Powerful and interpretable control of false discoveries in two-group differential expression studies

Nicolas Enjalbert-Courrech, Pierre Neuvial

OXFORD UNIVERSITY PRESS | Bioinformatics

Powerful and interpretable control of false discoveries in two-group differential expression studies

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2022-0352.R2
Category:	Original Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Enjalbert-Courrech, Nicolas; Institut de Mathématiques de Toulouse Neuvial, Pierre; Institut de Mathématiques de Toulouse
Keywords:	Algorithms, Gene expression, Microarray data analysis, Multiple testing, Next-generation sequencing, Statistics

"output" — 2022/9/7 — 8:27 — page 1 — #1

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Gene Expression

Subject Section

Powerful and interpretable control of false discoveries in two-group differential expression studies

Nicolas Enjalbert-Courrech [1,*] and Pierre Neuvial [1,*]

[1] Institut de Mathématiques de Toulouse;
UMR 5219, Université de Toulouse, CNRS
UPS, F-31062 Toulouse Cedex 9, France.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The standard approach for statistical inference in differential expression (DE) analyses is to control the False Discovery Rate (FDR). However, controlling the FDR does not in fact imply that the proportion of false discoveries is upper bounded. Moreover, no statistical guarantee can be given on subsets of genes selected by FDR thresholding. These known limitations are overcome by post hoc inference, which provides guarantees of the number of proportion of false discoveries among arbitrary gene selections. However, post hoc inference methods are not yet widely used for DE studies.

Results: In this paper, we demonstrate the relevance and illustrate the performance of adaptive interpolation-based post hoc methods for two-group DE studies. First, we formalize the use of permutation-based methods to obtain sharp confidence bounds that are adaptive to the dependence between genes. Then, we introduce a generic linear time algorithm for computing post hoc bounds, making these bounds applicable to large-scale two-group DE studies. The use of the resulting Adaptive Simes bound is illustrated on a RNA sequencing study. Comprehensive numerical experiments based on real microarray and RNA sequencing data demonstrate the statistical performance of the method.

Availability: A cross-platform open source implementation within the R package sanssouci is available at `https://sanssouci-org.github.io/sanssouci/`.

Contact: pierre.neuvial@math.univ-toulouse.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online. Rmarkdown vignettes for the differential analysis of microarray and RNAseq data are available from the package.

1 Introduction

Two-sample comparison problems are ubiguitous in genomics. The most classical example is the case of differential expression studies, where the goal is to pinpoint genes (or transcripts) whose average expression level differ significantly between two known populations, based on a sample of expression measurements from individuals from these populations. A classical strategy to identify *differentially expressed* ("DE") genes is to test, for each gene, the null hypothesis that its average expression is identical in both populations. DE genes are then defined as those passing some significance threshold, after accounting for the fact that many tests are performed simultaneously.

The state of the art approach to large-scale multiple testing is to control the False Discovery Rate (FDR). Introduced by Benjamini and Hochberg (1995), the FDR is the expected proportion of wrongly selected genes (false positives) among all selected genes. The most widely used method to control FDR is the Benjamini-Hochberg (BH) procedure, which has been shown to control FDR when the hypotheses corresponding to the non-differentially expressed genes are independent or satisfy a specific type of positive dependence called PRDS (Benjamini and Yekutieli, 2001). PRDS is widely accepted as a reasonable assumption in differential gene expression (DGE) studies and in genomic studies in general, see e.g.

"output" — 2022/9/7 — 8:27 — page 2 — #2

Goeman and Solari (2014). However, there exist two major caveats to the practical use and interpretation of FDR in genomics. Let us assume that we have obtained a list R of genes called DE by a FDR-controlling procedure applied at level q.

Practical use: FDR of gene subsets is not controlled. As noted by Goeman and Solari (2011), the statement $\mathrm{FDR}(R) \leq q$ applies to the list R only, and no further statistically valid inference can generally be made on other gene lists. However, a common practice is to manually curate this list by adding or subtracting genes, based on some external or priori knowledge (such as the knowledge of gene sets or pathways). A typical example is the case of volcano plots (Cui and Churchill, 2003), where one selects those genes passing both a significance threshold and a threshold on the fold change (difference of average gene expression on the log scale), see Figure 3. Ebrahimpoor and Goeman (2021) have recently shown in an extensive simulation study that this type of double filtering strategy yields inflated false discovery rates.

Interpretation: FDR control is not FDP control. The statement $\mathrm{FDR}(R) \leq q$ is often misinterpreted as "the proportion of false discoveries (FDP) in R is less than q". In fact, the FDP is a *random* quantity, and $\mathrm{FDR}(R) \leq q$ only implies that the *average FDP over hypothetical replications* of the same genomic experiment and p-value thresholding procedure, is upper bounded by q. This distinction would not matter much if gene expressions were statistically independent: indeed, as the number m of tests tend to infinity, the FDP concentrates to the corresponding FDR with a typical parametric convergence rate: $m^{-1/2}$ (Neuvial, 2008). However, as the dependence increases, the FDP distribution becomes strongly asymmetric and heavy-tailed, as reported by Korn *et al.* (2004), and further illustrated in Neuvial (2020, Fig. 2.1).

The notion of *post hoc inference* has been introduced by Goeman and Solari (2011) to address these limitations. Building on earlier works by Genovese and Wasserman (2006), Goeman and Solari (2011) have obtained confidence bounds for the FDP in *arbitrary, multiple and possibly data-driven subsets of hypotheses* using the theory of closed testing (Marcus *et al.*, 1976). In practice, *Simes post hoc bounds* are recommended in Goeman *et al.* (2019), as they are valid under the PRDS assumption and can be calculated efficiently. Simes post hoc bounds have recently been popularized in genomics by Ebrahimpoor and Goeman (2021), but also in neuroimaging studies by Rosenblatt *et al.* (2018), where this approach has been called "All-resolutions inference" (ARI).

Despite their very attractive theoretical properties, post hoc methods are not yet widely known and used for addressing multiple testing situations in genomics, where controlling FDR via the BH procedure remains standard. Two possible reasons for this situation are that contrary to the BH procedure for FDR control, the Simes post hoc bound for post hoc inference is (i) typically conservative in genomic applications, and (ii) its construction based on closed testing may be difficult to understand for practitioners.

An alternative construction of post hoc bounds that has been proposed in Blanchard *et al.* (2020) and further explored in Blanchard *et al.* (2021); Durand *et al.* (2020). This strategy can yield sharper bounds, by an adaptation to the statistical dependency between tests using permutations, and to the sparsity of the signal using a step-down principle. The main goal of the present paper is to popularize the use of the post hoc bounds introduced in Blanchard *et al.* (2020) in the context of two-group DE studies. Accordingly, the main contributions of this paper can be summarized as follows:

1. Providing a short and self-contained introduction to interpolation-based post hoc inference (Section 2) and to the use of the permutation-based calibration methods introduced in Blanchard *et al.* (2020) for DE studies (Section 3);
2. Proving that generic interpolation-based post hoc bounds can be computed in linear time (Section 4);
3. Applying the resulting "Adaptive Simes" method to a specific RNA seq DE study, to illustrate that it yields more interpretable results than those derived from FDR control, and sharper bounds than Simes post hoc bounds (Section 5);
4. Assessing the statistical performance of the method (control of the target risk, and statistical power) for DE studies via comprehensive numerical experiments based on real genomic data, both for microarray and sequencing data sets (Section 6).

Altogether, the results presented in this paper illustrate that subtantial gains in power can be achieved with respect to state-of-the-art post hoc bounds in the case of two-group DE studies, without sacrificing computational efficiency.

These developments are implemented in the R package `sanssouci` available from `https://sanssouci-org.github.io/sanssouci/`. The R code used for the numerical experiments is available from `https://github.com/sanssouci-org/IIDEA-method-paper`). References to the Supplementary data (text, figures, algorithms) are prefixed by "S-" throughout the paper.

2 Interpolation-based post hoc inference

We consider a DE study with m features. These features are called genes for simplicity, but the methods described below are also applicable more generally. For now, we only assume that a p-value is available to test the differential expression of each gene. The vector of p-values is denoted by $(p_1, \ldots, p_m)$. More specfic assumptions on how these p-values are obtained are given in Section 3.

2.1 Objective: post hoc bounds

For a given subset S of genes called DE, we denote by $\mathrm{FP}(S)$ the number of false positives in S, that is, the number of genes in S that are not truly DE. Our goal is to find a function $\overline{\mathrm{FP}}_\alpha$ such that with high probability, $\overline{\mathrm{FP}}_\alpha(S)$ is larger than the number of false positives in S:

$$\mathbb{P}\left(\forall S, \mathrm{FP}(S) \leq \overline{\mathrm{FP}}_\alpha(S)\right) \geq 1 - \alpha\,. \tag{1}$$

Following Goeman and Solari (2011), a function $\overline{\mathrm{FP}}_\alpha$ satisfying (1) will be called an α-level *post hoc upper bound on the number of false positives*. Post hoc inference can be equivalently formulated in terms of upper bounds on the FDP: $\overline{\mathrm{FDP}}_\alpha(S) = \overline{\mathrm{FP}}_\alpha(S)/\,|S|$, or in terms of lower bounds on the number or proportion of true positives: $\overline{\mathrm{TP}}_\alpha(S) = |S| - \overline{\mathrm{FP}}_\alpha(S)$, $\overline{\mathrm{TDP}}_\alpha(S) = \overline{\mathrm{TP}}_\alpha(S)/\,|S|$.

2.2 Strategy: JER control and interpolation

The bounds studied in this paper rely on a multiple testing risk called the Joint Error Rate (JER) and introduced in Blanchard *et al.* (2020). Given a non-decreasing family of thresholds $\mathbf{t} = (t_k)_{k=1\ldots K}$,

$$\mathrm{JER}(\mathbf{t}) = \mathbb{P}\left(\exists k \in \{1, \ldots, K\} \,:\, q_k < t_k\right)\,, \tag{2}$$

where for $k = 1, \ldots, K$, q_k denotes the k-th smallest p-value among the set of truly non-DE genes (that is, true null hypotheses). A key result is that any family $\mathbf{t}$ such that $\mathrm{JER}(\mathbf{t}) \leq \alpha$ yields an associated α-level post hoc bound, by the following interpolation argument.

Proposition 1 (Interpolation-based post hoc bound Blanchard *et al.* (2020), Proposition 2.3). *If $\mathbf{t} = (t_k)_{1\leq k\leq K}$ controls JER at level α,*

then (1) is satisfied for the bound

$$\overline{\mathrm{FP}}_\alpha(S) = \min_{1\le k\le K}\left\{\sum_{i\in S}\mathbf{1}\{p_i \ge t_k\} + k - 1\right\}. \qquad (3)$$

For completeness and in order to emphasize on the simplicitly of the argument, a proof of Proposition 1 is given in Appendix S-1.2.

2.3 Simes post hoc bounds

An important example is the Simes family $\mathbf{t}^{\mathrm{S}}(\alpha)$, defined by $t_k^S(\alpha) = \alpha k/m$ for all k. The Simes (1986) inequality ensures that $\mathrm{JER}(\mathbf{t}^{\mathrm{S}}(\alpha)) \le \alpha$ as soon as the p-value family is PRDS Sarkar *et al.* (2008). As noted by Blanchard *et al.* (2020), the post hoc bound then derived by Proposition 1 coincides with the Simes post hoc bound introduced in Goeman and Solari (2011).

Although the Simes inequality is sharp when the p-values are independent, it is increasingly conservative as the dependence between tests gets stronger (Blanchard *et al.*, 2020, Table 1). The associated JER control and post hoc bound naturally inherit this conservativeness (as illustrated in the numerical experiments of Sections 5 and 6). In order to address this conservativeness issue, it is useful to note that for $\lambda > 0$, the JER of the Simes family $\mathbf{t}^S(\lambda)$ can be written as

$$\mathrm{JER}(\mathbf{t}^S(\lambda)) = \mathbb{P}\left(\min_{1\le k\le m}\frac{mq_k}{k} < \lambda\right). \qquad (4)$$

In view of (4), a natural idea in order to obtain a tight JER control is to select the largest λ such that $\mathrm{JER}(\mathbf{t}^S(\lambda)) \le \alpha$. This idea is the basis of the calibration method described in Section 3.

3 JER calibration by permutation

The JER defined in (2) only depends on the joint p-value distribution of true null hypotheses. Although this distribution is unknown in practice, in two-group DGE studies, it can be approximated by permuting the group labels. Accordingly, the first step of our calibration method is to build a $B \times m$ matrix P of permutation p-values: P_{bi} is the p-value of the test of gene i associated to the b-th permutation of the group labels. This is illustrated in the first panel of Figure 1.

The next steps of the calibration are best explained in the particular case of the Simes family. Indeed, by (4), $\mathrm{JER}(\mathbf{t}^S(\lambda))$ is the value of the cumulative distribution function of $\psi = \min_{1\le k\le m} mq_k/k$ at λ. Accordingly, the calibration method proceeds by calculating B samples from the "pivotal statistic" ψ, and the output is the quantile of order α of these statistics.

The method as described in Figure 1 covers not only the case of the Simes family, but any family $\tau(\lambda) = (\tau_k(\lambda))_{k=1\dots K}$ where the τ_k are invertible functions. Following Blanchard *et al.* (2020, 2021), such a family is called a *template*. A more formal description of this calibration algorithm is given in Algorithm S-1. Note that for simplicity, we have described here a "single-step" version of the calibration algorithm. We have also implemented a "step-down" version: it is a slightly more powerful algorithm that is also adaptive to the unknown proportion of true null hypotheses (Blanchard *et al.*, 2020, Proposition 4.5).

Validity. Theorem 1 in Blanchard *et al.* (2021) ensures that this calibration method yields $\mathrm{JER}(\lambda) \le \alpha$, for tests whose p-value for a given gene depends on the data only via its own expression values. In particular, this is the case for two-sample Student tests or Wilcoxon rank sum tests, which can be used for microarray and RNA sequencing (RNA-seq) DE studies, respectively. However, note that this permutation-based strategy is formally only valid for two-group comparisons with no adjustment factors.

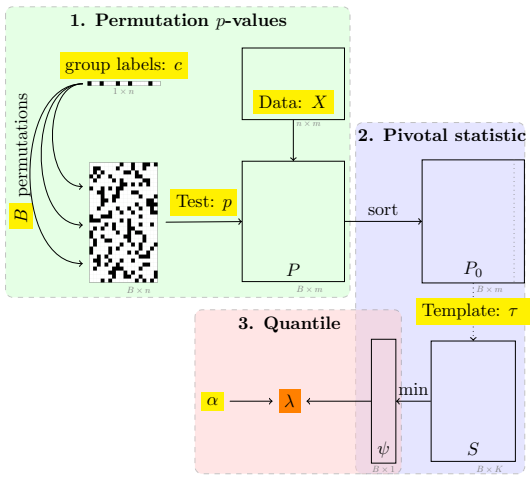

Fig. 1. Illustration of the three main steps of permutation-based JER calibration. The input data is in the form of a $n \times m$ gene expression matrix X and a binary vector c of n group labels, specifying which observations belong to each of the two populations to be compared. The parameters are the target JER level α and the number B of permutations, the p-value function to perform the test and the template τ.

More generally, the theory developed in Blanchard *et al.* (2020) is valid as soon as the joint distribution of the test statistics satisfies a randomization assumption (Romano and Wolf, 2005; Hemerik and Goeman, 2018). In the above case of two-sample tests, this is obtained via permutation of group labels. As noted in Blanchard *et al.* (2020), this assumption also holds for one-sample tests, where permutations at step 1 is replaced by sign-flips. It also holds when testing for the marginal independence between each gene's expression and a continuous outcome via a correlation test, as further explained and illustrated in Section S-6.

Let us recall that our methods relies on *group label permutations* in order to obtain statistically valid procedures that adapt to the dependency between genes observed in the data set at hand. While being theoretically valid regardless of the sample size n, such permutation techniques require a large enough sample sizeto appropriately learn gene dependencies. As a rule of thumb, we advocate the use of this method in studies with more than 5 samples per group.

Complexity. Assuming a linear time complexity $O(n)$ to perform the test of one single null hypothesis, the overall time complexity of the calibration method is $O(mB(n + \log(m)))$. Indeed, the most costly step is the calculation of P_0, which involves mB tests followed by B sorting operations on a vector of size m. The overall space complexity is $O(m(B + n))$.

Figure 1 also illustrates the modularity of Algorithm S-1, where the three main steps are highlighted in different colors. This modularity is important in practice. For example, it makes it possible to obtain the result for several values of α without re-computing the permutation matrix P_0. This modularity is also useful for the computational efficiency of the above-mentioned step-down version of the calibration algorithm.

4 Linear time interpolation-based post hoc bound

Post hoc bounds can be used for multiple gene selections S without compromising the corresponding error control. For post hoc inference to be applicable in practice $\overline{\mathrm{FP}}_\alpha(S)$ must be computed efficiently.

A naive implementation of the bound $\overline{\mathrm{FP}}_\alpha(S)$ defined in (3) would require s^2 operations (where s is the size of S) by performing a loop on both $k = 1, \dots, s$ and $i \in S$ in order to calculate $v_k(S) =$

$\sum_{i\in S}\mathbf{1}\{p_i \geq t_k\} + k - 1$ for all k. This induces a quadratic worst case time complexity $O(m^2)$, which is achieved when evaluating $\overline{\mathrm{FP}}_\alpha$ on the set of all genes. A quadratic time complexity for a single set is too slow for DE studies with $m \geq 10,000$. Moreover, a useful application of post hoc bounds is to build the false positive confidence curve associated to S, that is, all the bounds $\overline{\mathrm{FP}}_\alpha(S_i)$ for $i = 1 \dots s$, where S_i is the index set of the i smallest p-values in S. Using the above naive algorithm, this would require $O(s^3)$ operations, implying a cubic worst case time complexity $O(m^3)$ to build the false positive confidence curve associated to all hypotheses.

In contrast, Algorithm 1 computes $\overline{\mathrm{FP}}_\alpha(S)$ in linear time and space $O(s)$ for a given S. In fact, it even outputs the entire false positive confidence curve associated to S. For example, the largest set S such that $\overline{\mathrm{FDP}}_\alpha(S) \leq \gamma$ is then obtained in linear time and space for any user-defined γ. This complexity cannot be improved since the size of the output vector is s. The validity of algorithm 1 relies on the following formulation for $\overline{\mathrm{FP}}_\alpha(S_i)$.

Algorithm 1 Linear algorithm for interpolation-based post hoc bounds.

Require: $p = (p_{(1:S)}, \dots p_{(s:S)}), t = (t_1, t_2, \dots t_K)$
$\tau \leftarrow \mathtt{rep}(0, s)$
for $k \leftarrow 1$ to s **do**
  if $k \leq K$ **then**
    $\tau[k] \leftarrow t[k]$
  else
    $\tau[k] \leftarrow t[K]$
  end if
end for
$\kappa, r \leftarrow \mathtt{rep}(s, s)$
$k, i \leftarrow 1$
while $(k \leq s)$ & $(i \leq s)$ **do**
  if $(p[i] < t[k])$ **then**
    $\kappa[i] \leftarrow k - 1$ ▷ $\kappa[i] = |\{k/p[i] \geq t[k]\}|$
    $i \leftarrow i + 1$
  else
    $r[k] \leftarrow i - 1$ ▷ $r[k] = |\{i/p[i] < t[k]\}|$
    $k \leftarrow k + 1$
  end if
end while
$V, A, M \leftarrow \mathtt{rep}(0, s)$
for $k \leftarrow 1$ to s **do**
  $A[k] \leftarrow r[k] - (k - 1)$
  if $k > 1$ **then**
    $M[k] \leftarrow \max(M([k-1], A[k])$
  end if
end for
for $i \leftarrow 1$ to s **do**
  if $\kappa[i] > 1$ **then**
    $V[i] = \min(\kappa[i], i - M[\kappa[i]]))$
  end if
end for
return V

Proposition 2. *For $i \in \{1, \dots, s\}$, let S_i be the index set of the i-th smallest p-value in S, and $\kappa_i = \sum_{k=1\dots s}\mathbf{1}\{p_i \geq t_{k\wedge K}\}$. Then*

$$\overline{\mathrm{FP}}_\alpha(S_i) = \min\left(\kappa_i, \min_{1\leq k\leq \kappa_i} v_k(S) - (s - i)\right). \quad (5)$$

Proposition 2 is proved in Appendix S-1.3. The fact that $\overline{\mathrm{FP}}_\alpha(S_i)$ depends on i only via κ_i but not S_i in (5) is crucial for obtaining a linear time complexity. The properties of Algorithm 1 can be summarized as follows:

Corollary 1 (Validity and complexity of Algorithm 1). *Algorithm 1 returns the vector $(\overline{\mathrm{FP}}_\alpha(S_i))_{1\leq i\leq s}$ in $O(s)$ time and space complexity.*

Proof of Corollary 1. **Valididy.** The for loop at lines 1- 8 stores the thresholds $(t_{k\wedge K})$ for $k \in \{1, \dots, s\}$. The while loop at lines 11-19 outputs both $(\kappa_i)_{i\in S}$ and $(r_k)_{1\leq k\leq K}$, where $r_k = |\sum_{i\in S}\mathbf{1}\{p_i < t_k\}|$. Noting that $r_k = s - v_k(S) + (k-1)$, the **for** loop at lines 21-26 outputs $M_k = \max_{k'\leq k} s - v_{k'}(S)$, that is, $M_k = s - \min_{k'\leq k} v_{k'}(S)$. Thus, the **for** loop at lines 27-31 outputs $V_i = \overline{\mathrm{FP}}_\alpha(S_i)$ by Proposition 2.

Complexity. All the vectors stored within the algorithm are of size s, so the space complexity of Algorithm 1 is $O(s)$. For the time complexity, the $(\kappa_i)_i$ and $(r_k)_k$ are calculated within a single while loop of size s, in which exactly one of i or k is incremented at each step. The rest of the algorithm consists of two for loops of size s consisting of $O(1)$ operations.

5 Urothelial Bladder Carcinoma data set

In this section, we focus on an Urothelial Bladder Carcinoma (BLCA) RNA sequencing data set from the Cancer Genome Atlas Research Network *et al.* (2014, TCGA). This preprocessed data set is available from the R/Bioconductor package `curatedTCGAData`. Internally, this package itself relies on the R/Bioconductor package `RTCGAToolbox` to download TCGA data that have already been preprocessed by TCGA pipelines. For convenience, this data set has also been made available in the R package `sanssouci.data`. This data set consists of gene expression measurements for $n = 270$ patients, classified into two subgoups: stage II ($n_0 = 130$) and stage III ($n_1 = 140$). Bladder cancer stages range from 0 to IV, quantifying how much the cancer has spread. We have filtered out unexpressed genes, here defined as those for which raw expression counts were lower than 5 in at least 75% of the patients. This results in $m = 12,534$ genes. To identify DE genes between the stage II and stage III populations, we test for each gene the null hypothesis that the gene expression distribution is identical in the two populations. The calibration method described in Section 3 is performed using a Wilcoxon (1945) rank sum test (also known as Mann and Whitney (1947) test) with the Simes template, with $B = 1,000$ permutations and target risk (JER) set to $\alpha = 10\%$. The resulting method is called the **Adaptive Simes** method.

5.1 Confidence curves

In the absence of prior information on genes, a natural idea is to rank them by decreasing statistical significance. Post hoc methods provide confidence curves on the number (or proportion) of true positives (truly DE genes) among the most significant genes. Such curves are displayed in Figure 2 for the BLCA data set. The black lines in Figure 2 are $1-\alpha = 90\%$ confidence curves obtained by the Adaptive Simes method. Upper bounds on FDP and lower bounds on FP are displayed in the left and right panels, respectively. For reference, the corresponding curves obtained by ARI are displayed in gray; they are almost identical to the ones obtained from the original bound of Goeman and Solari (2011) (Equation (3)), see Section S-2.1.

Post hoc guarantees. Post hoc inference makes it possible to define DE genes as the largest set of genes for which the FDP bound is less than a user-given value q. The arbitrary choice $q = 0.1$ is illustrated in Figure 2, corresponding to the horizontal line in the left panel. The black lines in Figure 2 correspond to the set S of 1064 genes for which the adaptive Simes method ensures that $\mathrm{FDP}(S) \leq q$. This corresponds to at least $\overline{\mathrm{TP}}_\alpha(S) = 958$ true positives (since $1 - 958/1064 = 0.1$), as illustrated in the right panel.

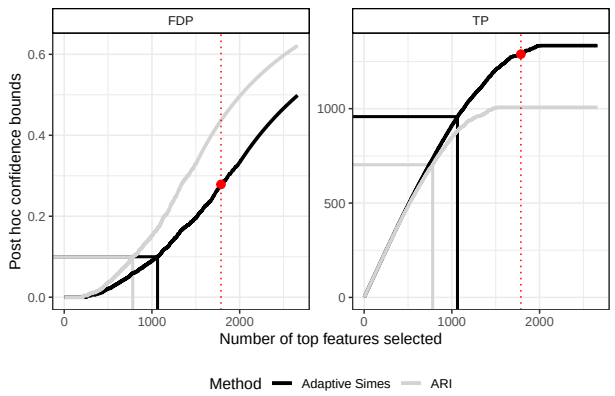

Fig. 2. 90% confidence curves on "top k" lists for the Urothelial bladder carcinoma data set. Left: upper bound on the False Discovery Proportion (FDP); right: lower bound on the number of true positives (TP). The adaptive Simes bound (black curves) outperforms ARI (gray curves). For reference, the set of $1,787$ genes called DE by the BH(0.05) procedure is represented by a dot.

Table 1. Post hoc bounds on BLCA data set for ARI and the proposed permutation method, for gene selections S illustrated in Figures 2 (target FDP$= 0.1$) and 3 (genes filtered by p-value).

	Confidence curve (Fig. 2)		Volcano plot (Fig. 3)	
	ARI	Adaptive Simes	ARI	Adaptive Simes
Number of genes in S	781	1064	569	569
$\overline{\mathrm{TP}}_\alpha(S)$	703	958	456	492
$\overline{\mathrm{FDP}}_\alpha(S)$	0.1	0.1	0.199	0.135

Adaptation to dependence. The above example also illustrates the increase in power obtained by Adaptive Simes thanks to the calibration described in Section 3. Indeed, for an identical statistical guaranty ($FDP \leq 0.1$), ARI yields a substantially smaller subset of 703 DE genes. More generally, the comparison between the black and gray curves in Figure 2 illustrates the gain in power obtained by using permutations methods to adapt to the dependence between genes.

Comparison to FDR control. For this data set, the BH procedure calls a set S of $1,787$ genes DE for a target FDR level of 0.05. As stated in Section 1, *the BH procedure does not provide guarantees on the FDP of these genes*, but only on their FDR, that is, the average FDP over hypothetical replications of the same genomic experiment and p-value thresholding procedure. (Note that this remark is not specific to the BH procedure: the same would be true for any FDR controlling procedure.) In contrast, the Adaptive Simes bound guarantees (with 90% confidence) that the number of true positives in S^{BH} is at least $1,289$, or, equivalently, that the corresponding FDP is less than 0.279.

5.2 Volcano plots

Volcano plots are a commonly used graphical representation of the results of a differential expression analysis (Cui and Churchill, 2003), illustrated in Figure 3. Each gene is represented in two dimensions by estimates of its effect size (or "fold-change", x axis) and significance (y axis). The fold change of a gene is generally defined as the difference between the average or median (log-scaled) gene expressions of the two compared groups. Its significance is quantified by $-\log_{10}(p)$-values for the test of its differential expression, where the "$-\log_{10}$", transformation ensures that large values of y correspond to genes which are likely to be differentially expressed.

As noted by Ebrahimpoor and Goeman (2021), post hoc inference makes it possible to select genes of interest based on both fold change and significance, without compromising the validity of the corresponding bounds. Moreover, even if Wilcoxon tests have been performed for the *calibration of the post hoc bounds*, our proposed post hoc bounds are still valid when relying on other statistics for the *selection genes of interest*. Figure 3 illustrates this idea by making a volcano plot based on the p-values and log-fold changes obtained from the limma-voom method of Law *et al.* (2014), which is implemented in the 'limma' package of Ritchie *et al.* (2015). In this example, the function $\overline{\mathrm{FP}}_\alpha$ defined in (3) depends on the Wilcoxon tests via the p-values $(p_i)_i$ and the thresholds (t_k) obtained at the calibration step, but it is statistically valid for arbitrary gene selections S. By construction, for a given selection size $|S|$, the tightest bound $\overline{\mathrm{FP}}_\alpha(S)$ corresponds to the set of the $|S|$ smallest Wilcoxon p-values. More generally, smaller bounds $\overline{\mathrm{FP}}_\alpha(S)$ will be obtained for selections S that consist of small Wilcoxon p-values. A quantitative comparison between the Wilcoxon and limma-voom p-value is provided in Figure S-2. It illustrates the coherence of the two methods for identifying DE genes in the settings considered.

An example selection of 569 genes is highlighted in in Figure 3. It corresponds to genes whose p-value is less than 10^{-3} and fold change larger than 0.5 in absolute value. The Adaptive Simes method ensures that with probability larger than 90%, the proportion of false discoveries (FDP) is less than 0.14. It also ensures that the FDP among the subset of 493 genes with positive fold change is less than 0.14, and that the FDP among the subset 76 of genes with negative fold change is less than 0.63. As already noted, the proposed bounds can be computed for multiple, arbitrary gene subsets (obtained e.g. by changing the p-value and fold change thresholds in Figure 3) without comprising their validity. Here again, the Adaptive Simes method yields tighter bounds than ARI, as illustrated in Table 1.

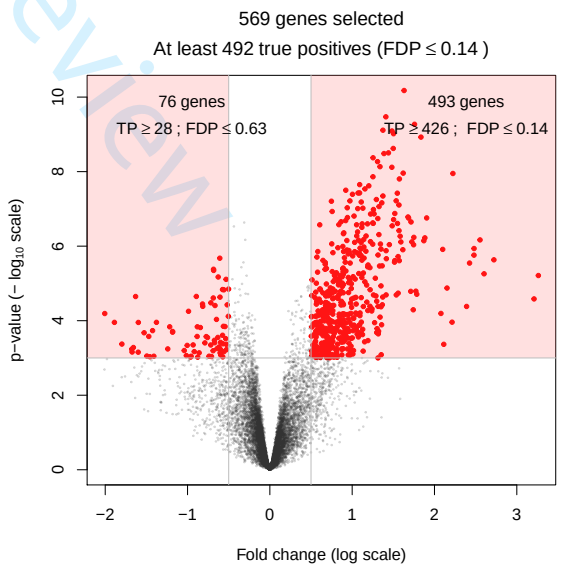

Fig. 3. Volcano plot for the urothelial bladder carcinoma data set. Each dot corresponds to a gene, represented by its fold change (x axis) and p-value (y axis) on the log scale. Fold changes and p-values were obtained by the limma-voom method Ritchie et al. (2015). The 569 genes with p-value less than 10^{-3} and fold change larger than 0.5 are highlighted. The Adaptive Simes method ensures that at least 492 of these genes are true positives.

"output" — 2022/9/7 — 8:27 — page 6 — #6

5.3 Influence of the number of permutations

The adaptive Simes method relies on random permutation of class labels. As such, running this method several times could lead to different results. Larger values of the number B of permutations are expected to give more stable results. However, this comes at a higher computational price, since the theoretical time complexity of the calibration step is linear in B, see Section 3. We have quantified the variability of the post hoc bounds as a function of B, by performing the calibration $1,000$ times for each $B \in \{100, 200, 500, 1000, 2000, 5000\}$. The results of these experiments are reported in Section S-2.3. For $B = 1,000$ (the default value in the `sanssouci` package), for the $1,787$ genes selected by the BH procedure at level $q = 0.05$, the FDP bound is between 0.23 and 0.37 for 99% of the $1,000$ replications of the calibration procedure. Choosing a larger value for B increases the precision of the post hoc bounds. In particular, with $B = 5,000$ the precision increases by a factor 2 (the FDP bound is between 0.26 and 0.33), at the price of an increased computation time. On a standard laptop, calibration takes $< 20s$ for $B = 1,000$ and $< 50s$ for $B = 5,000$ without parallelization. More generally, we observed that the empirical complexity of the calibration is slightly less than linear (see Figure S-4).

6 Statistical performance for DE studies

6.1 Existing post hoc inference methods

The first post hoc inference methods introduced in were not adaptive to the dependence between tests, since they were obtained from probabilistic inequalities:

- The **Simes** bound was first proposed in Goeman and Solari (2011) together with a quadratic algorithm ($(O(m^2))$). It has been implemented in the R package `cherry`.
- A slightly sharper version of the Simes bound has been introduced in Goeman *et al.* (2019), together with an algorithm of linearithmic complexity. This method is known as **ARI** for "All resolution inference" and implemented in the R package `hommel`.

The idea of using randomization to obtain sharp risk control is not new in the multiple testing literature. In particular, resampling or permutations have been used to control the Family-Wise Error Rate (FWER, Ge *et al.* (2003); Westfall and Young (1993)) and the k-FWER (Romano and Wolf, 2007). For post hoc inference:

- The **Adaptive Simes** method described in this paper exploits sign-flipping and permutation-based approaches introduced in Blanchard *et al.* (2020, 2021) in order to build post hoc bounds. It is implemented since 2017 in the R package `sanssouci`.
- A closely related approach called **pARI** has recently been proposed by Andreella *et al.* (2020) for the analysis of neuroimaging data. It is implemented in the R package `pARI`.

Both pARI and Adaptive Simes rely on the calibration method described in Section 3, combined with the interpolation bound (3). An important difference between Adaptive Simes (R package `sanssouci`) and pARI is that `sanssouci` implements the linearithmic time complexity algorithm described in Section 4. In contrast, the algorithm used in `pARI` to calculate the post hoc bound after calibration is the "naive" interpolation algorithm described in the beginning of Section 4, which has a quadratic complexity for a single set.

While `pARI` initially only implemented a single-step version, it has been updated after the initial submission of ths manuscript, so that both `sanssouci` and `pARI` now implement a step-down principle in order to adapt to the unknown quantity of signal (or, equivalently, to the proportion π_0 of true null hypotheses). These two step-down methods have the same goal but they are based on different principles. The experiments reported below show that our proposed step-down typically provides only marginal performance improvements on real data, due to the sparsity of the signals. Similar observations have been made for the step-down version of `pARI`, see Andreella *et al.* (2020), so we have not included this method in our comparisons.

The main features of existing post hoc bounds are summarized in Table 2.

		Adaptivity to:		Time
Method	R Package	dependence	π_0	complexity
Simes	cherry	NO	NO	quadratic
ARI	ARI	NO	YES	linear
permutation ARI	pARI	YES	YES	quadratic
Adaptive Simes	sanssouci	YES	YES	linear

Table 2. Main features of existing post hoc inference methods and software.

6.2 Evaluation framework

The mathematical validity of the post hoc bounds considered in this paper has been proved in Blanchard *et al.* (2020), where their numerical performance has also been illustrated by experiments on synthetic data. The goal of this section is to complement these results by numerical experiments based on gene expression data, which are more realistic for the purpose of DE studies.

Data set generation. In the absence of a gold standard data set where one would know which genes are truly DE or not, we created such data sets as follows. Starting from a $n \times m$ gene expression data set X, where each row corresponds to a gene and each column to an experiment or statistical observation, we have

1. partitioned the observations into two groups of size n_0 and n_1, such that $n_0 + n_1 = n$;
2. partitioned the genes into m_0 null genes and m_0 non-null genes, with $m_0 + m_1 = m$
3. modified the expression of the non-null genes in group 1 by shifting or scaling the corresponding submatrix of X of size $n_1 \times m_1$.

This process results in a perturbed gene expression data set Y where the null and non-null genes are known. Following Blanchard *et al.* (2020), we have quantified, for a set of such experiments, estimates of the risk (JER) and of the power of each method considered, for each value of the target risk α. The JER results are presented in Section 6.3. The power results, which are highly consistent with the JER results, are postponed to Supplementary Materials.

The empirical risk of a given method is estimated by the proportion of experiments for which the corresponding confidence curve on the false positives is now always below the actual number of false positives. This quantity is the empirical counterpart of the JER defined in (2), and can be compared to the target risk α: JER is empirically controlled if the empirical JER is lower than α, and the closer it is to α, the tighter JER control.

The parameters of such a numerical experiment are the proportion $\pi_0 = m_0/m$ of null genes, and a measure of distance (or signal to noise ratio) between null and non-null genes. Section 6.3 reports the numerical results obtained for RNA sequencing data. We have also performed the same type of experiments with microarray data. The results are similar, and they are reported in Section S-4.

A core feature of our proposed method is to use *group label permutations* in order to obtain statistically valid procedures that adapt to

the dependency between genes observed in the data set at hand. However, the number of distinct permutations is limited for lower sample sizes: for example, 252 distinct permutations are available for a comparison between two groups of size 5. We have evaluated the impact of the sample size on the performance of the methods in Section S-5. This can be done by down-sampling the BLCA data set and retaining only a smaller number of observations. Our experiments demonstrate that even for sample sizes less than 10, the Adaptive Simes method yields sharper bounds than its competitors. Although the Adaptive Simes method is formally valid regardless the sample size, we recommend using it in studies with more than 5 samples per group.

6.3 Results for bulk RNA sequencing data

Our starting point is the data set used in section 5. For this experiment we have selected only stage III samples, and performed the same filtering as in Section 5 on these samples only. We obtained a "null" data set (with no signal), consisting of 130 patients and $m = 12,418$ genes, after applying the same process as described in Section 5 for filtering out unexpressed genes. The parameters of the experiments are set as follows. The proportion of null genes is set to $\pi_0 \in \{0.8, 0.95, 1\}$. We have considered an multiplicative signal for differential expression: for each gene g among the m_1 non-null genes, the original expression values of g are multiplied by a constant s_g for n_1 of the n observations, where s_g is drawn uniformly between 1 and a signal to noise (SNR) parameter. The SNR value is set to 1 (no signal), 2 or 3 (weak to strong signal). We have used a two-sided Wilcoxon rank sum test for comparing the two groups.

The results are summarized by Figure 4, where the average empirical risk (JER) achieved across 1000 experiments is plotted (together with 95% confidence curves) against the target risk α for the methods described in Section 6.1. In particular, the single-step version of pARI is represented by the "Adaptive Simes (single step)" method.

Each panel corresponds to a combination of the parameters $\pi_0 \in \{0.8, 0.95, 1\}$ (in columns) and SNR $\in \{1, 2, 3\}$ (in rows). The JER is controlled for all methods and all parameter combinations, since all curves are below the diagonal. The risk for the Adaptive Simes methods is substantially closer to the target risk than for the parametric Simes methods (Simes and ARI). This illustrates the systematic gain in tightness provided by the calibration method described in Section 3. We also note that the gain obtained from the adaptation to π_0 is very small, except for situations with both high signal (SNR$= 3$) and low sparsity ($\pi_0 = 0.8$). This gain is negligible for $\alpha \leq 0.2$ in all situations. Indeed, the Simes and ARI methods are essentially indistinguishable from each other, and the same holds for the single-step and step-down Adaptive Simes methods. These results are also confirmed by those of the power assessment, which are given in Supplementary Materials.

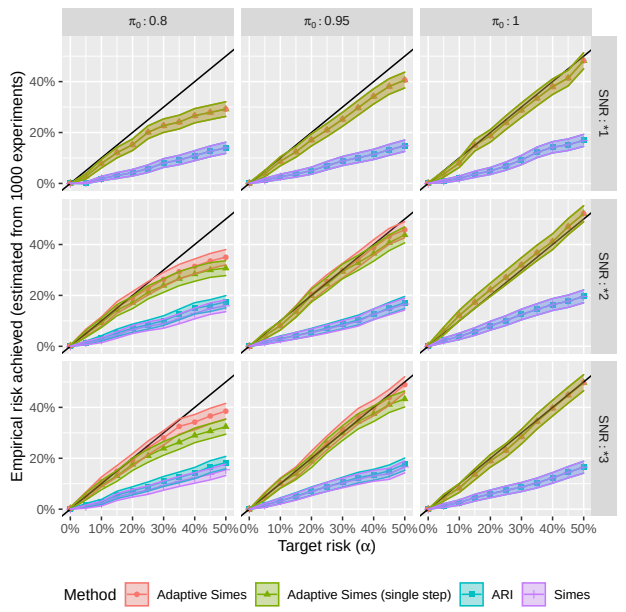

Fig. 4. Validity and compared tightness of the post hoc bounds on RNA-seq based numerical experiments. The average empirical JER achieved across 1000 experiments is plotted (together with 95% confidence curves) against the target risk α for all considered methods. Each panel corresponds to a combination of the parameters π_0 and SNR. The adaptive methods yield tighter risk control than parametric ones.

7 Discussion

This paper advocates for the use of post hoc inference in two-group DE studies, which provide more interpretable statistical guarantees than classical inference based on the False Discovery Rate. The methods proposed in this paper make it possible to obtain post hoc bounds that are both fast to compute, and powerful (in the sense of the proportion of true signal recovered). The resulting improvement over the state-of-the-art is illustrated by realistic numerical experiments based on RNA-seq and microarray data. These methods are implemented in the open-source R package `sanssouci`. The code used for the numerical experiments of this paper and to generate the figures is also provided. The default number of permutations used for the calibration of the method is set to $B = 1,000$ in the `sanssouci` package. Choosing a higher number of permutations will lead to an increased precision of the bounds, at the price of a higher computation time. Moreover, even though our proposed methods are theoretically valid regardless of the sample size, we recommend using them in studies with at least 5 samples per condition, so that the number of distinct group-label permutations is large enough to provide adaptation to the dependency between genes observed in the data set at hand. The methods proposed in this paper and their implementation in the R package `sanssouci` are generic, in the sense that they can be used with reference families (or *templates*) of arbitrary shape. The most natural choice is the Simes family (which corresponds to a linear template), as it is closely related both to FDR control and to the first post hoc bounds introduced by Goeman and Solari (2011). The resulting method, which is called the Adaptive Simes method, is used in the numerical experiments reported in this paper. An interesting perspective of this work is to compare the performance of other templates. Our experience in DE studies indicates that improving on the Simes family by changing the templat is challenging; similar conclusions have been reported in Andreella *et al.* (2020) for the analysis of fMRI data. Recent works in this field have shown the superiority of a fully non-parametric approach, whereby the entire family of templates (instead of a single parameter λ as in the present work) is learned from external data (Blain *et al.*, 2022). Applying this method to genomic data is another exciting perspective for the present work.

While this paper focuses on DE studies, these methods and our implementation are applicable to any practical situation involving multiple two-sample test, or one-sample tests, or tests of association with a continuous outcome. Such situations are frequent in genomics (differential expression, differential splicing, differential methylation) but also in neuroimaging, which is another field where post hoc inference methods have been introduced (Rosenblatt *et al.*, 2018). However, for studies have more complex designs such as multi-sample comparisons or studies including covariates, the calibration-based approach proposed here cannot be applied directly. This is a limitation of the Adaptive Simes method when compared to the state-of-the-art ARI method, which is applicable in any multiple testing framework where the Simes inequality holds. This

limitation is the current price to pay in order to obtain the substantial power gains that are illustrated numerically in this paper. Extensions of the present work to the problem of testing parameters of a general linear model is another interesting perspective that requires additional statistical developments.

Funding

This work has been supported by ANR-16-CE40-0019 (SansSouci), by Fondation Catalyses at Université Paul Sabatier, and by the Mission for Transversal and Interdisciplinary Initiatives (MITI) at CNRS through the DDisc project.

Acknowledgements

The authors warmly thank Alexandre Blain, Gilles Blanchard, Samuel Davenport and Etienne Roquain for insightful discussions about this work.

References

Andreella, A., Hemerik, J., Weeda, W., Finos, L., and Goeman, J. (2020). Permutation-based true discovery proportions for fMRI cluster analysis. *arXiv preprint arXiv:2012.00368*.

Benjamini, Y. and Hochberg, Y. (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, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, **29**(4), 1165–1188.

Blain, A., Thirion, B., and Neuvial, P. (2022). Notip: Non-parametric True Discovery Proportion control for brain imaging. *Neuroimage*, **260**.

Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *Annals of Statistics*, **48**(3), 1281–1303.

Blanchard, G., Neuvial, P., and Roquain, E. (2021). On agnostic post hoc approaches to false positive control. In X. Cui, T. Dickhaus, Y. Ding, and J. C. Hsu, editors, *Handbook of Multiple Comparisons*, Handbooks of Modern Statistical Methods. Chapman & Hall/CRC.

Cancer Genome Atlas Research Network *et al.* (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, **507**(7492), 315.

Cui, X. and Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, **4**(4), 210.

Durand, G., Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc false positive control for structured hypotheses. *Scandinavian Journal of Statistics*.

Ebrahimpoor, M. and Goeman, J. (2021). Inflated false discovery rate due to volcano plots: Problem and solutions. *Briefings in Bioinformatics*, **22**.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, **12**(1), 1–77.

Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, **101**(476), 1408–1417.

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, **26**(4), 584–597.

Goeman, J. J. and Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in medicine*, **33**(11), 1946–1978.

Goeman, J. J., Meijer, R. J., Krebs, T. J., and Solari, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, **106**(4), 841–856.

Hemerik, J. and Goeman, J. J. (2018). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**(1), 137–155.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, **124**(2), 379–398.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, **15**(2), 1–17.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**(3), 655–660.

Neuvial, P. (2008). Asymptotic properties of false discovery rate controlling procedures under independence. *Electron. J. Statist.*, **2**, 1065–1110. With corrigendum in EJS 2009(3):1083.

Neuvial, P. (2020). *Contributions to statistical inference from genomic data*. Habilitation thesis, Université Toulouse III. Avaiable from `https://tel.archives-ouvertes.fr/tel-02969229`.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7), e47.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, **100**(469), 94–108.

Romano, J. P. and Wolf, M. (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics*, **35**(4), 1378–1408.

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., and Goeman, J. J. (2018). All-resolutions inference for brain imaging. *Neuroimage*, **181**, 786–796.

Sarkar, S. K. *et al.* (2008). On the simes inequality and its generalization. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, pages 231–242. Institute of Mathematical Statistics.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**(3), 751–754.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, **1**(6), 80–83.