ORIGINAL PAPER

Hierarchical correction of p-values via an ultrametric 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), in functional Magnetic Resonance Imaging (fMRI), the goal is to identify voxels that are significantly activated in two different conditions (Cremers et al. 2017).

✉ Mahendra Mariadassou
  mahendra.mariadassou@inrae.fr

[1] LaMME, Université d'Évry val d'Essonne, 91000 Évry, France

[2] Enterome, 94-96 Avenue Ledru Rollin, 75011 Paris, France

[3] MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France

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, such as the popular Benjamini-Hochberg (BH) procedure, rely on independence, or some form of weak dependence, among the hypotheses, 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 remains valid under arbitrary dependency structures by replacing the BH procedure with the more conservative BY procedure of 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, as discussed in depth in Blanchard et al. (2020).

An alternative approach for dealing with multiple testing is to reduce the number of tests by aggregating certain hypotheses. Aggregation strategies vary and can be based on a priori knowledge (*e.g.* metabolic pathways, functional modules of genes) or on clustering algorithms (Sankaran and Holmes 2014; Renaux et al. 2020).

This article aims to take into account the dependencies between variables in order to offer a powerful statistical procedure of multiple testing. A hierarchical dependency structure between variables is assumed to be known up to certain constants. This assumption is common in our motivating example of microbiome studies (Sankaran and Holmes 2014; Xiao et al. 2017; Huang et al. 2021; Matsen IV and Evans 2013; Silverman et al. 2017), where the phylogeny is a natural hierarchical structure encoding similarities between variables (or namely species in that context). 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 then consider a three stage approach for our differential analysis procedure. The first stage reframes the initial problem as a linear regression problem that preserves the hierarchical structure. This linear problem is ill defined ($p \sim 2n$) and we therefore resort to an ℓ_1 penalized estimation of the mean of the Ornstein-Uhlenbeck process. The second stage produces asymptotically valid p-values. The output of ℓ_1 penalized

estimation produces are indeed biased and offer no theoretical guarantees about their asymptotic distribution; we therefore correct them using a debiasing procedure (Javanmard and Montanari 2013, 2014; Zhang and Zhang 2014) to compute valid p-values. The third and final stage controls the FDR of the overall procedure, using the tuning strategy of Javanmard et al. (2019).

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 hierarchical procedures in Sect. 2, introduce the model and statistical procedure in Sect. 3 and detail the computational steps in Sect. 4. The performances of the algorithm are assessed via simulations in Sect. 5. The use of the proposed model is illustrated in Sect. 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 a prior). 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; Renaux et al. 2020).

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 enables 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 procedures, 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 typically is 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.

2.3 Hierarchical testing

The Hierarchical FDR (hFDR) introduced by Yekutieli (2008) and implemented in the R package `structSSI` (Sankaran and Holmes 2014), proposes a top-down algorithm to sequentially reject hypotheses organized in a tree. The same approach is used in (Renaux et al. 2020) to select a group of variables arranged in a clustering tree. However, this approach suffers from some limitations, as shown in (Bichat et al. 2020; Huang et al. 2021). First, the algorithm in its 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 (user-defined) *a priori* FDR level and the structure of rejected nodes. This makes it difficult to calibrate the *a priori* FDR that would achieve a target *a posteriori* FDR and thus to compare it to other correction methods. Finally, it does not produce a corrected *p*-value, or *q*-value, per leaf, but only a *reject / no reject* decision and was shown in (Bichat et al. 2020) to perform no better than BH in many instances. Given all these drawbacks, we did not include the hFDR in our benchmark and use BH as a baseline instead.

`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 $\mathbf{C}_\rho = \left(\exp\left(-2\rho \mathbf{D}_{i,j} \right) \right)$ 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 q^{sf}. This method is implemented in the `StructFDR` package (Chen 2018).

`TreeclimbR` (Huang et al. 2021) 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 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 (also called drift) β_{ou}, selection strengh (also called mean reversion parameter) α_{ou} and variance of the white noise σ^2_{ou}, is a Gaussian process that satisfies the stochastic differential equation:

$$dW_t = -\alpha_{ou}(W_t - \beta_{ou})dt + \sigma_{ou}dB_t.$$

The important properties of OU processes are bounded variance and convergence to a stationary distribution centered on the optimal value β_{ou}, namely $W_t \xrightarrow{(d)} \mathcal{N}(\beta_{ou}, \sigma^2_{ou}/2\alpha_{ou})$ when $t \to \infty$. Thanks to those properties, OU processes have become a popular model applied in various subfields of biology, ranging from evolution of continuous traits, such as body mass (Freckleton et al. 2003), fitness (Lande 1976) or CpG enrichment in viral sequences (MacLean et al. 2021) to animal movement (Dunn and Gipson 1977) and epidemiology (Nåsell 1999). 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 leaves and n branches ($n = 2m - 1$ for binary trees). The internal nodes are labeled N_1 (the root) to N_{n-m} and the leaves T_1 to T_m. Let i be a node, W_i the value of the trait at that node and denote $pa(i)$ 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 $i \in \{T_1, \ldots, T_m\}$. For any pair of nodes (i, j), let t_{ij} be the time elapsed between the root and the most recent common ancestor of i and j and denote $d_{ij} = t_i - t_j - 2t_{ij}$ the distance in the tree between nodes i and j. The distribution of the trait at node i is given by:

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

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

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

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. Denote U the $m \times (n + m)$ incidence matrix of $\mathcal{T}$ with rows labeled by leaves ($i \in \{T_1, \ldots, T_m\}$) and columns labeled by inner nodes and leaves ($j \in \{N_1, \ldots, N_{n-m}, T_1, \ldots, T_m\}$), with entries defined as $U_{ij} = 1$ if and only if leaf i is in the subtree rooted at node j. Intuitively, column $U_{.j}$ encodes all leaves descending from node j and row $U_{i.}$ encodes all ancestors of leaf i. Denote Δ the dimension n column vector with entries defined as $\Delta_i = \beta_{\text{ou},i} - \beta_{\text{ou},pa(i)}$ where $i \in \{N_1, \ldots, N_{n-m}, T_1, \ldots, T_m\}$. Non-zero entries of Δ correspond to *shifts location*, nodes for which the optimal value $\beta_{\text{ou},i}$ differ from its parent's and their values to *shifts magnitude* (see Figure 2 for an example). Finally let Λ be the n diagonal matrix with diagonal entries $\Lambda_i = 1 - \exp(\alpha_{\text{ou}}(h - t_{pa(i)}))$ where $i \in \{N_1, \ldots, N_{n-m}, T_1, \ldots, T_m\}$. Straightforward computations (see Bastide et al. (2017) for detailed derivations) show that W is a gaussian vector with joint distribution:

$$W \sim \mathcal{N}(\mu, \Sigma) \quad \text{where} \quad \mu = U\Lambda\Delta \quad \text{and} \quad \Sigma_{ij} = \frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}} e^{-2\alpha_{\text{ou}}d_{ij}}(1 - e^{-2\alpha_{\text{ou}}t_{ij}}). \tag{2}$$

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 $\alpha_{\text{ou}}, \sigma_{\text{ou}}^2$ and shift vector
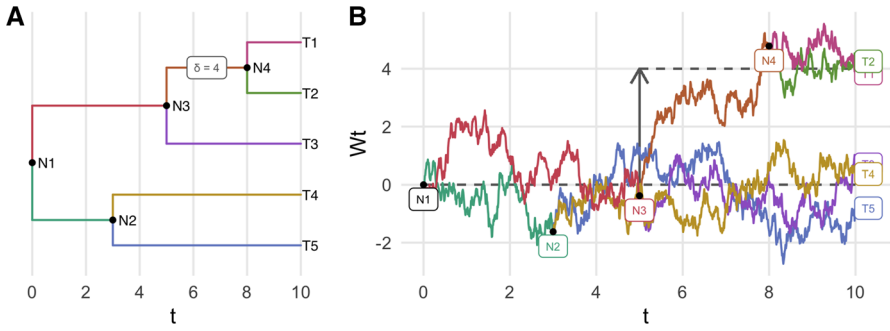
Fig. 1 **(A)** Phylogenetic tree with 5 leaves and 4 internal nodes (root N_1 included). A shift occurs on the branch leading to N_4. **(B)** Ornstein-Uhlenbeck process with shifts on the tree defined in the left panel. At each node, the process spawns two independent process with the same initial value. 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)

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

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

Δ, has been 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. Noting m_i^1 (resp. m_i^2) the median count (or relative abundance) of taxon i under condition 1 (resp. condition 2), we want to test $\mathcal{H}_{i0} : m_i^1 = m_i^2$ against $\mathcal{H}_{i1} : m_i^1 \neq m_i^2$ and assume that we have a testing procedure that outputs p-values, *e.g.* the Wilcoxon-Mann-Whitney test (Mann and Whitney 1947; Wilcoxon 1992). 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 $\mathcal{H}_{i0}$, so that $\mathfrak{z}_i \sim \mathcal{N}(0, 1)$. We also know that $\mathfrak{p}_i \preccurlyeq \mathcal{U}([0, 1])$ and thus $\mathfrak{z}_i \preccurlyeq \mathcal{N}(0, 1)$ under $\mathcal{H}_{i1}$. We could also test $\mathcal{H}_{i0} : m_i^1 = m_i^2$ against $\mathcal{H}_{i1} : m_i^1 < m_i^2$ or $\mathcal{H}_{i1} : m_i^1 > m_i^2$, we only require the procedure to output p-values that satisfy the previous distributional assumptions for these $\mathcal{H}_{i0}$ and $\mathcal{H}_{i1}$. Note that, even if the test statistic is itself a z-score before being transformed to a p-value, the z-score $\mathfrak{z}_i$ may differ from the *raw* test statistic z_i because of the intermediate p-value $\mathfrak{p}_i$. Indeed when considering the simple case of testing equality of means in two samples of size n, with gaussian distributions and known variance σ, the relation between $\mathfrak{z}_i$ and $z_i = \sqrt{n}(\hat{m}_i^1 - \hat{m}_i^2)/2\sigma$ is given by:

$$\mathfrak{z}_i = \Phi^{-1}(\mathfrak{p}_i) = \begin{cases} \Phi^{-1}(\Phi(z_i)) = z_i & \text{if } \mathcal{H}_{i1} : m_i^1 < m_i^2 \\ \Phi^{-1}(\Phi(1 - z_i)) = -z_i & \text{if } \mathcal{H}_{i1} : m_i^1 > m_i^2 \\ \Phi^{-1}(2\Phi(-|z_i|)) & \text{if } \mathcal{H}_{i1} : m_i^1 \neq m_i^2 \end{cases}$$

After transformation, the test can be thus always be reframed as one-sided on $\mathfrak{z}_i$: $\mathcal{H}_{i0} : E[\mathfrak{z}_i] = 0$ against $\mathcal{H}_{i1} : E[\mathfrak{z}_i] < 0$. We make two assumptions regarding the distribution of $\mathfrak{z}$.

(A1) Under $\mathcal{H}_{i1}$, $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$ where $\mu_i \leq 0$;
(A2) $\mathfrak{z}$ arises from a shifted Ornstein-Uhlenbeck process on an ultrametric tree $\mathcal{T}$ 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 negative 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 $\sigma_{\text{ou}}^2/2\alpha_{\text{ou}}(1 - 2e^{-2\alpha_{\text{ou}}h})$. As they correspond to marginal variances, this forces the equality $\sigma_{\text{ou}}^2 = (1 - 2e^{-2\alpha_{\text{ou}}h})/2\alpha_{\text{ou}}$ so that Σ depends only on α_{ou}, *i.e.* $\Sigma = \Sigma(\alpha_{\text{ou}})$. 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 leaves 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 constrained lasso penalty (Tibshirani 1996) :

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

where $\mathbb{R}_- = \{x \in \mathbb{R} \text{ s.t. } x \leq 0\}$.

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 define 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 $X = RT \in \mathbb{R}^{m \times n}$. Note that y is a whitened version of $\mathfrak{z}$, with independent components and spherical covariance matrix. This is a lasso problem with a convex feasability constraint on Δ. The optimisation algorithm used to solve this problem is detailed in Sect. 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^2_{ou}. 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 modified BIC criterion:

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

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. We use a modified BIC, where $\log(\log m) \log m$ replaces $\log m$, to account for the fact that m scales like n as suggested in Fan and Tang (2013).

3.2.2 Confidence intervals

Lasso procedures are known to produce biased estimators and do not 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 debias 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 $Y \sim \mathcal{N}_m(X\Delta, \sigma^2 I_m)$ 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):

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

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 considered 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 estimate is again a simple update of the initial scaled lasso estimator:

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

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

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

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

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

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:

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

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 $t_i = \frac{t_{i.}^T \hat{\Delta}}{\left(t_{i.}^T V t_{i.} \right)^{1/2}}$. Briefly, for FDR control at a given level α, let $t_{\max} = \sqrt{2 \log m - 2 \log \log m}$ and set:

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

where $R(t) = \sum_{i=1}^m 1_{\{t_i \leq -t\}}$ is the total number of rejections at threshold t, or $t^\star = \sqrt{2 \log m}$ if the previous expression is empty. Applying the procedure from Javanmard et al. (2019) strictly would replace $2m$ with m in the numerator, as we're considering one-sided tests instead of two-sided ones for μ_i. However, numerical analysis showed that the extra 2 led to better control of the FDR and we thus kept it. Hypothesis $\mathcal{H}_{i0}$ is rejected if $t_i \leq -t^\star$ or in term of q-values if

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

Since 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 summarises our procedure. We call it *zazou* for "**z**-scores **az O**rnstein-Uhlenbeck".

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 $\hat{\Delta}^{(\text{init})} = \hat{\Delta}_{\hat{\alpha},\hat{\lambda}}$
10: Update $\hat{\Delta}^{(\text{init})}$ according to (8) to debias it
11: Compute its covariance matrix $\hat{V}$ with (10)
12: Compute the vector p-values $\mathfrak{p}^{\text{ss}}$ of corrected with (11)
13: **return** Vector of corrected q-values $\mathfrak{q}^{\text{ss}}$ 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:

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

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.

Let X_{-j} (resp. Δ_{-j}) be the 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:

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

Let $I_+ = \{i : v_i > 0\}$ and $I_- = \{i : v_i < 0\}$ and denote $\theta_{max} = \min_{I_+}\{-u_i/v_i\}$ 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 only if (i) $\theta_{min} \leq \theta_{max}$ and (ii) for all i, $v_i = 0 \Rightarrow u_i \leq 0$, 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. Dozens of full-fledged testing pipelines are published each year, including some designed with omics data in mind. Since our goal is this study is to compare correction procedures rather than full testing 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 a 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 partitioning 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 Fig. 3. All correction methods have controlled 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.

The higher than intended FDR of `zazou` methods suggests that the problem of finding an adequate threshold for $\mathfrak{p}_i^{ss}$ is 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. 4 shows that `zazou` (both variants) has higher AUC than all other methods. As reported previously, TF and BH 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.

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 percentage points compared to no structure. The puzzling lack of AUC loss for the TF procedure is explained by an
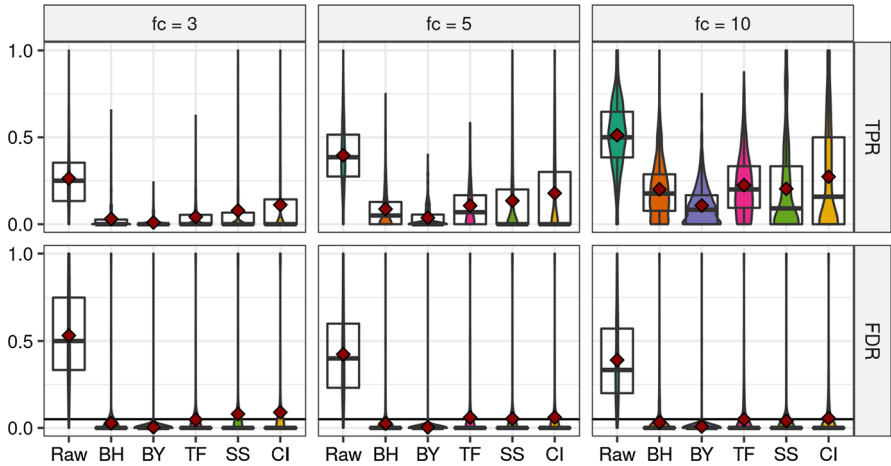
Fig. 3 Boxplots and average (red point) TPR and FDR across positive simulation settings. Each facet corresponds to a different fold-change (fc) and each boxplot is computed over 500 simulation replicates. All corrections control the FDR at the target level or slightly above but `zazou` (SS and CI) achieve higher TPR, especially for small fold changes
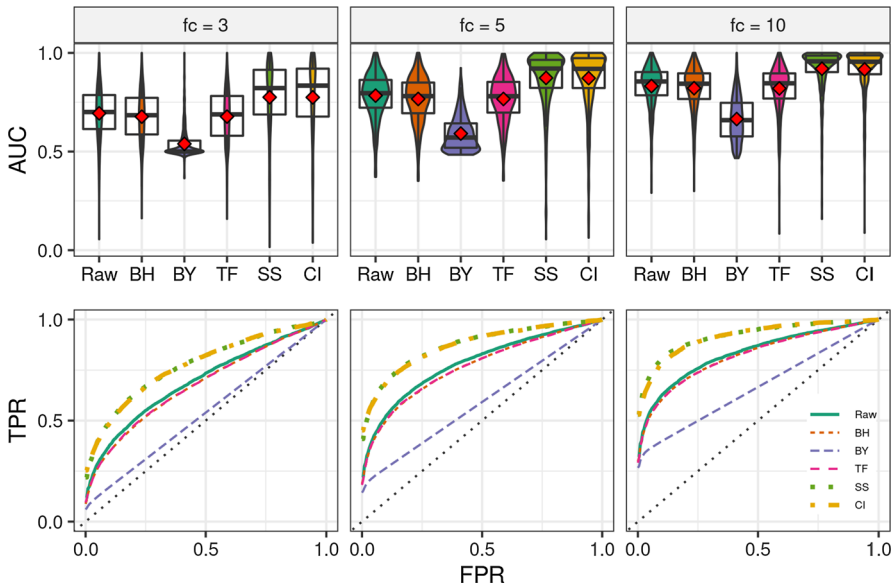

Fig. 4 AUC boxplots (top) and average ROC curves (bottom) across positive simulations settings. Facets correspond to fold-changes (fc). ROC curves are computed for each simulation and linearly interpolated over a fixed grid before being averaged. Each boxplot and each curve are computed over 500 replicates. In all settings, SS/CI have the highest AUC / ROC curve, followed by BH/TF while BY has the lowest values

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).
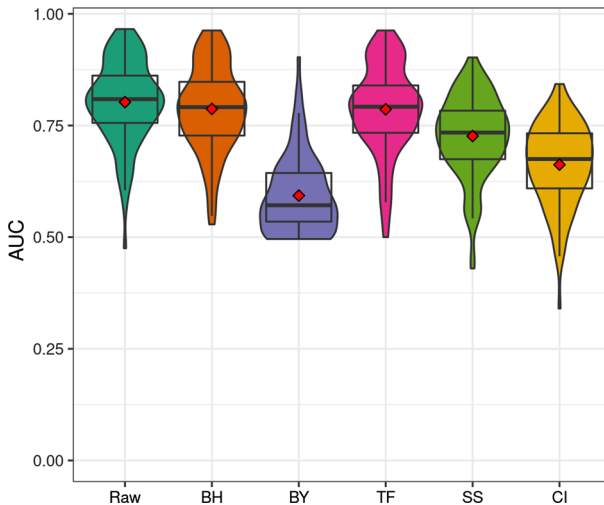
Fig. 5 AUC boxplots (computed over 500 replicates) in negative simulations. BH outperforms SS and CI, highlighting the cost of imposing a mispecified hierarchical structure

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 mimic 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 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 an ultrametric tree, and using the tree topology as a regularization parameter.
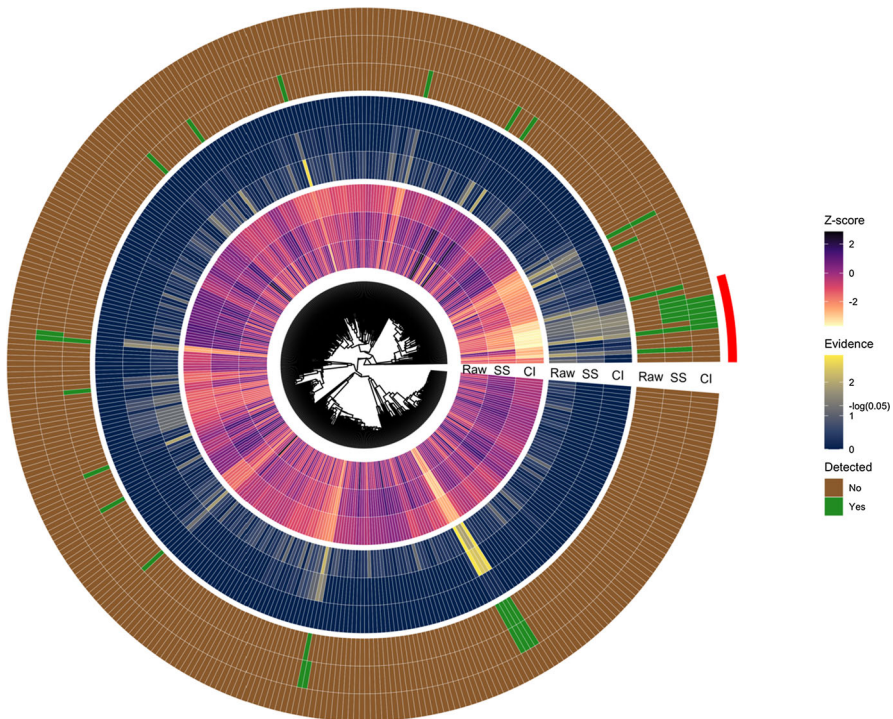
Fig. 6 Phylogeny of the 387 species from the Fidji dataset with associated *z*-scores (inner circle), evidence (middle circle) and detection status (outer circle) under different correction procedures. Species detected by `zazou` are generally close-by on the tree and often, but not always, detected by raw *p*-values. The red strip highlight the smoothing property of the procedure in a subtree where individual species are not detected when using independant univariate tests but are detected when accounting for the hierarchical structure

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 hypotheses 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 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 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

Ambroise C, Dehman A, Neuvial P, Rigaill G, Vialaneix N (2019) Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. Algorithms Mol Biol 14(1):22

Bastide P, Mariadassou M, Robin S (2017) Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree. J R Stat Soc Ser B (Stat Methodol) 79(4):1067–1093

Benjamini Y, Heller R (2007) False discovery rates for spatial signals. J Am Stat Assoc 102(480):1272–1281

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol) 57(1):289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 13:1165–1188

Bichat A, Plassais J, Ambroise C, Mariadassou M (2020) Incorporating phylogenetic information in microbiome differential abundance studies has no effect on detection power and fdr control. Front Microbiol 11:649. https://doi.org/10.3389/fmicb.2020.00649

Blanchard G, Neuvial P, Roquain E (2020) Post hoc confidence bounds on false positives using reference families. Ann Stat 48(3):1281–1303. https://doi.org/10.1214/19-AOS1847

Bland JM, Altman DG (1995) Multiple significance tests: the Bonferroni method. BMJ 310(6973):170

Ilana LB, Yilmaz S, Huang K, Xu L, Stacy DJ, Aaron PJ, Waisea N, Tamminen M, Smillie CS, Jennifer RW et al (2016) Mobile genes in the human microbiome are structured from global to individual scales. Nature 535(7612):435–439

Bush WS, Moore JH (2012) Genome-wide association studies. PLoS Comput Biol 8(12):e1002822

Chen J (2018) StructFDR: false discovery control procedure integrating the prior structure information. https://CRAN.R-project.org/package=StructFDR. R package version 1.3

Cremers HR, Wager TD, Yarkoni T (2017) The relation between statistical power and inference in fmri. PLoS ONE 12(11):e0184923

Dunn JE, Gipson PS (1977) Analysis of radio telemetry data in studies of home range. Biometrics 13:85–101

Eickhoff SB, Thirion B, Varoquaux G, Bzdok D (2015) Connectivity-based parcellation: critique and implications. Hum Brain Mapp 36(12):4771–4792

Fan Y, Tang CY (2013) Tuning parameter selection in high dimensional penalized likelihood. J R Stat Soc Ser B (Stat Method) 75(3):531–552

Freckleton RP, Harvey PH, Pagel M (2003) Bergmann's rule and body size in mammals. Am Nat 161(5):821–825

Fu WJ (1998) Penalized regressions: the bridge versus the lasso. J Comput Gr Stat 7(3):397–416

Goeman Jelle J, Livio Finos (2012) The inheritance procedure: multiple testing of tree-structured hypotheses. Stat Appl Genet Mol Biol 11(1):1–18

Huang R, Soneson C, Germain P-L, Schmidt TSB, Von Mering C, Robinson MD (2021) Treeclimbr pinpoints the data-dependent resolution of hierarchical hypotheses. Genome Biol 22(1):1–21

Javanmard Adel, Montanari Andrea (2013) Confidence intervals and hypothesis testing for high-dimensional statistical models. In: Advances in neural information processing systems, pp 1187–1195

Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 15(1):2869–2909

Javanmard A, Javadi H et al (2019) False discovery rate control via debiased lasso. Electron J Stat 13(1):1212–1253

Khabbazian M, Kriebel R, Rohe K, Ané C (2016) Fast and accurate detection of evolutionary shifts in Ornstein-Uhlenbeck models. Methods Ecol Evol 7(7):811–824

Kim KI, Roquain E, van de Wiel MA (2010) Spatial clustering of array cgh features in combination with hierarchical multiple testing. Stat Appl Genet Mol Biol 9(1):159

Lande R (1976) Natural selection and random genetic drift in phenotypic evolution. Evolution 30(2):314–334. https://doi.org/10.1111/j.1558-5646.1976.tb00911.x

MacLean OA, Lytras S, Weaver S, Singer JB, Boni MF, Lemey P, Kosakovsky PSL, Robertson DL (2021) Natural selection in the evolution of sars-cov-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol 19(3):e3001115

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 15:50–60

Matsen IV, Frederick A, Evans SN (2013) Edge principal components and squash clustering: Using the special structure of phylogenetic placement data for sample comparison. PLOS ONE 8(3):1–15. https://doi.org/10.1371/journal.pone.0056859

McLachlan G, Peel D (2000) Finite Mixture Models. Wiley, New York

McLachlan GJ, Do K-A, Ambroise C (2005) Analyzing Microarray Gene Expression Data, vol 422. Wiley, New York

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

Nåsell I (1999) On the time to extinction in recurrent epidemics. J R Stat Soc Ser B (Stat Methodol) 61(2):309–330. https://doi.org/10.1111/1467-9868.00178

Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB et al (2017) Accessible, curated metagenomic data through experimenthub. Nat Methods 14(11):1023

Renaux C, Buzdugan L, Kalisch M, Bühlmann P (2020) Hierarchical inference for genome-wide association studies: a view on methodology with software. Comput Stat 35(1):1–40

Sankaran K, Holmes S (2014) structssi: simultaneous and selective inference for grouped or hierarchically structured data. J Stat Softw 59(13):1

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C (2011) Metagenomic biomarker discovery and explanation. Genome Biol 12(6):1–18

Sesia M, Katsevich E, Bates S, Candès E, Sabatti C (2020) Multi-resolution localization of causal variants across the genome. Nat Commun 11(1):1–10

Silverman JD, Washburne AD, Mukherjee S, David LA (2017) A phylogenetic transform enhances analysis of compositional microbiota data. eLife. https://doi.org/10.7554/elife.21887

Sneath PHA, Sokal RR et al (1973) Numerical taxonomy. The principles and practice of numerical classification. Science 2:19

Sun T, Zhang C-H (2012) Scaled sparse linear regression. Biometrika 99(4):879–898. https://doi.org/10.1093/biomet/ass043

Tang Z-Z, Chen G, Alekseyenko AV, Li H (2017) A general framework for association analysis of microbial communities on a taxonomic tree. Bioinformatics 33(9):1278–1285

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol) 58(1):267–288

Tukey JW (1953) The problem of multiple comparisons. Mult Comp 2:39

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 98(9):5116–5121

Wilcoxon F (1992) Individual comparisons by ranking methods. In: Breakthroughs in statistics. Springer, pp 196–202

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

Yekutieli D (2008) Hierarchical false discovery rate-controlling methodology. J Am Stat Assoc 103(481):309–316

Zhang C-H, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc Ser B (Stat Methodol) 76(1):217–242