, depending on whether or not they harbor a specific mutation called “BCR/ABL”:

```
knitr::kable(table(colnames(dat)), col.names = c("Group name", "Frequency"))
```

Group name	Frequency
BCR/ABL	37
NEG	42

```
m <- nrow(dat)
```

The goal of this study is to understand the molecular differences at the gene expression level between the populations of BCR/ABL positive and negative (“NEG”) patients. For each patient, we observe a vector of $m = 9038$ gene expression values.

The most basic question to ask is:

> For which genes is there a difference in the mean expression level of the mutated and non-mutated population?

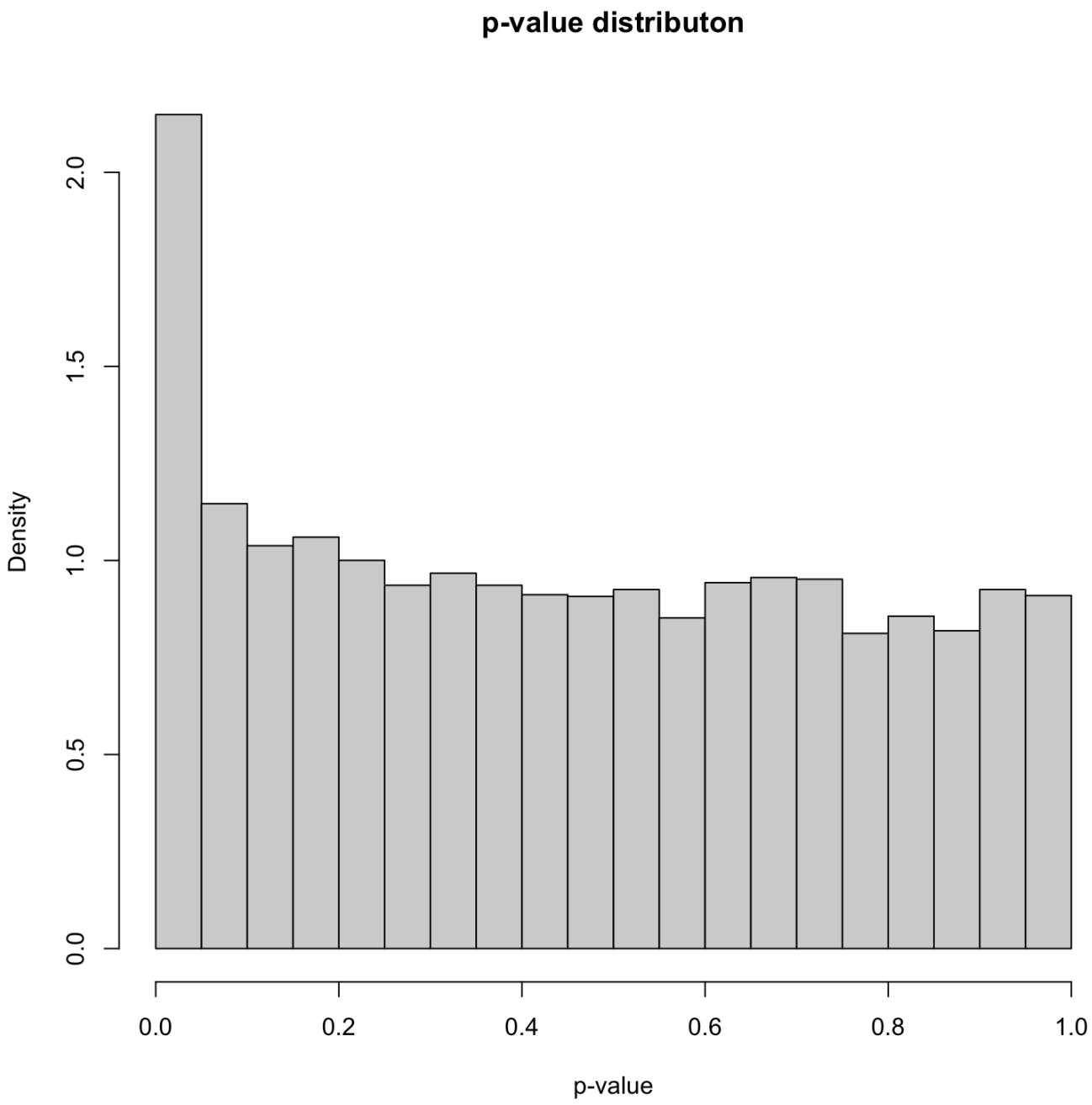
This question can be addressed by performing one statistical test of the *null hypothesis* of no difference between means for each gene, and to define *differentially expressed genes* as those passing some significance threshold.

Below, the Welch test for differential expression is applied to each gene. This can be done e.g. using the `sansSouci::rowWelchTests` function:

```
categ <- ifelse(colnames(dat) == "BCR/ABL", 1, 0) # map to 0/1
dex <- data.frame(rowWelchTests(dat, categ))
pval <- dex[["p.value"]]
```

We plot a histogram of the corresponding p-values:

```
hist(pval, probability = TRUE, breaks = 20,
     xlab = "p-value", main = "p-value distributon")
```



As expected, the distribution presents a large number of small p-values (which include signals, i.e. differentially expressed genes) mixed with uniformly distributed p-values (corresponding to non-differentially expressed genes).

Multiple testing correction: False Discovery Rate control

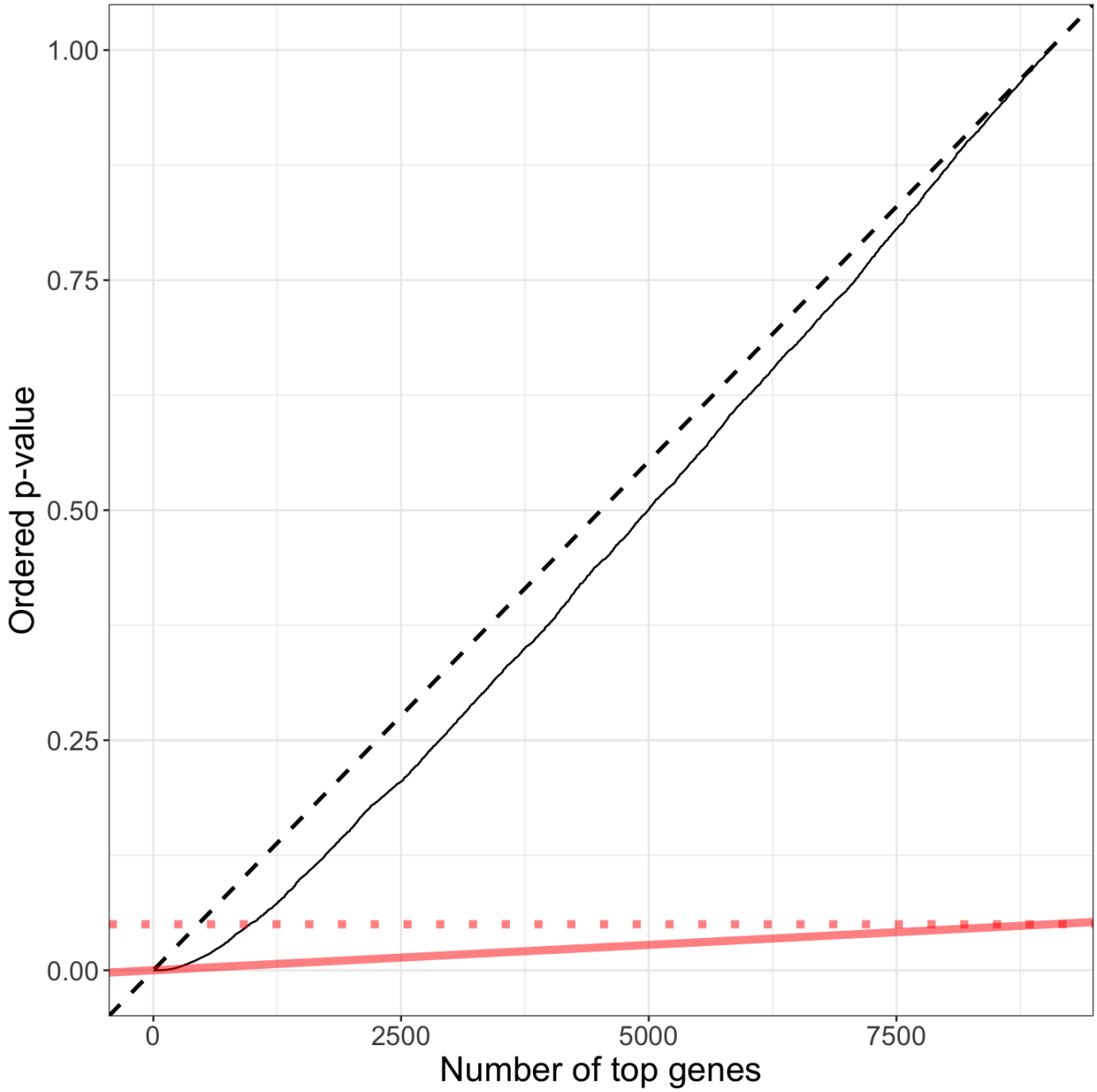
The state of the art approach to large-scale multiple testing is to control the False Discovery Rate (FDR), which is the expected proportion of wrongly selected genes (false positives) among all selected genes Benjamini and Hochberg (1995). The most widely used method to control this risk is the Benjamini-Hochberg (BH) procedure, which has been shown to control the FDR when the hypotheses corresponding to the non-differentially expressed genes are independent Benjamini and Hochberg (1995) or satisfy a specific type of positive dependence called Positive Regression Dependence on the Subset (PRDS) $\mathcal{H}_0$ of truly non-differentially expressed genes Benjamini and Yekutieli (2001).

```
q <- 0.05
adjp_BH <- p.adjust(pval, method = "BH")
dex$adjp <- adjp_BH
S_BH <- which(adjp_BH <= q)
nBH <- length(S_BH)
nBH
```

```
## [1] 150
```

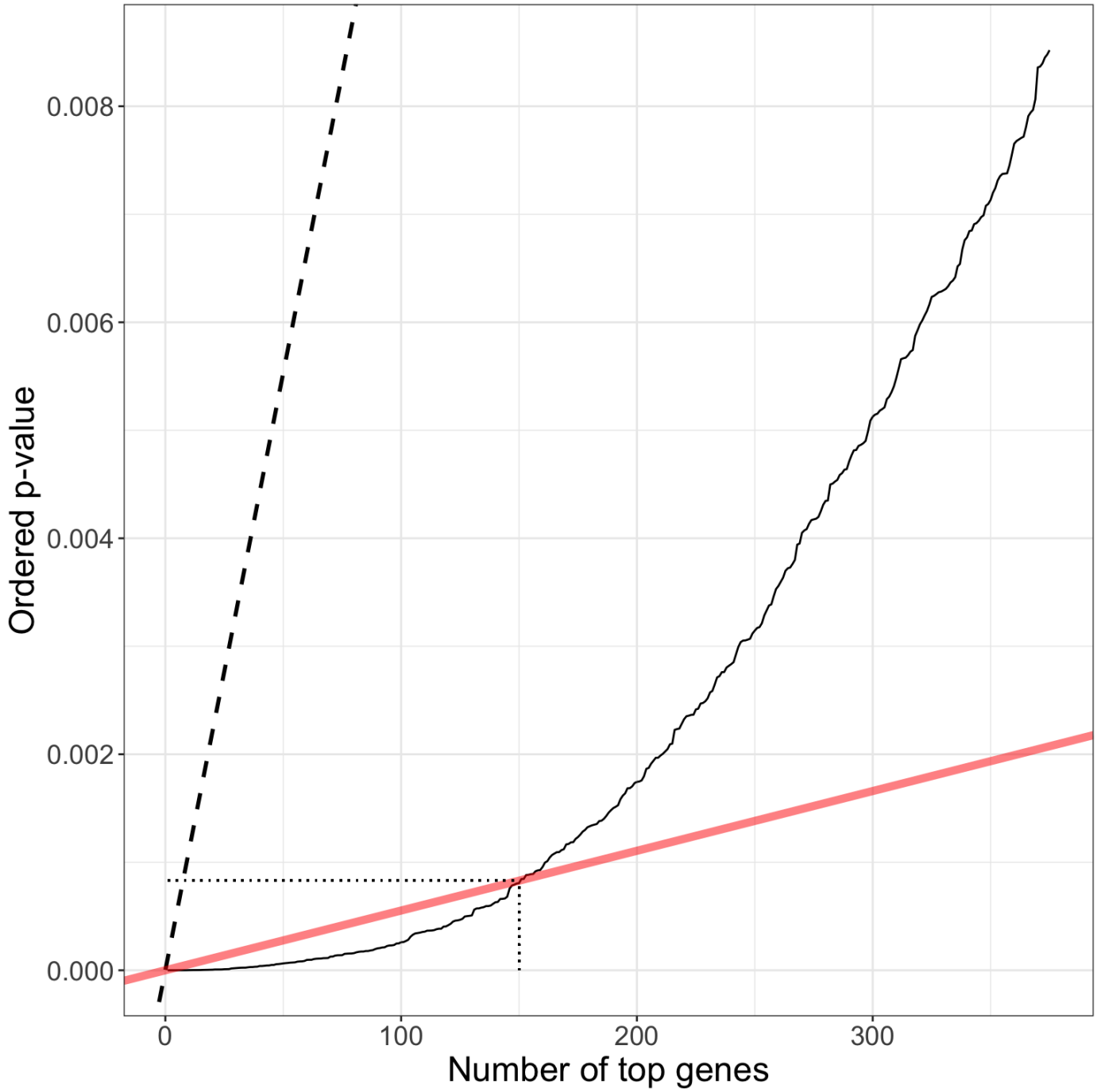
The application of the BH procedure at level $q = 0.05$ is illustrated in the figures below (all genes are displayed in the first one, second one is a zoom on the top genes):

```
my_col <- "#FF000080"
dexo <- dex[order(pval), ]  ## order genes by increasing p-values
dexo[["gene_order"]] <- 1:nrow(dex)

bh_plot <- ggplot(dexo, aes(x = gene_order, y = p.value)) +
  geom_line() +
  xlab("Number of top genes") + ylab("Ordered p-value") +
  geom_abline(slope = 1/m, intercept = 0, linetype = 2, size = 1) +
  geom_abline(slope = q/m, color = my_col, size = 2) +
  # geom_segment(aes(x = nBH, y = 0, yend = q*nBH/m, xend = nBH), linetype = "dotted") +
  # geom_segment(aes(x = 0, y = q*nBH/m, xend = nBH, yend = q*nBH/m), linetype = "dotted")
        +
  geom_abline(slope = 0, intercept = q, linetype = "dotted", color = my_col, size = 2) +
  theme_bw() +
  theme(axis.text = element_text(size = 14),
        axis.title = element_text(size = 18))
#geom_text(x = 0, y = q, label = expression(alpha), color = my_col)

bh_plot
```




```
xmax <- nBH*2.5
ymax <- dexo$p.value[xmax]
bh_plot +
  xlim(1, xmax) + ylim(0, ymax) +
  geom_segment(aes(x = nBH, y = 0, yend = q*nBH/m, xend = nBH), linetype = "dotted") +
  geom_segment(aes(x = 1, y = q*nBH/m, xend = nBH, yend = q*nBH/m), linetype = "dotted", co
        l = 1)
```

```
## Warning: Removed 8663 row(s) containing missing values (geom_path).
```



Caveats of FDR control

Caveat 1: FDR control prescribes a list of significant genes

In this data set, 150 genes are called differentially expressed at a False Discovery Rate (FDR) of $q = 0.05$. An investigator given such a list will generally be interested in refined and/or interpreting this list by combining it with on the problem at hand. For example, the investigator may not be interested in genes whose global expression level is small, or they may be interested in only a subset S of the list, or they may be interested in combining such an S with other genes which are below the significance threshold, but share some biological properties with the genes of S. Unfortunately, *FDR control on the list provides no guarantee on such user-refined gene lists*.

Caveat 2: Difficulty of interpretation of FDR control

The fact that 150 genes are called differentially expressed at a False Discovery Rate (FDR) of $q = 0.05$ does not mean that the proportion of false positives in this list, which is called the FDP for False Discovery Proportion, is less than 7.5. Indeed, the FDP is a *random* quantity, on which we can only have a probabilistic control. The FDR is defined as the *expected* FDP. Intuitively, FDR$\leq q$ means that the average FDP over hypothetical replications of the same genomic experiment and p-value thresholding procedure, is bounded by q. By construction, controlling FDR for a given data set does not give much information as to the FDP in this data set.

Post hoc inference

In order to address these important limitations, Goeman and Solari (2011) have popularized the concept of "post hoc inference." This approach elaborates on the theory of multiple testing in order to build confidence bounds on *arbitrary* subsets of hypotheses (subsets of genes in our context of differential gene expression).

Formally, let $\mathcal{H}$ be a set of m null hypotheses (one for each gene), and $\mathcal{H}_0$ be the (unknown) subset of true null hypotheses (non-differentially expressed genes). Then for $S \subset \mathcal{H}$, $|S \cap \mathcal{H}_0|$ is the number of false positives in S. With this notation, $\overline{V}$ is a post hoc upper bound at confidence level α if

$$\mathbb{P}(\forall S \subset \mathcal{H}, \quad |S \cap \mathcal{H}_0| \leq \overline{V}(S)) \geq 1 - \alpha$$

That is, there exists an event of probability $1 - \alpha$ such that *for any subset S of hypotheses* -possibly data-driven or cherry-picked by a user, the number of false positives in S is less than $V(S)$.

Post hoc inference may seem to be an excessively ambitious goal. Following earlier works by Genovese and Wasserman (2006), Goeman and Solari (2011) have proposed a general framework based on closed testing in order to build such bounds. In particular, they provide such a bound in the case where the tested hypotheses satisfy the PRDS assumption (that is, under the same assumptions as those under which FDR control is valid). We refer to Sarkar (1998) for a formal definition of PRDS. It was later shown by Blanchard, Neuvial, and Roquain (2020) that such post hoc bounds can be obtained as a consequence of the control of a multiple testing risk called the Joint Error Rate (JER). In particular, under PRDS, they recover the bound of Goeman and Solari (2011) under PRDS as a corollary of the Simes inequality Simes (1986), a probabilistic inequality that plays an important role in multiple testing.

In the remainder of this section, we introduce basic post hoc bounds and illustrate their application to differential expression studies. The next sections will introduce improved bounds (with respect to the original Goeman and Solari (2011) bounds) that are adaptive to dependence, and illustrate their applicationin the specific case of two-sample tests for differential gene expression studies.

Basic posthoc bounds

```
alpha <- 0.1
pvalo <- dexo$p.value
```

k_0-Bonferroni bound

For a fixed k_0, the generalized Bonferroni procedure consisting in rejecting the hypotheses in $R_{k_0} = \{i : p_i \leq \alpha k_0/m\}$ controls the k_0-Family-Wise Error Rate: it ensures that with probability larger than $1 - \alpha$, the number of false positives in R_{k_0} is not larger than $k_0 - 1$. As noted in Blanchard, Neuvial, and Roquain (2019), this leads to the post hoc bound:

$$V(S) = \sum_{i \in S} 1_{\{p_i \geq \alpha k_0/|S|\}} + k_0 - 1$$

As an application we calculate the bound associated to $k_0 = 100$ for $\alpha = 0.1$:

```
k0 <- 100
S <- 1:nBH
FP_k0 <- sum(pvalo[S] >= alpha*k0/m) + k0 - 1
FP_k0/nBH
```

```
## [1] 0.66
```

This implies that with probability larger than 0.9 the false discovery proportion among the genes selected by the BH procedure at level $q = 0.05$ is upper bounded by 0.66.

Simes bound

A more refined post hoc bound has been proposed by Goeman and Solari (2011) under the PRDS assumption. In the framework of Blanchard, Neuvial, and Roquain (2019) this bound is a direct consequence of the Simes (1986) inequality. It can be applied to the 150 rejections of the BH procedure as follows:

```
obj <- SansSouci(Y = dat, groups = categ)
res_Simes <- fit(obj, B = 0, family = "Simes", alpha = alpha) ## B=0 => no calibration!
FP_Simes <- predict(res_Simes, S_BH, what = "FP")
```

The Simes bound implies that with probability larger than 0.9, the false discovery proportion among the genes selected by the BH procedure at level $q = 0.05$ is upper bounded by 0.44. The Simes bound is sharper than the k_0-Bonferroni bound because it is obtained from a *joint* control of all k-FWER for all $k = 1, \dots, m$. The k_0-Bonferroni bound will therefore not be considered further in this vignette.

Tighter confidence bounds by adaptation to unknown dependence

As discussed in Blanchard, Neuvial, and Roquain (2020), the Simes bound has two major limitations, being a consequence of the Simes inequality:

- It is known to be valid only under certain positive dependence assumptions (PRDS) on the joint p-value distribution. Although the PRDS assumption is generally accepted in the case of differential expression studies (which justifies the application of the BH procedure itself), it has not been formally proved to hold in this case.
- It is not *adaptive* to the specific type of dependence at hand for a particular data set.

To bypass these limitations, Blanchard, Neuvial, and Roquain (2020) have proposed a randomization-based procedure known as λ-calibration, which yields tighter bounds that are adapted to the dependency observed in the data set at hand[1]. In the case of two-sample tests, this calibration can be achieved by permutation of class labels:

```
B <- 1000
res <- fit(obj, B = B, alpha = alpha, family = "Simes")
```

An alternative to the Simes/Linear reference family is the Beta reference family:

```
K <- 50
res_Beta <- fit(res, B = B, alpha = alpha, family = "Beta", K = K)
```

As expected from the theory, the post hoc bounds obtained after calibration by these methods is much tighter than the Simes bound:

```
resList <- list("Simes" = res_Simes,
            "Linear" = res,
            "Beta" = res_Beta)
names(resList)[3] <- sprintf("Beta (K=%s)", K)

bounds <- sapply(resList, predict, S_BH)
rownames(bounds) <- c("Lower bound on True Positives", "Upper bound on False Discovery Prop
        ortion")
knitr::kable(t(bounds), digits = 2)
```

	Lower bound on True Positives	Upper bound on False Discovery Proportion
Simes	85	0.43
Linear	116	0.23
Beta (K=50)	128	0.15

In the next two sections we illustrate the use of these improved bounds in order to build

- confidence curves for the true or false positives
- confidence statements for volcano plots

Confidence curves on "top-k" lists

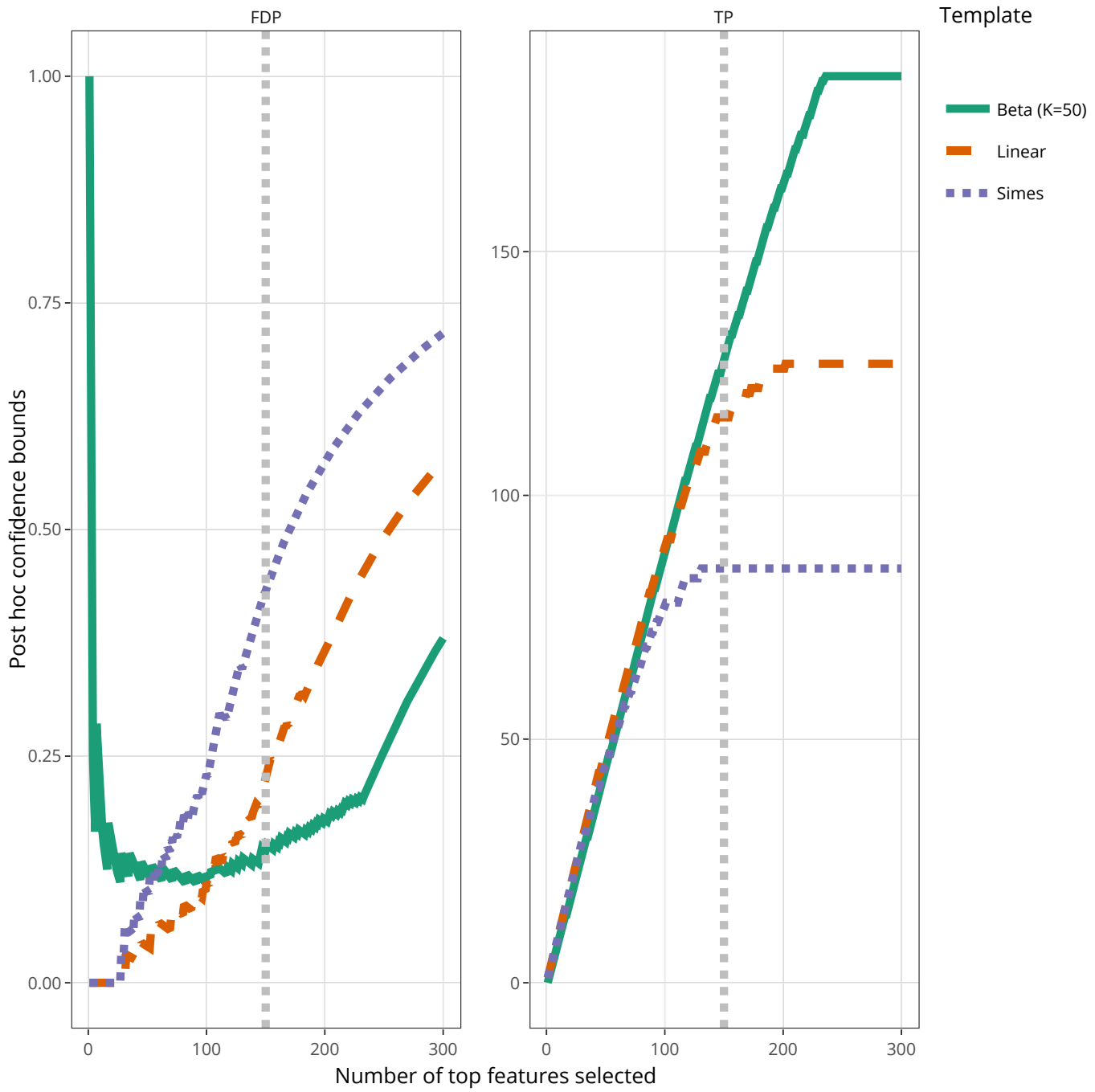
In the absence of prior information on genes, a natural idea is to rank them by decreasing statistical significance, and a natural question to ask is:

> Can we provide a lower confidence curve on the number (or proportion) of truly differentially expressed genes among the most significant genes?

We illustrate the use of post-hoc methods to provide this type of information. More specifcally, we build confidence statements on the number of true/false positives within the top k most significant genes in a differential gene expression study, where k may be defined by the user after seing the data, and multiple choices of k are allowed.

The confidence curves obtained by calibration of the Simes and Beta families can be compared graphically to the (parametric) Simes curve that can be obtained from Goeman and Solari (2011):

```
conf_bounds <- lapply(resList, predict, all = TRUE)
cols <- RColorBrewer::brewer.pal(length(conf_bounds), "Dark2")
p <- plotConfCurve(conf_bounds, xmax = 300, cols = cols)
p <- p + geom_vline(xintercept = nBH, color = "gray", linetype = "dotted", size = 1.5) +
  geom_line(size = 1.5)
ggplotly(p)
```



Both calibrated curves outperform the Simes curve in this example.

Volcano plots

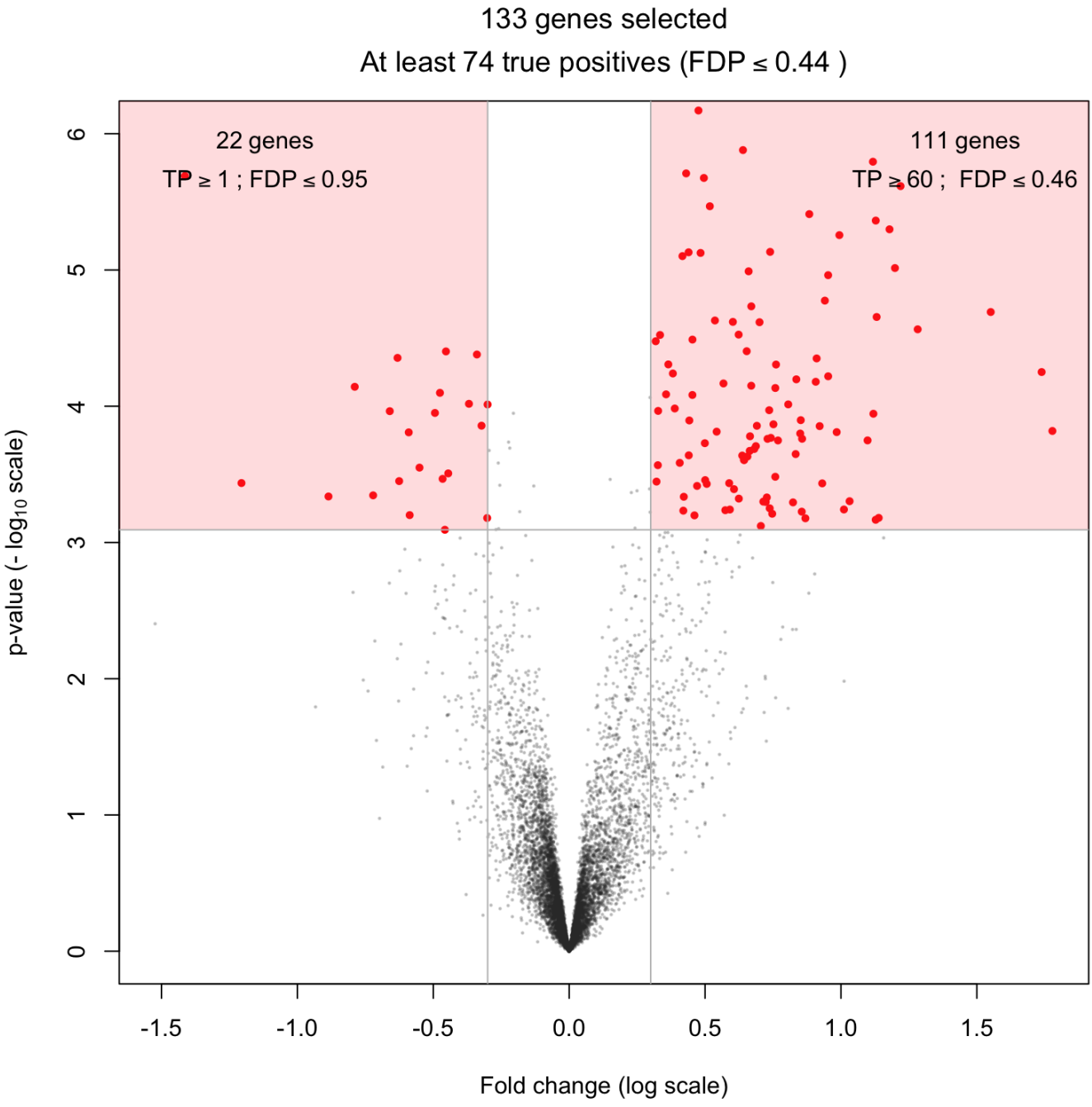
For an interactive volcano plot, see the volcano plot shiny application (https://shiny-iidea-sanssouci.apps.math.cnrs.fr/).

```
q <- 0.05
r <-  0.3
```

Let us assume that we are interested in genes selected by the BH procedure at level $q = 0.05$ and whose fold change is larger than $r = 0.3$ in absolute value. The "fold change" is defined as the difference between the expression means of the two groups compared; it is an estimate of the effect size of a gene. This double selection

corresponds to two sets of genes, with positive/negative fold change, which can be represented in the following plot:

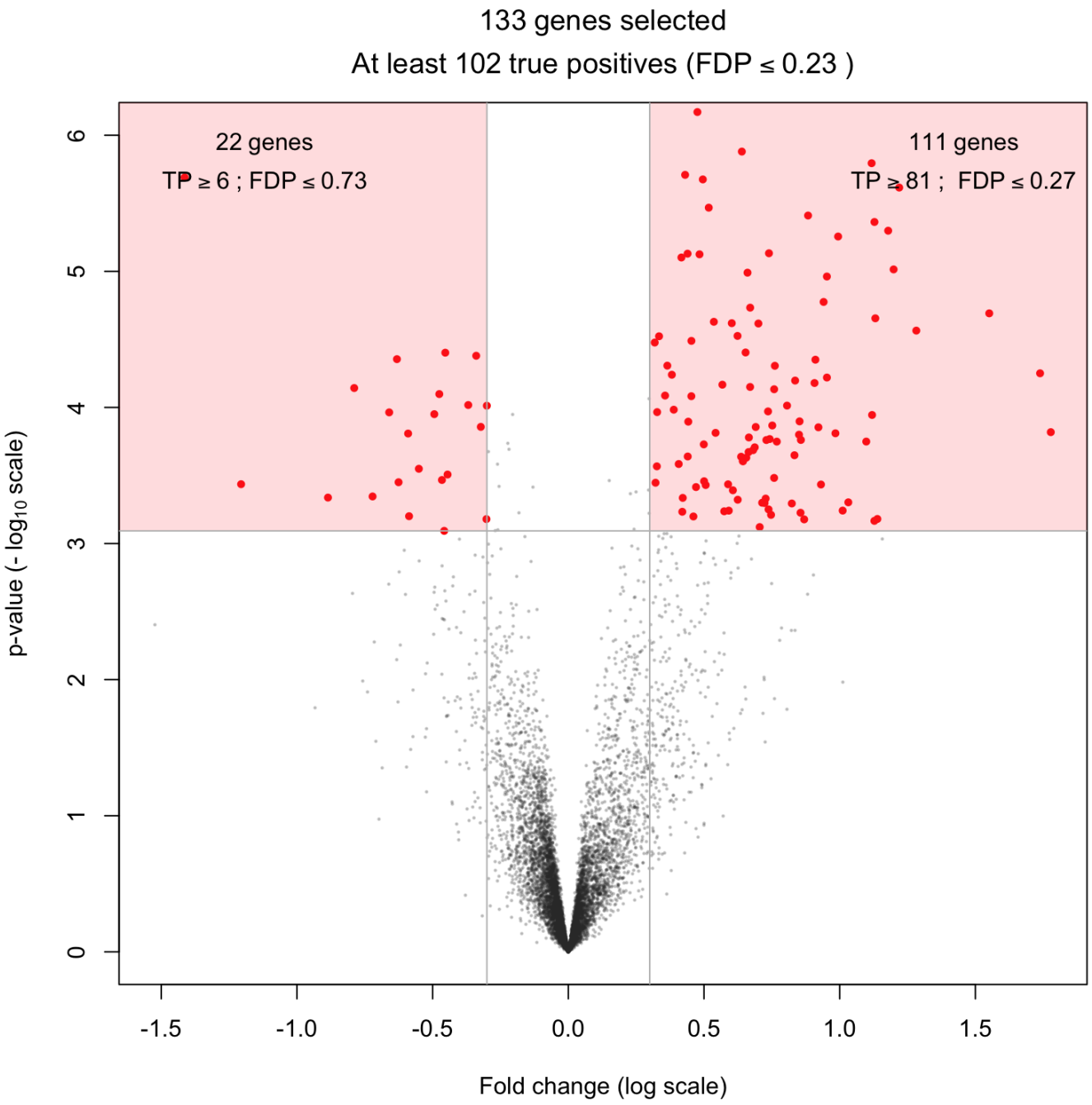
```
ylim <- c(0, 6)
volcanoPlot(res_Simes, q = q, r = r, ylim = ylim)
```



This type of plot is called a “volcano plot” Cui and Churchill (2003). Post hoc inference makes it possible to obtain statistical guarantees on selections such as the ones represented in the above figure.

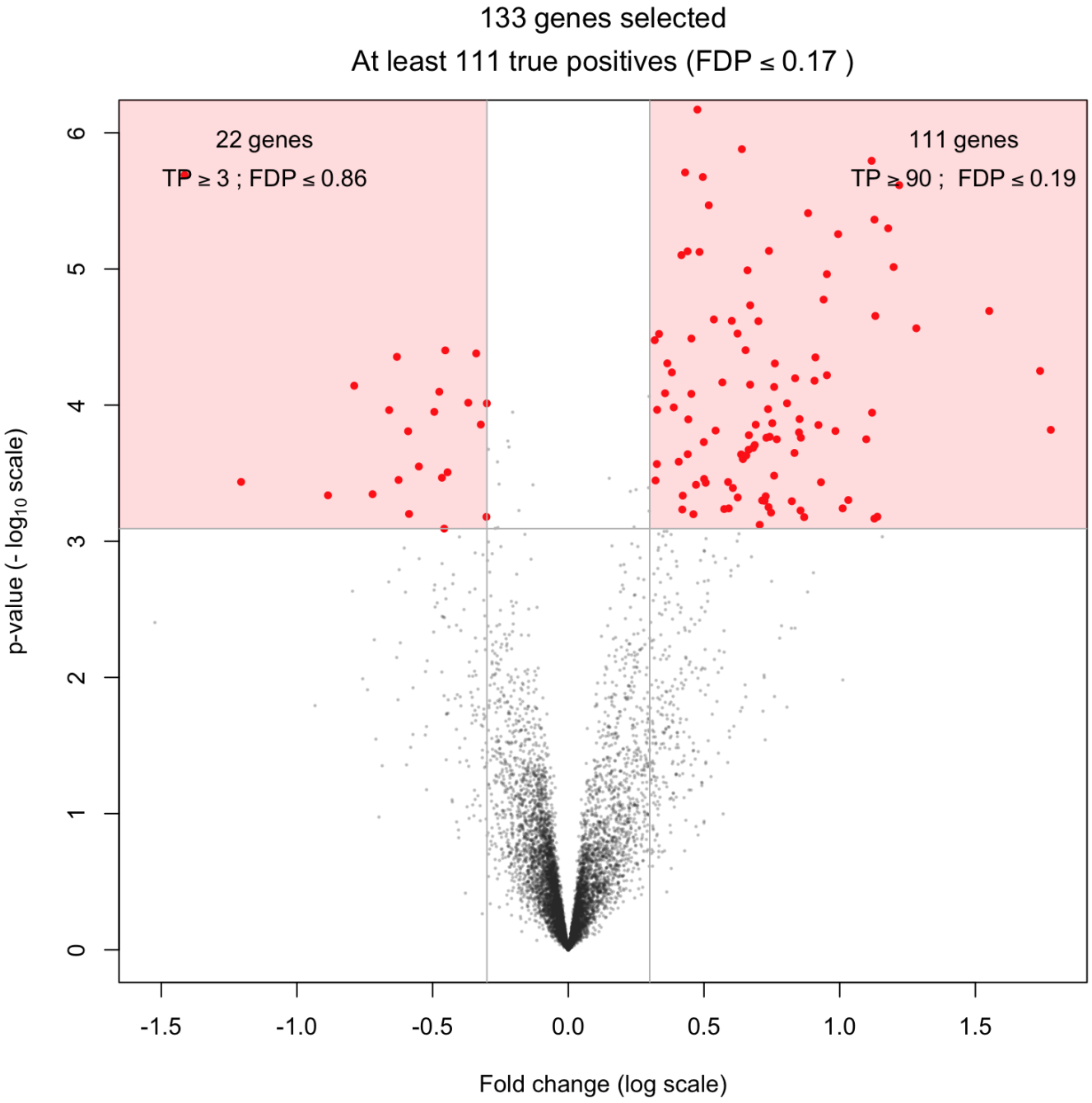
The substantial gain in power offered by the above-described calibration is illustrated as follows for the Simes reference family:

```
volcanoPlot(res, q = q, r = r, ylim = ylim)
```



and for the Beta reference family.

```
volcanoPlot(res_Beta, q = q, r = r, ylim = ylim)
```



The comparison between these bounds may be summarized by the following Table:

```
fc <- foldChanges(res)
S_pos <- which(fc >= r & adjp_BH <= q)
S_neg <- which(fc <= -r & adjp_BH <= q)
S_all <- union(S_pos, S_neg)

all_bounds <- function(S, resList) {
  c(length(S), sapply(resList, predict, S, "TP"))
}
tab <- rbind(all_bounds(S_pos, resList),
             all_bounds(S_neg, resList),
             all_bounds(S_all, resList))
plab <- paste("BH-adjusted p.value <", q)
lab <- c(paste(plab, "&", " fold change  > ", r),
         paste(plab, "&", " fold change  < ", -r),
         paste(plab, "&", "|fold change| > ", r))
tab <- cbind(lab, tab)
cap <- "Post hoc bounds on true positives in user-defined gene selections"
#knitr::kable(tab, caption = cap, format = "latex")
knitr::kable(tab, caption = cap)
```

Post hoc bounds on true positives in user-defined gene selections

lab		Simes	Linear	Beta (K=50)
BH-adjusted p.value < 0.05 & fold change > 0.3	111	60	81	90
BH-adjusted p.value < 0.05 & fold change < -0.3	22	1	6	3
BH-adjusted p.value < 0.05 & |fold change| > 0.3	133	74	102	111

Session information

```
sessionInfo()
```

```
## R version 4.0.3 (2020-10-10)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## Matrix products: default
## BLAS/LAPACK: /usr/local/miniconda/envs/computorbuild/lib/libopenblasp-r0.3.12.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] sansSouci_0.10.3 plotly_4.9.3     ggplot2_3.3.3
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.6         RColorBrewer_1.1-2 highr_0.9          pillar_1.6.0
##  [5] compiler_4.0.3     tools_4.0.3        digest_0.6.27      lattice_0.20-41
##  [9] viridisLite_0.4.0  jsonlite_1.7.2     evaluate_0.14      lifecycle_1.0.0
## [13] tibble_3.1.1       gtable_0.3.0       pkgconfig_2.0.3    rlang_0.4.10
## [17] Matrix_1.3-2       crosstalk_1.1.1    yaml_2.2.1         xfun_0.20
## [21] withr_2.4.2        stringr_1.4.0      dplyr_1.0.5        httr_1.4.2
## [25] knitr_1.31         generics_0.1.0     vctrs_0.3.7        htmlwidgets_1.5.3
## [29] grid_4.0.3         tidyselect_1.1.0   glue_1.4.2         data.table_1.14.0
## [33] R6_2.5.0           fansi_0.4.2        rmarkdown_2.7      farver_2.1.0
## [37] tidyr_1.1.3        purrr_0.3.4        magrittr_2.0.1     matrixStats_0.58.0
## [41] scales_1.1.1       ellipsis_0.3.1     htmltools_0.5.1.1  colorspace_2.0-0
## [45] labeling_0.4.2     utf8_1.2.1         stringi_1.5.3      lazyeval_0.2.2
## [49] munsell_0.5.0      crayon_1.4.1
```

References

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Benjamini, Yoav, and Daniel Yekutieli. 2001. “The Control of the False Discovery Rate in Multiple Testing Under Dependency.” *Annals of Statistics*, 1165–88.

Blanchard, Gilles, Pierre Neuvial, and Etienne Roquain. 2019. “On Agnostic Post Hoc Approaches to False Positive Control.” https://hal.archives-ouvertes.fr/hal-02320543 (https://hal.archives-ouvertes.fr/hal-02320543).

———. 2020. “Post Hoc Confidence Bounds on False Positives Using Reference Families.” *Annals of Statistics* 48 (3): 1281–1303. https://projecteuclid.org/euclid.aos/1594972818 (https://projecteuclid.org/euclid.aos/1594972818).

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber. 2010. “Independent filtering increases detection power for high-throughput experiments.” *PNAS*. https://doi.org/10.1073/pnas.0914005107/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.0914005107 (https://doi.org/10.1073/pnas.0914005107/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.0914005107).

Chiaretti, Sabina, Xiaochun Li, Robert Gentleman, Antonella Vitale, Kathy S Wang, Franco Mandelli, Robin Foa, and Jerome Ritz. 2005. “Gene Expression Profiles of b-Lineage Adult Acute Lymphocytic Leukemia Reveal Genetic Patterns That Identify Lineage Derivation and Distinct Mechanisms of Transformation.” *Clinical Cancer Research* 11 (20): 7209–19.

Cui, Xiangqin, and Gary A Churchill. 2003. “Statistical Tests for Differential Expression in cDNA Microarray Experiments.” *Genome Biol* 4 (4): 210.

Genovese, Christopher R, and Larry Wasserman. 2006. “Exceedance Control of the False Discovery Proportion.” *Journal of the American Statistical Association* 101 (476): 1408–17.
Goeman, Jelle J, and Aldo Solari. 2011. “Multiple Testing for Exploratory Research.” *Statistical Science* 26 (4): 584–97.
Hemerik, Jesse, Aldo Solari, and Jelle J Goeman. 2019. “Permutation-Based Simultaneous Confidence Bounds for the False Discovery Proportion.” *Biometrika* 106 (3): 635–49.
Sarkar, Sanat K. 1998. “Some Probability Inequalities for Ordered Mtp2 Random Variables: A Proof of the Simes Conjecture.” *Annals of Statistics*, 494–504.
Simes, R J. 1986. “An improved Bonferroni procedure for multiple tests of significance.” *Biometrika* 73 (3): 751–54.

---

1. A closely related approach has been proposed by Hemerik, Solari, and Goeman (2019)↩