Hierarchical correction of p-values via a tree running Ornstein-Uhlenbeck process

Antoine Bichat[1,2] ,   Christophe Ambroise[1],
Mahendra Mariadassou[3]

Abstract Statistical testing is classically used as an exploratory tool to search for association between a phenotype and many possible explanatory variables. This approach often leads to multiple testing under dependence.

We assume a hierarchical structure between tests via an Ornstein-Uhlenbeck process on a tree. The process correlation structure is used for smoothing the p-values. We design a penalized estimation of the mean of the Ornstein-Uhlenbeck process for p-value computation.

The performances of the algorithm are assessed via simulations. Its ability to discover new associations is demonstrated on a metagenomic dataset.

The corresponding R package is available from `https://github.com/abichat/zazou`.

Keywords multiple testing, Ornstein-Uhlenbeck process, lasso, debiasing, FDR control, metagenomic

1 Introduction

In many fields, statistical testing is classically used as an exploratory tool to look for the association between a variable of interest and many possible explanatory variables. For example, in transcriptomics, the link between a phenotype and the expression of tens of thousands of genes is tested (McLachlan et al., 2005), in Genome Wide Association Studies (GWAS) the association between millions of markers and a phenotype is tested (Bush and Moore, 2012),

Electronic address: `abichat@enterome.com`; Corresponding author

1: Université Paris-Saclay, CNRS, Univ. Évry, Laboratoire de Mathématiques et Modélisation d'Évry 91037, Évry, France.
2: Enterome, 94-96 Avenue Ledru Rollin, 75011 Paris, France.
3: MaIAGE, INRAE, Université Paris-Saclay, 78350, Jouy-en-Josas, France.

in functional Magnetic Resonance Imaging (fMRI), the goal is to identify voxels that are significantly activated in two diffrent conditions (Cremers et al., 2017).

This problem of multiple comparisons dates back to the work of Tukey (Tukey, 1953). It has since been the subject of abundant literature and aims at controlling a probability of error of some sort. Most of the literature focus on the control of the Familiy Wise Error Rate (FWER) (Bland and Altman, 1995), being the probability of at least one false discovery among detections, or of the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995), defined as the expected proportion of false positives among detections.

Most of the correction procedures for controlling FWER or FDR rely on independence, or some form of weak dependence, among the hypothesis, which is rarely observed in practice. Multiple testing under dependence is a difficult problem occurring in many fields. In transcriptomics, differential analysis has to deal with gene expressions that are often highly correlated. When performing GWAS, the linkage desiquilibrium imposes a strong spatial dependence between markers, and in Functional Magnetic Resonance Imaging (fMRI), two spatially close voxels have often comparable activation.

The control of the FDR via the popular Benjamini-Hochberg procedure remains valid under arbitrary dependency structures (Benjamini and Yekutieli, 2001). However, based on results obtained from simulated datasets, it is obvious that there is a substantial loss of power when the real dependency structure is not taken into account.

An alternative approach for dealing with multiple testing is to reduce the number of tests by aggregating certain hypotheses. Aggregation strategies vary and can based on a priori knowledge (*e.g.* metabolic pathways, functional modules of genes) or on clustering algorithms.

In this paper we assume that a hierarchical structure exists between variables and is known up to some constants. The hypotheses tested can then be organized in a tree structure which captures correlations at different scales of observation. This type of hierarchical structure is observable in transcriptomics differential analysis, where gene expressions can easily be represented by a hierarchy based on gene expression correlation. In GWAS and fMRI, spatial dependence also proves to be very suitable for hierarchical modeling (Ambroise et al., 2019; Eickhoff et al., 2015; Sesia et al., 2020).

We propose to model the hierarchical structure of the multiple tests through an Ornstein-Uhlenbeck process on a tree. The process correlation structure is used for smoothing the p-values, after conversion to z-scores, similarly to the algorithm proposed in Xiao et al. (2017) but with an explicit underlying model. We resort to an ℓ_1 penalized estimation of the mean of the Ornstein-Uhlenbeck process, followed by a debiasing procedure (Javanmard and Montanari, 2013, 2014; Zhang and Zhang, 2014) for p-value computation. Eventually, we use a tuning proposed by Javanmard et al. (2019) to control the FDR.

The selection strength of the Ornstein-Uhlenbeck process and the penalty parameter are hyperparameters of our model, whose selection is achieved via a Bayesian Information Criterion (BIC). We provide some background on hier-

archical procedures in Section 2, introduce the model and statistical procedure in Section 3 and detail the computational steps in Section 4. The performances of the algorithm are assessed via simulations in Section 5. The use of the proposed model is illustrated in Section 6, where we demonstrate its ability to discover novel associations in a metagenomic dataset.

2 Background

2.1 Examples of multiple testing strategies

A classic example in genomics consists in grouping the markers according to whether they belong to the same genes (aggregation by an *a priori*). The genes can then be grouped according to their similarity, computed for example from expression profiles. Kim et al. (2010) have, for example, proposed a hierarchical testing strategy controlling the FWER in a hierarchical manner, by testing clusters of genes, then individual genes associated with a phenotype with the goal of finding genomic regions associated with a specific type of cancer. This type of top-down approach uses the concept of sequential rejection principle (Goeman and Finos, 2012; Meinshausen, 2008).

fMRI is another domain where tests are aggregated: neighboring voxels that are highly correlated are aggregated into a single voxel cluster. Benjamini and Heller (2007) propose an adaptation of the False Discovery Rate (FDR) to allow for cluster-level multiple testing for fMRI data.

Ad hoc aggregating methods for multiple testing also exist in Metagenomics. LEfSe (Segata et al., 2011) performs a bottom up approach where a factorial Kruskal-Wallis rank sum test is applied to each feature with respect to a class factor, followed by a pairwise Wilcoxon test, and a linear discriminant analysis. MiLineage (Tang et al., 2017) performs multivariate tests concerning multiple taxa in a lineage to test the association of lineages to a phenotypic outcome.

2.2 Independence assumption

The assumption of independence of tests is convenient as it provides for both exact analyses and simple error bounds for classical procedures (Benjamini and Hochberg, 1995, e.g.). It is however unrealistic in practice. In many fields, including all the previous examples, measurements typically exhibit strong correlations. Some correction procedure, like the one proposed by Benjamini and Yekutieli (2001), make few assumptions while guaranteeing control of the FDR. Those general guarantees come with a high cost in terms of statistical power: the nominal FDR much smaller that the target, resulting in many FN. Permutation procedures are an appealing alernative that can automatically adapt to the dependence structure of the p-values (Tusher et al., 2001) but may fail when confronted to unbalanced design or correlated data. Knowledge of the correlation structure can be leveraged to increase the power while still

controling the FDR below a given target. Several approaches have been developed along those lines when the tests are organized along a hierarchical structure, typically encoded in a tree.

115   2.3 Hierarchical testing

The Hierarchical FDR (Yekutieli, 2008), implemented in the R package `structSSI` (Sankaran and Holmes, 2014), proposes a top-down algorithm to sequentially reject hypotheses organized in a tree. However, the algorithm suffers from some limitations (Bichat et al., 2020; Huang et al., 2020). First, the algorithm in its
120   vanilla formulation commonly fails to move down on the tree because of failure to reject the topmost node. Second, it only controls for an *a posteriori* FDR level, which is a complex function of the *a priori* FDR level and the structure of rejected nodes. Finally, it does not produce a corrected p-value, or q-value, per tip, but only a *reject / no reject* decision. Given these drawbacks, we did
125   not consider the Hierarchical FDR in our benchmark.

`StructFDR` (Xiao et al., 2017) was developed for metagenomics Differential Abundance Testing (DAT) and relies on z-scores / p-values smoothing followed by permutation correction. Given any taxa-wise DAT procedure, p-values $\mathfrak{p}$ are first computed for all m taxa (*i.e.* leaves of the tree) and then transformed to z-scores $\mathfrak{z}$. The tree is used to compute a distance matrix $(\mathbf{D}_{i,j})$ and then turned into a correlation matrix $\mathbf{C}_\rho = (\exp(-2\rho\mathbf{D}_{i,j}))$ between taxa using a Gaussian kernel. The z-scores are then smoothed using the following hierarchical model:

$$\mathfrak{z} \mid \mu \sim \mathcal{N}_m\left(\mu, \sigma^2\mathbf{I}_m\right),$$

$$\mu \sim \mathcal{N}_m\left(\gamma\mathbf{1}_m, \tau^2\mathbf{C}_\rho\right),$$

where μ captures the effect size of each taxa and $\mathfrak{z}$ is a noisy observation of μ. The maximum a posteriori estimator μ^* of μ is given by

$$\mu^* = \left(\mathbf{I}_m + k\mathbf{C}_\rho^{-1}\right)^{-1}\left(k\mathbf{C}_\rho^{-1}\gamma\mathbf{1}_m + \mathfrak{z}\right) \quad \text{where} \quad k = \sigma^2/\tau^2.$$

The FDR is controlled by means of a resampling procedure to estimate the distribution of μ^* under H_0 and estimate adjusted p-values $\mathfrak{q}^{\text{sf}}$. This method is implemented in the `StructFDR` package (Chen, 2018).

`TreeclimbR` (Huang et al., 2020) is a bottom-up approach also developed for metagenomics DAT but with a broader scope. It relies on aggregating abundances at each node of the tree (understood as a cluster of taxa) and performing a test to compute one p-value per node (compared one test per leaf for `StructFDR`). The main idea is then to use those p-values to compute a score for node i

$$U_i(t) = \left|\frac{\sum_{k \in B(i)} \mathfrak{s}_k \mathbb{1}_{\{\mathfrak{p}_k \leq t\}}}{\#B(i)}\right|$$

where $B(i)$ is the set of descendants of node i, $\mathfrak{p}_k$ and $\mathfrak{s}_k \in \{-1, -1\}$ are the
130   p-value of the node k and the sign of the associated effect, and t is a tuning

parameter. A node i will be considered as candidate if $U_i(t) \simeq 1$ and $\mathfrak{p}_i < \alpha$. This ensure that all descendants are (i) significant at level t with (ii) effects of coherent sign. At the end, multiplicity correction is only done on nodes (including leaves) that do not descend from another candidate.

3 Models and algorithms

Our correction methods assumes that p-values, or rather z-scores, evolve according to an Ornstein-Uhlenbeck process on a tree. We thus use the corresponding correlation structure to decorrelate the z-scores and, in turn, the p-values. This is similar in spirit to the smoothing algorithm of Xiao et al. (2017) but we derive our procedure from first principles and explicit assumptions. We first remind a few properties of Ornstein-Uhlenbeck processes before proceeding to our model and procedure.

3.1 Ornstein-Uhlenbeck process on a tree

An Ornstein-Uhlenbeck (OU) process (W_t) with optimal value β_{ou}, selection strengh α_{ou} and drift parameter σ_{ou} is a Gaussian process that satisfies the stochastic differential equation:

$$\mathrm{d}W_t = -\alpha_{\mathrm{ou}}(W_t - \beta_{\mathrm{ou}})\mathrm{d}t + \sigma_{\mathrm{ou}}\mathrm{d}B_t.$$

The important properties of OU processes are bounded variance and convergence to a stationary distribution centered on the optimal value β_{ou}, namely $W_t \xrightarrow{(d)} \mathcal{N}(\beta_{\mathrm{ou}}, \sigma_{\mathrm{ou}}^2/2\alpha_{\mathrm{ou}})$ when $t \to \infty$. Thanks to those properties, OU processes have become popular to model the evolution of continuous traits, such as body mass (Freckleton et al., 2003). They naturally emerge as the continuous limit of broad range of discrete-time evolution models (Lande, 1976). Ornstein-Uhlenbeck processes can be readily adapted to tree-like structures as illustrated in Fig. 1.

Formally, we consider a rooted ultrametric tree $\mathcal{T}$ with m tips and n branches ($n = 2m - 1$ for binary trees). The internal nodes are labeled N_1 (the root) to N_{n-m} and the tips T_1 to T_m. Let i be a node, W_i the value of the trait at that node and note $pa(i)$ its unique parent. By convention, we set $t_{N_1} = 0$ and assume $W_{N_1} = 0$. The branch leading to i from $pa(i)$ is denoted b_i and has length $l_i = t_i - t_{pa(i)}$ where t_i is the time elapsed between the root and node i. Since the tree is ultrametric, $t_i = h$ for all $i \in \{T_1, \ldots, T_n\}$. For any pair of nodes (i, j), note t_{ij} the time elapsed between the root and the most recent common ancestor of i and j and note $d_{ij} = t_i - t_j - 2t_{ij}$ the distance in the tree between nodes i and j. The distribution of the trait at node i is given by:

$$W_i | W_{pa(i)} \sim \mathcal{N}\left(\lambda_i W_{pa(i)} + (1 - \lambda_i)\beta_{\mathrm{ou},i}, \frac{\sigma_{\mathrm{ou}}^2}{2\alpha_{\mathrm{ou}}}(1 - \lambda_i^2)\right) \qquad (1)$$

where $\lambda_i = \exp(-\alpha_{\mathrm{ou}} l_i)$ and $\beta_{\mathrm{ou},i}$ is the optimal value on branch i. Remark that the process mean value does not immediately shift to $\beta_{\mathrm{ou},i}$ but lags behind it with a shrinkage parameter controlled by $1 - \lambda_i$. If $\beta_{\mathrm{ou},i} = 0$ for all i, straightforward computations show that $W = (W_{T_1}, \ldots, W_{T_m})$ is a gaussian vector with distribution

$$W \sim \mathcal{N}(0, \Sigma) \quad \text{where} \quad \Sigma_{ij} = \frac{\sigma_{\mathrm{ou}}^2}{2\alpha_{\mathrm{ou}}} e^{-2\alpha_{\mathrm{ou}} d_{ij}} (1 - e^{-2\alpha_{\mathrm{ou}} t_{ij}}).$$

When, the optimal value can shift on a branch (*e.g.* the branch b_{N_4} leading to N_4 in Fig. 1), the mean vector of W is a slightly more complex and depends on both the tree topology and the location and magnitude of the shifts. Note U the $m \times (n+m)$ incidence matrix of $\mathcal{T}$ with rows labeled by tips ($i \in \{T_1, \ldots, T_n\}$) and columns labeled by inner nodes and tips ($j \in \{N_1, \ldots, N_m, T_1, \ldots, T_n\}$), with entries defined as $U_{ij} = 1$ if and only if tip i is in the subtree rooted at node j. Intuitively, column $U_{\cdot j}$ encodes all tips descending from node j and row $U_{i \cdot}$ encodes all ancestors of tip i. Note Δ the dimension n column vector with entries defined as $\Delta_i = \beta_{\mathrm{ou},i} - \beta_{\mathrm{ou},pa(i)}$ where $i \in \{N_1, \ldots, N_m, T_1, \ldots, T_n\}$. Non null entries of Δ correspond to *shifts location*, nodes for which the optimal value $\beta_{\mathrm{ou},i}$ differ from its parent's and their values to *shifts magnitude* (see Figure 2 for an example). Finally note Λ the $n+m$ diagonal matrix with diagonal entries $\Lambda_i = 1 - \exp(\alpha_{\mathrm{ou}}(h - t_{pa(i)}))$ where $i \in \{N_1, \ldots, N_m, T_1, \ldots, T_n\}$. Straightforward computations (see Bastide et al. (2017) for detailed derivations) show that W is a gaussian vector with joint distribution:

$$W \sim \mathcal{N}(\mu, \Sigma) \quad \text{where} \quad \mu = U\Lambda\Delta \quad \text{and} \quad \Sigma_{ij} = \frac{\sigma_{\mathrm{ou}}^2}{2\alpha_{\mathrm{ou}}} e^{-2\alpha_{\mathrm{ou}} d_{ij}} (1 - e^{-2\alpha_{\mathrm{ou}} t_{ij}}). \tag{2}$$
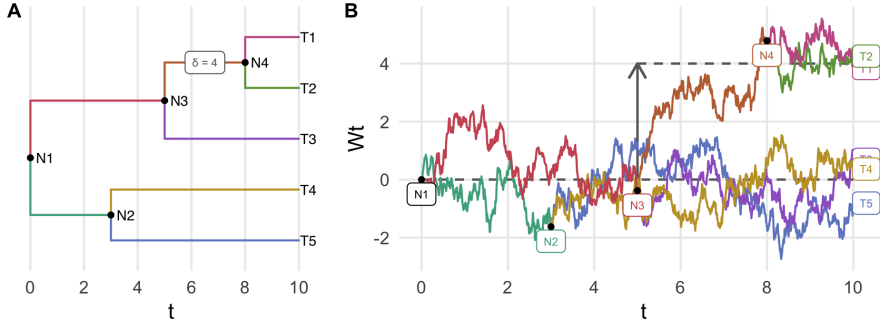
Fig. 1 **(A)** Phylogenetic tree with 5 tips and 4 internal nodes (root N_1 included). A shift occurs on the branch leading to N_4. **(B)** Ornstein-Uhlenbeck process with shifts on the tree defined in the left panel. At each node, the process spawns two independent process with the same initial value. The shifts on the optimal value on the branch leading to N_4 results in a different mean value for N_4 and all its offsprings (T_1 and T_2).

$$
T = \begin{array}{c} \\ T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \end{array}
\begin{array}{ccccccccc}
N_1 & N_2 & N_3 & N_4 & T_1 & T_2 & T_3 & T_4 & T_5 \\
\left(\begin{array}{ccccccccc}
1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right)
\end{array}
\qquad
\Delta = \begin{array}{c}
b_{N_1} \\ b_{N_2} \\ b_{N_3} \\ b_{N_4} \\ b_{T_1} \\ b_{T_2} \\ b_{T_3} \\ b_{T_4} \\ b_{T_5}
\end{array}
\left(\begin{array}{c}
0 \\ 0 \\ 0 \\ \delta \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{array}\right)
\qquad
\mu = \begin{array}{c}
\mu_{T_1} \\ \mu_{T_2} \\ \mu_{T_3} \\ \mu_{T_4} \\ \mu_{T_5}
\end{array}
\left(\begin{array}{c}
\delta \Lambda_{N_4} \\ \delta \Lambda_{N_4} \\ 0 \\ 0 \\ 0
\end{array}\right)
$$

Fig. 2 Incidence matrix T, shift vector Δ and mean vector μ associated with Fig. 1. $\Lambda_{N_4} = 1 - e^{\alpha_{\mathrm{ou}}(h - t_{N_3})}$ is the shrinkage parameter from equation (1).

When $\mathcal{T}$ is known, the matrix $T = U\Lambda$ is completely specified up to parameter α_{ou}. The shifted Ornstein-Uhlenbeck model, with parameters α_{ou}, σ_{ou}^2 and shift vector Δ, has beeen used (Bastide et al., 2017; Khabbazian et al., 2016) to find adaptive events, modeled as non zero values in Δ, in the evolution of continuous traits of interest (turtle shell size, great monkey brain shape, etc). In this work, we apply the same mathematical framework to the joint distribution of p-values transformed to z-scores.

3.2 Procedure

We show here how to use the previously described Ornstein-Uhlenbeck process to incorporate the tree structure $\mathcal{T}$ in the correction of the p-values vector $\mathfrak{p}$.

*Framework.*We first convert the p-values to z-scores using the quantile function Φ^{-1} of the standard gaussian:

$$
\mathfrak{z} = \Phi^{-1}(\mathfrak{p}).
$$

Provided the use of a correct statistical test, we known that $\mathfrak{p}_i \sim \mathcal{U}([0,1])$ under H_0, so that $\mathfrak{z}_i \sim \mathcal{N}(0,1)$. We also know that $\mathfrak{p}_i \preccurlyeq \mathcal{U}([0,1])$ under H_1. We make two assumptions regarding the distribution of $\mathfrak{z}$.

(A1) Under H_1, $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$ where $\mu_i \leq 0$;
(A2) $\mathfrak{z}$ arises from a shifted Ornstein-Uhlenbeck process on a $\mathcal{T}$ with parameters α_{ou}, Δ_{ou} and Δ.

Assumption (A1) is very classic when working with z-scores (McLachlan and Peel, 2000): finding the alternative hypotheses is equivalent to finding the non-zeros entries of μ. Assumption (A2) allows us to specify the joint distribution of $\mathfrak{z}$ as:

$$
\mathfrak{z} \sim \mathcal{N}_m(\mu, \Sigma) \tag{3}
$$

where Σ is fully specified by the parameters σ_{ou} and α_{ou}. Note that the diagonal coefficients of Σ are all equal to $\sigma_{\mathrm{ou}}^2 / 2\alpha_{\mathrm{ou}}(1 - 2e^{-2\alpha_{\mathrm{ou}}h})$. As they correspond

to marginal variances, this forces the equality $\sigma_{\mathrm{ou}}^2 = (1 - 2e^{-2\alpha_{\mathrm{ou}}h})/2\alpha_{\mathrm{ou}}$ so that Σ depends only on α_{ou}, *i.e.* $\Sigma = \Sigma(\alpha_{\mathrm{ou}})$. Finally, the decompositon $\mu = T\Delta$, where T acts as a phylogenetic design matrix, ensures that alternative hypotheses are likely to form clades, *i.e.* groups of tips obtained by cutting a single branch in the tree.

This framework allows us to use $\mathcal{T}$ as a prior structure in the mean vector μ and variance matrix Σ and to recast the hypothesis testing problem as a regression problem.

3.2.1 Parameter Estimation

*Estimation of $\hat{\mu}$.*Assume first that Σ, or equivalently α_{ou}, is known. Our main goal is to estimate the negative components of μ. To leverage the known tree structure, we use the decomposition $\mu = T\Delta$ and estimate μ by means of Δ. Since Δ has dimension n compared to dimension m for μ, we force $\hat{\Delta}$ to be sparse using a lasso penalty (Tibshirani, 1996) :

$$\hat{\Delta} = \operatorname*{argmin}_{\Delta \in \mathbb{R}^n \text{ s.t. } T\Delta \in \mathbb{R}^m_-} \frac{1}{2} \|\mathfrak{z} - T\Delta\|^2_{\Sigma^{-1},2} + \lambda\|\Delta\|_1. \tag{4}$$

Intuitively, the decomposition together with the ℓ_1 penalty works as a nested group lasso penalty for the components of μ, where the groups correspond to clades of $\mathcal{T}$, while the constraint $T\Delta \in \mathbb{R}^m_-$ forces components of μ to be non positive. For compacity, we note the feasible set $\mathcal{D} = \{\Delta \in \mathbb{R}^n \text{ s.t. } T\Delta \in \mathbb{R}^m_-\}$. Finally, we use the Cholesky decomposition $\Sigma^{-1} = R^T R$ to simplify the problem into the very well studied optimisation problem:

$$\hat{\Delta} = \operatorname*{argmin}_{\Delta \in \mathcal{D}} \frac{1}{2} \|y - X\Delta\|^2_2 + \lambda\|\Delta\|_1 \tag{5}$$

with $y = R\mathfrak{z} \in \mathbb{R}^m$ and $X = RT \in \mathbb{R}^{m \times n}$. This is a lasso problem with a convex feasability constraint on Δ. The optimisation algorithm used to solve this problem is detailed in 4.

*Estimation of $\hat{\Sigma}$ and tuning of λ.*Remember first that Σ is completely determined by α_{ou} because of the link between α_{ou} and σ_{ou}^2. There are no closed-form expression for the maximum likelihood estimator of α_{ou}. We therefore resort to numerical optimisation. To tune the parameter λ, we test several values to estimate models with different sparsity levels and select the best one using a BIC criterion.

$$(\hat{\alpha}_{\mathrm{ou}}, \hat{\lambda}) = \operatorname*{argmin}_{\alpha > 0, \lambda \geq 0} \|\mathfrak{z} - T\Delta_{\alpha,\lambda}\|^2_{\Sigma^{-1}(\alpha),2} + \log|\Sigma(\alpha)| + \|\Delta_{\alpha,\lambda}\|_0 \log m \tag{6}$$

where $\Delta_{\alpha,\lambda}$ is the solution of problem (4) for $\Sigma(\alpha)$ and λ. In practice, α and λ vary in a bidimensional grid and we select the values that minimize the objective.

3.2.2 Confidence intervals

Lasso procedures are known to produce biased estimators and do no return
confidence intervals for the point estimate $\hat{\mu}_i$. Instead of simply returning all
negative components of $\hat{\mu} = T\hat{\Delta}$, we first unbias the estimates and construct
confidence intervals for the components of Δ, and in turn of $\hat{\mu}$, using the
debiasing procedure of Javanmard and Montanari (2013, 2014); Zhang and
Zhang (2014).

Debiasing. All debiasing procedures assume a model $Y \sim \mathcal{N}_m\left(X\Delta, \sigma^2 I_m\right)$ and
require both an initial estimator $\hat{\Delta}^{(\text{init})}$ of Δ and $\hat{\sigma}$ of σ. We use the scaled
lasso (Sun and Zhang, 2012) with the same negativity constraint as in (4):

$$\left(\hat{\Delta}^{(\text{init})}, \hat{\sigma}\right) = \operatorname*{argmin}_{\Delta \in \mathcal{D}, \sigma > 0} \frac{\|y - X\Delta\|_2^2}{2\sigma m} + \frac{\sigma}{2} + \lambda_{scaled}\|\Delta\|_1. \tag{7}$$

Problem (7) can be solved efficiently by iterating between updates of (i) $\hat{\sigma}$ using
the closed-form expression $\hat{\sigma} = \|y - X\hat{\Delta}\|_2/\sqrt{m}$ and (ii) of $\hat{\Delta}$ by solving the
constrained lasso problem (5) with tuning parameter $\lambda_{scaled} = \lambda m \hat{\sigma}$. Debiasing
is achieved by the corrected update:

$$\hat{\Delta}_j = \hat{\Delta}_j^{(\text{init})} + \frac{\langle s_j, y - X\hat{\Delta}^{(\text{init})}\rangle}{\langle s_j, x_j\rangle}. \tag{8}$$

where the s_j form a score-system (SS). Intuitively, s_j should form a relaxed
orthogonalization of x_j against other column-vectors of X. The s_j are used to
decorrelate the estimators. We used the strategy of Zhang and Zhang (2014)
and take the residuals of a lasso regression of x_j against X_{-j}. We also con-
sidered the alternative debiasing strategy of Javanmard and Montanari (2013,
2014), which is based on a pseudo-inverse of $\hat{\Sigma} = \frac{X^T X}{m}$. Their debiased esti-
mate is again a simple update of the initial scaled lasso estimator:

$$\hat{\Delta} = \hat{\Delta}^{(\text{init})} + \frac{1}{m} S X^T \left(y - X\hat{\Delta}^{(\text{init})}\right)$$

but the decorrelation matrix S is computed in a so-called colwise inverse ap-
proach (CI), by inverting $\hat{\Sigma}$ in a columnwise fashion. Column s_j is solution of
the optimization problem :

$$\begin{cases} s_j = \operatorname{argmin}_{s \in \mathbb{R}^n} \ s^T \hat{\Sigma} s \\ \text{s.t. } \|\hat{\Sigma} s - e_j\|_\infty \leq \gamma. \end{cases} \tag{9}$$

where e_j is the jth canonical vector and $\gamma \geq 0$ is a slack hyperparameter. If γ
is too small, the problem is not feasible (unless $\hat{\Sigma}$ is non singular). If γ is too
large, the unique solution is $s_j = 0$.

*Confidence Interval.*Zhang and Zhang (2014) showed that asymptotically $\hat{\Delta} \sim \mathcal{N}(\Delta, V)$ with the covariance matrix V defined by

$$v_{ij} = \hat{\sigma}^2 \frac{\langle s_i, s_j \rangle}{\langle s_i, x_i \rangle \langle s_j, x_j \rangle}.$$ (10)

Similarly, the columnwise-inverse estimator of Javanmard and Montanari (2013) has asymptotic distribution $\mathcal{N}(\Delta, V)$ with variance matrix $V = S\hat{\Sigma}S^T/m$. For both procedures, the bilateral confidence interval at level α for $\hat{\Delta}_j$ is

$$IC_\alpha(\hat{\Delta}_j) = \left[\hat{\Delta}_j \pm \phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{v_{jj}} \right].$$

Note that the estimator of the ith component of μ can be written $\hat{\mu}_i = t_{i.}^T \hat{\Delta}$ with $t_{i.}^T$ the ith row of T. Its unilateral confidence intervals at level α is thus given by $\left[-\infty, \hat{\mu}_i + \sqrt{t_{i.}^T V t_{i.}}\, \phi^{-1}(1 - \alpha) \right]$. We can thus simply check whether 0 falls in the interval to test $\mathcal{H}_{i0} : \{\mu_i = 0\}$ versus $\mathcal{H}_{i1} : \{\mu_i < 0\}$ at level α or compute the p-value of the one-sided test as:

$$\mathfrak{p}_i^{\text{ss}} = \Phi\left(\frac{t_{i.}^T \hat{\Delta}}{\left(t_{i.}^T V t_{i.}\right)^{1/2}} \right).$$ (11)

3.2.3 FDR control

The debiasing procedure achieves marginally consistent interval estimation of the shifts Δ but additional care is required to control the FDR when testing all components of μ simultaneously. We use the procedure proposed in Javanmard et al. (2019), which is specific to debiased lasso estimators, and relies on the t-scores $\mathfrak{t}_i = \frac{t_{i.}^T \hat{\Delta}}{\left(t_{i.}^T V t_{i.}\right)^{1/2}}$. Briefly, for FDR control at a given level α, note $t_{\max} = \sqrt{2\log m - 2\log\log m}$ and set:

$$t^\star = \inf\left\{ 0 \le t \le t_{\max} : \frac{2m(1 - \Phi(t))}{R(t) \vee 1} \le \alpha \right\}$$

where $R(t) = \sum_{i=1}^m 1_{\{t_i \le -t\}}$ is the total number of rejections at threshold t, or $t^\star = \sqrt{2\log m}$ if the previous expression is non finite. Hypothesis $\mathcal{H}_{i0}$ is rejected if $\mathfrak{t}_i \le -t^\star$ or in term of q-values if

$$\mathfrak{q}_i^{\text{ss}} := \frac{\mathfrak{p}_i^{\text{ss}} \alpha}{\Phi(-t^\star)} \le \alpha.$$ (12)

Since $\mathfrak{t}$ itself depends on α, the corrected p-values depend on α, unlike in the standard BH procedure, where they only depend on the order statistics.

3.2.4 Algorithm

The algorithm 1 summarise our procedure.

Algorithm 1 Zazou procedure

1: Compute the vector $\mathfrak{p}$ of raw p-values
2: Transform it to the vector z of raw z-scores
3: **for** values of α and λ varying in a grid **do**
4:     Compute Σ, R, y and X
5:     Compute $\hat{\Delta}_{\alpha,\lambda}$ and $\hat{\sigma}_{\alpha,\lambda}$ by solving (7)
6:     Compute the BIC criterion (6)
7: **end for**
8: Select parameter values $\hat{\alpha}$ and $\hat{\lambda}$ that minimize the BIC
9: Set $\hat{\Delta}^{(\mathrm{init})} = \hat{\Delta}_{\hat{\alpha},\hat{\lambda}}$
10: Update $\hat{\Delta}^{(\mathrm{init})}$ according to (8) to debias it
11: Compute its covariance matrix $\hat{V}$ with (10)
12: Compute the vector p-values $\mathfrak{p}^{\mathrm{ss}}$ of corrected with (11)
13: **return** Vector of corrected q-values $\mathfrak{q}^{\mathrm{ss}}$ computed from (12) for a target FDR level α.

---

4 Sign-constrained lasso

Our inference procedure is based on very standard estimates but requires to solve the following constrained lasso problem:

$$\hat{\Delta} = \operatorname*{argmin}_{\Delta \text{ s.t. } T\Delta \in \mathbb{R}^m_-} \frac{1}{2} \|y - X\Delta\|_2^2 + \lambda\|\Delta\|_1.$$

For arbitrary vector y and matrices X and T. This a convex problem as both the objective function and feasibility set are convex. We therefore adapt the shooting algorithm (Fu, 1998), an iterative algorithm used to solve the standard lasso by looping over coordinates and solving simpler unidimensional problem, to our constrained problem.

Note X_{-j} (resp. Δ_{-j}) matrix X (resp. vector Δ) deprived of its jth column (resp. j^{th} coordinate). We can isolate Δ_j in (5) and decompose the objective as $\|y - X\Delta\|_2^2 + \lambda|\Delta| = \|y - z_j - x_j\Delta_j\|_2^2 + \lambda|\Delta_j| + \lambda\|\Delta_{-j}\|_1$ where $z_j = X_{-j}\Delta_{-j} \in \mathbb{R}^m$. We can likewise decompose $T\Delta = u_j + v_j\Delta_j$ where $u_j = T_{-j}\Delta_{-j} \in \mathbb{R}^m$ and $v_j = t_j$. When updating Δ_j, we can thus consider the simpler univariate problem in θ:

$$\begin{cases} \operatorname*{argmin}_{\theta \in \mathbb{R}} h(\theta) = \dfrac{1}{2}\|y - z - x\theta\|_2^2 + \lambda|\theta| \\ \text{s.t. } u + v\theta \le 0. \end{cases} \tag{13}$$

Let $I_+ = \{i : v_i > 0\}$ and $I_- = \{i : v_i < 0\}$ and note $\theta_{\max} = \min_{I_+}\{-u_i/v_i\}$ and $\theta_{\min} = \max_{I_-}\{-u_i/v_i\}$ with the usual conventions that $\max(\emptyset) = -\infty$ and $\min(\emptyset) = +\infty$. Problem (13) is feasible if only if (i) $\theta_{\min} \le \theta_{\max}$ and (ii) for all i, $v_i = 0 \Rightarrow u_i \le 0$, in which case the feasible region is $[\theta_{\min}, \theta_{\max}]$. Computing the subgradient $\partial h(\theta)$ of h and looking for values θ such that $0 \in \partial h(\theta)$ leads to the usual shrinked estimates:

$$\begin{cases} \frac{(y-z)^T x + \lambda}{x^T x} & \text{if } (y - z)^T x < -\lambda, \\ \frac{(y-z)^T x - \lambda}{x^T x} & \text{if } (y - z)^T x > \lambda, \\ 0 & \text{if } |(y - z)^T x| < \lambda. \end{cases}$$

By convexity of h, the solution of (13) can be found by projecting the previous unconstrained minimum to the feasibility set. If problem (13) is feasible, its solution is thus given by

$$
\theta^\star = \begin{cases} P_\mathcal{I}\left(\frac{(y-z)^T x + \lambda}{x^T x}\right) & \text{if } (y-z)^T x < -\lambda, \\ P_\mathcal{I}\left(\frac{(y-z)^T x - \lambda}{x^T x}\right) & \text{if } (y-z)^T x > \lambda, \\ P_\mathcal{I}(0) & \text{if } |(y-z)^T x| < \lambda, \end{cases}
$$

where $P_\mathcal{I} : u \mapsto \max(\theta_{\min}, \min(u, \theta_{\max}))$ is the projection of u on the segment $\mathcal{I} = [\theta_{\min}, \theta_{\max}]$.

5 Synthetic Data

5.1 Metagenomics

Metagenomics data are made up of three components. The first component is the count or abundance matrix $X = (x_{ij})$, with $1 \leq i \leq m$ and $1 \leq j \leq p$, which represents the quantity of taxa i in sample j. The second component is a set of sample covariates, such as disease status, environmental conditions, group, etc. The final component is a phylogenetic tree which captures the shared evolutionary history of all taxa. When performing DAT, we are interested in taxa whose abundance is significantly associated to a covariate.

Most DAT procedures proceed with univariate tests (one test per species) followed by a correction procedure. In the synthetic datasets, we consider discrete covariates only. Since our goal is to compare correction procedures, we use Wilcoxon or Kruskall-Wallis tests, which are classical and widespread non parametric tests in Metagenomics.

5.2 Simulations

*Simulation scheme.* We use the following simulation scheme:

1. start with an homogeneous dataset,
2. assign each sample to group A or B at random
3. select differentially abundant taxa in a phylogenetically consistent manner (diffentially abundant taxa)
4. apply a fold-change to the observed abundance of diffentially abundant taxa in group B.

This non-parametric simulation scheme was previously used in Bichat et al. (2020). We considered two variants for step 3, respectively called *positive* and *negative*. In the negative variant, differentially abundant taxa were selected randomly across the tree, so that the phylogeny is not informative. In the positive variant, taxa are instead selected in a phylogenetically consistent manner. Formally, the phylogeny was first used to compute the cophenetic (Sneath

et al., 1973) distance matrix between taxa. A partioning around medoids algorithm was then used to create cluster of related species. One or more clusters were then picked at random and all species in those clusters were selected as differentially abundant.

For each fold-change (fc ∈ {3, 5, 10}), 500 simulated datasets were created, with a proportion of differentially abundant species ranging from 3 % to 35 %. For each simulation, we corrected p-values using no correction (raw), BH procedure (bh), BY procedure (by), `StructFDR` (tf) or our procedure with either score system (ss) or colwise inverse debiasing (ci), targeting in all instances a 5% FDR level. We compared the 6 procedures in terms of True Positive Rate (TPR), nominal FDR and AUC (Area Under the Curve).

*Positive simulations.*The results of positive simulations (*i.e.* where the phylogeny is informative) are shown in Figure 3. All correction methods have control the FDR at the target rate or below when the fold change is larger than 5. For smaller fold changes, both SS and CI variations of zazou exhibit nominal FDR slightly above the target level (up to 9% in the worst case). In all settings, BY had the lowest TPR, whereas TF was comparable to vanilla BH, in line with results of Bichat et al. (2020). Finally, zazou (both SS and CI variations) had the best overall TPR, with largest gains observed in the lowest fold-change setting.
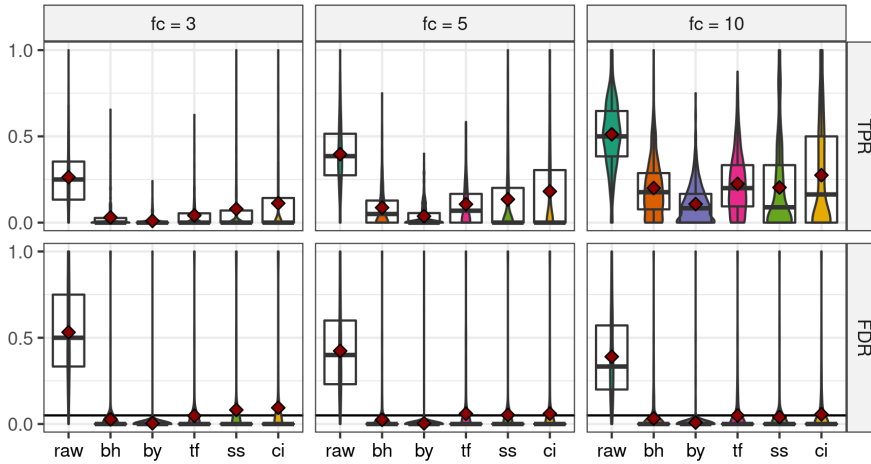
Fig. 3 Boxplots and average (red point) TPR and FDR across positive simulation settings. Each facet corresponds to a different fold-change and each boxplot is computed over 500 simulation replicates. All correction control the FDR at the target level or slightly above but zazou (SS and CI) achieve higher TPR, especially for small fold changes.

The higher than intended FDR of zazou methods suggests that the problem of finding an adequate threshold for $\mathfrak{p}_i^{ss}$ not completely solved by Javanmard

et al. (2019) procedure. To assess the performance of zazou in a threshold-independent manner, we also compared the AUC of all procedures. Fig. ref-fig:aucroc shows that zazou (both variants) has higher AUC than all other methods. As reported previously, TF and BH have are at the same level and BY has the lowest ROC curve. Focus on the beginning of left hand side side of the curve shows that zazou is more efficient starting from the first discoveries.
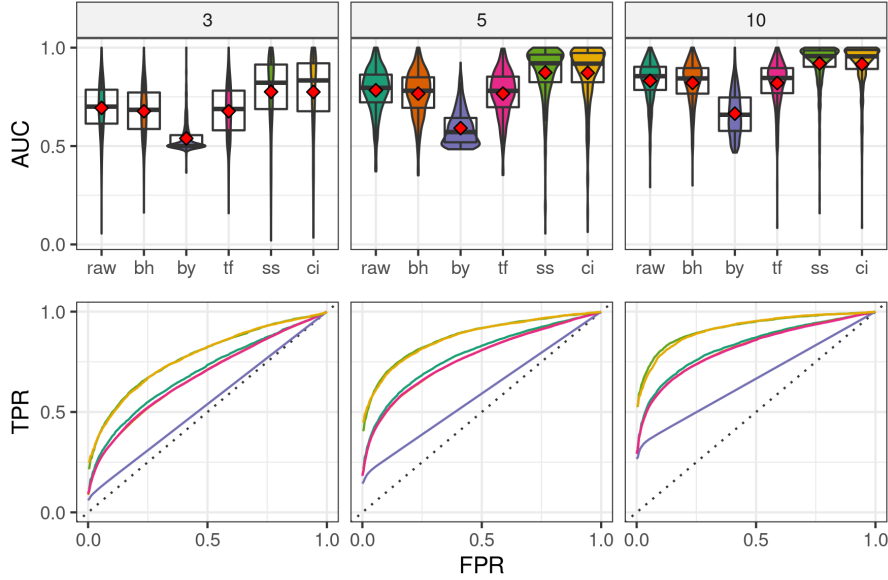
Fig. 4 AUC boxplots (top) and average ROC curves (bottom) across positive simulations settings. Facets correspond to fold-changes. ROC curves are computed for each simulation and linearly interpolated over a fixed grid before being averaged. Each boxplot and each curve are computed over 500 replicates. In all settings, SS/CI have the highest AUC / ROC curve, followed by BH/TF while BY has the lowest values.

*Negative simulations.*The negative simulations are designed to assess the robustness of our algorithm with respect to uninformative phylogenies, or equivalently mispecified hierarchies. Fig. 5 shows that, as expected, standard BH outperforms competing methods (in terms of AUC) when the tree is mispecified. Forcing an inadequate tree structure results in AUC losses ranging from 15 to 20 points compared to no structure. The puzzling lack of AUC loss for the TF procedure is explained by an implementation trick: TF always performs BH correction in parallel to its hierarchical procedure and falls back to BH when the hierarchical procedure detects much fewer species than BH (Bichat et al., 2020; Xiao et al., 2017).
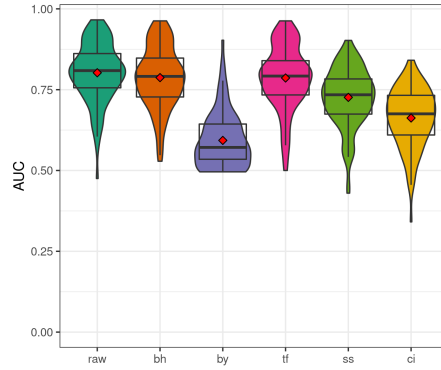
Fig. 5 AUC boxplots (computed over 500 replicates) in negative simulations. BH outperforms SS and CI, highlighting the cost of imposing a mispecified hierarchical structure.

6 Application

We use our zazou procedure on a gut microbiota dataset from the Fiji Islands (Brito et al., 2016; Pasolli et al., 2017) to identify species that are differentially abundant between adults and children. The data sets consists in the abundances of $p = 387$ species among $n = 146$ islanders, split into 112 adults and 34 children.

To mimick the simulation study, we used Wilcoxon tests for the univariate tests. Without correction, 21 species were detected as differentially abundant at the 5% level. None of them remained significant after correction by BH, BY, TreeFDR (tf) or treeclimbR. By contrast, zazou detected differentially abundant species with both desparsification methods: 17 for SS and 6 for CI. Fig. 6 shows that they are not a strict subset of the 21 detected with no correction. Smoothing salvages some species that are closely related to one of the 21 without being significant on their own (red box in the figure). It also illustrate some numerical problems associated with colwise-inverse debiasing, which is highly sensitive to the choice of the slack hyperparameter γ. The window of relevant values for γ is narrow and too large or too small values γ respectively lead to no correction or a faulty p-value correction.

7 Conclusion

In this work, we introduced `zazou`, a new method for correcting p values in a hierarchical context. `zazou` is based on recasting the testing problem as a regression problem, under the framework of stochastic processes on tree, and using the tree topology as a regularization parameter.

It outperforms competing methods, hierarchical (TreeFDR, TreeclimbR) or not (BH, BY) in terms of AUC but this does not translate immediately to superior results in terms of FDR and TPR. The threshold for rejecting hy-
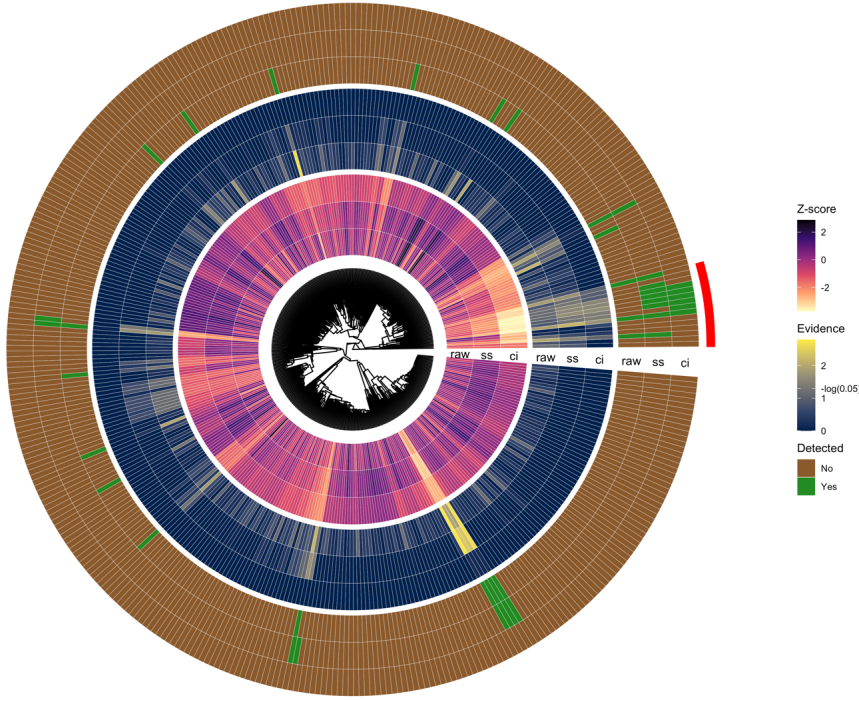
Fig. 6 Phylogeny of the 387 species from the Fidji dataset with associated z-scores (inner circle), evidence (middle circle) and detection status (outer circle) under different correction procedures. Species detected by zazou are generally close-by on the tree and often, but not always, detected by raw p-values. The red shape highlight species that are not detected but univariate tests but related enough to detected species to be salvaged by the hierarchical procedure.

potheses is turned out to be quite difficult to calibrate while controling the FDR and warrants further work.

There are several other parts of the procedure that are not as powerful as
380 expected. First, the BIC step used to select λ and in turn the number of shifts tends to choose models with very few shifts, and sometimes even none. In such instances, the relevance of the debiasing step is limited. Second, the correction procedure proposed by Javanmard et al. (2019) is too conservative for our purpose. It was indeed developed to control both the FDR and the directional
385 FDR (*i.e.* proportion of Type S errors, where the effect size have the wrong sign, in the discoveries) whereas we only need to control the former. For both these steps, specific developments taking into account the sign constraint on $\hat{\mu}$ and the structure of the topology matrix of tree $\mathcal{T}$ could lead to better performances for zazou.

References

Christophe Ambroise, Alia Dehman, Pierre Neuvial, Guillem Rigaill, and Nathalie Vialaneix. Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. *Algorithms for Molecular Biology*, 14(1):22, 2019.

Paul Bastide, Mahendra Mariadassou, and Stéphane Robin. Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1067–1093, 2017.

Yoav Benjamini and Ruth Heller. False discovery rates for spatial signals. *Journal of the American Statistical Association*, 102(480):1272–1281, 2007.

Yoav Benjamini and Yosef Hochberg. 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, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

Antoine Bichat, Jonathan Plassais, Christophe Ambroise, and Mahendra Mariadassou. Incorporating phylogenetic information in microbiome differential abundance studies has no effect on detection power and fdr control. *Frontiers in Microbiology*, 11:649, 2020. ISSN 1664-302X. doi: 10.3389/fmicb.2020. 00649. URL https://www.frontiersin.org/article/10.3389/fmicb. 2020.00649.

J Martin Bland and Douglas G Altman. Multiple significance tests: the bonferroni method. *Bmj*, 310(6973):170, 1995.

Ilana L Brito, S Yilmaz, K Huang, Liyi Xu, Stacy D Jupiter, Aaron P Jenkins, Waisea Naisilisili, M Tamminen, CS Smillie, Jennifer R Wortman, et al. Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 535(7612):435–439, 2016.

William S Bush and Jason H Moore. Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012.

Jun Chen. *StructFDR: False Discovery Control Procedure Integrating the Prior Structure Information*, 2018. URL https://CRAN.R-project.org/ package=StructFDR. R package version 1.3.

Henk R Cremers, Tor D Wager, and Tal Yarkoni. The relation between statistical power and inference in fmri. *PloS one*, 12(11):e0184923, 2017.

Simon B Eickhoff, Bertrand Thirion, Gaël Varoquaux, and Danilo Bzdok. Connectivity-based parcellation: Critique and implications. *Human brain mapping*, 36(12):4771–4792, 2015.

Robert P. Freckleton, Paul H. Harvey, and Mark Pagel. Bergmann's rule and body size in mammals. *The American Naturalist*, 161(5):821–825, May 2003. doi: 10.1086/374346. URL https://doi.org/10.1086/374346.

Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

Jelle J Goeman and Livio Finos. The inheritance procedure: multiple testing of tree-structured hypotheses. *Statistical applications in genetics and molecular biology*, 11(1):1–18, 2012.

Ruizhu Huang, Charlotte Soneson, Pierre-Luc Germain, Thomas Schmidt, Christian von Mering, and Mark Robinson. treeclimbr pinpoints the data-dependent resolution of hierarchical hypotheses. 06 2020. doi: 10.1101/2020. 06.08.140608.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional statistical models. In *Advances in Neural Information Processing Systems*, pages 1187–1195, 2013.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Adel Javanmard, Hamid Javadi, et al. False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1):1212–1253, 2019.

Mohammad Khabbazian, Ricardo Kriebel, Karl Rohe, and Cécile Ané. Fast and accurate detection of evolutionary shifts in ornstein–uhlenbeck models. *Methods in Ecology and Evolution*, 7(7):811–824, 2016.

Kyung In Kim, Etienne Roquain, and Mark A van de Wiel. Spatial clustering of array cgh features in combination with hierarchical multiple testing. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

Russell Lande. Natural Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution*, 30(2):314–334, June 1976. doi: 10.1111/j.1558-5646. 1976.tb00911.x. URL https://doi.org/10.1111/j.1558-5646.1976. tb00911.x.

G McLachlan and D Peel. Finite mixture models.,(john wiley & sons: New york.). 2000.

Geoffrey J McLachlan, Kim-Anh Do, and Christophe Ambroise. *Analyzing microarray gene expression data*, volume 422. John Wiley & Sons, 2005.

Nicolai Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278, 2008.

Edoardo Pasolli, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, Faizan Malik, Marcel Ramos, Jennifer B Dowd, et al. Accessible, curated metagenomic data through experimenthub. *Nature methods*, 14(11):1023, 2017.

Kris Sankaran and Susan Holmes. structssi: simultaneous and selective inference for grouped or hierarchically structured data. *Journal of statistical software*, 59(13):1, 2014.

Nicola Segata, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S Garrett, and Curtis Huttenhower. Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6):1–18, 2011.

Matteo Sesia, Eugene Katsevich, Stephen Bates, Emmanuel Candès, and Chiara Sabatti. Multi-resolution localization of causal variants across the genome. *Nature communications*, 11(1):1–10, 2020.

Peter HA Sneath, Robert R Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 09 2012. ISSN 0006-3444. doi: 10.1093/biomet/ass043. URL `https://doi.org/10.1093/biomet/ass043`.

Zheng-Zheng Tang, Guanhua Chen, Alexander V Alekseyenko, and Hongzhe Li. A general framework for association analysis of microbial communities on a taxonomic tree. *Bioinformatics*, 33(9):1278–1285, 2017.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

John Wilder Tukey. The problem of multiple comparisons. *Multiple comparisons*, 1953.

Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

Jian Xiao, Hongyuan Cao, and Jun Chen. False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics*, 33(18):2873–2881, 2017.

Daniel Yekutieli. Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103(481):309–316, 2008.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.