On agnostic post hoc approaches to false positive control

Gilles Blanchard, Pierre Neuvial, Etienne Roquain

On agnostic post hoc approaches to false positive control

Gilles Blanchard

Universität Potsdam, Institut für Mathematik
Karl-Liebknecht-Straße 24-25 14476 Potsdam, Germany

Laboratoire de Mathématiques d'Orsay,
Université Paris-Sud, CNRS,
Université Paris-Saclay, 91405 Orsay Cedex, France
e-mail: gilles.blanchard@math.u-psud.fr

Pierre Neuvial

Institut de Mathématiques de Toulouse;
UMR 5219, Université de Toulouse, CNRS
UPS IMT, F-31062 Toulouse Cedex 9, France
e-mail: pierre.neuvial@math.univ-toulouse.fr

Etienne Roquain

Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation, LPSM,
4, Place Jussieu, 75252 Paris cedex 05, France
e-mail: etienne.roquain@upmc.fr

Abstract: This document is a book chapter which gives a partial survey on post hoc approaches to false positive control.

Contents

Classical approaches to multiple testing grant control over the amount of false positives for a specific method prescribing the set of rejected hypotheses. On the other hand, in practice many users tend to deviate from a strictly prescribed multiple testing method and follow ad-hoc rejection rules, tune some parameters by hand, compare several methods and pick from their results the one that suits them best, etc. This will invalidate standard statistical guarantees because of the selection effect. To compensate for any form of such "data snooping", an approach which has garnered significant interest recently is to derive "user-agnostic", or post hoc, bounds on the false positives valid uniformly over all possible rejection sets; this allows arbitrary data snooping from the user. In this chapter, we start from a common approach to post hoc bounds taking into account the p-value level sets for any candidate rejection set, and explain how to calibrate the bound under different assumption concerning the distribution of p-values. We then build towards a general approach to this problem using a family of candidate rejection subsets (call this a reference family) together with associated bounds on the number of false positives they contain, the latter holding uniformly over the family. It is then possible to interpolate from this reference family to find a bound valid for any candidate rejection subset. This general program encompasses in particular the p-value level sets considered initially in the chapter; we illustrate its interest in a different context where the reference subsets are fixed and spatially structured. These methods are then applied to a genomic example (differential gene expression), and a neuromaging example (functional Magnetic Resonance Imaging). Code vignettes to reproduce these examples using the R [21] package `sansSouci` [19] are provided as Supplementary Materials[1]. In this chapter, all references are gathered in Section 11.

1. A motivating example

Differential gene expression studies in cancerology aim at identifying genes whose activity differs significantly between two (or more) cancer populations, based on a sample of measurements from individuals from these populations. The activity of a gene is usually quantified by its level of expression in the cell. We consider here a microarray data set[2] consisting of expression measurements for more than $12,000$ genes for biological samples from $n = 79$ individuals with B-cell acute lymphoblastic leukemia (ALL). A subset of cardinal $n_1 = 37$ of these individuals harbor a specific mutation called BCR/ABL, while the remaining $n_2 = 42$ do not. One of the goals of this study is to identify, from

[1] The corresponding source files are available from the package web site: `https://pneuvial.github.io/sanssouci/`.

[2] Taken from Chiaretti *et. al.*, *Clinical cancer research*, 11(20):7209–7219, 2005.

this sample, those genes for which there is a difference in the mean expression level between the mutated and non-mutated populations. This question can be addressed, after relevant data preprocessing, by performing a statistical test of equality in means for each gene. A classical approach is then to derive a list of "differentially expressed" genes (DEG) as those passing a FDR correction by the Benjamini-Hochberg (BH) procedure at a user-defined level. This is illustrated by Figure 1 for the Leukemia data set, where 163 genes are called "differentially expressed" at FDR level $q = 0.05$.

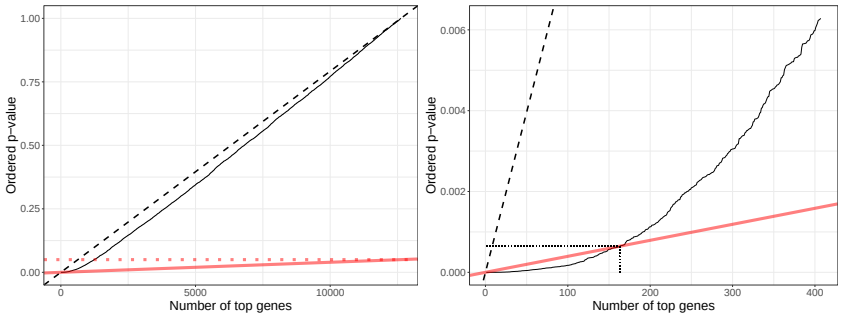

FIG 1. *Left: sorted p-values for the Leukemia data set (thin black solid line). Right: zoom on the smallest 400 p-values. Dashed line: $y = x/m$; bold pink solid line: $y = qx/m$ (for $q = 0.05$), whose intersection with the p-value curve determines the rejections of the BH procedure at level q. Pink dotted line: $y = q$. Here, 163 genes are declared as differentially expressed.*

2. Setting and basic assumptions

Let us observe a random variable X with distribution P belonging to some model $\mathcal{P}$. Consider m null hypotheses $H_{0,i} \subset \mathcal{P}$, $i \in \mathbb{N}_m = \{1, \dots, m\}$, for P. We denote $\mathcal{H}_0(P) = \{i \in \mathbb{N}_m : P \text{ satisfies } H_{0,i}\}$ the set of true null hypotheses and $\mathcal{H}_1(P) = \mathbb{N}_m \backslash \mathcal{H}_0(P)$ its complement. We assume that a p-value $p_i(X)$ is available for each null hypothesis $H_{0,i}$, for each $i \in \mathbb{N}_m$.

We introduce the following assumptions on the p-values and the distribution P, that will be useful in the sequel:

$$\forall i \in \mathcal{H}_0(P),\ \ \forall t \in [0,1],\ \ \mathbb{P}\left(p_i(X) \leq t\right) \leq t; \tag{Superunif}$$

$$\{p_i(X)\}_{i \in \mathcal{H}_0(P)} \text{ is a family of independent variables, which is independent of } \{p_i(X)\}_{i \in \mathcal{H}_1(P)}. \tag{Indep}$$

3. From confidence bounds ...

Consider some fixed deterministic $S \subset \mathbb{N}_m$. A $(1-\alpha)$-confidence bound $V = V(X)$ for $|S \cap \mathcal{H}_0(P)|$, the number of false positives in S, is such that

$$\forall P \in \mathcal{P}, \qquad \mathbb{P}_{X \sim P}\Big(|S \cap \mathcal{H}_0(P)| \leq V\Big) \geq 1 - \alpha.$$

A first example is given by the k_0-Bonferroni bound $V(X) = \sum_{i\in S} \mathbf{1}\{p_i(X) \geq \alpha k_0/|S|\}+ k_0-1$, for some fixed $k_0 \in \mathbb{N}_m$ such that $k_0 \leq |S|$ (otherwise the bound is trivial). The coverage probability is ensured under (Superunif) by the Markov inequality:

$$\begin{aligned}
&\mathbb{P}(|S\cap\mathcal{H}_0(P)| \geq V+1)\\
&\qquad\leq \mathbb{P}\bigg(|S\cap\mathcal{H}_0(P)| \geq \sum_{i\in S\cap\mathcal{H}_0(P)} \mathbf{1}\{p_i(X)\geq \alpha k_0/|S|\} + k_0\bigg)\\
&\qquad= \mathbb{P}\bigg(\sum_{i\in S\cap\mathcal{H}_0(P)} \mathbf{1}\{p_i(X)< \alpha k_0/|S|\} \geq k_0\bigg)\\
&\qquad\leq \frac{|S\cap\mathcal{H}_0(P)|\alpha k_0/|S|}{k_0} \leq \alpha.
\end{aligned}$$

However, in practice, S is often chosen by the user and possibly depends on the same data set, then denoted $\widehat{S}$ to emphasize this dependence; it typically corresponds to items of potential strong interest. The most archetypal example is when $\widehat{S}$ consists of the s_0 smallest p-values $p_{(1:m)},\dots,p_{(s_0:m)}$, for some fixed value of $s_0 \in \mathbb{N}_m$. In that case, it is easy to check that the above bound does not have the correct coverage: for instance, when the p-values are i.i.d. $U(0,1)$ and $\mathcal{H}_0(P) = \mathbb{N}_m$, we have for $k_0 \leq s_0$ (that is, when the bound is informative),

$$\begin{aligned}
\mathbb{P}(|\widehat{S}\cap\mathcal{H}_0(P)| \leq V) &= \mathbb{P}\Big(k_0 - 1 + \textstyle\sum_{i\in\hat{S}}\mathbf{1}\{p_i(X)\geq \alpha k_0/s_0\} \geq s_0\Big)\\
&= \mathbb{P}\Big(\textstyle\sum_{i\in\hat{S}}\mathbf{1}\{p_i(X)< \alpha k_0/s_0\} \leq k_0-1\Big)\\
&= \mathbb{P}\left(p_{(k_0:m)}(X)\geq \alpha k_0/s_0\right)\\
&= \mathbb{P}(\beta(k_0, m-k_0+1) \geq \alpha k_0/s_0),
\end{aligned}$$

where $\beta(k_0, m-k_0+1)$ denotes the usual beta distribution with parameters k_0 and $m-k_0+1$. For instance, taking $s_0 = 10$, $k_0 = 5$, $\alpha = 0.05$ and $m = 500$, the latter is approximately equal to 0.005, while the intended target is $1-\alpha = 0.95$.

This phenomenon is often referred to as the selection effect: after some data driven selection, the probabilities change and thus the usual statistical inferences are not valid.

4. ... to post hoc bounds

To circumvent the selection effect, one way is to aim for a function $V(X,\cdot) : S \subset \mathbb{N}_m \mapsto V(X,S) \in \mathbb{N}$ (denoted by $V(S)$ for short) satisfying

$$\forall P\in\mathcal{P}, \qquad \mathbb{P}_{X\sim P}\Big(\forall S\subset\mathbb{N}_m,\ |S\cap\mathcal{H}_0(P)| \leq V(S)\Big) \geq 1-\alpha, \tag{1}$$

that is, a $(1-\alpha)$ confidence bound that is valid *uniformly* over all subsets $S \subset \mathbb{N}_m$. As a result, for any particular algorithm $\widehat{S}$, inequality (1) entails

$\mathbb{P}\big(|\hat{S} \cap \mathcal{H}_0(P)| \leq V(\hat{S})\big) \geq 1-\alpha$, and thus does not suffer from the selection effect. Such a bound will be referred to as a $(1-\alpha)$-*post hoc confidence bound* throughout this chapter, "post hoc" meaning that the set S can be chosen after having seen the data, and possibly using the data several times.

As a first example, the k_0-*Bonferroni post hoc bound* is

$$V^{k_0\mathrm{Bonf}}(S) = |S| \wedge \left(\sum_{i \in S} \mathbf{1}\left\{p_i(X) \geq \alpha k_0/m\right\} + k_0 - 1 \right). \tag{2}$$

Following the same reasoning as above, it has a coverage at least $1-\alpha$ under (Superunif):

$$\begin{aligned}
&\mathbb{P}(\exists S \subset \mathbb{N}_m \,:\, |S \cap \mathcal{H}_0(P)| \geq V^{k_0\mathrm{Bonf}}(S) + 1) \\
&\qquad \leq \mathbb{P}\left(\exists S \subset \mathbb{N}_m \,:\, \sum_{i \in S \cap \mathcal{H}_0(P)} \mathbf{1}\left\{p_i(X) < \alpha k_0/m\right\} \geq k_0 \right) \\
&\qquad = \mathbb{P}\left(\sum_{i \in \mathcal{H}_0(P)} \mathbf{1}\left\{p_i(X) < \alpha k_0/m\right\} \geq k_0 \right) \\
&\qquad \leq \frac{|\mathcal{H}_0(P)|\alpha k_0/m}{k_0} \leq \alpha.
\end{aligned}$$

Remark 0.1 *Compared to the k_0-Bonferroni confidence bound of Section 3, α has been replaced by $\alpha|S|/m$, so that the post hoc bound is much more conservative than a (standard, non uniform, S fixed) confidence bound when $|S|/m$ gets small, which is well expected. This scaling factor is the price paid here to make the inference post hoc. We will see in Sections 5 and 8 that it can be diminished when considering bounds of a different nature.*

Coming back to the motivating example of Section 1, if we choose $k_0 = 100$, the k_0-Bonferroni post-hoc bound (2) ensures that with probability at least 90%, the number of false positives among the 163 genes selected by the BH procedure at level $q = 0.05$ is upper bounded by 99.

Example 0.1 *For $k_0 = 1$, when the p-values are i.i.d. $U(0,1)$ and $\mathcal{H}_0(P) = \mathbb{N}_m$, the coverage probability of the k_0-Bonferroni post hoc bound is equal to $(1-\alpha/m)^m = e^{m\log(1-\alpha/m)}$, which is very close to $1-\alpha$ when α is small.*

The Bonferroni post hoc bound, while it is valid under no assumption on the dependence structure of the p-value family, may be conservative, in the sense that $V(S)$ will be large for many subsets S. For instance, one has $V^{k_0\mathrm{Bonf}}(S) = |S|$ (trivial bound) for all the sets S such that $S \subset \{i \in \mathbb{N}_m \,:\, p_i(X) > \alpha k_0/m\}$.

The Bonferroni bound can be further improved under some dependence restriction, with the *Simes post hoc bound:*

$$V^{\mathrm{Sim}}(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} \mathbf{1}\left\{p_i(X) \geq \alpha k/m\right\} + k - 1 \right\} = \min_{1 \leq k \leq |S|} \{V^{k\mathrm{Bonf}}(S)\}. \tag{3}$$

Its coverage can be computed as follows (using arguments similar as above):

$$\begin{aligned}
&\mathbb{P}(\exists S \subset \mathbb{N}_m \,:\, |S \cap \mathcal{H}_0(P)| \geq V^{\mathrm{Sim}}(S) + 1) \\
&\leq \mathbb{P}\left(\exists S \subset \mathbb{N}_m \,,\, \exists k \in \{1, \ldots, m\} \,:\, \sum_{i \in S \cap \mathcal{H}_0(P)} \mathbf{1}\left\{p_i(X) < \alpha k/m\right\} \geq k\right) \\
&= \mathbb{P}\left(\exists k \in \{1, \ldots, |\mathcal{H}_0(P)|\} \,:\, p_{(k:\mathcal{H}_0(P))} < \alpha k/m\right). \qquad (4)
\end{aligned}$$

Under (Superunif) and (Indep), this is lower than or equal to $\alpha|\mathcal{H}_0(P)|/m \leq \alpha$ by using the *Simes inequality*. More generally, the Simes post-hoc bound is valid in any setting where the Simes inequality holds. This is the case under a specific positive dependence assumption called Positive Regression Dependency on a Subset of hypotheses (PRDS), which is also the assumption under which the Benjamini-Hochberg (BH) procedure has been shown to control the false discovery rate (FDR).

While it uses more stringent assumptions, $V^{\mathrm{Sim}}(S)$ can be much less conservative than $V^{k_0\mathrm{Bonf}}$. For instance, if $S = \{i \in \mathbb{N}_m \,:\, 5\alpha/m \leq p_i(X) < 10\alpha/m\}$, we have $V^{5\mathrm{Bonf}}(S) = |S|$ and $V^{\mathrm{Sim}}(S) \leq |S| \wedge 9$, which can lead to a substantial improvement. Coming back to the motivating example of Section 1, the Simes post-hoc bound (3) ensures that with probability at least 90%, the number of false positives among the 163 genes selected by the BH procedure at level $q = 0.05$ is upper bounded by 78.

From Example 0.2 below, the Simes bound has a nice graphical interpretation: $|S| - V^{\mathrm{Sim}}(S)$ can be interpreted as the smallest integer u for which the shifted line $v \mapsto \alpha(v-u)/m$ is strictly below the ordered p-value curve, see Figure 2.

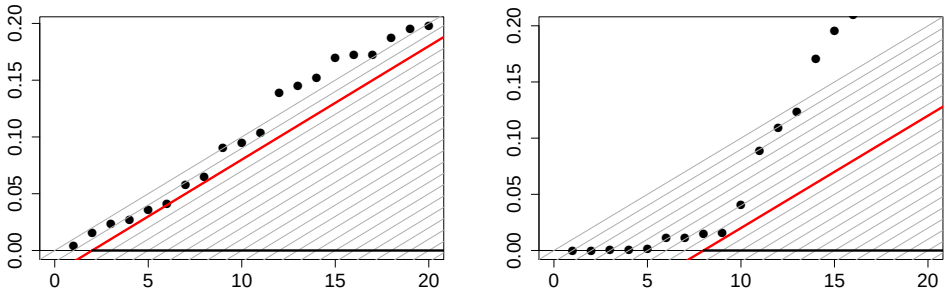

FIG 2. *Illustration of the Simes post hoc bound* (3) *according to the expression* (5)*, for two subsets of* $\mathbb{N}_m$ *(left display/right display), both of cardinal* 20 *and for* $m = 50$*. The level is* $\alpha = 0.5$ *(taken large only for illustration purposes). Black dots: sorted p-values in the respective subsets. Lines: thresholds* $k \in \{u+1, \ldots, |S|\} \mapsto \alpha(k-u)/m$ *(in bold red for* $u = |S| - V(S)$*, in light gray otherwise). The post hoc bound* $V^{Sim}(S)$ *corresponds the length of the bold line on the X-axis.*

Example 0.2 *Starting from (3) and writing $|S| - V^{Sim}(S) \leq u$ for some u, one can show that for all $S \subset \mathbb{N}_m$, $|S| - V^{Sim}(S)$ is equal to*

$$\min\{u \in \{0, \ldots, |S|\} \, : \, \forall v \in \{u+1, \ldots, |S|\} \, : \, p_{(v:S)} \geq \alpha(v-u)/m\}, \qquad (5)$$

where $p_{(1:S)}, \ldots, p_{(|S|:S)}$ denote the ordered p-values of $\{p_i(X), i \in S\}$.

Example 0.3 *In Figure 2, we have $V^{Sim}(S) = 18$ (resp. $V^{Sim}(S) = 12$) in the left (resp. right) situation. Instead, $V^{k_0 Bonf}$ for $k_0 = 7$ is equal to 18 (resp. 16).*

The Simes post hoc bound (3) has, however, some limitations: first, the coverage is only valid when the Simes inequality holds. This imposes restrictive conditions on the model used, which are rarely met or provable in practice. As noted above, the same caveat applies to the BH procedure.

Second, even in that case, the bound does not incorporate the dependence structure, which may yield conservativeness (see Example 0.4 below). Finally, this bound intrinsically compares the ordered p-values to the threshold $k \mapsto \alpha k/m$ (possibly shifted). We can legitimately ask whether taking a different threshold (called template below) does not provide a better bound.

Example 0.4 *Consider the case $\mathcal{H}_0(P) = \mathbb{N}_m$, for which m is even, and denote $\overline{\Phi}$ the upper-tail distribution function of a standard $\mathcal{N}(0,1)$ variable. Consider the one-sided testing situation where $p_i = \overline{\Phi}(X_1)$, $1 \leq i \leq m/2$ and $p_i = \overline{\Phi}(X_2)$, $m/2+1 \leq i \leq m$, for a 2-dimensional Gaussian vector (X_1, X_2) that is centered, with covariance matrix having 1 as diagonal elements and $\rho \in [-1, 1]$ as off-diagonal elements. In this case, one can show that the coverage probability of the Simes post hoc bound is equal to*

$$\alpha/2 + \int_{\alpha/2}^{\alpha} \overline{\Phi}\left(\frac{\overline{\Phi}^{-1}(\alpha) - \rho\overline{\Phi}^{-1}(w)}{(1-\rho^2)^{1/2}}\right) dw + \int_{\alpha}^{\infty} \overline{\Phi}\left(\frac{\overline{\Phi}^{-1}(\alpha/2) - \rho\overline{\Phi}^{-1}(w)}{(1-\rho^2)^{1/2}}\right) dw \qquad (6)$$

The above quantity is displayed in Figure 3 for $\alpha = 0.2$, as a function of ρ.

5. Threshold-based post hoc bounds

This section presents the λ-calibration method, which allows to derive more accurate threshold-based post hoc bounds under mild assumptions. This is of major interest from a practical perspective, since these assumptions are met in the two-sample multiple testing setting, which is often encountered in applications.

Let us consider bounds of the form

$$V^{\lambda}(S) = \min_{1 \leq k \leq |S|} \left\{ \sum_{i \in S} \mathbf{1}\left\{p_i(X) \geq t_k(\lambda)\right\} + k - 1 \right\}, \quad \lambda \in [0,1], \qquad (7)$$

where $t_k(\lambda)$, $\lambda \in [0,1]$, $1 \leq k \leq m$, is a family of functions, called a template. A template can be seen as a spectrum of curves, parametrized by λ. We focus here on the two following examples:

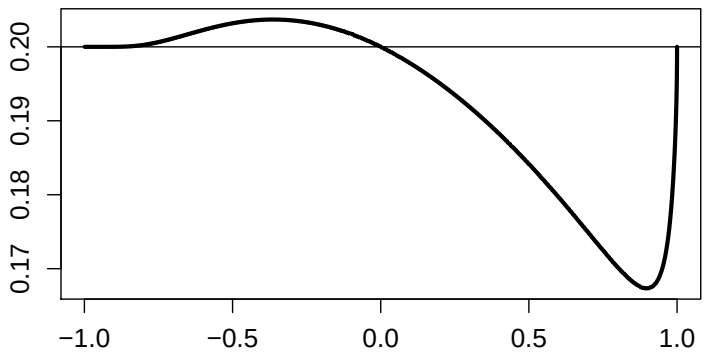

FIG 3. *Coverage of the Simes post hoc bound* (6) *in the setting of Example 0.4 as a function of* ρ *and for* $\alpha = 0.2$.

- Linear template: $t_k(\lambda) = \lambda k/m$, $t_k^{-1}(y) = ym/k$;
- Beta template: $t_k(\lambda) =$$\lambda$-quantile of $\beta(k, m-k+1)$, $t_k^{-1}(y) = \mathbb{P}(\beta(k, m-k+1) \leq y)$.

An illustration for the above templates is provided in Figure 4.

For a fixed template, the idea is now to choose one of these curves, that is, one value of the parameter $\lambda = \lambda(\alpha)$, so that the overall coverage is larger than $1-\alpha$. Following exactly the same reasoning as the one leading to (4), we obtain

$$\begin{aligned}
&\mathbb{P}(\exists S \subset \mathbb{N}_m \,:\, |S \cap \mathcal{H}_0(P)| \geq V^\lambda(S) + 1) \\
&\qquad \leq \mathbb{P}\left(\exists k \in \{1, \dots, |\mathcal{H}_0(P)|\} \,:\, p_{(k:\mathcal{H}_0(P))} < t_k(\lambda)\right) \qquad (8) \\
&\qquad = \mathbb{P}\left(\min_{k \in \{1,\dots,|\mathcal{H}_0(P)|\}} \left\{t_k^{-1}(p_{(k:\mathcal{H}_0(P))})\right\} < \lambda\right), \qquad (9)
\end{aligned}$$

by letting $t_k^{-1}(y) = \max\{x \in [0,1] \,:\, t_k(x) \leq y\}$ the generalized inverse of t_k (in general, this is valid provided that for all $k \in \{1, \dots, m\}$, $t_k(0) = 0$ and $t_k(\cdot)$ is non-decreasing and left-continuous on $[0,1]$, as in the case of the two above examples). What remains to be done is thus to calibrate $\lambda = \lambda(\alpha, X)$ such that the quantity (9) is below α.

Several approaches can be used for this. It is possible that for the model under consideration, the joint distribution of $(p_i(X))_{i \in \mathcal{H}_0(P)}$ is equal to the restriction of some known, fixed distribution on $[0,1]^{\mathbb{N}_m}$ to the coordinates of $\mathcal{H}_0(P)$ (this is a version of the so-called subset-pivotality condition). It is met under condition (Indep), but it is also possible that the dependence structure of the p-values is known (for example, in genome-wide association studies, the

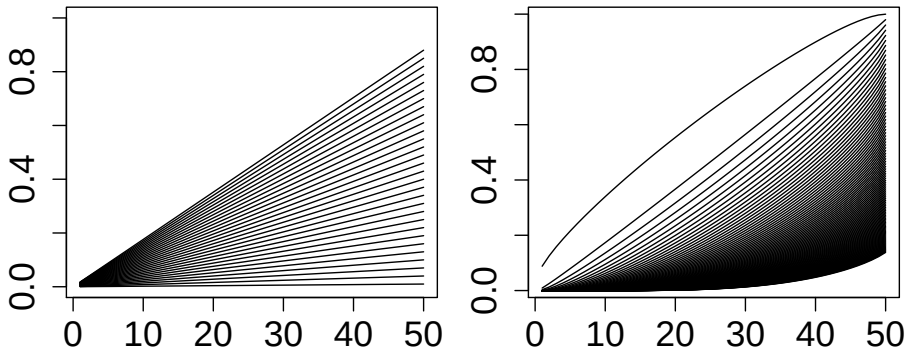

FIG 4. *Curves $k \mapsto t_k(\lambda)$ for a wide range of λ values. Left: linear template. Right: Beta template.*

structure and strength of linkage disequilibrium can be tabulated from previous studies and give rise to a precise dependence model). In such a situation, the calibration of $\lambda = \lambda(\alpha, X)$ can be obtained either by exact computation, numerical approximation or Monte-Carlo approximation under the full null.

Another situation of interest, on which we focus for the remainder of this section, is when the null corresponds to an invariant distribution with respect to a certain group of data transformations, which is the setting for (generalized) permutation tests, allowing for the use of an exact randomization technique. More precisely, assume the existence of a finite transformation group $\mathcal{G}$ acting onto the observation set $\mathcal{X}$. By denoting $p_{\mathcal{H}_0}(x)$ the null p-value vector $(p_i(x))_{i\in\mathcal{H}_0(P)}$ for $x \in \mathcal{X}$, we assume that the joint distribution of the transformed null p-values is invariant under the action of any $g \in \mathcal{G}$, that is,

$$\forall P \in \mathcal{P},\ \ \forall g \in \mathcal{G},\ \ (p_{\mathcal{H}_0}(g'.X))_{g'\in\mathcal{G}} \sim (p_{\mathcal{H}_0}(g'.g.X))_{g'\in\mathcal{G}}, \tag{Rand}$$

where $g.X$ denotes X that has been transformed by g.

Let us consider a (random) $B-$tuple $(g_1, g_2, \ldots, g_B)$ of $\mathcal{G}$ (for some $B \geq 2$), where g_1 is the identity element of $\mathcal{G}$ and $g_2, \ldots, g_B$ have been drawn (independently of the other variables) as i.i.d. variables, each being uniformly distributed on $\mathcal{G}$. Now, let for all $x \in \mathcal{X}$, $\Psi(x) = \min_{1\leq k\leq m}\left\{t_k^{-1}\left(p_{(k:m)}(x)\right)\right\}$ and consider $\lambda(\alpha, X) = \Psi_{(\lfloor \alpha B\rfloor+1)}$ where $\Psi_{(1)} \leq \Psi_{(2)} \leq \cdots \leq \Psi_{(B)}$ denote the ordered sample $(\Psi(g_j.X), 1 \leq j \leq B)$. The following result holds.

Theorem 0.1 *Under* (Rand)*, for any deterministic template, the bound $V^{\lambda(\alpha,X)}$ is a post hoc bound of coverage $1-\alpha$. This level is to be understood as a joint probability with respect to the data and the draw of the group elements $(g_i)_{2\leq i\leq B}$.*

As a case in point, let us consider a two-sample framework where

$$X = (X^{(1)}, \ldots, X^{(n_1)}, X^{(n_1+1)}, \ldots, X^{(n_1+n_2)}) \in (\mathbb{R}^m)^n$$

is composed of $n = n_1 + n_2$ independent m-dimensional real random vectors with $X^{(j)}$, $1 \leq j \leq n_1$, i.i.d. $\mathcal{N}(\theta^{(1)}, \Sigma)$ (case) and $X^{(j)}$, $n_1 + 1 \leq j \leq n$, i.i.d. $\mathcal{N}(\theta^{(2)}, \Sigma)$ (control). Then we aim at testing the null hypotheses $H_{0,i}$: "$\theta_i^{(1)} = \theta_i^{(2)}$", simultaneously for $1 \leq i \leq m$, without knowing the covariance matrix Σ. Consider any family of p-values $(p_i(X))_{1 \leq i \leq m}$ such that $p_i(X)$ only depends on the i-th coordinate $(X_i^{(j)})_{1 \leq j \leq n}$ of the observations (e.g., based on difference of the coordinate means of the two groups). Note that $p_{\mathcal{H}_0}(X)$ is thus a measurable function of $(X_i^{(j)})_{i \in \mathcal{H}_0, 1 \leq j \leq n}$. Now, the group $\mathcal{G}$ of permutations of $\{1, \ldots, n\}$ is naturally acting on $\mathcal{X} = (\mathbb{R}^m)^n$ via the permutation of the individuals: for all $\sigma \in \mathcal{G}$,

$$\sigma.X = (X^{(\sigma(1))}, \ldots, X^{(\sigma(n_1))}, X^{(\sigma(n_1+1))}, \ldots, X^{(\sigma(n))}).$$

Since the variables $(X_i^{(1)})_{i \in \mathcal{H}_0}, \ldots, (X_i^{(n)})_{i \in \mathcal{H}_0}$ are i.i.d., it is clear that (Rand) holds in this case.

The practical interest of Theorem 0.1 is illustrated in Section 9 for the differential gene expression study introduced in Section 1, and in Section 10 for functional Magnetic Resonance Imaging (fMRI) data. These numerical results demonstrate that substantial gains in power may be obtained by λ-calibration: in both cases, the lower bounds on the number of true positives are two to three times higher than with the classical Simes bounds.

An illustration of the above λ-calibration method is provided in Figure 5 in the case where $\Sigma = I_m$,

$$p_i(X) = 2\left(1 - \Phi\left(s_{n_1,n_2}^{-1}\left|n_2^{-1}\sum_{j=n_1+1}^{n_1+n_2} X_i^{(j)} - n_1^{-1}\sum_{j=1}^{n_1} X_i^{(j)}\right|\right)\right),$$

for $s_{n_1,n_2} = (n_1^{-1} + n_2^{-1})^{1/2}$ and using a Beta template. In the left panel (full null), we have $\theta^{(1)} = \theta^{(2)} = 0$, so that $\mathcal{H}_0(P) = \mathbb{N}_m$. In the right panel (half of true nulls), we have $\theta_i^{(1)} = \theta_i^{(2)} = 0$ for $1 \leq i \leq m/2$ and $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = \delta / s_{n_1,n_2}$ for $m/2 + 1 \leq i \leq m$, for some $\delta > 0$, so that $\mathcal{H}_0(P) = \{1, \ldots, m/2\}$. Following expression (8), $k \mapsto t_k(\lambda(\alpha, X))$ is the highest Beta curve such that at most $B\alpha$ orange curves have a point situated below it. This also shows that the above λ-calibration is slightly more severe when part of the data follows the alternative distribution. This is a commonly observed phenomenon: although the permutation approach is valid even when part of the null hypotheses are false, their inclusion in the permutation procedure tends to yield test statistics that exhibit more variation under permutation, thus inducing more conservativeness in the calibration.

6. Reference families

We cast the previous bounds in a more general setting, where $(1 - \alpha)$–post hoc bounds are explicitly based on a *reference family* with some *joint error rate* (JER

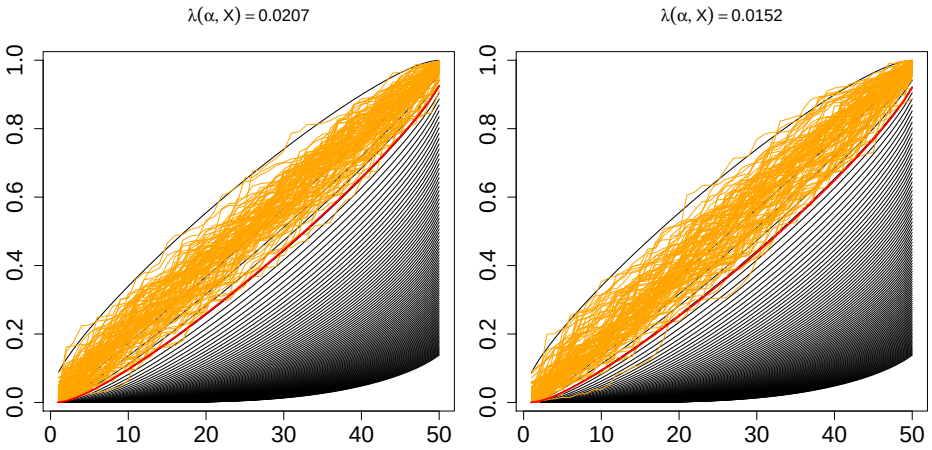

FIG 5. *Illustration of the $\lambda = \lambda(\alpha, X)$ calibration method on one realization of the data X. Black curves: Beta template $k \mapsto t_k(\lambda)$ for some range of λ values. Orange curves: ordered p-values (after permutation) $k \mapsto p_{(k:m)}(g_j.X)$ for $1 \leq j \leq B = 1000$. Bold red curve: $k \mapsto t_k(\lambda(\alpha, X))$. Left panel : full null, right panel : half of true nulls (see text). (Parameters $m = 50$, $\alpha = 0.2$, $n_1 = 50$, $n_2 = 50$, $\delta = 3$.)*

in short) controlling property. This general point of view offers more flexibility and allows us to consider post hoc bounds of a different nature, as for instance those incorporating a spatial structure, see Section 8.

In general, a reference family is defined by a collection $\mathfrak{R} = \big((R_1(X), \zeta_1(X)), \ldots, (R_K(X), \zeta_K(X))\big)$, where the R_k's are data-dependent subsets of $\mathbb{N}_m$ and the ζ_k's are data dependent integer numbers (we will often omit the dependence in X to ease notation). The reference family $\mathfrak{R}$ is said to control the JER at level α if

$$\forall P \in \mathcal{P}, \qquad \mathbb{P}_{X\sim P}(\forall k \in \mathbb{N}_K \ : \ |R_k(X) \cap \mathcal{H}_0| \leq \zeta_k(X)) \geq 1 - \alpha. \tag{10}$$

Markedly, (10) is similar to (1), but restricted to some subsets R_k, $k \in \mathbb{N}_K$. The rationale behind this approach is that, while the choice of S is let completely free in (1) (to accommodate any choice of the practitioner), the choice of the R_k's and ζ_k's in (10) is done by the statistician and is part of the procedure. Once we obtain a reference family $\mathfrak{R}$ satisfying (10), we obtain a post hoc bound by interpolation:

$$V^*_{\mathfrak{R}}(S) = \max\{|S \cap A| \ : \ A \subset \mathbb{N}_m, \forall k \in \mathbb{N}_K, |R_k \cap A| \leq \zeta_k\}, \quad S \subset \mathbb{N}_m\,. \tag{11}$$

We call $V^*_{\mathfrak{R}}$ the *optimal post hoc bound* (built upon the reference family $\mathfrak{R}$). Computing the bound $V^*_{\mathfrak{R}}(S)$ can be time-consuming, it actually has NP-hard complexity in a general configuration. We can introduce the following com-

putable relaxations: for $S \subset \mathbb{N}_m$,

$$\overline{V}_{\mathfrak{R}}(S) = \min_{k \in \mathbb{N}_K} \left(|S \setminus R_k| + \zeta_k \right) \wedge |S|; \tag{12}$$

$$\widetilde{V}_{\mathfrak{R}}(S) = \left(\sum_{k \in \mathbb{N}_K} |S \cap R_k| \wedge \zeta_k + \left| S \setminus \bigcup_{k \in \mathbb{N}_K} R_k \right| \right) \wedge |S|. \tag{13}$$

One can easily check that $V^*_{\mathfrak{R}}(S) \leq \overline{V}_{\mathfrak{R}}(S)$ and $V^*_{\mathfrak{R}}(S) \leq \widetilde{V}_{\mathfrak{R}}(S)$ for all $S \subset \mathbb{N}_m$. Moreover, provided that (10) holds, $V^*_{\mathfrak{R}}$, $\overline{V}_{\mathfrak{R}}$ and $\widetilde{V}_{\mathfrak{R}}$ are all valid $(1-\alpha)$–post hoc bounds. The details are left to the reader.

In addition, the following result shows that the relaxed versions coincide with the optimal bound if the reference sets have some special structure:

Lemma 0.1

- *In the nested case, that is, $R_k \subset R_{k+1}$, for $1 \leq k \leq K-1$, we have $\overline{V}_{\mathfrak{R}} = V^*_{\mathfrak{R}}$;*
- *In the disjoint case, that is, $R_k \cap R_{k'} = \emptyset$ for $1 \leq k \neq k' \leq K$, we have $\widetilde{V}_{\mathfrak{R}} = V^*_{\mathfrak{R}}$.*

We can briefly revisit the post-hoc bounds of the previous sections in this general framework. The k_0-Bonferroni post hoc bound (2) derives from the one-element reference family $(R = \{i \in \mathbb{N}_m : p_i(X) < \alpha k_0/m\}, \zeta = k_0 - 1)$. The Simes post hoc bound (3) derives from the reference family comprising the latter reference sets for all $k_0 \in \mathbb{N}_m$. More generally, the threshold-based post hoc bounds V^λ of the form (7) are equal to the optimal bound $V^*_{\mathfrak{R}}$ with $R_k = \{i \in \mathbb{N}_m : p_i(X) < t_k(\lambda)\}$ and $\zeta_k = k-1$, $k \in \mathbb{N}_m$(indeed, these reference sets are nested, so that $V^*_{\mathfrak{R}} = \overline{V}_{\mathfrak{R}}$).

How to choose a suitable reference family in general? A general rule of thumb is to choose the reference sets R_k of the same qualitative form as the sets S for which the bound is expected to be accurate. For instance, the Simes post hoc bound will be more accurate for sets S with the smallest p-values. In Section 8, we will choose reference sets R_k with a spatial structure, which will produce a post hoc bound more tailored for spatially structured subsets S.

7. Case of a fixed single reference set

It is useful to focus first on the case of a single fixed (non-random) reference set R_1, with (random) ζ_1 satisfying (10), that is,

$$\mathbb{P}(|\mathcal{H}_0(P) \cap R_1| \leq \zeta_1(X)) \geq 1 - \alpha.$$

(In contrast with the k_0-Bonferroni bound (2) where ζ was fixed and R variable, here R_1 is fixed and ζ_1 is variable.) In other words, $\zeta_1(X)$ is a $(1-\alpha)$–confidence bound of $|\mathcal{H}_0(P) \cap R_1|$. Several examples of such $\zeta_1(X)$ can be built, under various assumptions.

Example 0.5 *For $R_1 \subset \mathbb{N}_m$ fixed, the following bounds are $(1-\alpha)$–confidence bounds for $|\mathcal{H}_0(P) \cap R_1|$:*

- *under* (Superunif)*, for some fixed $t \in (0, \alpha)$,*

$$\zeta_1(X) = |R_1| \wedge \left\lfloor \sum_{i \in R_1} \mathbf{1}\left\{p_i(X) > t\right\}/(1 - t/\alpha) \right\rfloor, \tag{14}$$

  where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x (this is a simple application of the Markov inequality).
- *under* (Superunif) *and* (Indep)*,*

$$\zeta_1(X) = |R_1| \wedge \min_{t \in [0,1)} \left\lfloor \frac{C}{2(1-t)} + \left(\frac{C^2}{4(1-t)^2} + \frac{\sum_{i \in R_1} \mathbf{1}\{p_i(X) > t\}}{1-t} \right)^{1/2} \right\rfloor^2, \tag{15}$$

  where $C = \sqrt{\frac{1}{2}\log\left(\frac{1}{\alpha}\right)}$ (this can deduced by using the DKW inequality, that is, for any integer $n \geq 1$, for $U_1, \ldots, U_n$ i.i.d. $U(0,1)$, we have $\sup_{t \in [0,1]} \left\{ n^{-1} \sum_{i=1}^n \mathbf{1}\{U_i > t\} - (1-t) \right\} \geq -\sqrt{\log(1/\lambda)/(2n)}$ with probability at least $1 - \lambda$).

In addition to the two above bounds (14) and (15), we can elaborate another bound in the generalized permutation testing framework (Rand), as described in Section 5. Applying the result of that section, the following bound is also valid:

$$\zeta_1(X) = \min_{1 \leq k \leq |R_1|} \left\{ \sum_{i \in R_1} \mathbf{1}\left\{p_i(X) \geq t_k(\lambda(\alpha, X))\right\} + k - 1 \right\}, \tag{16}$$

where $t_k(\lambda)$ denotes the λ-quantile of a $\beta(k, |R_1| - k + 1)$ distribution and $\lambda(\alpha, X) = \Psi_{(\lfloor \alpha B \rfloor + 1)}$, where $\Psi_{(1)} \leq \Psi_{(2)} \leq \cdots \leq \Psi_{(B)}$ denote the ordered sample $(\Psi(g_j.X), 1 \leq j \leq B)$ for which $\Psi(x) = \min_{1 \leq k \leq |R_1|} \left\{ t_k^{-1}\left(p_{(k:|R_1|)}(x)\right) \right\}$ (see the λ-calibration method of Section 5).

Once a proper choice of $\zeta_1(X)$ has been done, the optimal post hoc bound can be computed as follows: for any $S \subset \mathbb{N}_m$, $V^*_{\mathfrak{R}}(S) = \overline{V}_{\mathfrak{R}}(S) = \widetilde{V}_{\mathfrak{R}}(S) = |S \cap R_1^c| + \zeta_1(X) \wedge |S \cap R_1|$. When S is large and does not contain very small p-values, this bound can be sharper than the Simes bound. For instance, let us consider the single reference family $R_1 = \mathbb{N}_m$ and $\zeta_1(X)$ as in (15) (choosing $t = 1/2$). For S such that $S \subset \{i \in \mathbb{N}_m : p_i(X) > \alpha|S|/m\}$, we have $V^{\mathrm{Sim}}(S) = |S|$ and $V^*_{\mathfrak{R}}(S) = |S| \wedge \zeta_1(X) \leq |S| \wedge 2\left(\log\left(\frac{1}{\alpha}\right) + 2\sum_{i \in \mathbb{N}_m} \mathbf{1}\left\{p_i(X) > 1/2\right\}\right)$. The latter can be smaller than $|S|$ when many p-values are below $1/2$.

Finally, while the case of a single reference set can be considered as an elementary example, the bounds developed in this section will be useful in the next section, for which several fixed reference sets R_k are considered, and thus several (random) ζ_k should be designed.

8. Case of spatially structured reference sets

We consider here the case where the null hypotheses $H_{0,i}$, $1 \leq i \leq m$, have a spatial structure, and we are interested in obtaining accurate bounds on $|S \cap \mathcal{H}_0(P)|$ for subsets S of the form $S = \{i \in \mathbb{N}_m \ : \ i_0 \leq i \leq j_0\}$, for some $1 \leq i_0 < j_0 \leq m$.

In that case, it is natural to choose R_k formed of contiguous indices. To be concrete, consider reference sets consisting of disjoint intervals of the same size : assume $m = Ks$ for some integers $K > 0$ and $s > 0$ and let

$$R_k = \{(k-1)s + 1, \ldots, ks\}, \ k \in \mathbb{N}_K. \tag{17}$$

When each of these regions is considered in isolation, Section 7 suggested several approaches (in the appropriate settings (Superunif), (Indep) or (Rand)) of a specific form $\zeta_k(X) = f(R_k, \alpha, X)$, to underline the dependence of $\zeta_k(X)$ in R_k and α. By using a simple union bound, it is then straightforward to show that the JER control (10) is satisfied for

$$\zeta_k(X) = f(R_k, \alpha/K, X), \ k \in \mathbb{N}_K. \tag{18}$$

When the reference regions R_k are disjoint as in the example (17) above, we can use the proxy $\widetilde{V}_{\mathfrak{R}}(S)$ (see (13)) which is known to coincide with the optimal bound $V^*_{\mathfrak{R}}(S)$. This gives rise to a post hoc bound that accounts for the spatial structure of the data.

Example 0.6 *In the case where $\zeta_1(X) = f(R_1, \alpha, X)$ is given by (14) ($t = \alpha^2$, $K < 1/\alpha$), we obtain*

$$\zeta_k(X) = |R_k| \wedge \left\lfloor \sum_{i \in R_k} \mathbf{1}\left\{p_i(X) > \alpha^2\right\}/(1 - \alpha K) \right\rfloor.$$

Note that this bound quickly increases as the size of the family K increases. By contrast, when $\zeta_1(X) = f(R_1, \alpha, X)$ is given by (15), one can derive

$$\zeta_k(X) = |R_k| \wedge \min_{t \in [0,1)} \left\lfloor \frac{}{2(1-t)} + \left(\frac{C^2}{4(1-t)^2} + \frac{\sum_{i \in R_k} \mathbf{1}\{p_i(X) > t\}}{1-t} \right)^{1/2} \right\rfloor^2,$$

for $C = \sqrt{\frac{1}{2}\log\left(\frac{K}{\alpha}\right)}$. The size of the family K appears here in a logarithmic term, which makes this bound less sensitive to the parameter K.

When considering the reference regions defined by segments (17), we have to prescribe a scale (s here, the size of the segments). It is possible to extend this to a multi-scale approach, choosing overlapping reference intervals R_k at different resolutions arranged in a tree structure, where parent sets are formed by taking union of (disjoint) children sets taken at a finer resolution. Furthermore, the proxy (13) has to be replaced by a more elaborate one, minimizing over all

possible multi-scale partitions made of such reference regions. This can still be computed efficiently by exploiting the the tree structure. Doing so, the post hoc bound will be more scale adaptive to sets S with possibly various sizes. The price to pay lies in the cardinality K of the family, which gets larger. However, this does not necessarily make the corresponding bound much larger, as Example 0.6 shows when using the bound (15), since the level α only enters it logarithmically.

9. Application to differential gene expression studies

In this section, we illustrate how the post hoc inference framework introduced in the preceding sections can be applied to the case of differential gene expression introduced in Section 1 to build confidence envelopes for the proportion of false positives (Section 9.1), and to obtain bounds on data-driven sets of hypotheses (Section 9.2), and on sets of hypotheses defined by an a priori structure (Section 9.3). These numerical results were obtained using the R package `sansSouci`, version 0.9.0. A Rmarkdown vignette [3] to reproduce results and plots from this section is provided as Supplementary Material.

9.1. Confidence envelopes

In absence of specific prior information on relevant subsets of hypotheses to consider, it is natural to focus on subsets consisting of the most significant hypotheses. Specifically, we define the $k-$th p-value level set S_k as the set of the k most significant hypotheses, corresponding to the p-values $(p_{(1:m)}, p_{(2:m)}, \ldots, p_{(k:m)})$, and consider post hoc bounds associated to S_k for $k \in \mathbb{N}_m$. Figure 6 provides *post hoc confidence envelopes* for the ALL data set, for $\alpha = 0.1$. While $(1-\alpha)$-lower confidence bounds on the proportion of false positives $\left\{\left(k, \overline{V}(S_k)/\left|S_k\right|\right) : k \in \mathbb{N}_m\right\}$ are displayed in the left panel, $(1-\alpha)$-upper confidence bounds on the number of true positives of the form $\left\{\left(k, |S_k| - \overline{V}(S_k)\right) : k \in \mathbb{N}_m\right\}$ are shown in the right panel.

The confidence envelopes are built from the Simes bound (3) (long-dashed purple curve), and from two bounds obtained from Theorem 0.1 by λ-calibration using $B = 1,000$ permutation of the sample labels, based on the two templates introduced in Section 5: the dashed orange curve corresponds to the linear template with $K = m$, and the solid green curve to the Beta template with $K = 50$. Note that Assumption (Rand) holds because we are in the two-sample framework described after Theorem 0.1.

The vertical line in Figure 6 corresponds to the 163 genes selected by the BH procedure at level 5%. The Simes bound ensures that the FDP of this subset is not larger than 0.48. As noted above concerning the BH procedure, we have a priori no guarantee that this bound is valid, because such multiple two-sample testing situations have not been shown to satisfy the PRDS assumption under

[3] See https://github.com/rstudio/rmarkdown

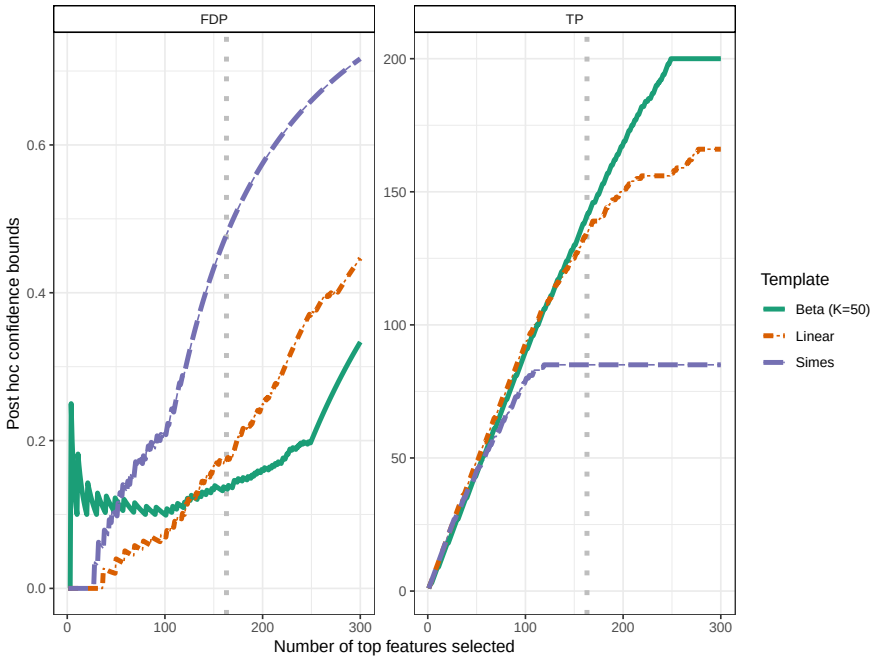

FIG 6. *Confidence bounds on the proportion of false positives (left) and on the number of true positives (right) for the Leukemia data set. Reference families: Simes reference family (long-dashed purple curve), linear template after λ-calibration (dashed orange curve), and Beta template after λ-calibration (solid green curve).*

which the Simes inequality is valid[4]. In contrast, the λ-calibrated bounds built by permutation are by construction valid here. Moreover, both are much sharper than the Simes bound while the λ-calibrated bound using the linear template is twice smaller, ensuring FDP< 0.23, and even smaller for the Beta template with $K = 50$. The bound obtained by λ-calibration of the linear template is uniformly sharper that the original Simes bound (3), which corresponds to $\lambda = \alpha$. This illustrates the adaptivity to dependence achieved by λ-calibration. The bound obtained from the Beta template is less sharp for p-value level sets S_k of cardinal less than $k = 120$, and then sharper. This is consistent with the shape of the threshold functions displayed in Figure 4.

9.2. Data-driven sets

A common practice in the biomedical literature is to only retain, among the genes called significant after multiple testing correction, those whose "fold change" exceeds a prescribed level. The fold change is the ratio between the mean expression levels of the two groups. With the notation of Section 5, the fold-change of gene i is given by $\Delta_i = \overline{X}_i^{(2)}/\overline{X}_i^{(1)}$, where $\overline{X}_i^{(1)} = n_1^{-1}\sum_{j=1}^{n_1} X_i^{(j)}$ and $\overline{X}_i^{(2)} = n_2^{-1}\sum_{j=1}^{n_2} X_i^{(j)}$.

This is illustrated by Figure 7, where each gene is represented as a point in the $(\log(\text{fold change}), -\log(p))$ plan. This representation is called a "volcano plot" in the biomedical literature. Among the 163 genes selected by the BH procedure at level 0.05, 151 have an absolute log fold change larger than 0.3. As FDR is not preserved by selection, FDR controlling procedures provide no statistical guarantee on such data-driven lists of hypotheses.

In contrast, the post hoc bounds proposed in this chapter are valid for such data-driven sets. The two shaded boxes in Figure 7 correspond to the data-driven subsets $S^{\mathrm{BH}} \cap S^-$ and $S^{\mathrm{BH}} \cap S^+$, where S^{BH} is the set of 163 genes selected by the BH procedure at level 0.05, $S^- = \{i \in \mathbb{N}_m, \log(\Delta_i) < -0.3\}$ and $S^+ = \{i \in \mathbb{N}_m, \log(\Delta_i) > +0.3\}$. The post hoc bounds on the number of true positives in $S^{\mathrm{BH}} \cap S^+, S^{\mathrm{BH}} \cap S^-$ and $S^{\mathrm{BH}} \cap (S^+ \cup S^-)$ obtained by the Simes bound and by the λ-calibrated linear and Beta templates are given in Table 1. Both λ-calibrated bounds are more informative than the Simes bound, in the sense that they provide a higher bound on the number of true confidence. Moreover, they have proven $(1-\alpha)$-coverage, whereas the coverage of the Simes bound is a priori unknown for multiple two-sample tests. None of the two λ-calibrated bounds dominates the other one, which is in line with the fact that the linear template is well-adapted to situations with smaller p-value level sets than the Beta template.

Finally, we also note that the bound obtained for $S^+ \cup S^-$ is systematically larger than the sum of the two individual bounds, which, again, is in accordance with the theory.

[4]In this particular case, λ-calibration with the linear template yields $\lambda(\alpha) > \alpha$, which a posteriori implies that the Simes inequality was indeed valid.

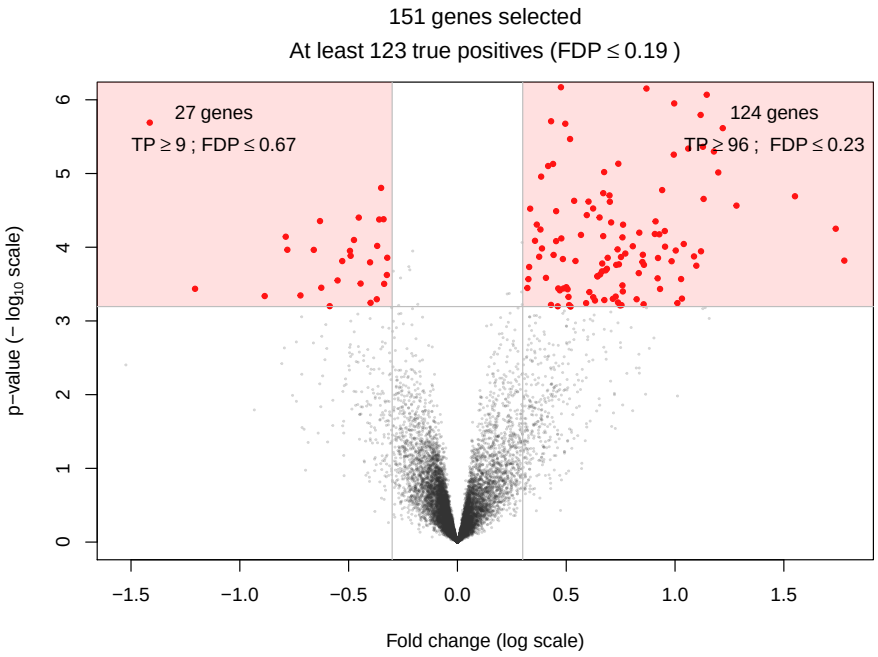

FIG 7. *Post-hoc inference for volcano plots. Each dot corresponds to a gene in the (fold change, p-value space) on a logarithmic scale. Bold red dots corresponds to 151 genes that (i) are selected by the BH procedure at level* $q = 0.05$*, and (ii) have a absolute log fold change larger than* 0.3.

S	$\lvert S\rvert$	Simes	Linear	Beta(K=50)
$S^{\mathrm{BH}} \cap S^-$	124	62	96	103
$S^{\mathrm{BH}} \cap S^+$	27	1	9	7
$S^{\mathrm{BH}} \cap (S^+ \cup S^-)$	151	79	123	130

TABLE 1

Post hoc bounds on the number of true positives in $S^{\mathrm{BH}} \cap S^+$, $S^{\mathrm{BH}} \cap S^-$ and $S^{\mathrm{BH}} \cap (S^+ \cup S^-)$ obtained by the post hoc bounds displayed in Figure 6.

9.3. Structured reference sets

In this section we give an example of application of the bounds mentioned in Section 8. Our biological motivation is the fact that gene expression activity can be clustered along the genome.

The m individual hypotheses are naturally partitioned into 23 subsets, each corresponding to a given chromosome. Within each chromosome, we consider sets of $s = 10$ successive genes as in (17). Hence, we focus on a reference family with the following elements

$$R_{c,k} = \{(k-1)s + 1, \dots, \min(ks, m_c)\}, \quad k \in \mathbb{N}_{K_c}, \quad c \in \{1, \dots, 23\},$$

where, in chromosome c, m_c denotes the number of genes, $K_c = \lceil m_c/s \rceil$ the number of corresponding regions. In addition, for each (c, k) we use $\zeta_{c,k}(X) = f(R_{c,k}, \alpha_c/K_c, X)$ coming from the union bound (18) in combination with the device (15) and $\alpha_c = \alpha m_c/m$. This choice accounts for a union bound over all the chromosomes. As shown in Example 0.5, $\zeta_{c,k}(X)$ is a valid upper confidence bound for $|\mathcal{H}_0(P) \cap R_{c,k}|$ under (Superunif) and (Indep). In this genomic example, (Indep) may not hold, so we have in fact no formal guarantee that this bound is valid. Therefore, the results obtained below are merely illustrative of the approach and may not have biological relevance.

We report the results for chromosome $c = 19$, which contains $m_c = 626$ genes. In this particular case, we obtain trivial bounds $\zeta_{c,k}(X) = |R_{c,k}|$ for all $k \in \mathbb{N}_{K_c}$. Therefore, the proxy $\tilde{V}^*_{\mathfrak{R}}$ defined in (13) for disjoint sets does not identify any signal for this chromosome. However, non-trivial bounds can be obtained via the multi-scale approach briefly mentioned in Section 8. The idea is to enrich the reference family by recursive binary aggregation of the neighboring $R_{c,k}$. The total number of elements in this family is less than $2K_c$. In our example, it turns out that (15) yields 6 true discoveries in the interval $R_{17:24}$ and 1 true discovery in the interval $R_{53:54}$, where we have denoted

$$R_{u:v} = \bigcup_{u \le k \le v} R_{c,k}.$$

This is illustrated by Figure 8 where the individual p-values are displayed (on the $-\log_{10}$ scale) as a function of their order on chromosome 19. The sets $R_{17:24}$ and $R_{53:54}$ are highlighted in orange, with the corresponding number of true discoveries marked in each region. We obtain a non-trivial bound not because of the large effect of any individual gene, but because of the presence

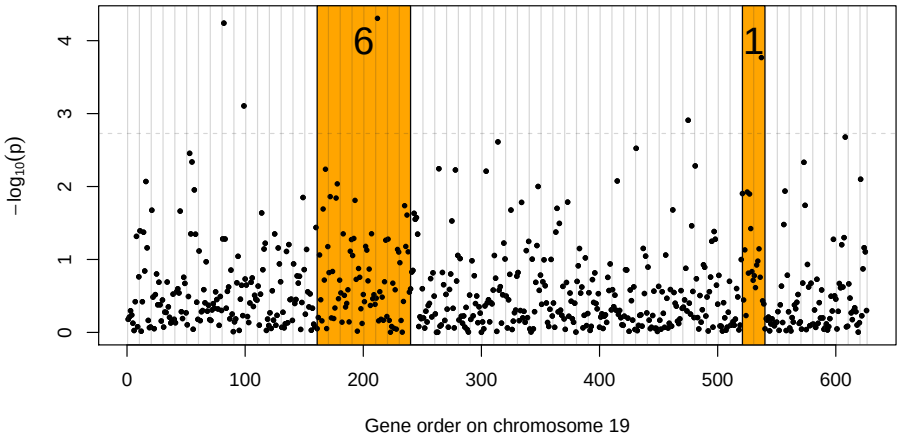

FIG 8. *Evidence of locally-structured signal on chromosome 19 detected by the bound* (15). *The numbers correspond to the lower bound on the false positives in each of the highlighted regions.*

of sufficiently many moderate effects. In particular, in the rightmost orange region in Figure 8, the distribution of $-\log_{10}(p)$ is shifted away from 0 when compared to the rest of chromosome 19. In comparison, we obtain trivial bounds $\overline{V}_{\mathfrak{R}}(R_{53:54}) = |R_{53:54}| = 2s$ and $\overline{V}_{\mathfrak{R}}(R_{17:24}) = |R_{17:24}| = 8s$ from (12) both for the linear or the Beta template. These numerical results illustrate the interest of the bounds introduced in Section 8 in situations where one expects the signal to be spatially structured.

10. Application to fMRI studies

We focus on the problem of detecting brain regions whose activity is significantly different between two motor tasks performed by subjects: left versus right click. The fMRI data have been extracted from the Localizer data set[5]. A Rmarkdown vignette to reproduce results and plots from this section is provided as Supplementary Material.

10.1. Confidence envelopes

As in Section 9, we begin by constructing confidence envelopes for top-k feature lists. Figure 9 provides *post hoc confidence envelopes* for the Localizer data set, for $\alpha = 0.1$. While $(1-\alpha)$-lower confidence bounds on the proportion of false positives $\left\{\left(k, \overline{V}(S_k)/|S_k|\right) : k \in \mathbb{N}_m\right\}$ are displayed in the left panel, $(1-\alpha)$-upper confidence bounds on the number of true positives of the form $\left\{\left(k, |S_k| - \overline{V}(S_k)\right) : k \in \mathbb{N}_m\right\}$ are shown in the right panel.

[5]Orfanos, D. P. *et al. Neuroimage*, 181:786–796 (2017).

The confidence envelopes are built from the Simes bound (3) (long-dashed purple curve), and from two bounds obtained from Theorem 0.1 by λ-calibration using $B = 1,000$ permutation of the sample labels, based on the two templates introduced in Section 5: the dashed orange curve corresponds to the linear template with $K = m$, and the solid green curve to the Beta template with $K = 500$. Assumption (Rand) holds because we are in the two-sample framework described after Theorem 0.1. The Simes bound is also called All Resolution Inference (ARI) in that context, see Supplementary Material 2 for more details and references.

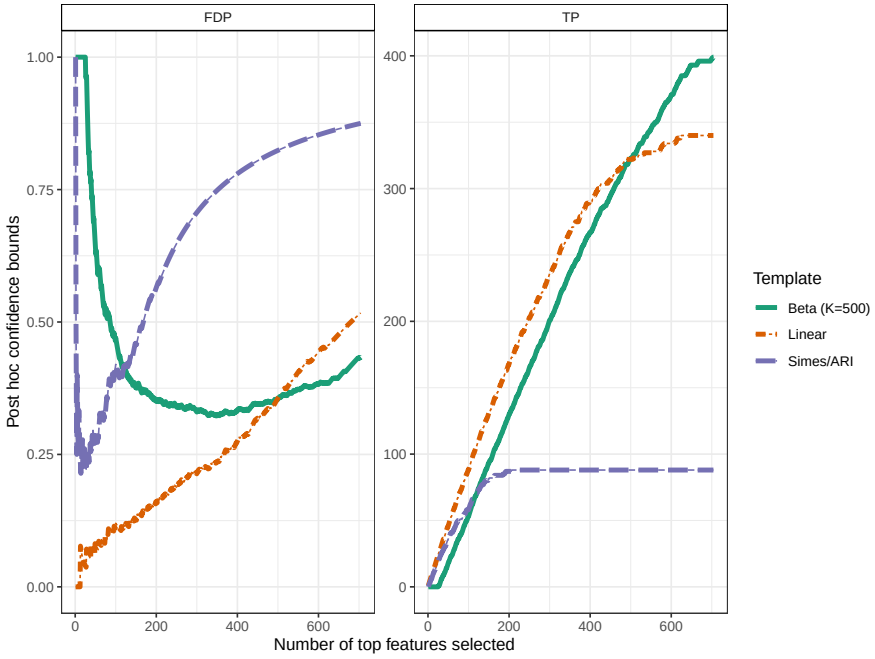

FIG 9. *Confidence bounds on the proportion of false positives (left) and on the number of true positives (right) for the fMRI data set. Reference families: Simes reference family (long-dashed purple curve), linear template after λ-calibration (dashed orange curve), and Beta template after λ-calibration (solid green curve).*

The results are qualitatively similar as for the genomic example given in Section 9. Both permutation-based post hoc bounds are much sharper than the Simes bound, illustrating the adaptivity to dependence achieved by λ-calibration.

10.2. Post hoc bounds on brain atlas areas

The goal in this section is to calculate post hoc bound on user-defined brain regions. One definition of such regions is given by the Harvard-Oxford brain

atlas[6]. We have calculated the post hoc bounds associated to each of these 48 areas: by definition, these are confidence bounds valid *simultaneously* for all areas. In this particular example, evidence of signal is obtained for three of these atlases, as summarized in Table 2. For the same target risk level ($\alpha = 0.1$),

	area	size	Simes/ARI	Linear	Beta (K=500)
7	Precentral Gyrus	6640	79	254	281
17	Postcentral Gyrus	5030	69	202	196
19	Supramarg. Gyrus (ant. div.)	2017	0	3	0

TABLE 2
Lower bounds on the number of true positives in 3 brain regions.

both permutation-based bounds are much less conservative than the classical Simes/ARI bound, showing that permutation-based approaches are able to adapt to unknown dependency.

11. Bibliographical notes

The material exposed in this chapter is mainly a digested account of the article [3]. The seminal work [10] introduced the idea of false positive bounds for arbitrary rejection sets. It started from the idea of building a confidence set on the set of null hypotheses $\mathcal{H}_0(P)$, and introduced the concepts of *augmentation procedure* and *inversion procedure*. The latter consists in building a confidence set based on the inversion of tests for $\mathcal{H}_0(P) = \mathcal{A}$ for all $\mathcal{A} \subset \mathbb{N}_m$. The former starts from a set R with controlled k-familywise error rate, and the proposed associated post hoc bound is (10) (for the one-element reference family $(R, \zeta = k - 1)$). The name *augmentation* refers to a similar idea found in [7]. The relaxation (10) can in this sense be called "generalized augmentation procedure". A post hoc bound for an arbitrary rejection set based on a closed test principle was proposed in [11]. It can also be seen as a reformulation of the inversion procedure of [10]. Post-hoc bounds over a large class of reference families extracted from classical FDR control procedures combined with martingale techniques were recently proposed in [15]. The principle of the graphical representation used in Figure 2 to visualize the Simes inequality-based bound originates from J. Goeman.

The use of generalized permutation procedures in a multiple testing framework has been explored in several landmark works [27, 23, 18, 7, 12, 14]. The subset-pivotality condition has been defined in [27]. Assumption (Rand) has been introduced in [13] and is a weaker version of the randomization hypothesis of [23]. The phenomenon of conservativeness in the permutation-based calibration mentioned at the end of Section 5, when not all the null hypotheses are true, can be in part alleviated by using a step-down principle (see [23] for a seminal work on this topic and [3] for more details on this approach in the specific setting considered here). The choice of the size K of the reference family, which can be crucial in practice, is also discussed in [3].

[6] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases

Multiple testing for spatially structured hypotheses is in itself a very active and broad area of research. It has been specifically considered in conjunction with post-hoc bounds in [17]. The use of the reference family approach for post-hoc bounds in combination with spatially structured hypotheses has been studied in [8], where the notion of tree- (or forest-)structured reference regions is introduced, along with an efficient algorithm to compute the optimal bound $V^*_{\mathfrak{R}}$ in this setting.

The Simes inequality [25] is a particularly nice and elegant theoretical device with manifold applications in multiple testing which is still a very active research area, see, e.g., [4, 5, 9]. The DKW inequality with optimal constant was proved in [16]. The Benjamini-Hochberg (BH) procedure has been introduced in [2], where it is also proved to control the false discovery rate (FDR). A huge literature on FDR control has followed this seminal paper.

The data used for the genomics application in Section 9 are taken from [6]. The fact that the signal is clustered along the genome is motivated by previous studies showing possible links between gene expression and DNA copy number changes or other regulation mechanisms [22, 26]. The data used for the neuroimaging application in Section 10 were obtained from the Brainomics/Localizer database [20] and the Harvard-Oxford brain atlas[7], using the Python package `nilearn` [1]. The ARI method mentioned in that section [24] corresponds to the Simes post hoc bound of [11].

Acknowledgements

We would like to thank referees for their helpful comments. PN would like to thank Alexandre Blain who contributed to the application to fMRI studies, and Bertrand Thirion for constructive feedback on this application. This work has been supported by ANR-16-CE40-0019 (SansSouci), ANR-17-CE40-0001 (BASICS) and by the GDR ISIS through the "projets exploratoires" program (project TASTY). GB acknowledges the support from the german DFG under the Collaborative Research Center SFB-1294 "Data Assimilation". GB and PN acknowledge the support from the Franco-German University through the binational Doktorandenkolleg CDFA 01-18.

Supplementary Materials

Vignette "Permutation-based post hoc inference for differential gene expression studies":

This Rmarkdown vignette[8] demonstrates how the R package `sansSouci` may be used to obtain post hoc confidence bounds on false positives in the case of differential gene expression analysis. In particular, it contains the R code to reproduce Figures 1, 6 and 7, and Table 1.

[7] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases

[8] Available at https://pneuvial.github.io/sanssouci/articles/post-hoc_differential-expression.html.

Vignette: "Permutation-based post hoc inference for fMRI studies"

This Rmarkdown vignette[9] demonstrates how the R package `sansSouci` may be used to obtain post hoc confidence bounds on false positives in the case of functional Magnetic Resonance Imaging (fMRI) studies. In particular, in contains R code to reproduce Figure 9 and Table 2.

References

[1] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.

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

[4] H. W. Block, T. H. Savits, J. Wang, and S. K. Sarkar. The multivariate-t distribution and the Simes inequality. *Statist. Probab. Lett.*, 83(1):227–232, 2013.

[5] T. Bodnar and T. Dickhaus. On the Simes inequality in elliptical models. *Ann. Inst. Statist. Math.*, 69(1):215–230, 2017.

[6] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, K. S. Wang, F. Mandelli, R. Foa, and J. Ritz. 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):7209–7219, 2005.

[7] S. Dudoit and M. J. van der Laan. *Multiple testing procedures with applications to genomics*. Springer Series in Statistics. Springer, New York, 2008.

[8] G. Durand, G. Blanchard, P. Neuvial, and E. Roquain. Post hoc false positive control for spatially structured hypotheses. *Scandinavian journal of Statistics*, 2020.

[9] H. Finner, M. Roters, and K. Strassburger. On the Simes test under dependence. *Statist. Papers*, 58(3):775–789, 2017.

[10] C. R. Genovese and L. Wasserman. Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.*, 101(476):1408–1417, 2006.

[11] J. J. Goeman and A. Solari. Multiple testing for exploratory research. *Statist. Sci.*, 26(4):584–597, 2011.

[12] J. Hemerik and J. Goeman. Exact testing with random permutations. *TEST*, 27(4):811–825, 2018.

[9] Available at https://pneuvial.github.io/sanssouci/articles/post-hoc_fMRI.html.

[13] J. Hemerik and J. J. Goeman. False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017.

[14] J. Hemerik, A. Solari, and J. J. Goeman. Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, to appear.

[15] E. Katsevich and A. Ramdas. Simultaneous high-probability bounds on the false discovery proportion in structured, regression, and online settings, 2018. arXiv preprint 1803.06790.

[16] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990.

[17] R. J. Meijer, T. J. Krebs, and J. J. Goeman. A region-based multiple testing method for hypotheses ordered in space or time. *Statistical Applications in Genetics and Molecular Biology*, 14(1):1–19, 2015.

[18] N. Meinshausen. False discovery control for multiple tests of association under general dependence. *Scand. J. Statist.*, 33(2):227–237, 2006.

[19] P. Neuvial, B. Sadacca, G. Blanchard, G. Durand, and E. Roquain. sanssouci: Post hoc multiple testing inference, 2020. R package version 0.9.0.

[20] D. P. Orfanos, V. Michel, Y. Schwartz, P. Pinel, A. Moreno, D. Le Bihan, and V. Frouin. The brainomics/localizer database. *Neuroimage*, 144:309–314, 2017.

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[22] F. Reyal, N. Stransky, I. Bernard-Pierrot, A. Vincent-Salomon, Y. de Rycke, P. Elvin, A. Cassidy, A. Graham, C. Spraggon, Y. Désille, A. Fourquet, C. Nos, P. Pouillart, H. Magdelénat, D. Stoppa-Lyonnet, J. Couturier, B. Sigal-Zafrani, B. Asselain, X. Sastre-Garau, O. Delattre, J. P. Thiery, and F. Radvanyi. Visualizing chromosomes as transcriptome correlation maps: evidence of chromosomal domains containing co-expressed genes - a study of 130 invasive ductal breast carcinomas. *Cancer Research*, 65(4):1376–1383, Feb. 2005.

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

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

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

[26] N. Stransky, C. Vallot, F. Reyal, I. Bernard-Pierrot, S. G. D. de Medina, R. Segraves, Y. de Rycke, P. Elvin, A. Cassidy, C. Spraggon, A. Graham, J. Southgate, B. Asselain, Y. Allory, C. C. Abbou, D. G. Albertson, J.-P. Thiery, D. K. Chopin, D. Pinkel, and F. Radvanyi. Regional copy number-independent deregulation of transcription in cancer. *Nature Genetics*, 38:1386–1396, 2006.

[27] P. H. Westfall and S. S. Young. *Resampling-Based Multiple Testing*. Wiley, 1993.