université PARIS-SACLAY

Discovering multi-scale metagenomic signatures through hierarchical organization of species

PhD defense

Antoine BICHAT
LaMME – Enterome
December 9, 2020

C. Ambroise

M. Mariadassou

J. Plassais

F. Strozzi

université PARIS-SACLAY

Laboratoire de Mathématiques et Modélisation d'Évry LaMME

MaIAGE

enterome

Context

Microbiota

Ecological community of microorganisms that reside in an environmental niche.

Microbiota

Ecological community of microorganisms that reside in an environmental niche.

For human gut

- 10^{14} bacterial cells in one gut...
- ... weighting $2\ kg$.
- More than 1500 different species.
- More than 10 millions unique genes.
- Helpful for nutrient absorption and anti inflammatory properties.

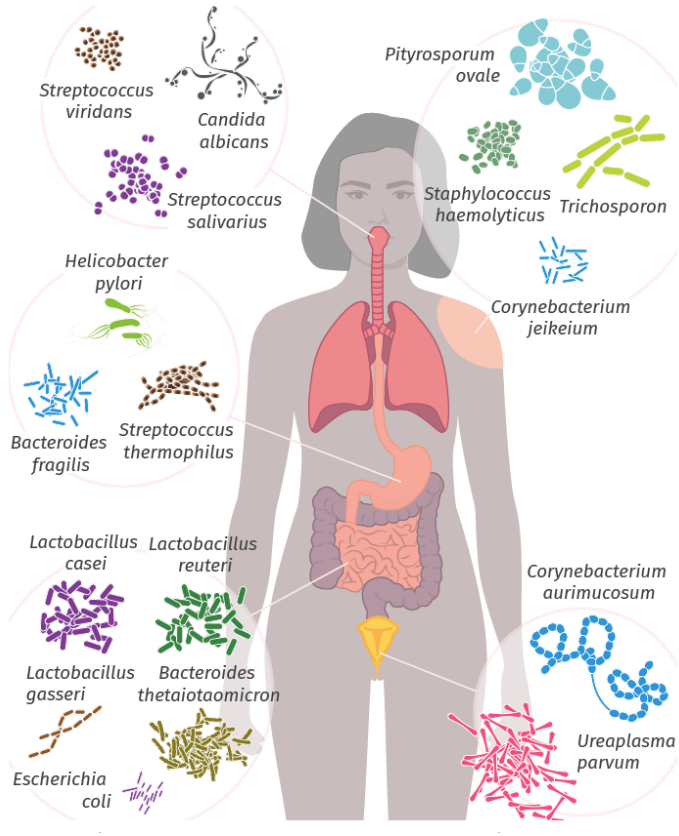
Ley Peterson, et al. [LPG06]; Flint Scott, et al. [Fli+12]; Canani Di Costanzo, et al. [Can+11]

Microbiota

Ecological community of microorganisms that reside in an environmental niche.

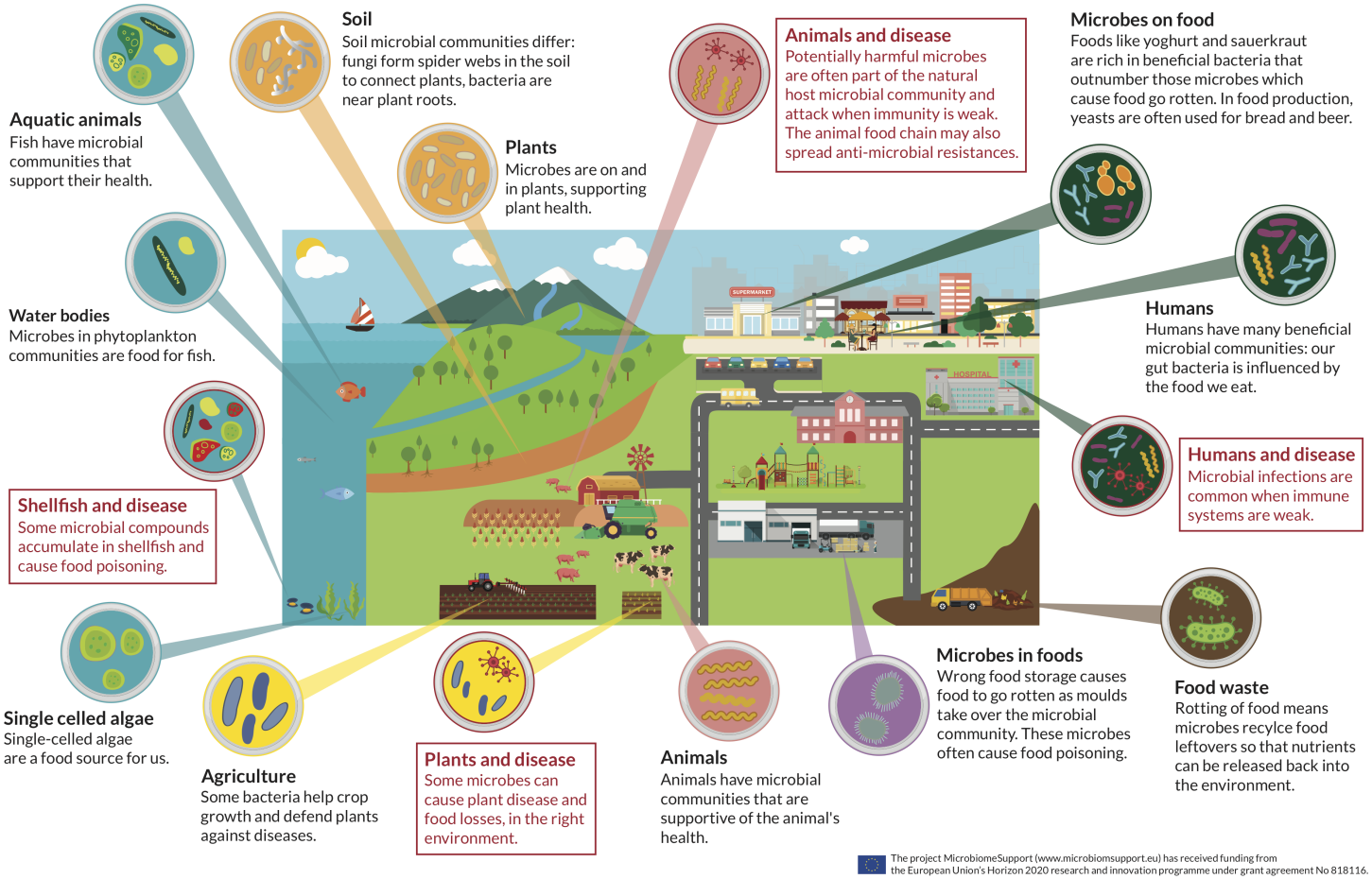
For human gut

- 10^{14} bacterial cells in one gut...
- ... weighting 2 kg.
- More than 1500 different species.
- More than 10 millions unique genes.
- Helpful for nutrient absorption and anti inflammatory properties.



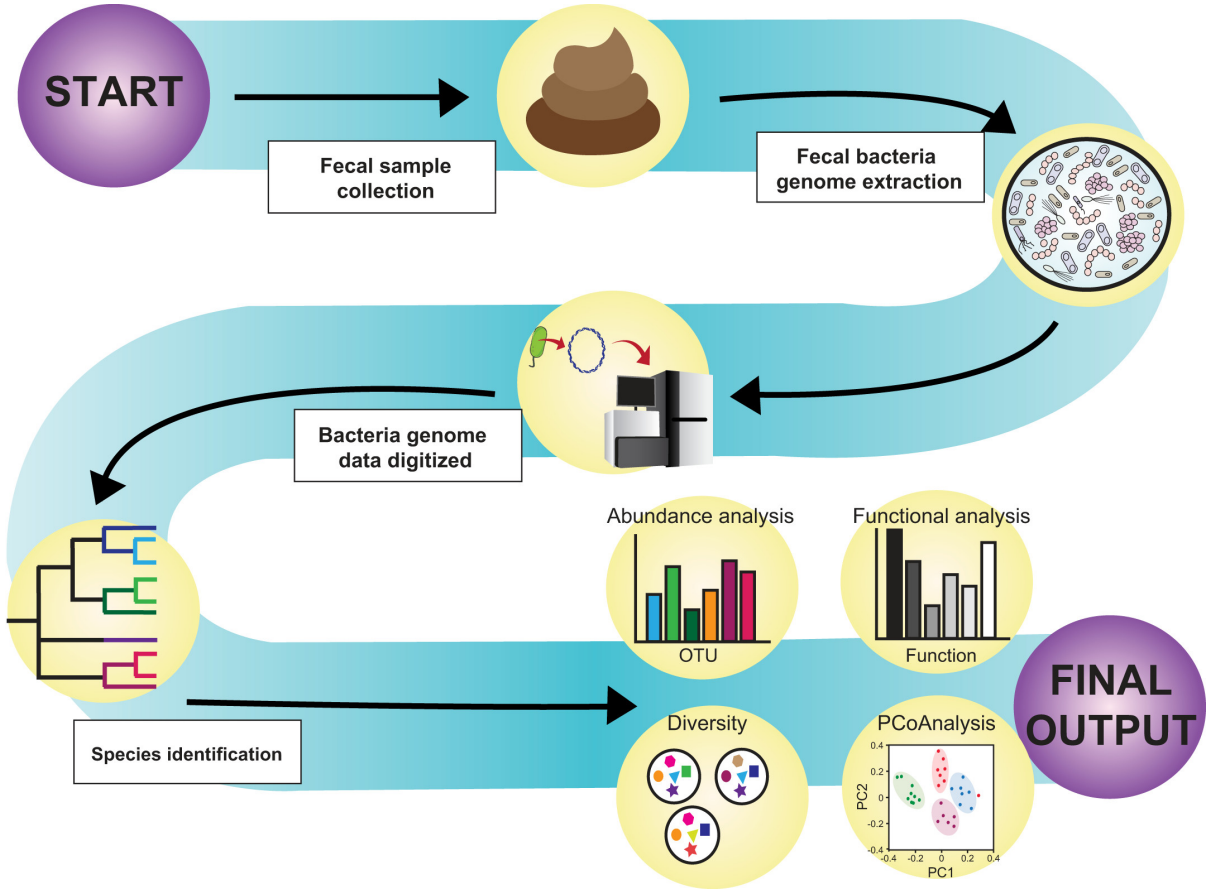
Ley Peterson, et al. [LPG06]; Flint Scott, et al. [Fli+12]; Canani Di Costanzo, et al. [Can+11] – lelivrescolaire.fr

Microbiota everywhere



The project MicrobiomeSupport (www.microbiomsupport.eu) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 818116.

Sequencing



Cryan O'Riordan, et al. [Cry+19]

Abundance table

- Matrix with m taxa and p samples.
- Count or compositional data, with inflation in zero.
- Correlation between abundances.

Abundance table

- Matrix with m taxa and p samples.
- Count or compositional data, with inflation in zero.
- Correlation between abundances.

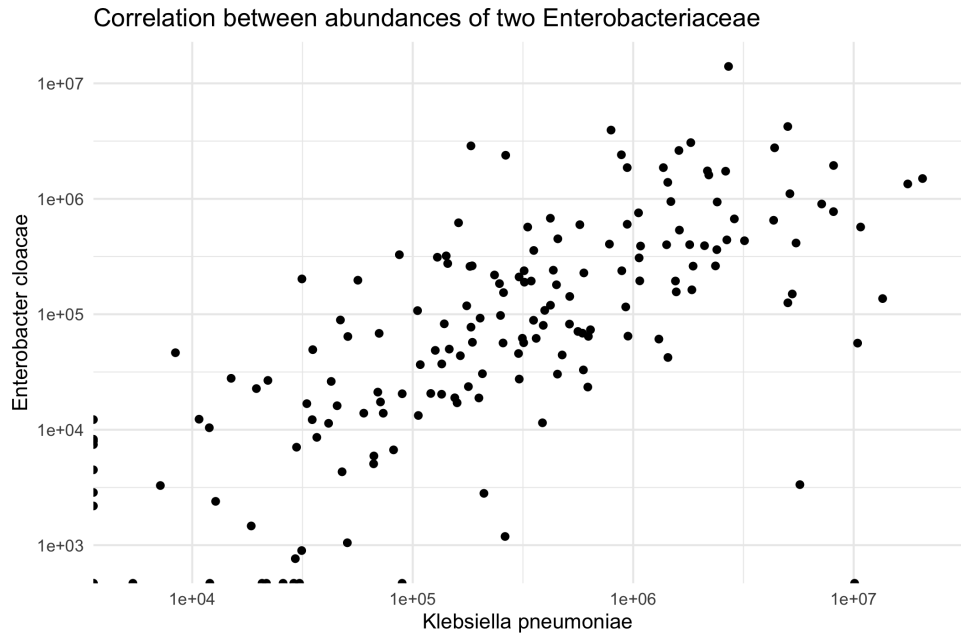
	S1	S2	S3	S4	S5	S6
Akkermansia muciniphila	0	934850	247037	0	181167	0
Bacteroides dorei	89893	2567192	648639	0	0	8498
Bifidobacterium longum	376119	0	30671	0	1292193	1830318
Enterobacter cloacae	756064	12234	142652	2186	238175	535786
Escherichia coli	256424	5551041	9905179	76052	70234	79805
Klebsiella pneumoniae	1057187	0	515407	0	887678	1620535
Megamonas hypermegale	0	0	0	0	0	27317
Streptococcus anginosus	0	0	0	985	0	19134

Brito Yilmaz, et al. [Bri+16]

Abundance table

- Matrix with m taxa and p samples.
- Count or compositional data, with inflation in zero.
- Correlation between abundances.

	S1	S2	S3	S4	S5	S6
Akkermansia muciniphila	0	934850	247037	0	181167	0
Bacteroides dorei	89893	2567192	648639	0	0	8498
Bifidobacterium longum	376119	0	30671	0	1292193	1830318
Enterobacter cloacae	756064	12234	142652	2186	238175	535786
Escherichia coli	256424	5551041	9905179	76052	70234	79805
Klebsiella pneumoniae	1057187	0	515407	0	887678	1620535
Megamonas hypermegale	0	0	0	0	0	27317
Streptococcus anginosus	0	0	0	985	0	19134



Brito Yilmaz, et al. [Bri+16]

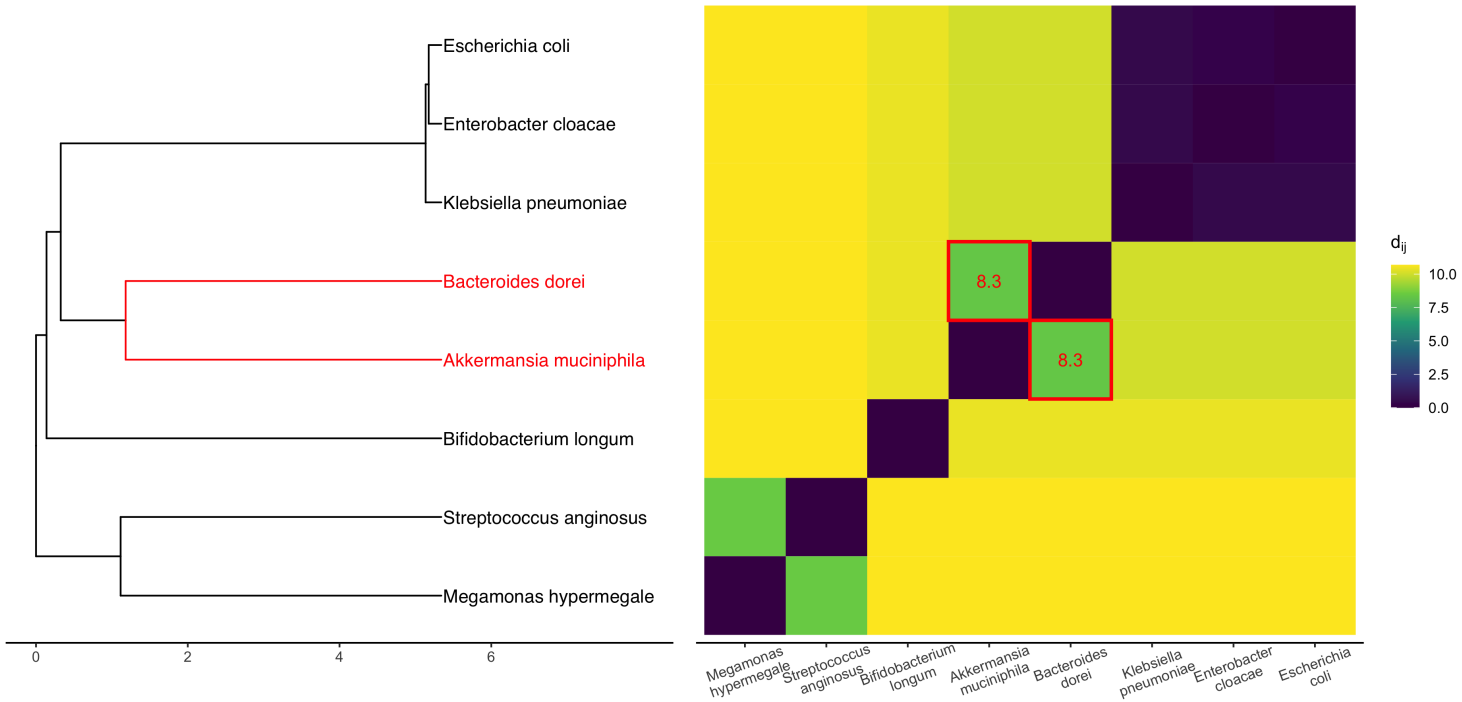
Phylogeny

- Tree with m leaves which describes the evolutionary history of the taxa.

Brito Yilmaz, et al. [Bri+16]

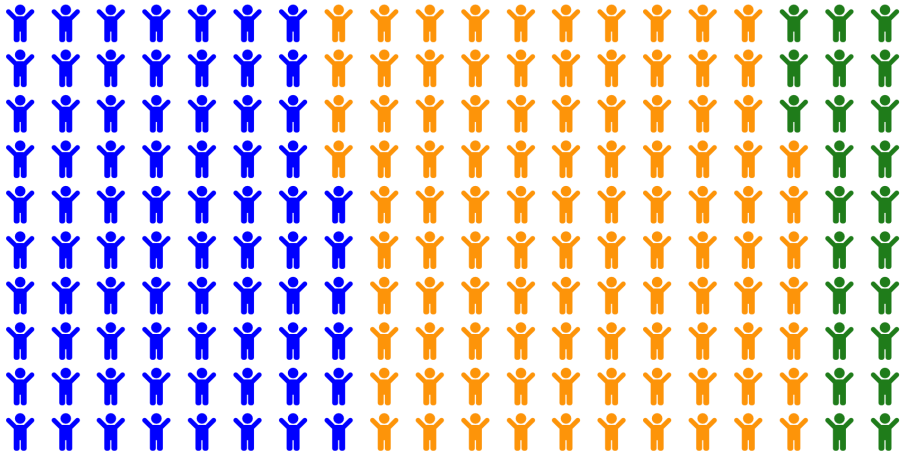
Phylogeny

- Tree with m leaves which describes the evolutionary history of the taxa.
- Associated with a patristic distance matrix $\left(d_{i,j}\right)_{i,j}$.



Brito Yilmaz, et al. [Bri+16]

Differential abundance studies



Used to find **associations** between microbiota and

- diet [Dav+14],
- birth mode [Bok+16],
- age [Yat+12],
- pet owning [Kat+20],
- tobacco [Ops+16],
- antibiotics [Pal+18],
- Crohn's disease [Mor+12],
- cirrhosis [Qin+14]
- schizophrenia [Zhe+19]...

Classical approach

Vector $\mathfrak{p} \in [0, 1]^m$ of p-values, computed on each taxa independently.

Wilcoxon and Kruskal-Wallis non-parametric tests can be used.

Classical approach

Vector $\mathfrak{p} \in [0, 1]^m$ of p-values, computed on each taxa independently.

Wilcoxon and Kruskal-Wallis non-parametric tests can be used.

A taxon i is detected differentially abundant if $\mathfrak{p}_i < \alpha$.

Each taxon fall into one class of the confusion matrix:

		True condition	
		Positive	Negative
Detection	Positive	TP	FP
	Negative	FN	TN

Classical approach

Vector $\mathfrak{p} \in [0,1]^m$ of p-values, computed on each taxa independently.

Wilcoxon and Kruskal-Wallis non-parametric tests can be used.

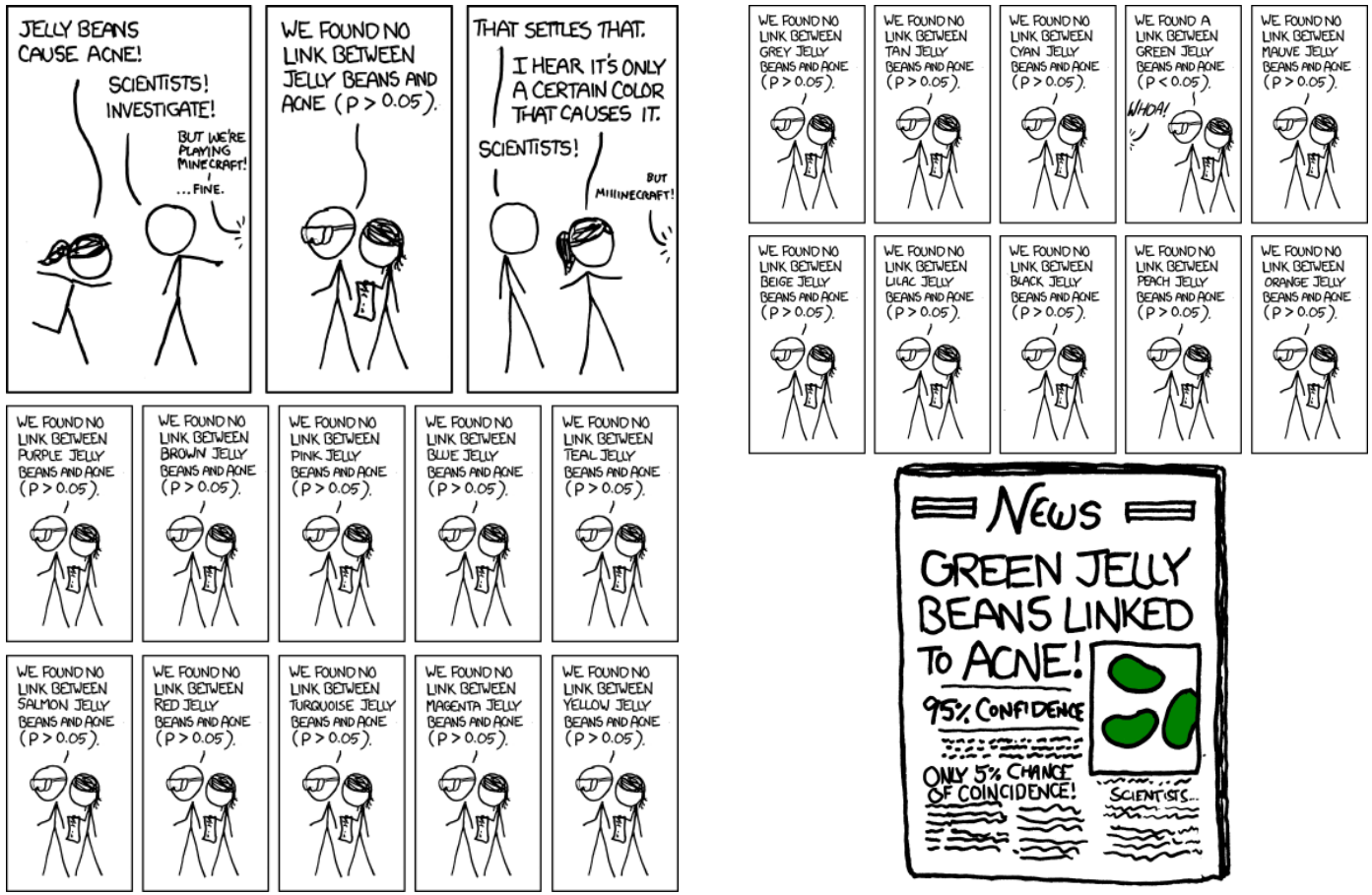
A taxon i is detected differentially abundant if $\mathfrak{p}_i < \alpha$.

Each taxon fall into one class of the confusion matrix:

		True condition	
		Positive	Negative
Detection	Positive	TP	FP
	Negative	FN	TN

Under $\mathbb{H}_0$, $\mathfrak{p}_i \sim \mathcal{U}([0,1])$.

Multiple testing problem



XKCD

TPR and FDR

		True condition: Positive	True condition: Negative
Detection	Positive	TP	FP
	Negative	FN	TN

True Positive Rate:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

False Discovery Rate:

$$\mathrm{FDP} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}, \mathrm{FDR} = \mathbb{E}[\mathrm{FDP}].$$

Multiple testing problem correction

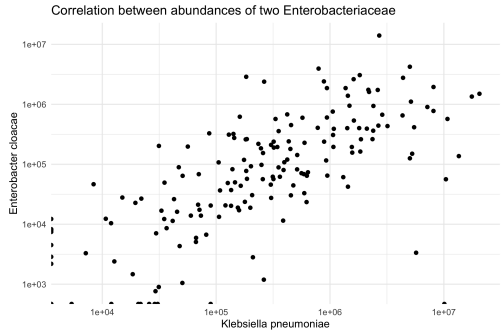
Correction with Benjamini-Hochberg procedure to respect an *a priori* FDR : q^{bh}.

Benjamini and Hochberg [BH95]

Multiple testing problem correction

Correction with Benjamini-Hochberg procedure to respect an *a priori* FDR : q^{bh}.

But it assumes independence between taxa and it is not respected.

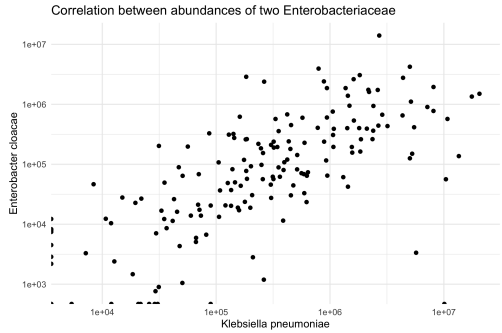


Benjamini and Hochberg [BH95]

Multiple testing problem correction

Correction with Benjamini-Hochberg procedure to respect an *a priori* FDR : q^{bh}.

But it assumes independence between taxa and it is not respected.



One can use Benjamini-Yekutieli correction which does not make any assumption about dependence between taxa but

- it's too conservative,
- we want to correct explicitly for correlation between taxa.

Benjamini and Hochberg [BH95]; Benjamini and Yekutieli [BY01]

Incorporation of hierarchical information

Goal

- Correct explicitly for correlation between taxa.

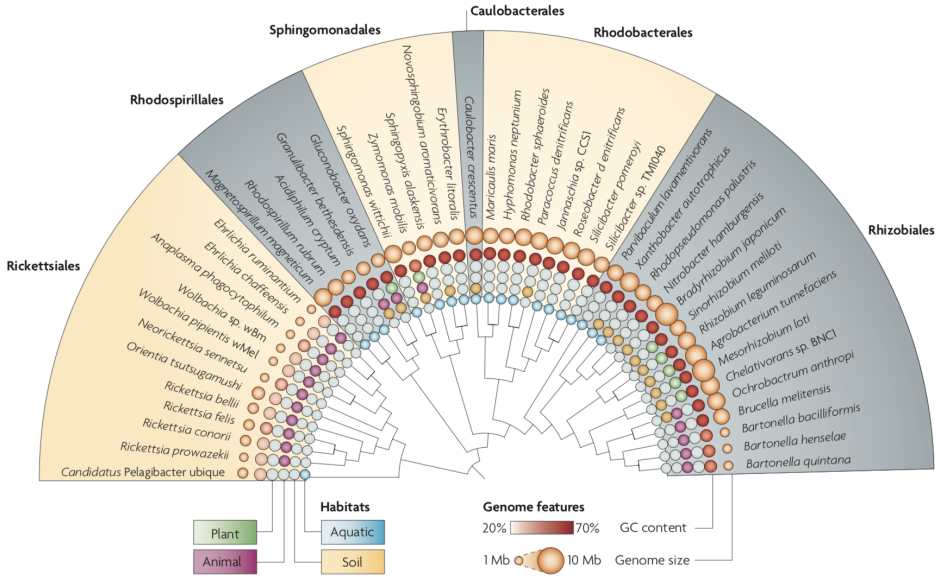
Goal

- Correct explicitly for correlation between taxa.
- Increase power.

Goal

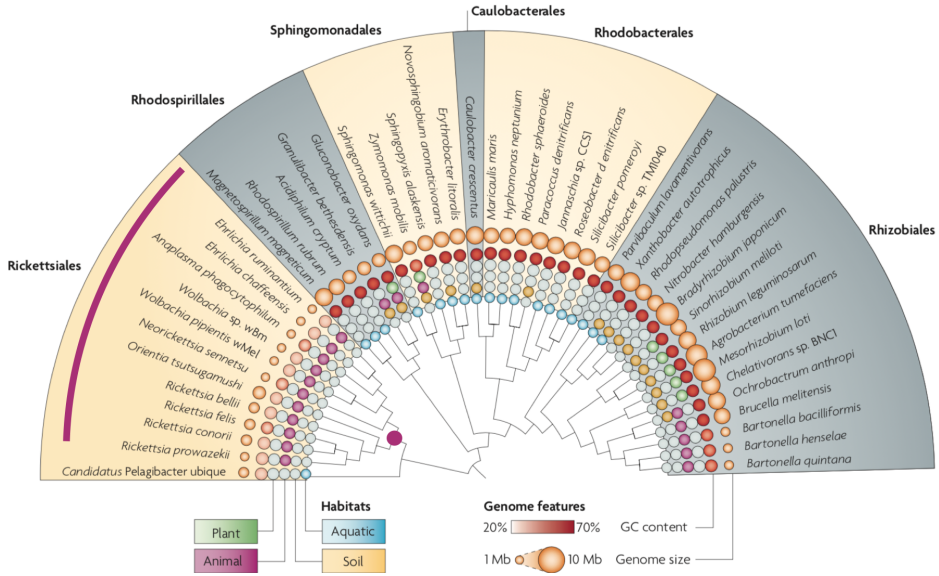
- Correct explicitly for correlation between taxa.
- Increase power.
- Keep FDR under a desired level.

Rationale



Philippot Andersson, et al. [Phi+10]

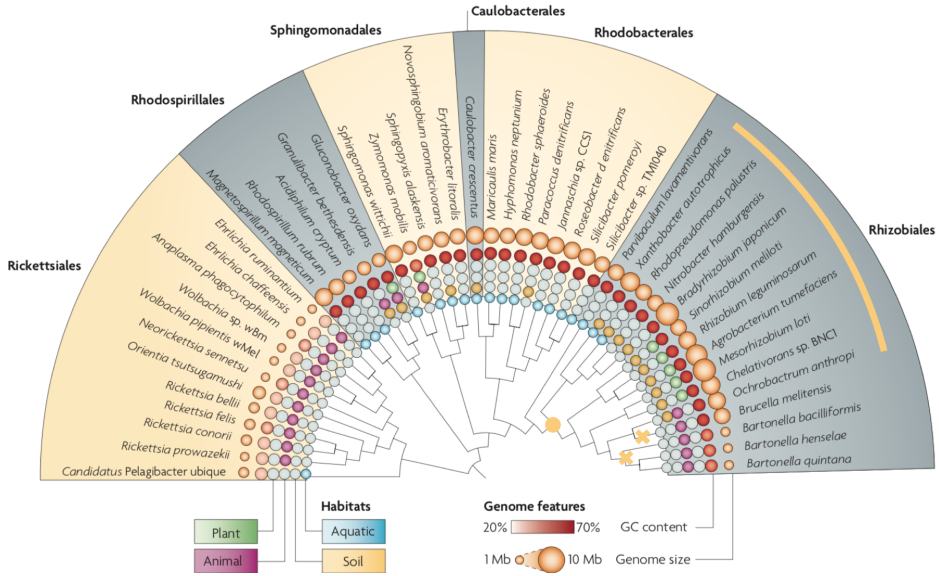
Rationale



These species are associated with animals.

Philippot Andersson, et al. [Phi+10]

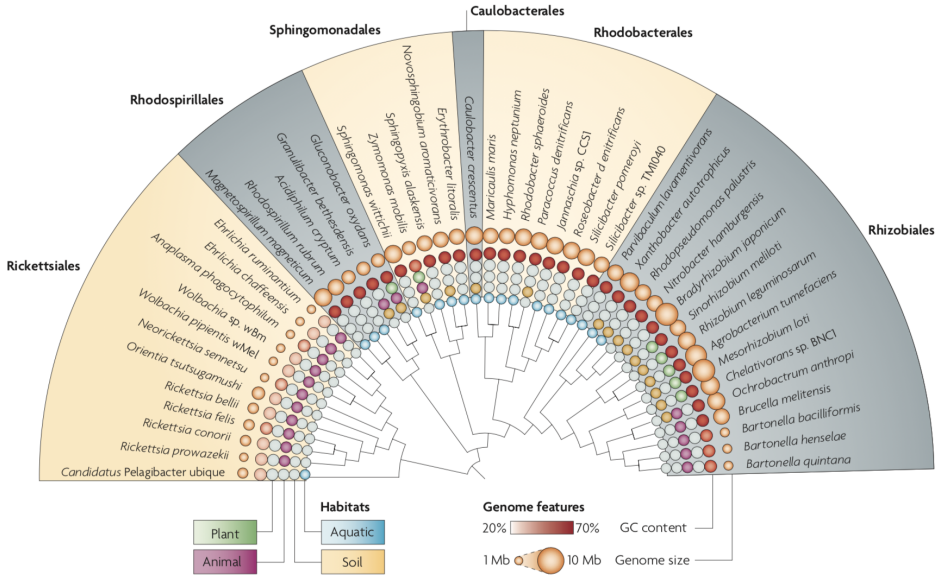
Rationale



These species are associated with soil.

Philippot Andersson, et al. [Phi+10]

Rationale



Already used in

- Hierarchical FDR [Yek08; SH14],
- **TreeFDR [XCC17]**.

Philippot Andersson, et al. [Phi+10]

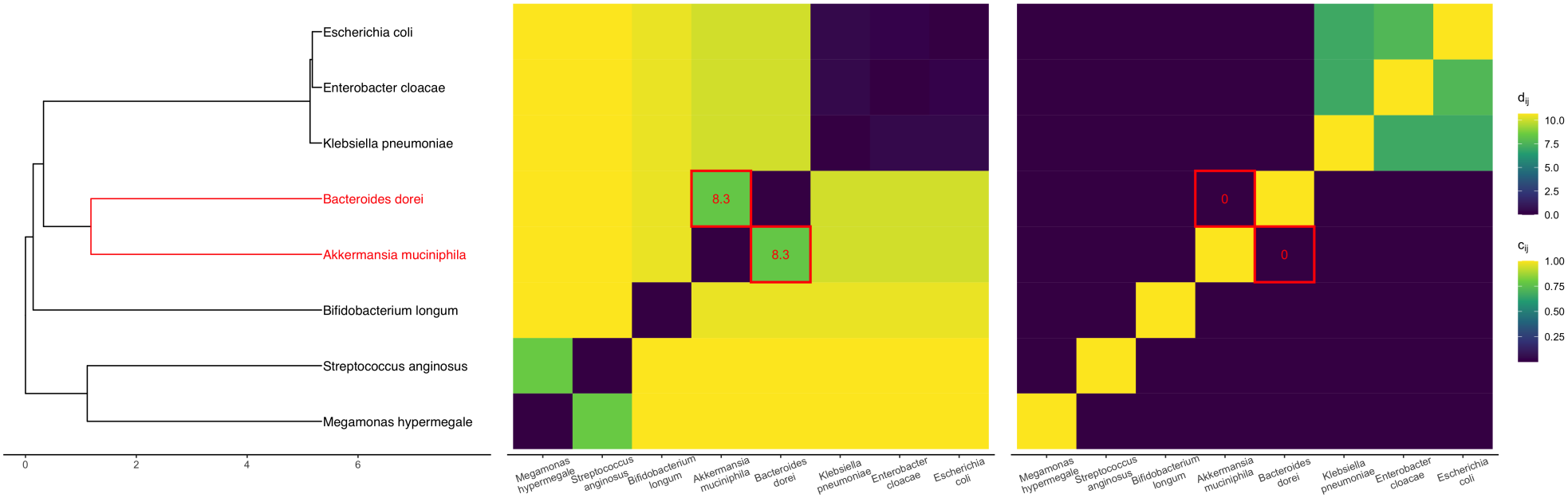
z-scores smoothing with TreeFDR

$\mathfrak{z} = \Phi^{-1}(\mathfrak{p})$ is the vector of observed z-scores and μ the vector of "true" z-scores.

z-scores smoothing with TreeFDR

$\mathfrak{z} = \Phi^{-1}(\mathfrak{p})$ is the vector of observed z-scores and μ the vector of "true" z-scores.

Assume a hierarchical model: $\mathfrak{z} \mid \mu \sim \mathcal{N}_m(\mu, \sigma^2 \mathbf{I}_m)$ and $\mu \sim \mathcal{N}_m(\gamma \mathbf{1}_m, \tau^2 C_\rho)$ with $C_\rho = \left(\exp\left(-2\rho d_{i,j}\right)\right)_{i,j}$.



z-scores smoothing with TreeFDR

Then $\mathcal{z} \sim \mathcal{N}_m\left(\gamma \mathbf{1}_m, \tau^2 C_\rho + \sigma^2 \mathbf{I}_m\right)$ by Bayes formula and the maximum a posteriori gives

$$\mu^* = \left(\mathbf{I}_m + k^2 C_\rho^{-1}\right)\left(k^2 C_\rho^{-1}\gamma \mathbf{1}_m + \mathcal{z}\right),$$

with $k = \frac{\sigma}{\tau}$ and ρ_0 hyperparameters to optimize.

z-scores smoothing with TreeFDR

Then $\mathcal{Z} \sim \mathcal{N}_m\left(\gamma \mathbf{1}_m, \tau^2 C_\rho + \sigma^2 \mathbf{I}_m\right)$ by Bayes formula and the maximum a posteriori gives

$$\mu^* = \left(\mathbf{I}_m + k^2 C_\rho^{-1}\right)\left(k^2 C_\rho^{-1} \gamma \mathbf{1}_m + \mathcal{Z}\right),$$

with $k = \frac{\sigma}{\tau}$ and ρ_0 hyperparameters to optimize.

At the end, a multiple testing correction by resampling is done on smoothed values.

Which tree?

- Phylogeny? Taxonomy?

😀 Proxy for correlations at high-level niches.

☹️ Not so much for low-level niches?

☹️ Not available every time.

Which tree?

- Phylogeny? Taxonomy?

😀 Proxy for correlations at high-level niches.

☹ Not so much for low-level niches?

☹ Not available every time.

- Correlation tree?

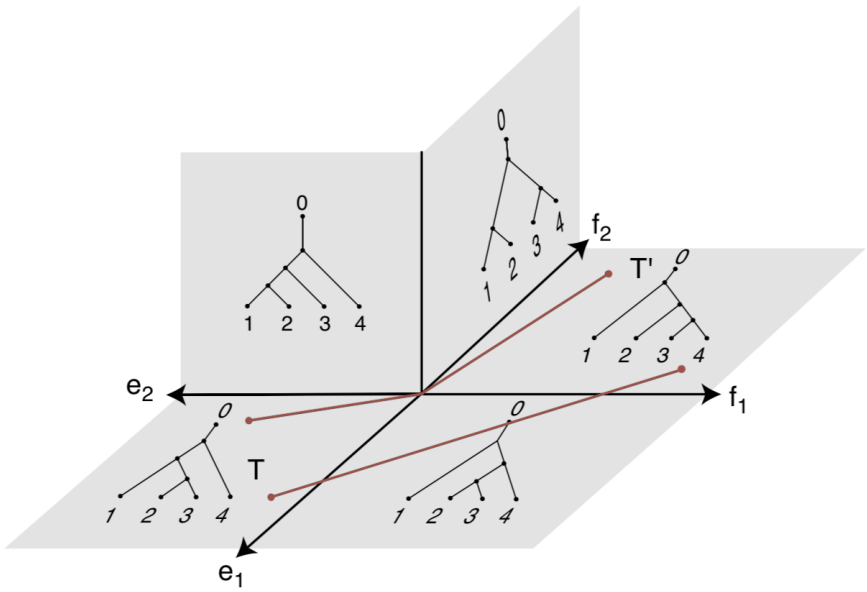
😀 Actual correlation between taxa.

😀 Always available.

☹ Risk of overfitting.

Bichat Plassais, et al. [Bic+20]

Billera-Holmes-Vogtmann distance



- Each tree is mapped into a space composed by merged orthants.
- An orthant corresponds to a topology.
- The BHV distance is the length of the unique shortest path between the trees on treespace.
- It can be computed with a $O\left(m^4\right)$ algorithm.

Billera Holmes, et al. [BHV01]

Quantifying distance between trees

- Trees of primary interest:
  - correlation tree on original data,
  - phylogeny.

Quantifying distance between trees

- Trees of primary interest:
  - correlation tree on original data,
  - phylogeny.
- What is the confidence region for the correlation tree?
  - correlation trees on boostrapped data (resampling on samples).

Quantifying distance between trees

- Trees of primary interest:
  - correlation tree on original data,
  - phylogeny.
- What is the confidence region for the correlation tree?
  - correlation trees on boostrapped data (resampling on samples).
- Are trees significantly closer than two random trees?
  - trees created by random shuffling of correlation tree tip labels,
  - trees created by random shuffling of phylogeny tip labels.

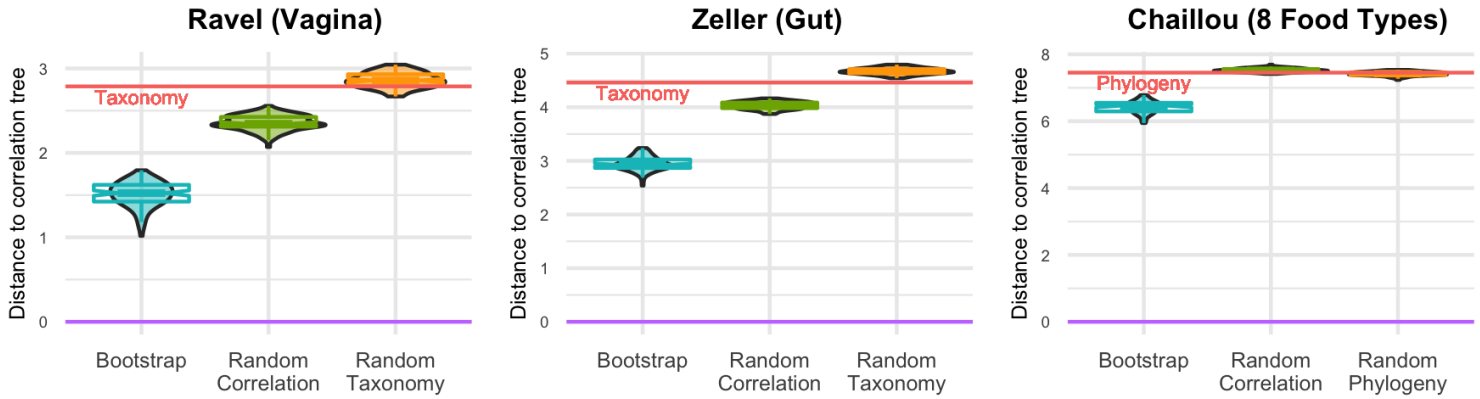
Quantifying distance between trees

- Trees of primary interest:
  - correlation tree on original data,
  - phylogeny.
- What is the confidence region for the correlation tree?
  - correlation trees on boostrapped data (resampling on samples).
- Are trees significantly closer than two random trees?
  - trees created by random shuffling of correlation tree tip labels,
  - trees created by random shuffling of phylogeny tip labels.

We compute all pairwise distances between these trees.

Comparisons between trees

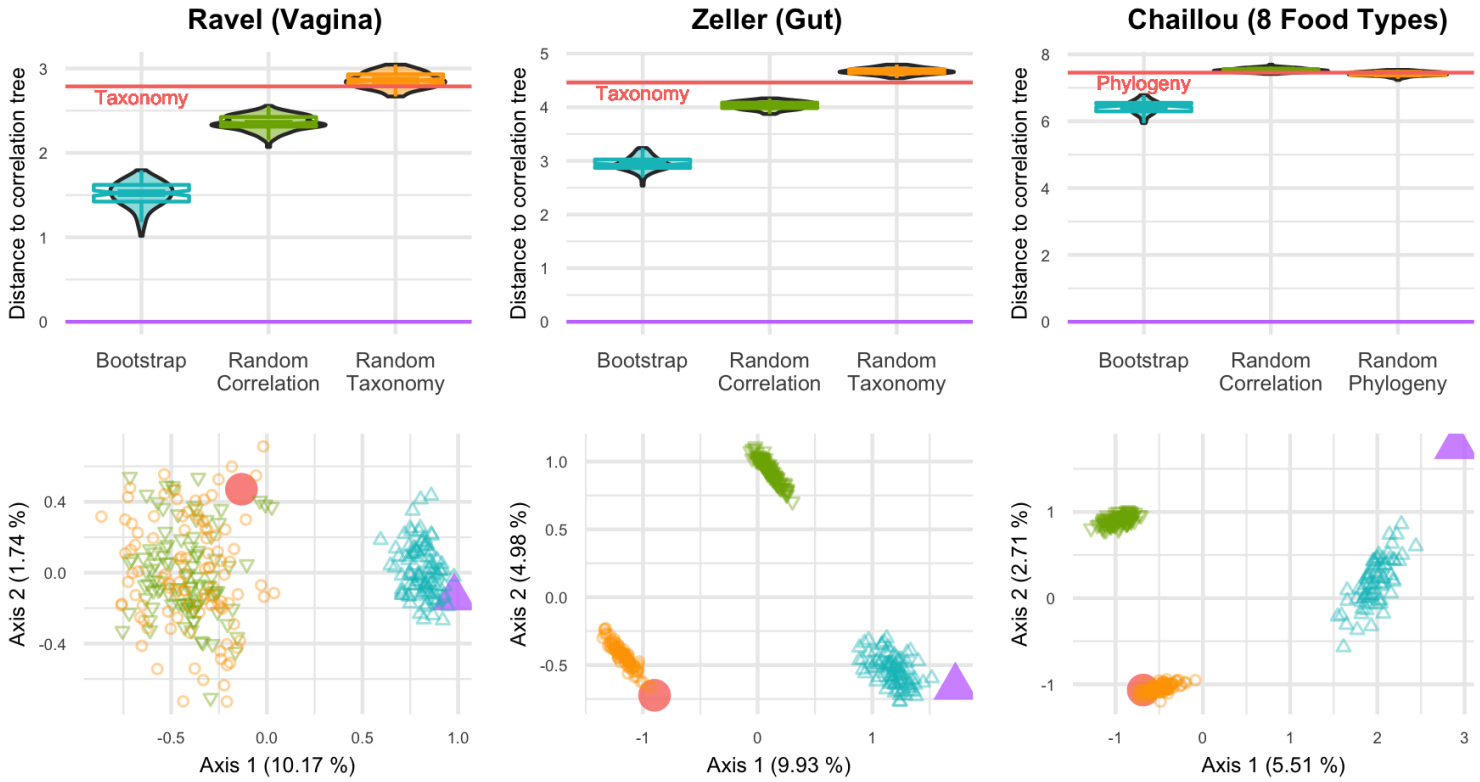
Neither phylogeny nor taxonomy is in the confident region of the correlation tree.



[Bic+20; Rav+11; Zel+14; Cha+15]

Comparisons between trees

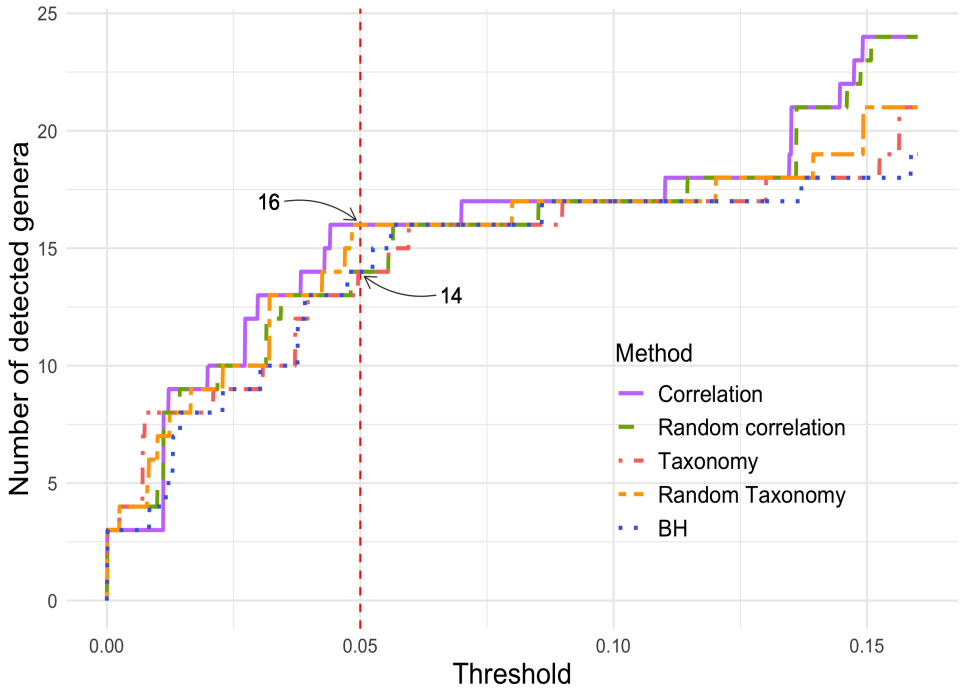
Neither phylogeny nor taxonomy is in the confident region of the correlation tree.



[Bic+20; Rav+11; Zel+14; Cha+15]

Impact of the tree with Zeller dataset

All hierarchies give highly similar results.



- 119 genera
- 199 patients
  - 66 healthy
  - 42 adenoma
  - 91 colorectal cancer

Bichat Plassais, et al. [Bic+20]; Zeller Tap, et al. [Zel+14]

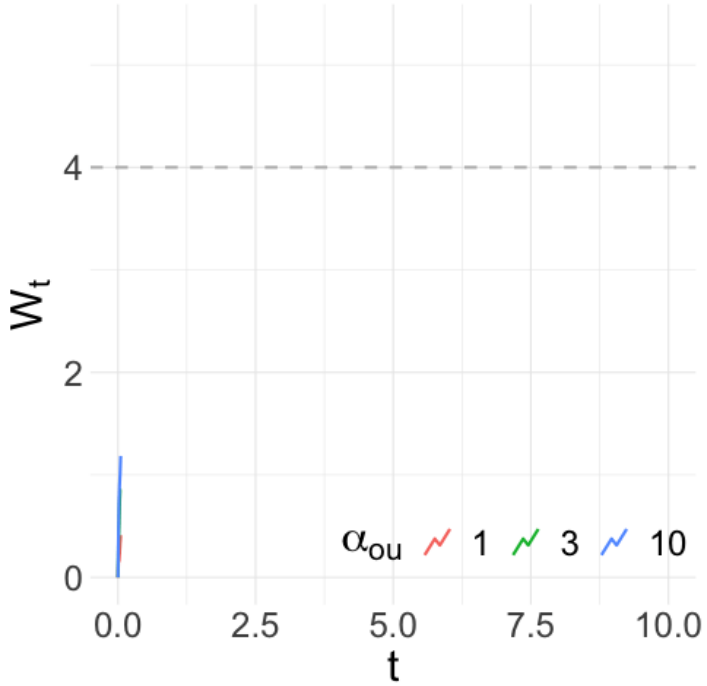
zazou

Z-scores AZ Ornstein-Uhlenbeck

Ornstein-Uhlenbeck process

An Ornstein-Uhlenbeck (OU) process with an optimal value of β_{ou} and a strength of selection $\alpha_{\text{ou}} > 0$ is a Gaussian process that satisfies the SDE

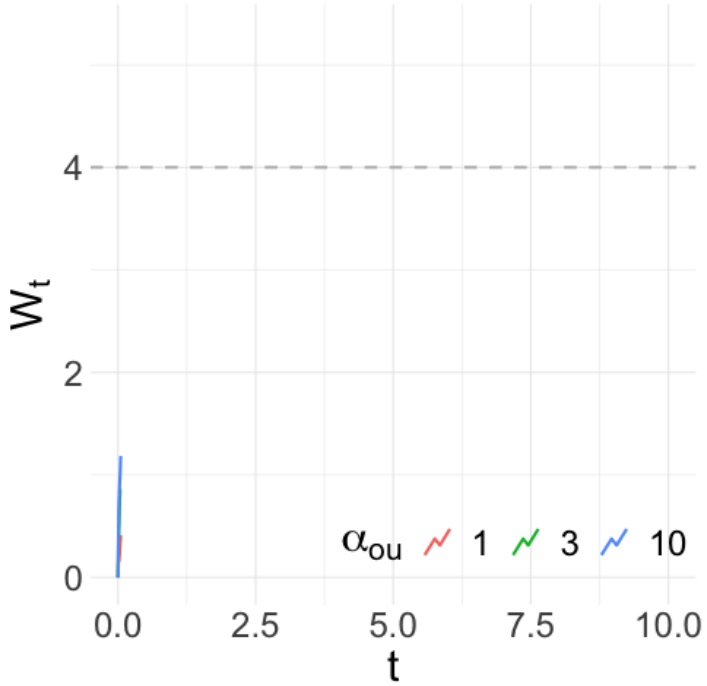
$$\mathrm{d}W_t = -\alpha_{\text{ou}}(W_t - \beta_{\text{ou}})\mathrm{d}t + \sigma_{\text{ou}}\mathrm{d}B_t.$$



Ornstein-Uhlenbeck process

An Ornstein-Uhlenbeck (OU) process with an optimal value of β_{ou} and a strength of selection $\alpha_{\text{ou}} > 0$ is a Gaussian process that satisfies the SDE

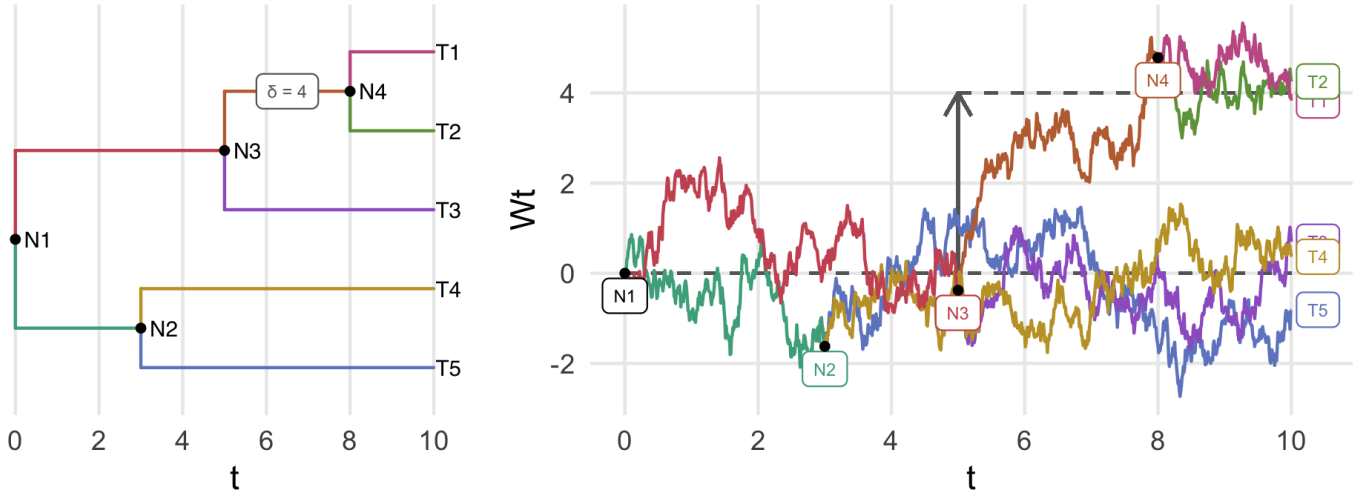
$$\mathrm{d}W_t = -\alpha_{\text{ou}}(W_t - \beta_{\text{ou}})\mathrm{d}t + \sigma_{\text{ou}}\mathrm{d}B_t.$$



W_t is Gaussian with bounded variance and

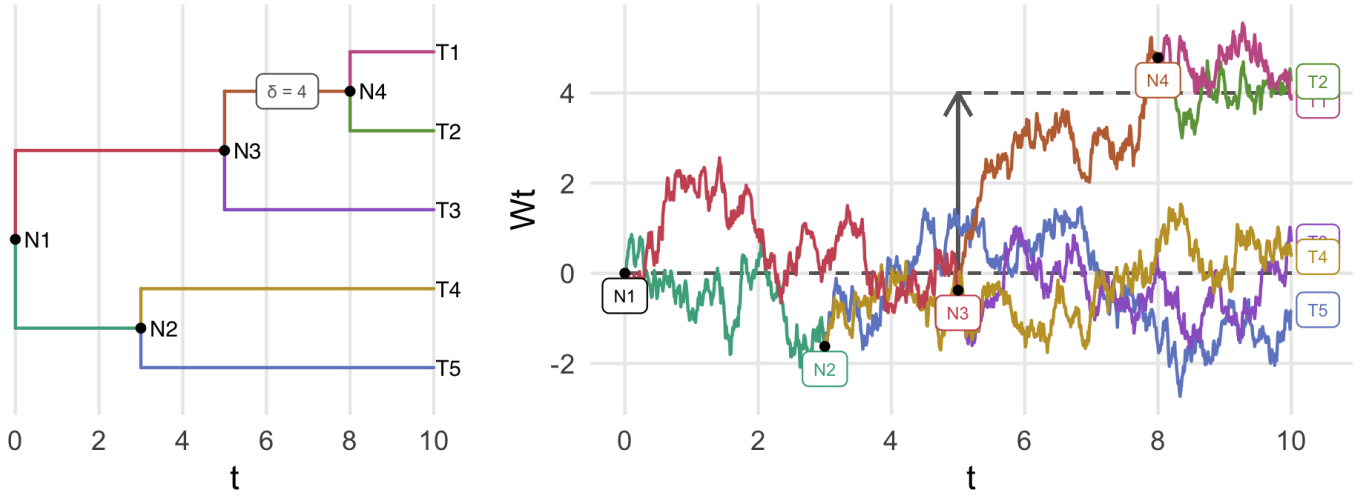
$$W_t \xrightarrow[t\to\infty]{} \mathcal{N}\left(\beta_{\text{ou}}, \frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}}\right).$$

OU process on a tree with shifts δ



Bastide Mariadassou, et al. [BMR17]

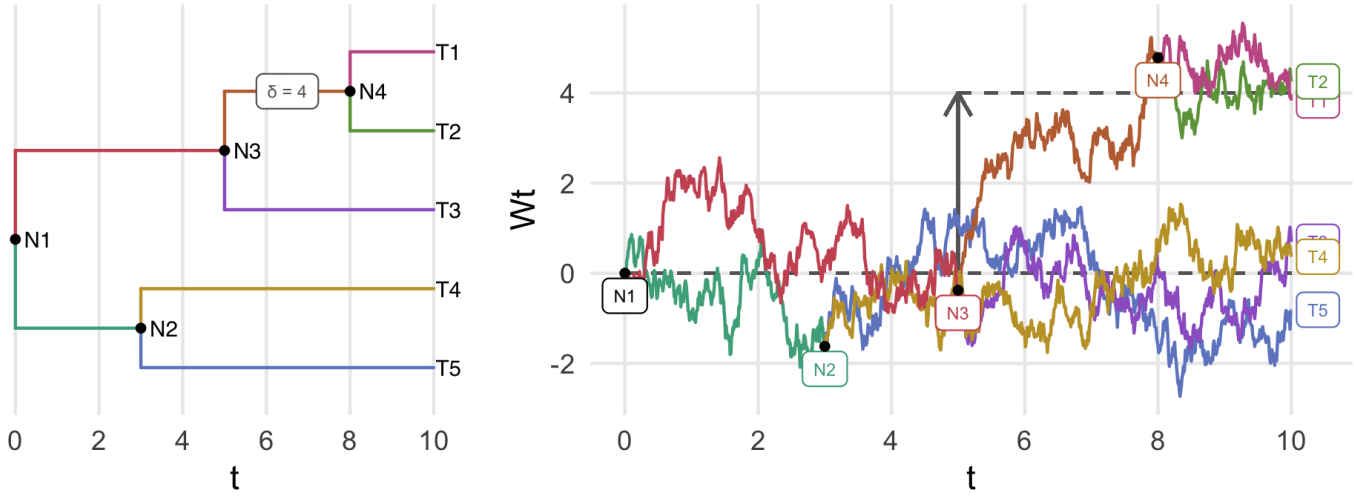
OU process on a tree with shifts δ



- On a branch, the process behaves like on $\mathbb{R}_+$.

Bastide Mariadassou, et al. [BMR17]

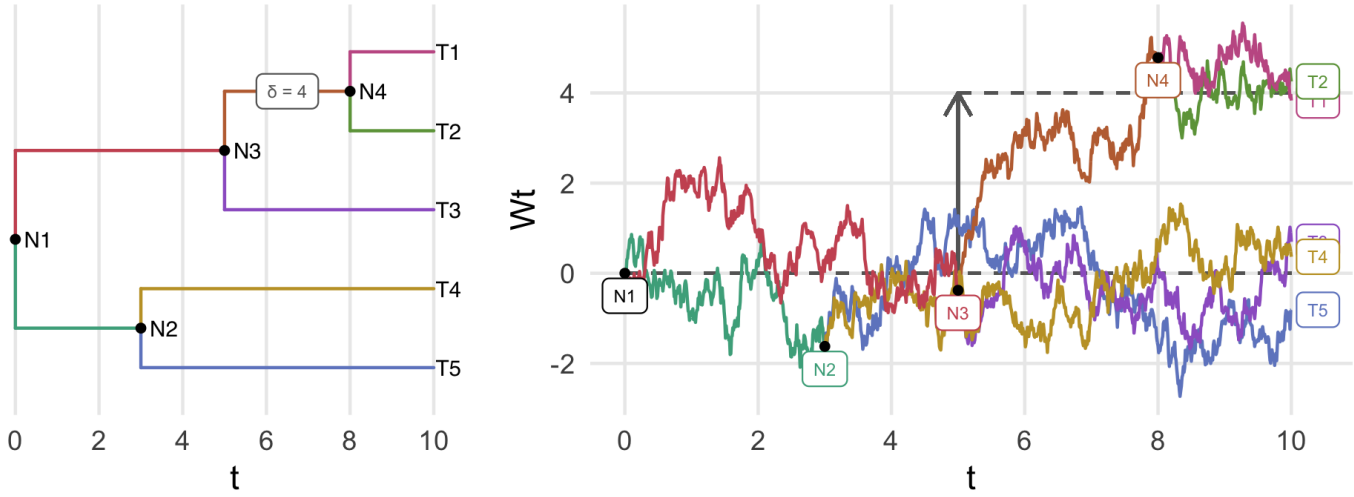
OU process on a tree with shifts δ



- On a branch, the process behaves like on $\mathbb{R}_+$.
- At each node, the process splits into two independent processes with the same initial value.

Bastide Mariadassou, et al. [BMR17]

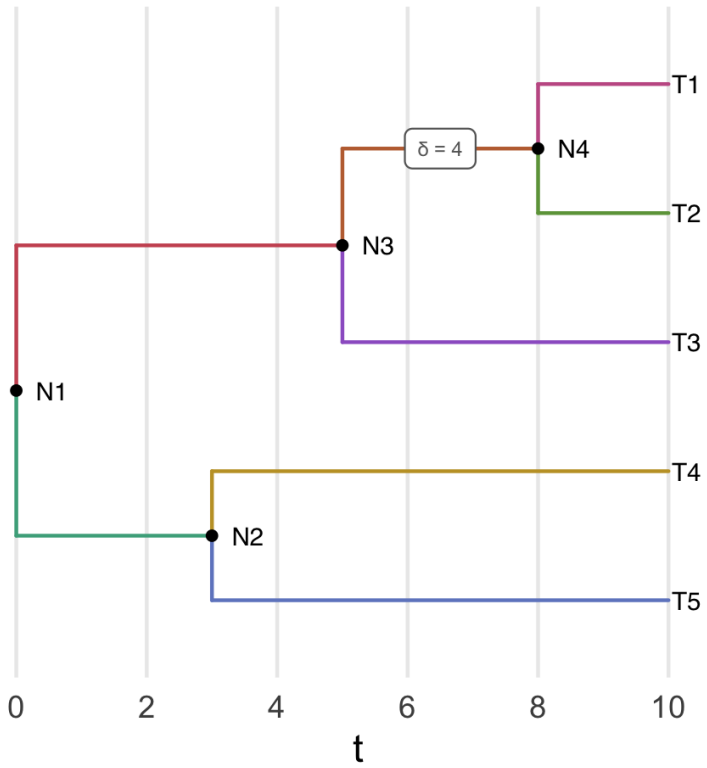
OU process on a tree with shifts δ



- On a branch, the process behaves like on $\mathbb{R}_+$.
- At each node, the process splits into two independent processes with the same initial value.
- The optimal value can shift at a node.

Bastide Mariadassou, et al. [BMR17]

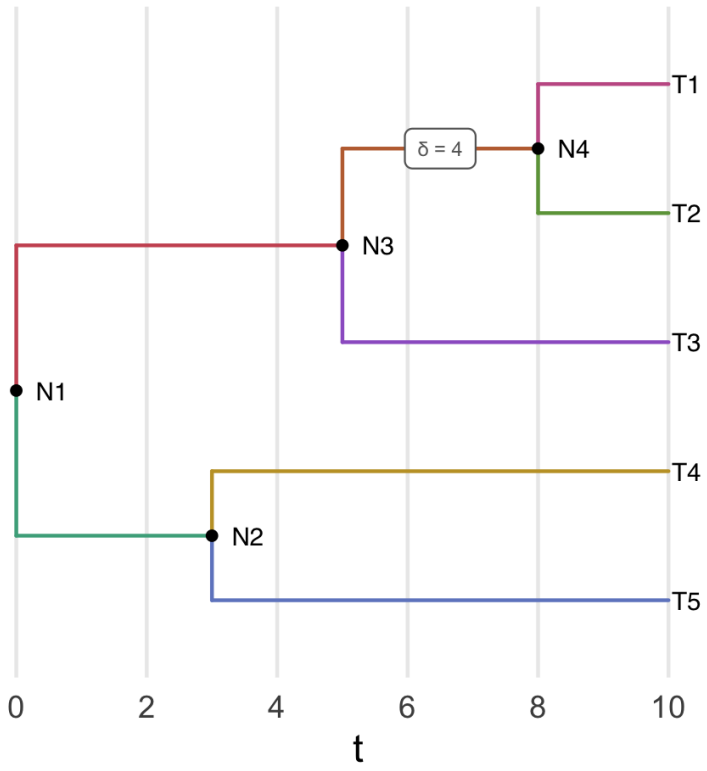
Incidence matrix and vector of shifts



$T = (\mathbb{1}_{\{i \in \mathrm{desc}(j)\}})_{ij} \in \{0,1\}^{m \times n}$ is the incidence matrix of the tree:

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Incidence matrix and vector of shifts



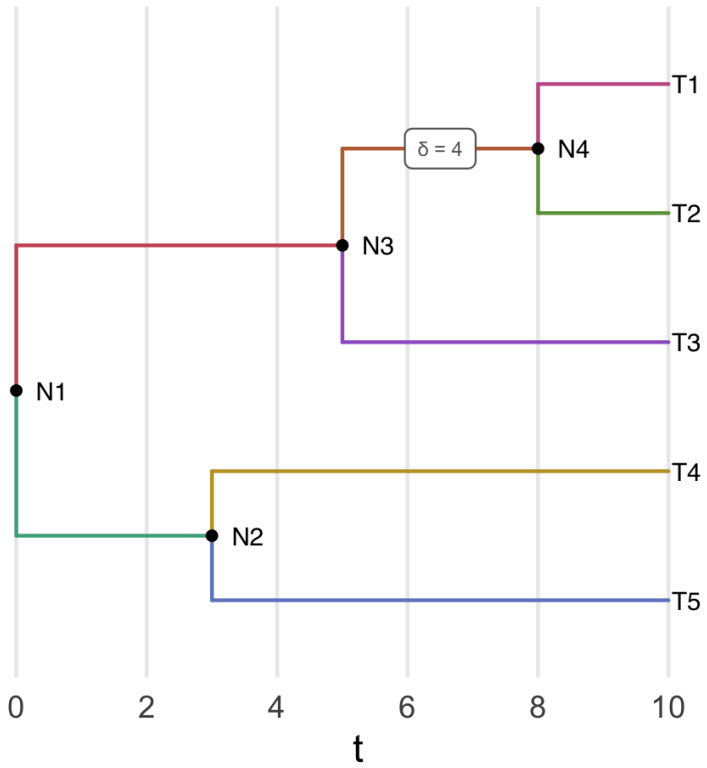
$T = (\mathbb{1}_{\{i \in \mathrm{desc}(j)\}})_{ij} \in \{0,1\}^{m \times n}$ is the incidence matrix of the tree:

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and δ the vector of shifts:

$$\begin{bmatrix} 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T.$$

Incidence matrix and vector of shifts



$T = (\mathbb{1}_{\{i \in \mathrm{desc}(j)\}})_{ij} \in \{0,1\}^{m \times n}$ is the incidence matrix of the tree:

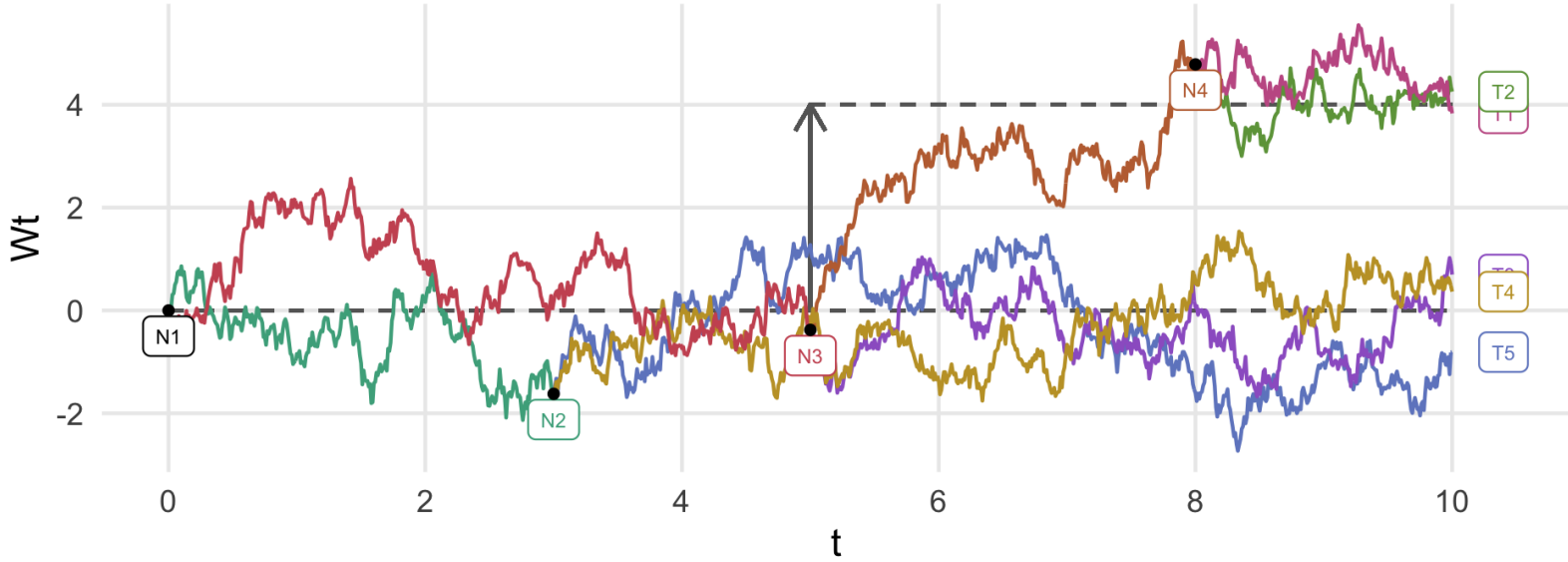
$$\begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and δ the vector of shifts:

$$\begin{bmatrix} 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T.$$

The product $T\delta$ is the vector of optimal values on leaves.

Random variables on leaves



The random variables on leaves are jointly Gaussian $\mathcal{N}_m(T\delta, \Sigma)$ with

$$\Sigma_{i,j} = \frac{\sigma^2_{\text{ou}}}{2\alpha_{\text{ou}}}\left(1 - e^{-2\alpha_{\text{ou}} t_{i,j}}\right) \times e^{-\alpha_{\text{ou}} d_{i,j}}.$$

First assumption

z is the realization of an OU on a tree with shifts δ.

First assumption

$\mathfrak{z}$ is the realization of an OU on a tree with shifts δ.

Then,

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

with $\mu = T\delta$ and Σ depends on α_{ou} and σ_{ou} by

$$\Sigma_{i,j} = \frac{\sigma^2_{\text{ou}}}{2\alpha_{\text{ou}}}\left(e^{-\alpha_{\text{ou}} d_{i,j}} - e^{-2\alpha_{\text{ou}} h}\right)$$

for a tree with total height h.

Second assumption

For a taxa i,

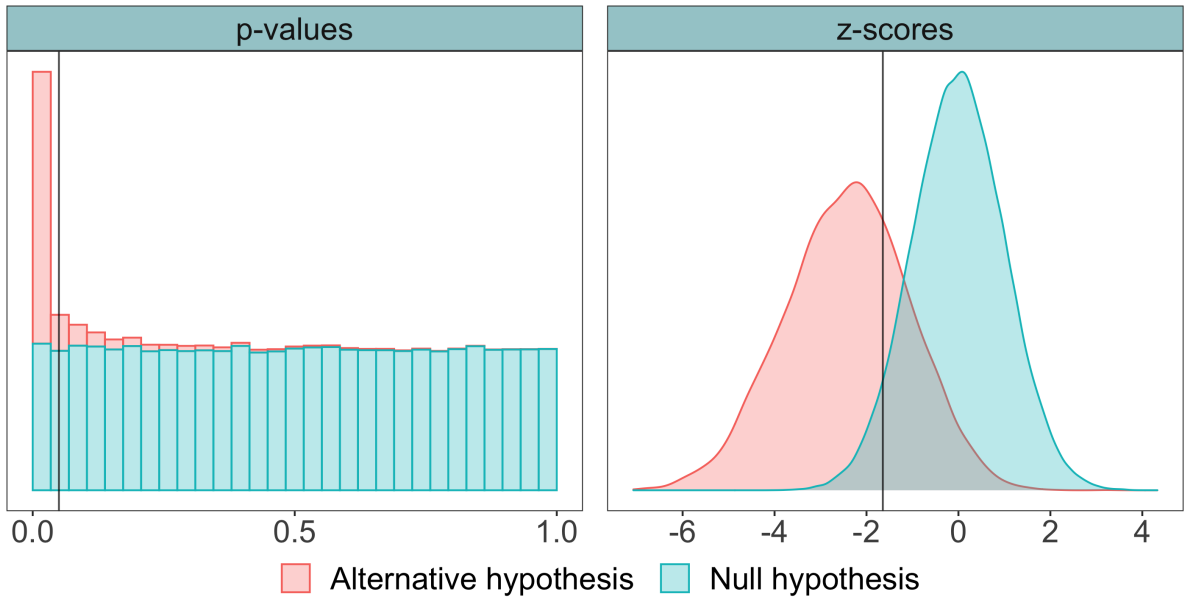
- if $\mathcal{H}_i \in \mathbb{H}_0$, $\mathfrak{p}_i \sim \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(0,1)$,
- if $\mathcal{H}_i \notin \mathbb{H}_0$, $\mathfrak{p}_i \preccurlyeq \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$ with $\mu_i < 0$.

McLachlan and Peel [MP04]

Second assumption

For a taxa i,

- if $\mathcal{H}_i \in \mathbb{H}_0$, $\mathfrak{p}_i \sim \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(0,1)$,
- if $\mathcal{H}_i \notin \mathbb{H}_0$, $\mathfrak{p}_i \preccurlyeq \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$ with $\mu_i < 0$.



McLachlan and Peel [MP04]

Second assumption

For a taxa i,

- if $\mathcal{H}_i \in \mathbb{H}_0$, $\mathfrak{p}_i \sim \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(0,1)$,
- if $\mathcal{H}_i \notin \mathbb{H}_0$, $\mathfrak{p}_i \preccurlyeq \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$ with $\mu_i < 0$.

Then,

$$\mathfrak{z} \sim \mathcal{N}_m\left(\mu \in \mathbb{R}_-^m, \Sigma\right).$$

One will find **differentially abundant** taxa by finding the **non-zero elements** of μ.

This impose $\Sigma_{i,i} = 1$ so $\sigma_{\text{ou}} = \frac{2\alpha_{\text{ou}}}{1-e^{-2\alpha_{\text{ou}}h}}$.

McLachlan and Peel [MP04]

Estimation of μ

With Σ known, a naive ML estimator gives

$$\hat{\mu} = \underset{\mu \in \mathbb{R}^m}{\operatorname{argmin}} \|\mathfrak{z} - \mu\|^2_{\Sigma^{-1},2}.$$

Estimation of μ

With Σ known, a naive ML estimator gives

$$\hat{\mu} = \underset{\mu \in \mathbb{R}_-^m}{\operatorname{argmin}} \|z - \mu\|^2_{\Sigma^{-1},2}.$$

To take the tree into account, $\hat{\mu} = T\hat{\delta}$ with

$$\hat{\delta} = \underset{\delta \in \mathbb{R}^n / T\delta \in \mathbb{R}_-^m}{\operatorname{argmin}} \|z - T\delta\|^2_{\Sigma^{-1},2}.$$

Estimation of μ

With Σ known, a naive ML estimator gives

$$\hat{\mu} = \underset{\mu \in \mathbb{R}_-^m}{\operatorname{argmin}} \|\mathfrak{z} - \mu\|^2_{\Sigma^{-1},2}.$$

To take the tree into account, $\hat{\mu} = T\hat{\delta}$ with

$$\hat{\delta} = \underset{\delta \in \mathbb{R}^n / T\delta \in \mathbb{R}_-^m}{\operatorname{argmin}} \|\mathfrak{z} - T\delta\|^2_{\Sigma^{-1},2}.$$

To add hierarchically coherent sparsity in our estimate

$$\hat{\delta} = \underset{\delta \in \mathbb{R}^n / T\delta \in \mathbb{R}_-^m}{\operatorname{argmin}} \|\mathfrak{z} - T\delta\|^2_{\Sigma^{-1},2} + \lambda\|\delta\|_1.$$

Estimation of μ (bis)

By Cholesky decomposition, $\Sigma^{-1} = R^T R$

$$\begin{aligned}
\|z - T\delta\|^2_{\Sigma^{-1},2} &= (z - T\delta)^T \Sigma^{-1} (z - T\delta) \\
&= (z - T\delta)^T R^T R (z - T\delta) \\
&= (Rz - RT\delta)^T (Rz - RT\delta) \\
&= (y - X\delta)^T (y - X\delta) = \|y - X\delta\|_2^2
\end{aligned}$$

with $y = Rz$ and $X = RT$.

Estimation of μ (bis)

By Cholesky decomposition, $\Sigma^{-1} = R^T R$

$$\begin{aligned}
\|\mathfrak{z} - T\delta\|^2_{\Sigma^{-1},2} &= (\mathfrak{z} - T\delta)^T \Sigma^{-1} (\mathfrak{z} - T\delta) \\
&= (\mathfrak{z} - T\delta)^T R^T R (\mathfrak{z} - T\delta) \\
&= (R\mathfrak{z} - RT\delta)^T (R\mathfrak{z} - RT\delta) \\
&= (y - X\delta)^T (y - X\delta) = \|y - X\delta\|_2^2
\end{aligned}$$

with $y = R\mathfrak{z}$ and $X = RT$.

Finally, we fall back to a constrained lasso problem:

$$\hat{\delta} = \underset{\delta \in \mathbb{R}^n / T\delta \in \mathbb{R}^m_-}{\operatorname{argmin}} \|y - X\delta\|_2^2 + \lambda \|\delta\|_1.$$

Numerical resolution

The previous problem could be numerically solved by the shooting algorithm, **iterating unidirectional updates** form the associated problem:

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

Fu [98]

Numerical resolution

The previous problem could be numerically solved by the shooting algorithm, **iterating unidirectional updates** form the associated problem:

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

But Σ and λ are not yet known.

Fu [98]

Estimation of Σ and choice of λ

$\widehat{\Sigma} = \left(\frac{e^{-\hat{\alpha}_{\text{ou}} d_{ij}} - e^{-2\hat{\alpha}_{\text{ou}} h}}{1 - e^{-2\hat{\alpha}_{\text{ou}} h}} \right)_{i,j}$ is determined by $\hat{\alpha}_{\text{ou}}$.

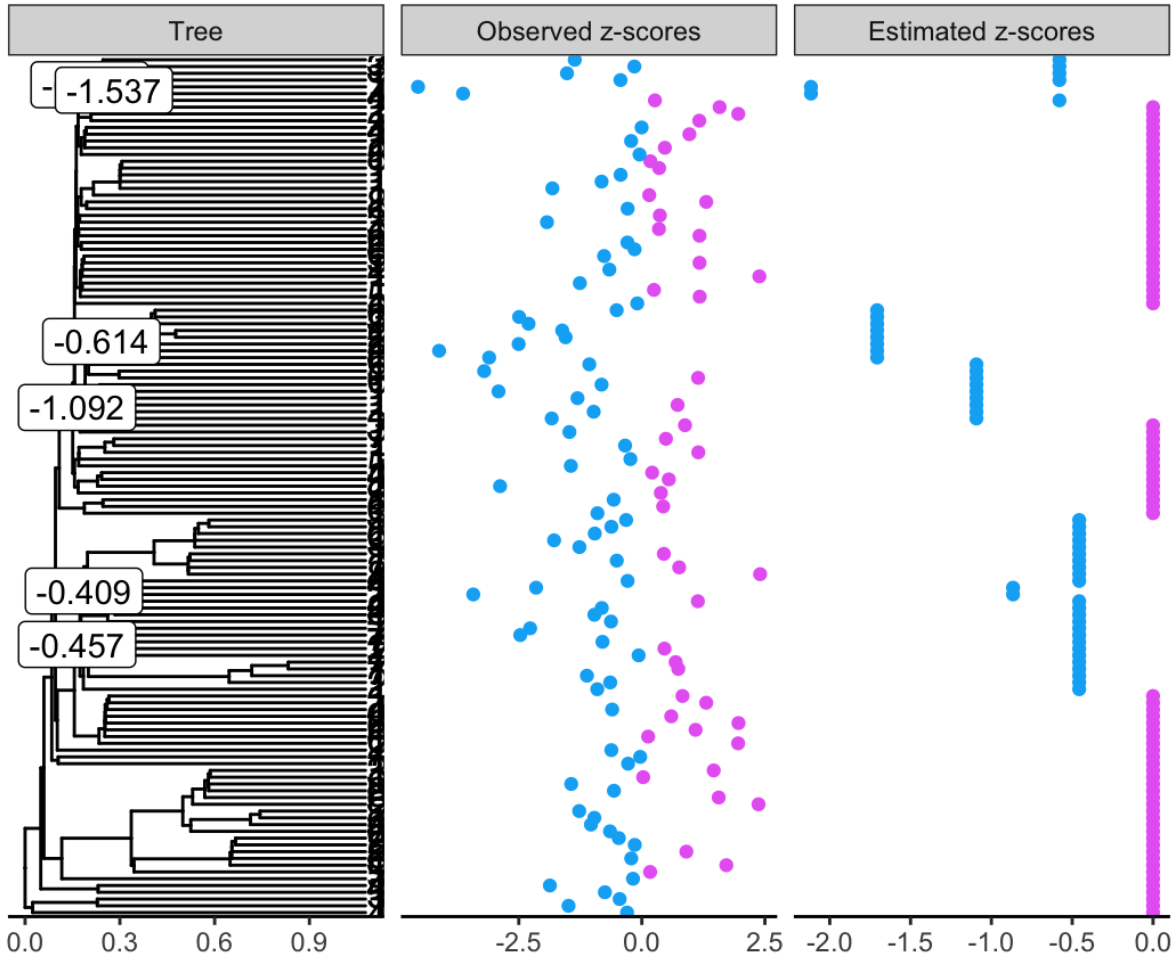
Estimation of Σ and choice of λ

$\widehat{\Sigma} = \left(\frac{e^{-\hat{\alpha}_{\text{ou}} d_{ij}} - e^{-2\hat{\alpha}_{\text{ou}} h}}{1 - e^{-2\hat{\alpha}_{\text{ou}} h}} \right)_{i,j}$ is determined by $\hat{\alpha}_{\text{ou}}$.

The optimal $(\alpha_{\text{ou}}, \lambda)$, is chosen on a bidimensional grid as the argmin of the BIC

$$\|\mathfrak{z} - T\delta_{\alpha_{\text{ou}},\lambda}\|^2_{\Sigma(\alpha_{\text{ou}})^{-1},2} + \log|\Sigma(\alpha_{\text{ou}})| + \|\delta_{\alpha_{\text{ou}},\lambda}\|_0 \log m.$$

Effect of hierarchical smoothing



Find non zero values

Estimation from lasso provides biased estimators without confidence intervals.

We need confidence intervals on $\hat{\delta}$ and $\hat{\mu}$.

Find non zero values

Estimation from lasso provides biased estimators without confidence intervals.

We need confidence intervals on $\hat{\delta}$ and $\hat{\mu}$.

Use of a debiasing procedure

- **score system (ss) [ZZ14],**
- column-wise inverse (ci) [JM13; JM14].

Scaled lasso

The debiasing procedure requires a **initial joint estimator** of $\delta^{(\text{init})}$ and its associated standard error σ.

This can be done with a scaled lasso

$$\left(\hat{\delta}^{(\text{init})}, \hat{\sigma}\right) = \underset{\delta,\sigma}{\text{argmin}} \frac{\|y - X\delta\|_2^2}{2\sigma n} + \frac{\sigma}{2} + \lambda\|\delta\|_1.$$

Sun and Zhang [SZ12]

Score system

The **score system** $S \in \mathbb{R}^{n \times p}$ associated with X and y where s_j is the residual of the (classical) lasso regression of y against X_{-j}:

$$s_j = y - \delta_{\text{lasso}}^{-j} X_{-j}.$$

S is as **weak orthogonalisation** of X.

Debiasing procedure

From the initial estimator $\hat{\delta}_j^{(\text{init})}$ of the scaled lasso, one can do a **one-step correction**

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

Debiasing procedure

From the initial estimator $\hat{\delta}_j^{(\text{init})}$ of the scaled lasso, one can do a **one-step correction**

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

Asymptotically, $\hat{\delta} \sim \mathcal{N}_n(\delta, V)$ with

$$v_{i,j} = \hat{\sigma} \frac{\langle s_i, s_j \rangle}{\langle s_i, x_i \rangle \langle s_j, x_j \rangle}.$$

Debiasing procedure

From the initial estimator $\hat{\delta}_j^{(\text{init})}$ of the scaled lasso, one can do a **one-step correction**

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

Asymptotically, $\hat{\delta} \sim \mathcal{N}_n(\delta, V)$ with

$$v_{i,j} = \hat{\sigma} \frac{\langle s_i, s_j \rangle}{\langle s_i, x_i \rangle \langle s_j, x_j \rangle}.$$

Then the **bilateral confidence interval** for a shift $\hat{\delta}_j$ is

$$\left[\hat{\delta}_j \pm \phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{v_{j,j}} \right].$$

Smoothed p-values

To have the **unilateral hierarchically smoothed p-values** $\mathfrak{p}^{\mathrm{h}}$, we need to propagate the shifts with the incidence matrix T

$$\mathfrak{p}_i^{\mathrm{h}} = \Phi\left(\frac{t_{i.}^T \hat{\delta}}{\left(t_{i.}^T V t_{i.}\right)^{1/2}}\right)$$

with $t_{i.}$ the i^{th} row of T.

Multiple testing correction

Let $\mathrm{t}_i = \frac{t_{i.}^T \hat{\delta}}{\left(t_{i.}^T V t_{i.}\right)^{1/2}}$ be the t-scores, $t_{\max} = \sqrt{2\log m - 2\log\log m}$ and

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

with $R(t) = \sum_{i=1}^m \mathbb{1}_{\{\mathrm{t}_i \leq -t\}}$.

Javanmard Javadi, et al. [JJ19]

Multiple testing correction

Let $\mathrm{t}_i = \frac{t_{i.}^T \hat{\delta}}{\left(t_{i.}^T V t_{i.}\right)^{1/2}}$ be the t-scores, $t_{\max} = \sqrt{2\log m - 2\log\log m}$ and

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

with $R(t) = \sum_{i=1}^m \mathbb{1}_{\{\mathrm{t}_i \leq -t\}}$.

One reject if $\mathrm{t}_i \leq -t^\star$ and the associated **hierarchical q-values** are

$$\mathrm{q}_i^{\mathrm{h}} = \frac{\mathrm{p}_i^{\mathrm{h}} \alpha}{\Phi(-t^\star)}.$$

Javanmard Javadi, et al. [JJ19]

zazou overview

zazou is a recipe with four steps:

- a modeling of z-scores by an Ornstein-Uhlenbeck process on a tree with shifts,

zazou overview

zazou is a recipe with four steps:

- a modeling of z-scores by an Ornstein-Uhlenbeck process on a tree with shifts,
- a constrained scaled lasso to estimate a sparse distribution of the shifts,

zazou overview

zazou is a recipe with four steps:

- a modeling of z-scores by an Ornstein-Uhlenbeck process on a tree with shifts,
- a constrained scaled lasso to estimate a sparse distribution of the shifts,
- a debiasing procedure for the scaled lasso,

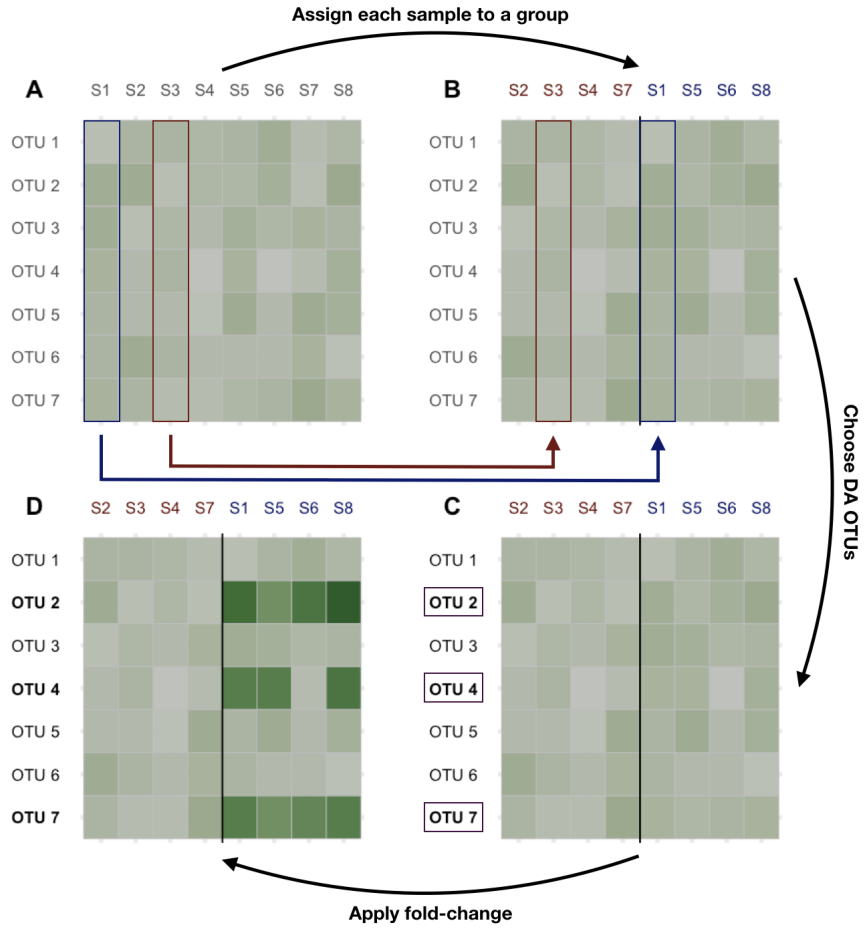
zazou overview

zazou is a recipe with four steps:

- a modeling of z-scores by an Ornstein-Uhlenbeck process on a tree with shifts,
- a constrained scaled lasso to estimate a sparse distribution of the shifts,
- a debiasing procedure for the scaled lasso,
- a debiased lasso designed multiple testing procedure.

Evaluation of zazou

Simulations



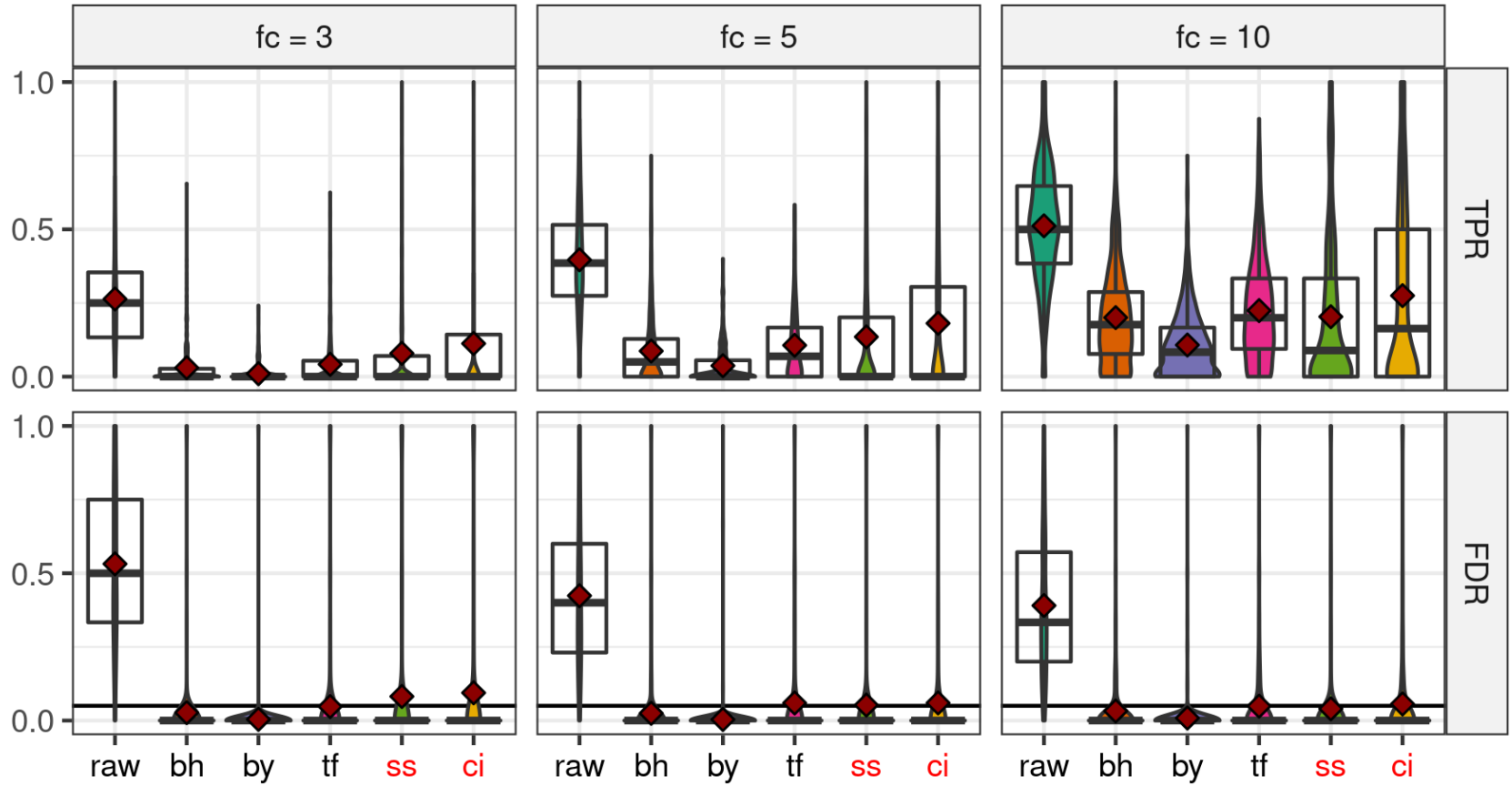
The choice of differentially abundant taxa is done in a hierarchically consistent manner.

- $m = 127, p = 49$ split into two groups.
- Fold-changes of 3, 5 and 10.
- 5 estimated proportions of π_1.
- 100 replicates for each parameters.

Bichat Plassais, et al. [Bic+20]

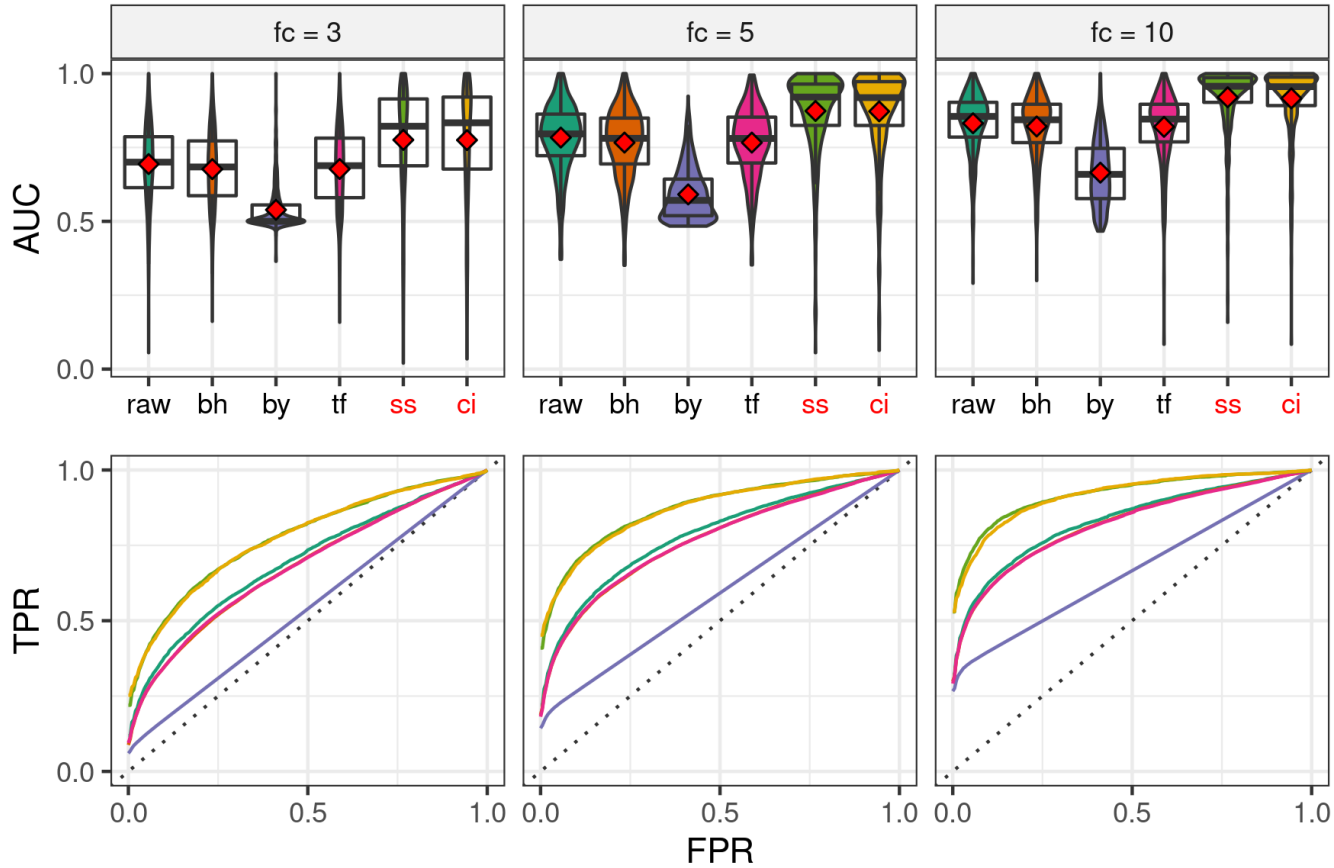
Results

zazou increases the TPR but does not always control the FDR.



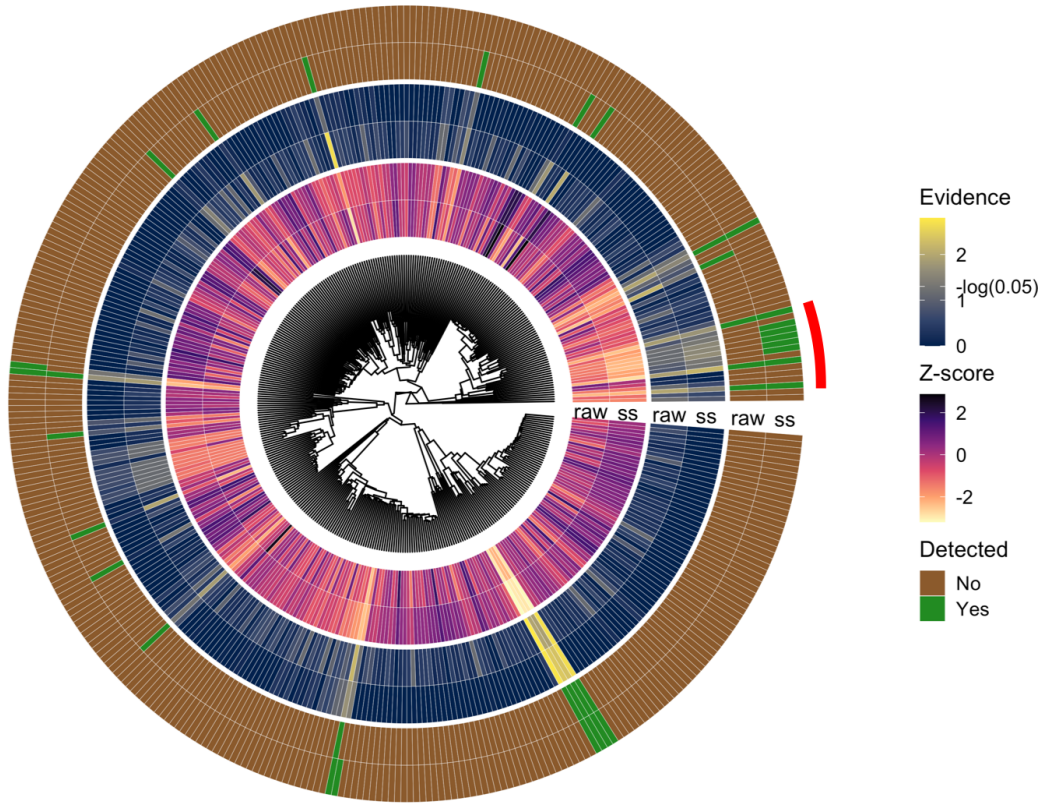
Results

zazou has better AUC and ROC curves.



Association with age

zazou identifies phylogenetically coherent taxa.



Brito Yilmaz, et al. [Bri+16]

Conclusions

Conclusions

- Phylogeny does not capture the true structure of the data.

Conclusions

- Phylogeny does not capture the true structure of the data.
- Previous methods (hFDR, TreeFDR) does not work in practice.

Conclusions

- Phylogeny does not capture the true structure of the data.
- Previous methods (hFDR, TreeFDR) does not work in practice.
- zazou algorithm improves performances a bit ...

Conclusions

- Phylogeny does not capture the true structure of the data.
- Previous methods (hFDR, TreeFDR) does not work in practice.
- zazou algorithm improves performances a bit ...
- ... but the choice of the rejection threshold could be improved.

Conclusions

- Phylogeny does not capture the true structure of the data.
- Previous methods (hFDR, TreeFDR) does not work in practice.
- zazou algorithm improves performances a bit ...
- ... but the choice of the rejection threshold could be improved.
- R packages `correlationtree` and `zazou` are available on GitHub.

Conclusions

- Phylogeny does not capture the true structure of the data.
- Previous methods (hFDR, TreeFDR) does not work in practice.
- zazou algorithm improves performances a bit ...
- ... but the choice of the rejection threshold could be improved.
- R packages `correlationtree` and `zazou` are available on GitHub.
- Articles published in Frontiers in Microbiology and submitted in Statistics and Computing.

Outlooks

- More theoretical work on zazou is required.

Outlooks

- More theoretical work on zazou is required.
- `zazou` could be speed up.

Outlooks

- More theoretical work on zazou is required.
- `zazou` could be speed up.
- Use hierarchical information during testing step, and not only correction step.

Outlooks

- More theoretical work on zazou is required.
- `zazou` could be speed up.
- Use hierarchical information during testing step, and not only correction step.
- Use the framework of prediction instead of association.

Thanks!

Manuscript

abichat.github.io

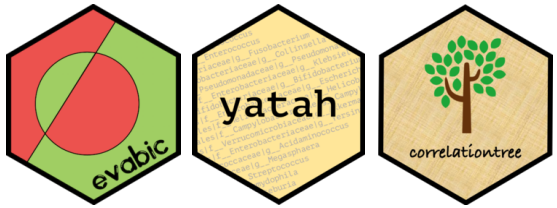
antoinebichat

@_abichat

@abichat

université
PARIS-SACLAY

R packages



evabic evaluates the performance of binary classifiers. It can compute compute 19 different measures with tidy outputs.

yatah provides functions to manage taxonomy when lineages are described like `k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria`.

correlationtree computes correlation trees.

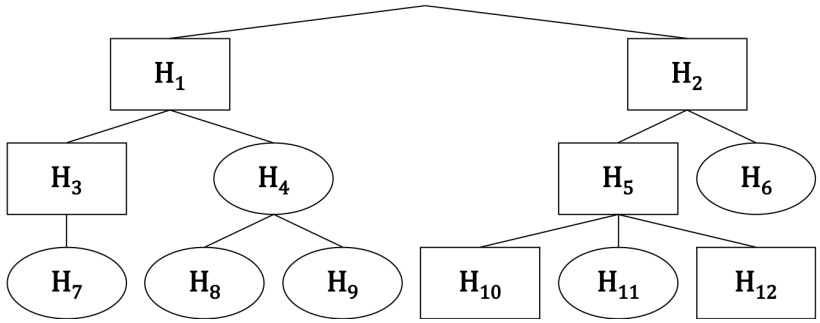
zazou implements the zazou procedure.

hadamardown contains the Université Paris-Saclay PhD manuscript template for bookdown.

Hierarchical FDR

This procedure inscreases statistical power by lessening the number of test to do with a descending method:

- test the family $\mathcal{T}_0$,
- if node t is rejected, test $\mathcal{T}_t = \{H_i \mid \mathrm{Par}(i) = t\}$ with a BH procedure at level α.

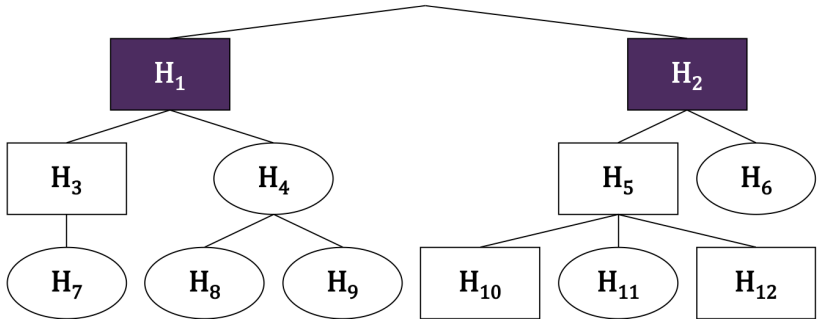


Yekutieli [Yek08]; Sankaran and Holmes [SH14]

Hierarchical FDR

This procedure inscreases statistical power by lessening the number of test to do with a descending method:

- test the family $\mathcal{T}_0$,
- if node t is rejected, test $\mathcal{T}_t = \{H_i \mid \mathrm{Par}(i) = t\}$ with a BH procedure at level α.

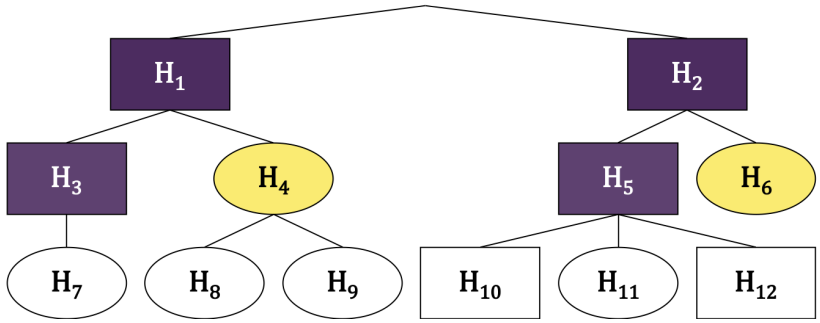


Yekutieli [Yek08]; Sankaran and Holmes [SH14]

Hierarchical FDR

This procedure inscreases statistical power by lessening the number of test to do with a descending method:

- test the family $\mathcal{T}_0$,
- if node t is rejected, test $\mathcal{T}_t = \{H_i \mid \mathrm{Par}(i) = t\}$ with a BH procedure at level α.

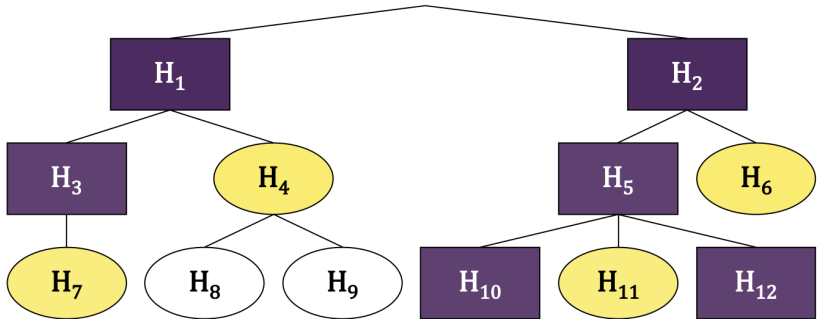


Yekutieli [Yek08]; Sankaran and Holmes [SH14]

Hierarchical FDR

This procedure inscreases statistical power by lessening the number of test to do with a descending method:

- test the family $\mathcal{T}_0$,
- if node t is rejected, test $\mathcal{T}_t = \{H_i \mid \mathrm{Par}(i) = t\}$ with a BH procedure at level α.

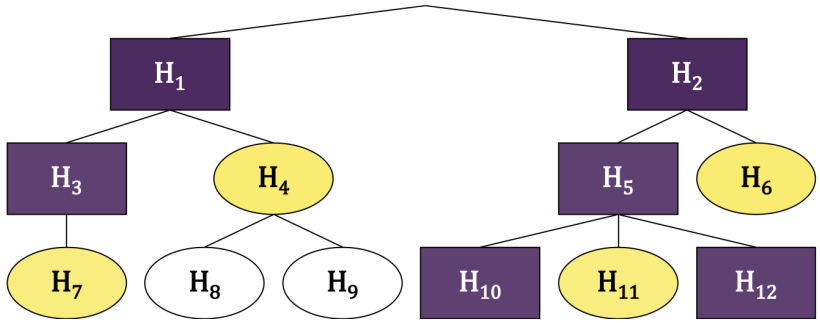


Yekutieli [Yek08]; Sankaran and Holmes [SH14]

Hierarchical FDR

This procedure inscreases statistical power by lessening the number of test to do with a descending method:

- test the family $\mathcal{T}_0$,
- if node t is rejected, test $\mathcal{T}_t = \{H_i \mid \mathrm{Par}(i) = t\}$ with a BH procedure at level α.

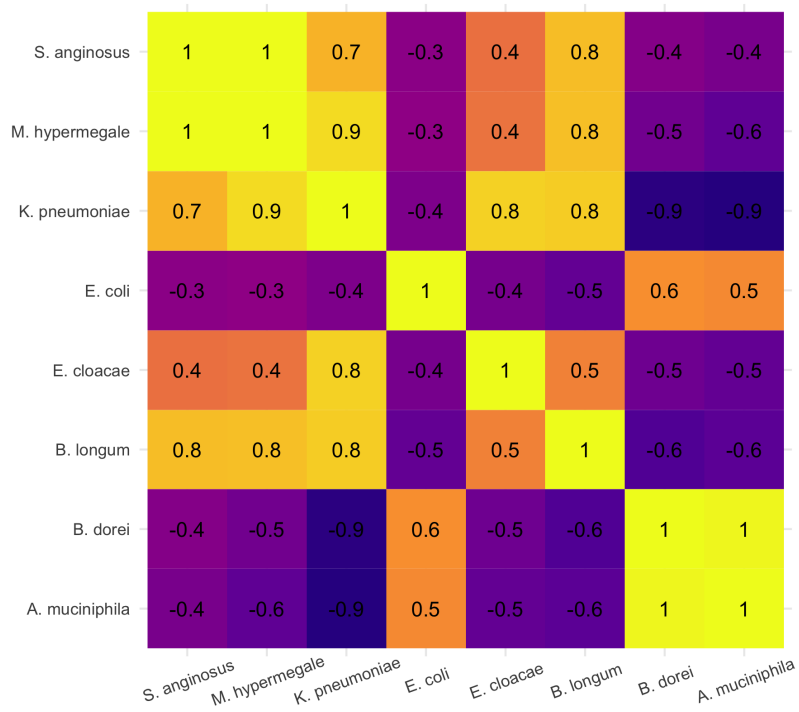


The FDR is controlled at level $\alpha' = 1.44 \times \alpha \times \frac{\#\text{discoveries} + \#\text{families tested}}{\#\text{discoveries} + 1}$.

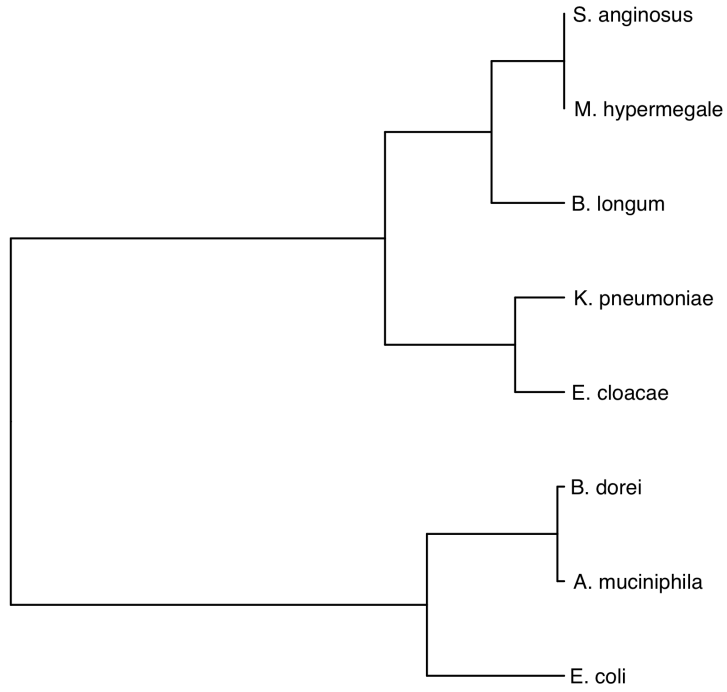
Yekutieli [Yek08]; Sankaran and Holmes [SH14]

Correlation tree

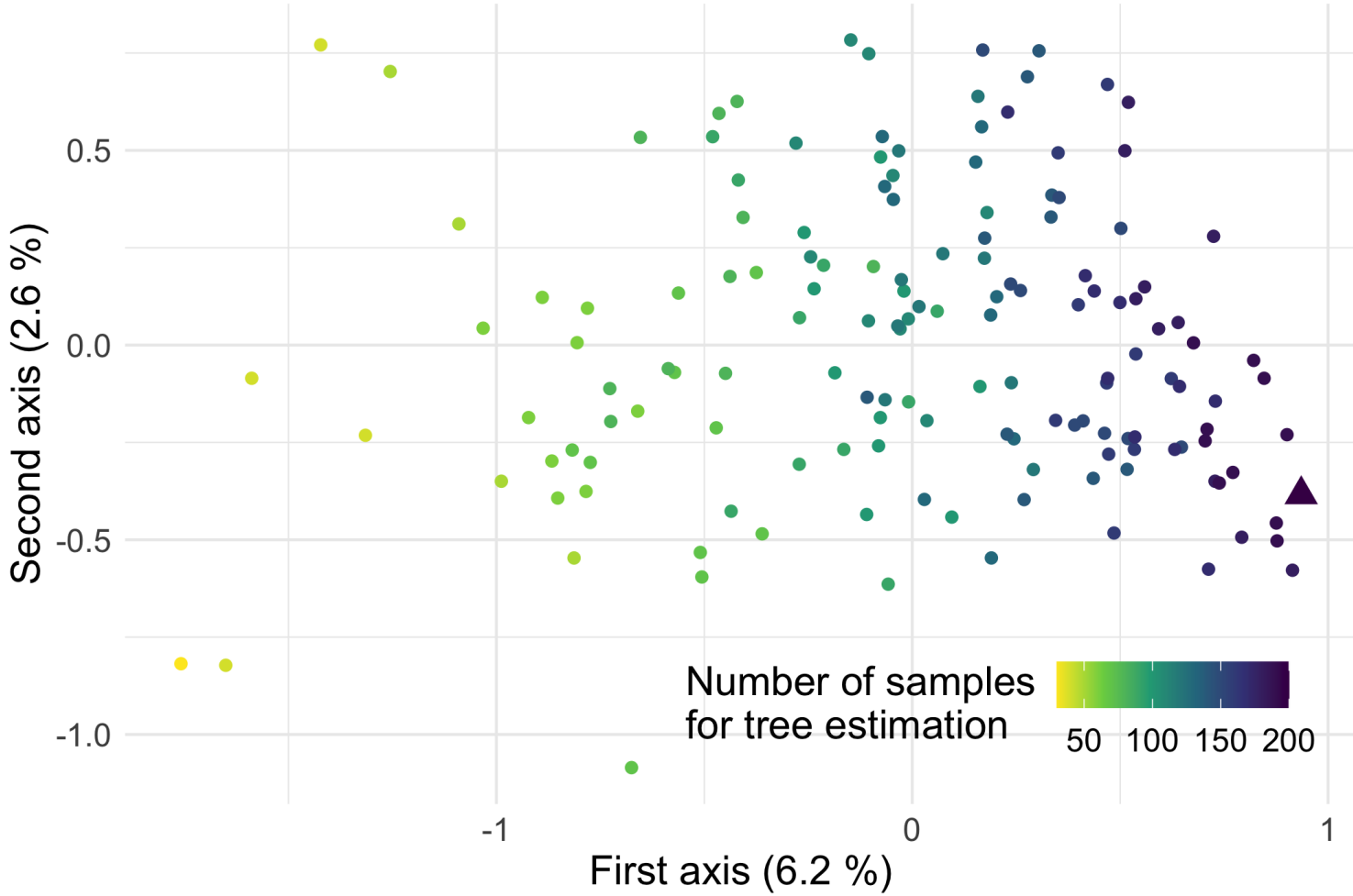
Matrix of pairwise correlation C



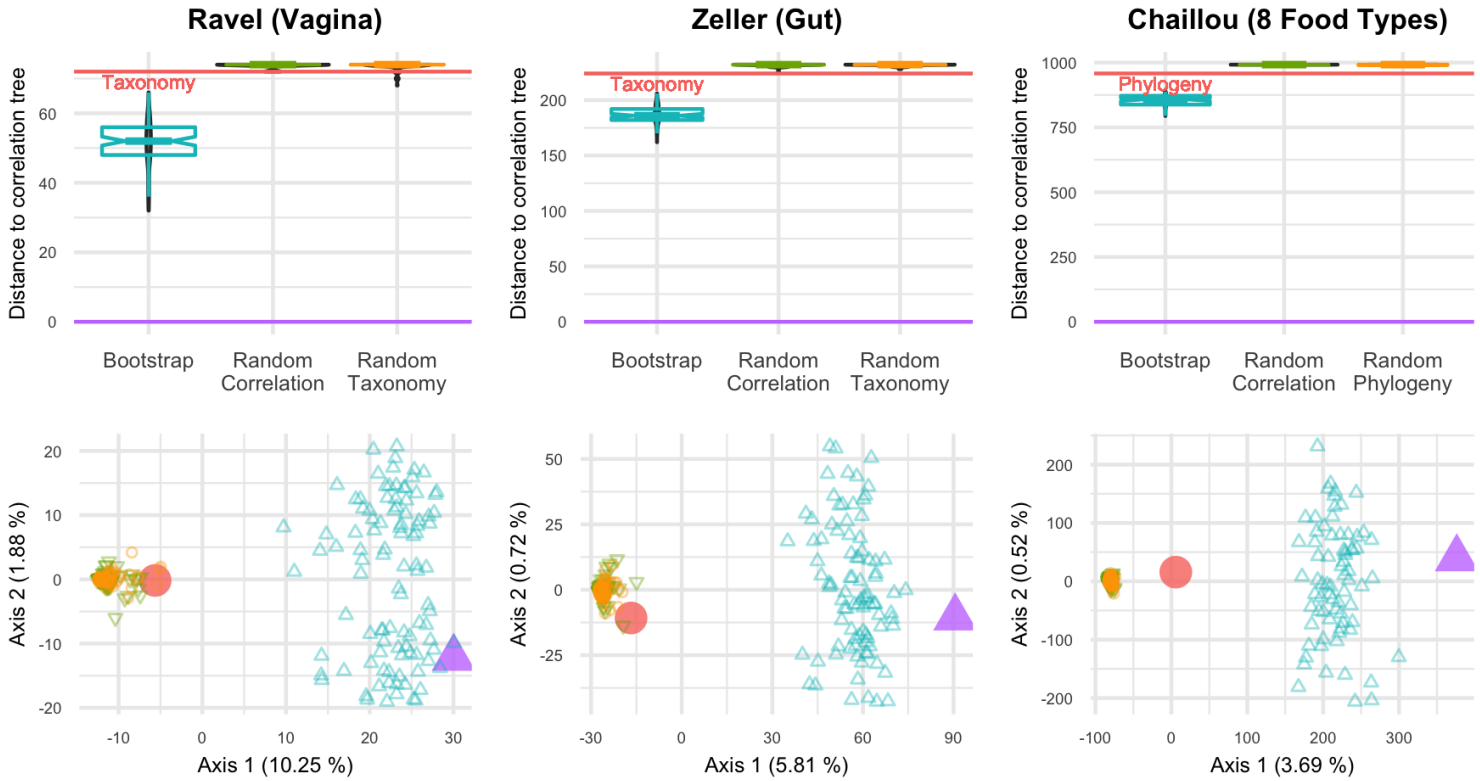
Hierarchical clustering on $1 - C$



Convergence to the correlation tree



With Robinson-Foulds distance



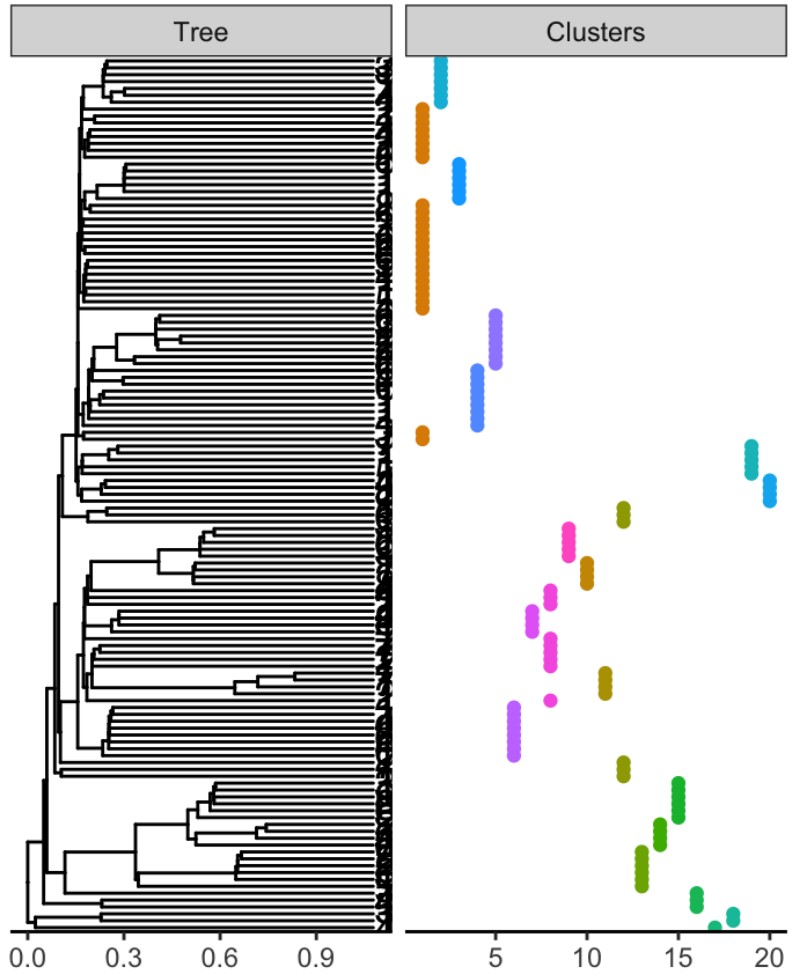
Robinson and Foulds [RF81]

Hierarchically consistent taxa

> The choice of differentially abundant taxa is done in a hierarchically consistent manner.

A partitioning around medoids (PAM) algorithm on the patristic distances matrix $\left(d_{i,j}\right)_{i,j}$ is use for this purpose.

1 to 5 clusters among 20 are selected to be differentially abundant.

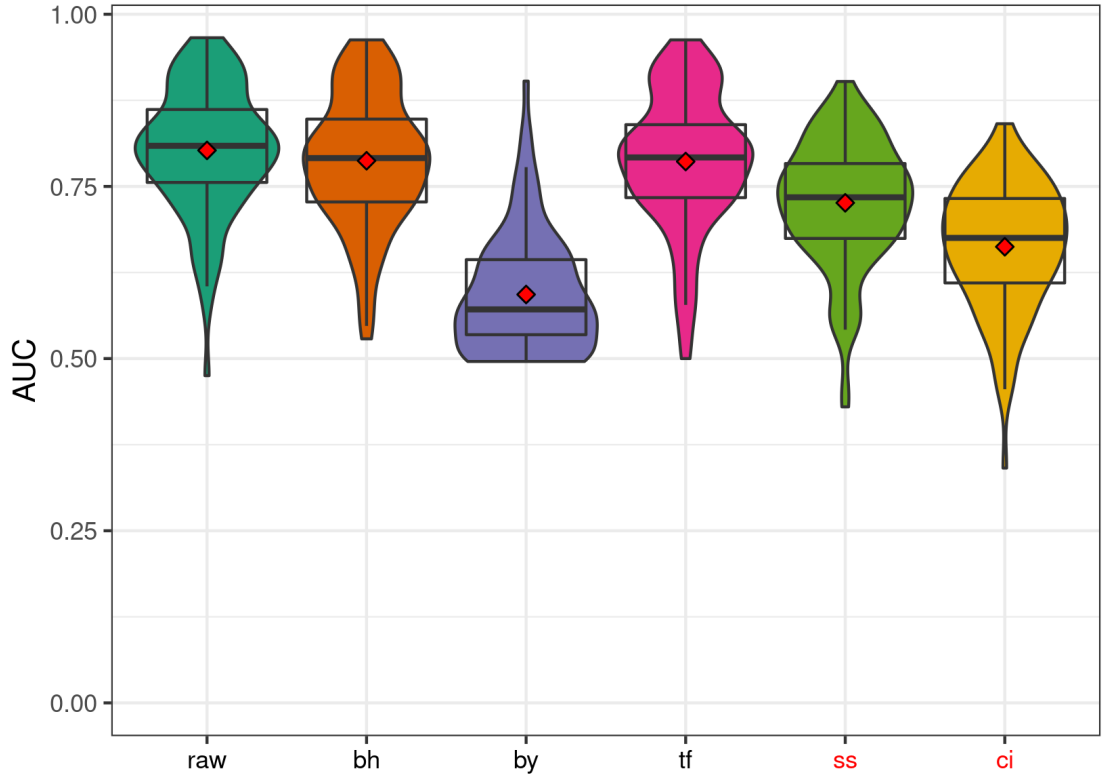


Quantitative results about FDP

(in %)	FDR			P(FDP > 5%)		
method	fc = 3	fc = 5	fc = 10	fc = 3	fc = 5	fc = 10
bh	2.69	2.46	3.23	4.40	7.20	12.20
by	0.40	0.36	0.85	0.40	0.60	2.20
tf	4.75	5.96	4.90	8.72	14.58	18.06
ss	8.17	5.26	3.99	12.40	8.40	8.80
ci	9.41	5.98	5.63	14.80	12.60	18.60

Negative simulations

When differentially abundant taxa are chosen uniformly.



References

Bastide, P., M. Mariadassou, and S. Robin. "Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree Series B Statistical methodology". (2017).

Benjamini, Y. and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289-300.

Benjamini, Y. and D. Yekutieli. "The control of the false discovery rate in multiple testing under dependency". In: *Annals of statistics* (2001), pp. 1165-1188.

Bichat, A., J. Plassais, C. Ambroise, and M. Mariadassou. "Incorporating Phylogenetic Information in Microbiome Differential Abundance Studies Has No Effect on Detection Power and FDR Control". In: *Frontiers in Microbiology* 11 (2020), p. 649. ISSN: 1664-302X.

References

Billera, L. J., S. P. Holmes, and K. Vogtmann. "Geometry of the space of phylogenetic trees". In: *Advances in Applied Mathematics* 27.4 (2001), pp. 733-767.

Bokulich, N. A., J. Chung, T. Battaglia, N. Henderson, M. Jay, H. Li, A. D. Lieber, F. Wu, G. I. Perez-Perez, Y. Chen, and others. "Antibiotics, birth mode, and diet shape microbiome maturation during early life". In: *Science translational medicine* 8.343 (2016), pp. 343ra82-343ra82.

Brito, I. L., S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, M. Tamminen, C. Smillie, J. R. Wortman, and others. "Mobile genes in the human microbiome are structured from global to individual scales". In: *Nature* 535.7612 (2016), pp. 435-439.

Canani, R. B., M. Di Costanzo, L. Leone, M. Pedata, R. Meli, and A. Calignano. "Potential beneficial effects of butyrate in intestinal and extraintestinal diseases". In: *World journal of gastroenterology: WJG* 17.12 (2011), p. 1519.

References

Chaillou, S., A. Chaulot-Talmon, H. Caekebeke, M. Cardinal, S. Christieans, C. Denis, M. H. Desmonts, X. Dousset, C. Feurer, E. Hamon, and others. "Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage". In: *The ISME journal* 9.5 (2015), pp. 1105-1118.

Cryan, J. F., K. J. O'Riordan, C. S. Cowan, K. V. Sandhu, T. F. Bastiaanssen, M. Boehme, M. G. Codagnone, S. Cussotto, C. Fulling, A. V. Golubeva, and others. "The microbiota-gut-brain axis". In: *Physiological reviews* 99.4 (2019), pp. 1877-2013.

David, L. A., C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, and others. "Diet rapidly and reproducibly alters the human gut microbiome". In: *Nature* 505.7484 (2014), pp. 559-563.

Flint, H. J., K. P. Scott, S. H. Duncan, P. Louis, and E. Forano. "Microbial degradation of complex carbohydrates in the gut". In: *Gut microbes* 3.4 (2012), pp. 289-306.

References

Fu, W. J. "Penalized regressions: the bridge versus the lasso". In: *Journal of computational and graphical statistics* 7.3 (1998), pp. 397-416.

Javanmard, A., H. Javadi, and others. "False discovery rate control via debiased lasso". In: *Electronic Journal of Statistics* 13.1 (2019), pp. 1212-1253.

Javanmard, A. and A. Montanari. "Confidence intervals and hypothesis testing for high-dimensional regression". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 2869-2909.

--- "Confidence intervals and hypothesis testing for high-dimensional statistical models". In: *Advances in Neural Information Processing Systems*. 2013, pp. 1187-1195.

References

Kates, A. E., O. Jarrett, J. H. Skarlupka, A. Sethi, M. Duster, L. Watson, G. Suen, K. Poulsen, and N. Safdar. "Household Pet Ownership and the Microbial Diversity of the Human Gut Microbiota". In: *Frontiers in Cellular and Infection Microbiology* 10 (2020), p. 73.

Ley, R. E., D. A. Peterson, and J. I. Gordon. "Ecological and evolutionary forces shaping microbial diversity in the human intestine". In: *Cell* 124.4 (2006), pp. 837-848.

McLachlan, G. J. and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Morgan, X. C., T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, and others. "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment". In: *Genome biology* 13.9 (2012), p. R79.

References

Opstelten, J. L., J. Plassais, S. W. van Mil, E. Achouri, M. Pichaud, P. D. Siersema, B. Oldenburg, and A. C. Cervino. "Gut microbial diversity is reduced in smokers with Crohn's disease". In: *Inflammatory bowel diseases* 22.9 (2016), pp. 2070-2077.

Palleja, A., K. H. Mikkelsen, S. K. Forslund, A. Kashani, K. H. Allin, T. Nielsen, T. H. Hansen, S. Liang, Q. Feng, C. Zhang, and others. "Recovery of gut microbiota of healthy adults following antibiotic exposure". In: *Nature microbiology* 3.11 (2018), pp. 1255-1265.

Philippot, L., S. G. Andersson, T. J. Battin, J. I. Prosser, J. P. Schimel, W. B. Whitman, and S. Hallin. "The ecological coherence of high bacterial taxonomic ranks". In: *Nature Reviews Microbiology* 8.7 (2010), pp. 523-529.

Qin, N., F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. Le Chatelier, J. Yao, L. Wu, and others. "Alterations of the human gut microbiome in liver cirrhosis". In: *Nature* 513.7516 (2014), pp. 59-64.

References

Ravel, J., P. Gajer, Z. Abdo, G. M. Schneider, S. S. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, and others. "Vaginal microbiome of reproductive-age women". In: *Proceedings of the National Academy of Sciences* 108.Supplement 1 (2011), pp. 4680-4687.

Robinson, D. F. and L. R. Foulds. "Comparison of phylogenetic trees". In: *Mathematical biosciences* 53.1-2 (1981), pp. 131-147.

Sankaran, K. and S. Holmes. "structSSI: simultaneous and selective inference for grouped or hierarchically structured data". In: *Journal of statistical software* 59.13 (2014), p. 1.

Sun, T. and C. Zhang. "Scaled sparse linear regression". In: *Biometrika* 99.4 (2012), pp. 879-898.

References

Xiao, J., H. Cao, and J. Chen. "False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing". In: *Bioinformatics* 33.18 (2017), pp. 2873-2881.

Yatsunenko, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, and others. "Human gut microbiome viewed across age and geography". In: *nature* 486.7402 (2012), pp. 222-227.

Yekutieli, D. "Hierarchical false discovery rate-controlling methodology". In: *Journal of the American Statistical Association* 103.481 (2008), pp. 309-316.

Zeller, G., J. Tap, A. Y. Voigt, S. Sunagawa, J. R. Kultima, P. I. Costea, A. Amiot, J. Böhm, F. Brunetti, N. Habermann, and others. "Potential of fecal microbiota for early-stage detection of colorectal cancer". In: *Molecular systems biology* 10.11 (2014), p. 766.

References

Zhang, C. and S. S. Zhang. "Confidence intervals for low dimensional parameters in high dimensional linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014), pp. 217-242.

Zheng, P., B. Zeng, M. Liu, J. Chen, J. Pan, Y. Han, Y. Liu, K. Cheng, C. Zhou, H. Wang, and others. "The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-GABA cycle and schizophrenia-relevant behaviors in mice". In: *Science advances* 5.2 (2019), p. eaau8317.

université PARIS-SACLAY

Discovering multi-scale metagenomic signatures through hierarchical organization of species

PhD defense

Antoine BICHAT
LaMME – Enterome
December 9, 2020

C. Ambroise

M. Mariadassou

J. Plassais

F. Strozzi

université PARIS-SACLAY

Laboratoire de Mathématiques et Modélisation d'Évry LaMME

MaIAGE

enterome

Context

Microbiota

Ecological community of microorganisms that reside in an environmental niche.

Microbiota

Ecological community of microorganisms that reside in an environmental niche.

For human gut

- 10^{14} bacterial cells in one gut...
- ... weighting $2\ kg$.
- More than 1500 different species.
- More than 10 millions unique genes.
- Helpful for nutrient absorption and anti inflammatory properties.

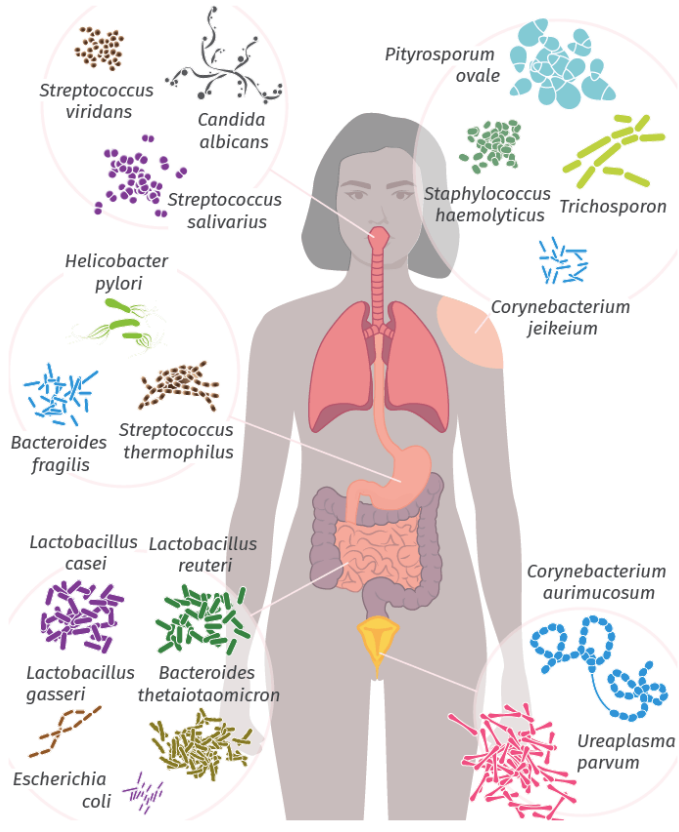
Ley Peterson, et al. [LPG06]; Flint Scott, et al. [Fli+12]; Canani Di Costanzo, et al. [Can+11]

Microbiota

Ecological community of microorganisms that reside in an environmental niche.

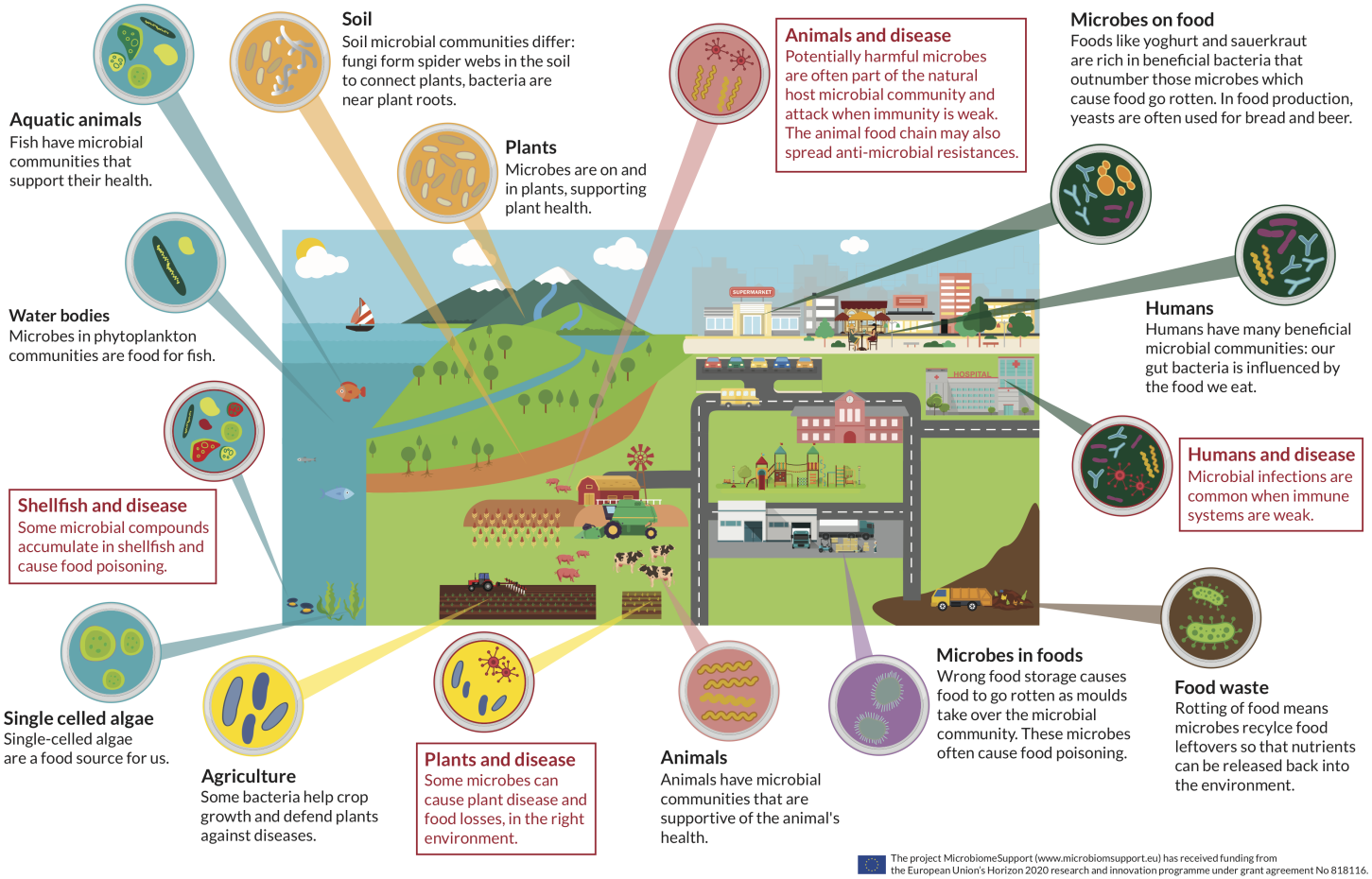
For human gut

- 10^{14} bacterial cells in one gut...
- ... weighting 2 kg.
- More than 1500 different species.
- More than 10 millions unique genes.
- Helpful for nutrient absorption and anti inflammatory properties.



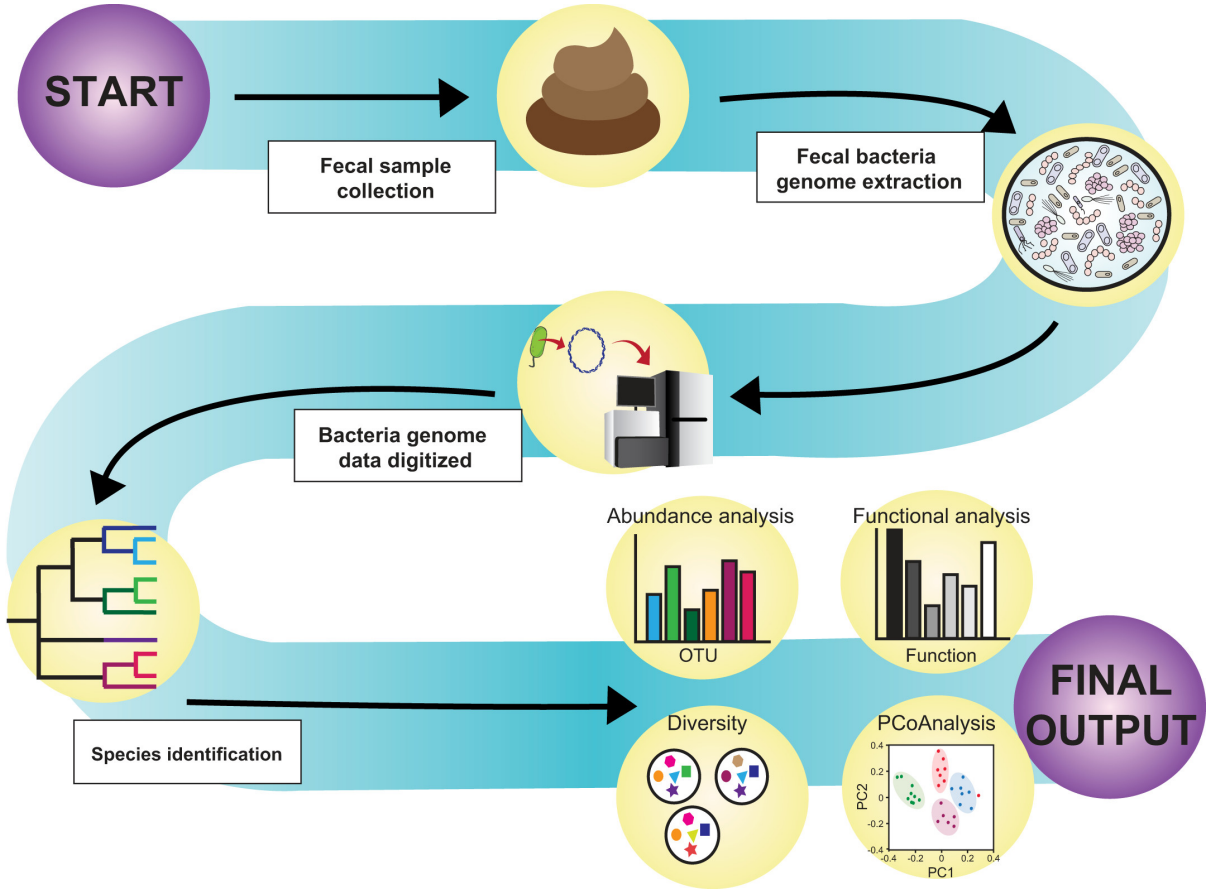
Ley Peterson, et al. [LPG06]; Flint Scott, et al. [Fli+12]; Canani Di Costanzo, et al. [Can+11] – lelivrescolaire.fr

Microbiota everywhere



The project MicrobiomeSupport (www.microbiomsupport.eu) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 818116.

Sequencing



Cryan O'Riordan, et al. [Cry+19]

Abundance table

- Matrix with m taxa and p samples.
- Count or compositional data, with inflation in zero.
- Correlation between abundances.

Abundance table

- Matrix with m taxa and p samples.
- Count or compositional data, with inflation in zero.
- Correlation between abundances.

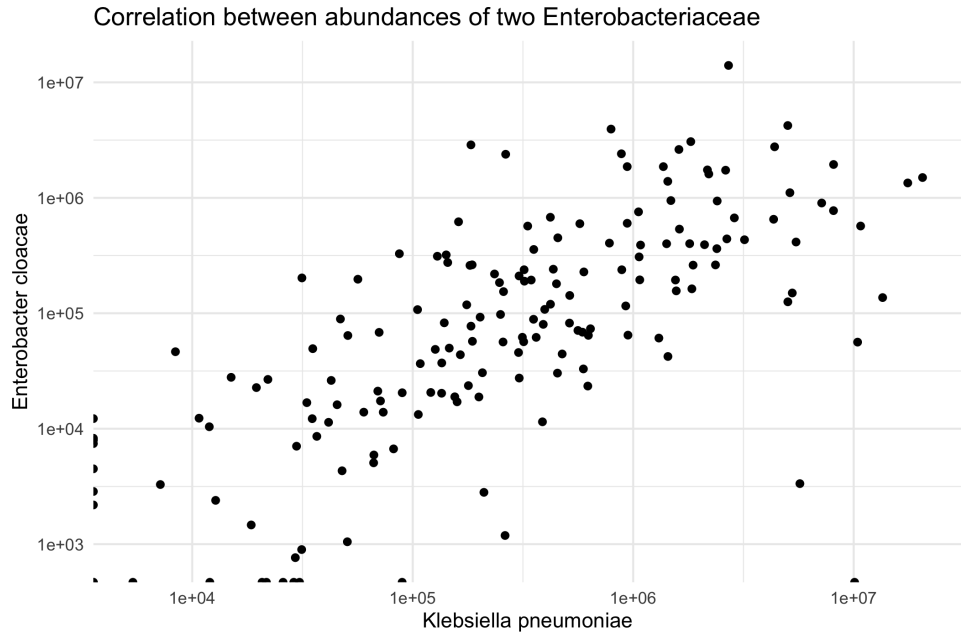
	S1	S2	S3	S4	S5	S6
Akkermansia muciniphila	0	934850	247037	0	181167	0
Bacteroides dorei	89893	2567192	648639	0	0	8498
Bifidobacterium longum	376119	0	30671	0	1292193	1830318
Enterobacter cloacae	756064	12234	142652	2186	238175	535786
Escherichia coli	256424	5551041	9905179	76052	70234	79805
Klebsiella pneumoniae	1057187	0	515407	0	887678	1620535
Megamonas hypermegale	0	0	0	0	0	27317
Streptococcus anginosus	0	0	0	985	0	19134

Brito Yilmaz, et al. [Bri+16]

Abundance table

- Matrix with m taxa and p samples.
- Count or compositional data, with inflation in zero.
- Correlation between abundances.

	S1	S2	S3	S4	S5	S6
Akkermansia muciniphila	0	934850	247037	0	181167	0
Bacteroides dorei	89893	2567192	648639	0	0	8498
Bifidobacterium longum	376119	0	30671	0	1292193	1830318
Enterobacter cloacae	756064	12234	142652	2186	238175	535786
Escherichia coli	256424	5551041	9905179	76052	70234	79805
Klebsiella pneumoniae	1057187	0	515407	0	887678	1620535
Megamonas hypermegale	0	0	0	0	0	27317
Streptococcus anginosus	0	0	0	985	0	19134



Brito Yilmaz, et al. [Bri+16]

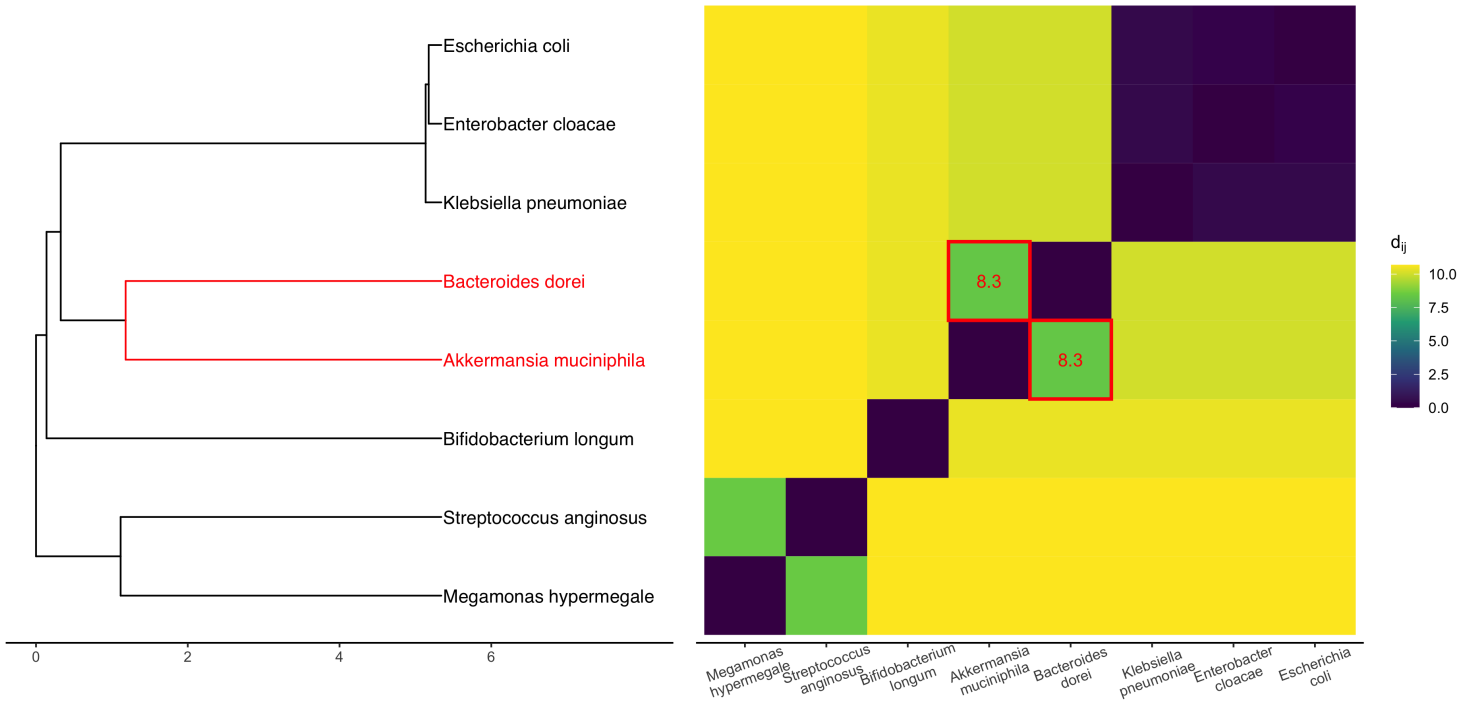
Phylogeny

- Tree with m leaves which describes the evolutionary history of the taxa.

Brito Yilmaz, et al. [Bri+16]

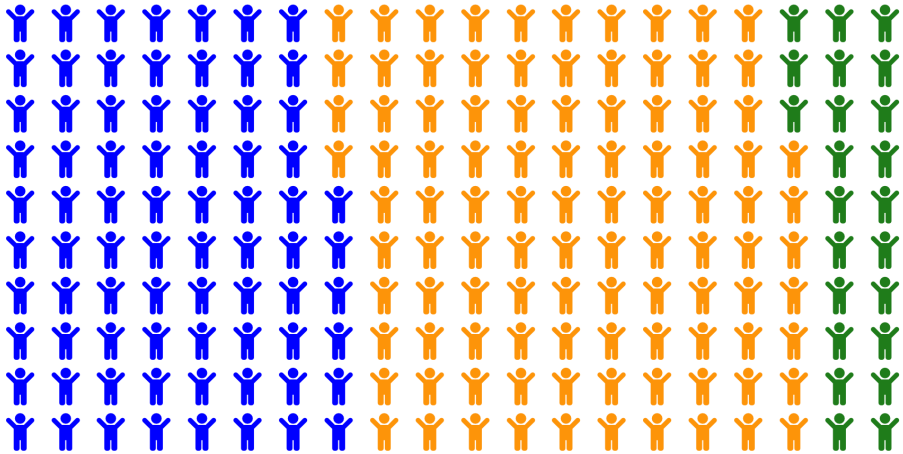
Phylogeny

- Tree with m leaves which describes the evolutionary history of the taxa.
- Associated with a patristic distance matrix $\left(d_{i,j}\right)_{i,j}$.



Brito Yilmaz, et al. [Bri+16]

Differential abundance studies



Used to find **associations** between microbiota and

- diet [Dav+14],
- birth mode [Bok+16],
- age [Yat+12],
- pet owning [Kat+20],
- tobacco [Ops+16],
- antibiotics [Pal+18],
- Crohn's disease [Mor+12],
- cirrhosis [Qin+14]
- schizophrenia [Zhe+19]...

Classical approach

Vector $\mathfrak{p} \in [0, 1]^m$ of p-values, computed on each taxa independently.

Wilcoxon and Kruskal-Wallis non-parametric tests can be used.

Classical approach

Vector $\mathfrak{p} \in [0,1]^m$ of p-values, computed on each taxa independently.

Wilcoxon and Kruskal-Wallis non-parametric tests can be used.

A taxon i is detected differentially abundant if $\mathfrak{p}_i < \alpha$.

Each taxon fall into one class of the confusion matrix:

		True condition	
		Positive	Negative
Detection	Positive	TP	FP
	Negative	FN	TN

Classical approach

Vector $\mathfrak{p} \in [0,1]^m$ of p-values, computed on each taxa independently.

Wilcoxon and Kruskal-Wallis non-parametric tests can be used.

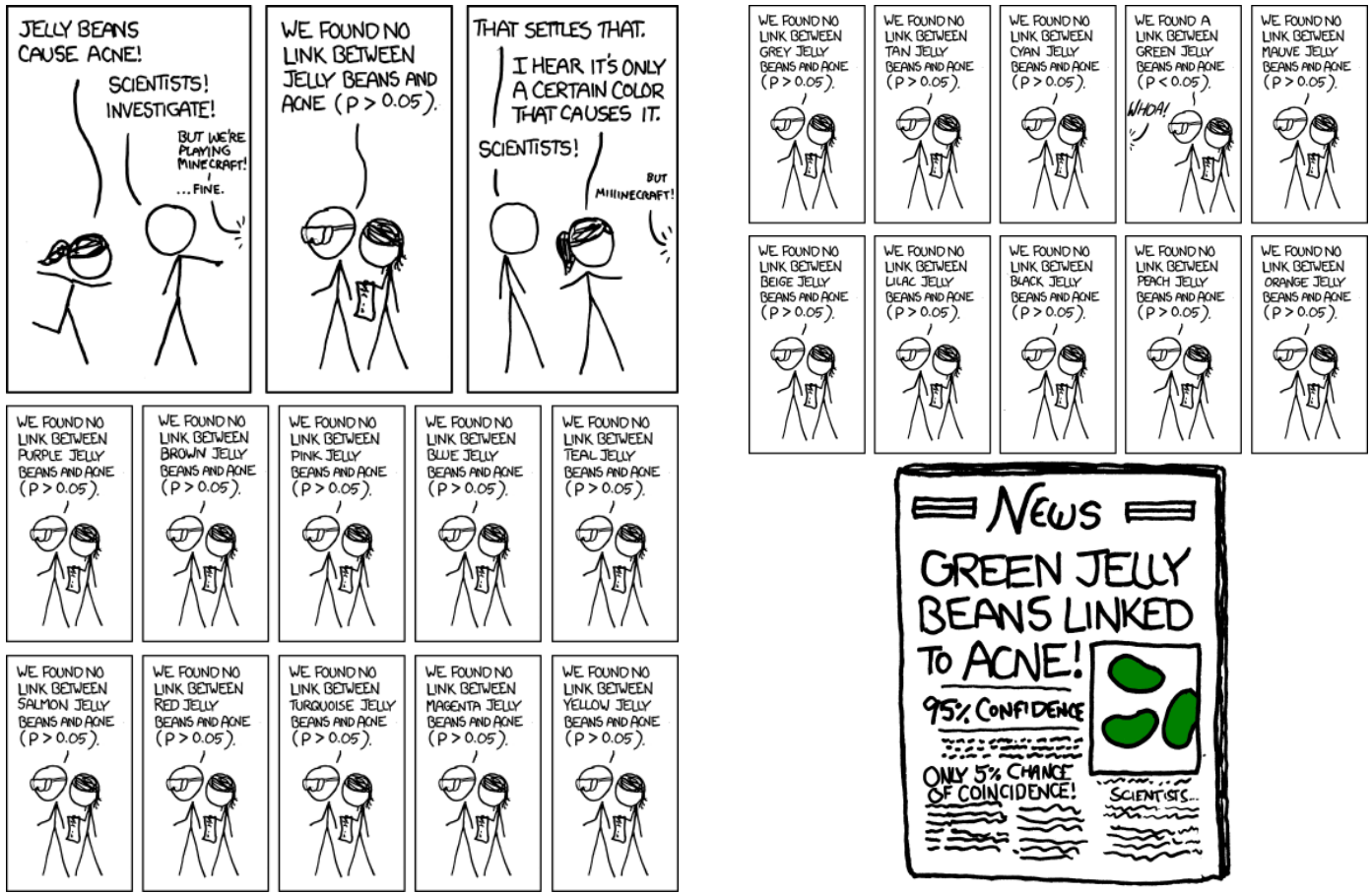
A taxon i is detected differentially abundant if $\mathfrak{p}_i < \alpha$.

Each taxon fall into one class of the confusion matrix:

		True condition	
		Positive	Negative
Detection	Positive	TP	FP
	Negative	FN	TN

Under $\mathbb{H}_0$, $\mathfrak{p}_i \sim \mathcal{U}([0,1])$.

Multiple testing problem



XKCD

TPR and FDR

		True condition: Positive	True condition: Negative
Detection	Positive	TP	FP
	Negative	FN	TN

True Positive Rate:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

False Discovery Rate:

$$\mathrm{FDP} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}, \mathrm{FDR} = \mathbb{E}[\mathrm{FDP}].$$

Multiple testing problem correction

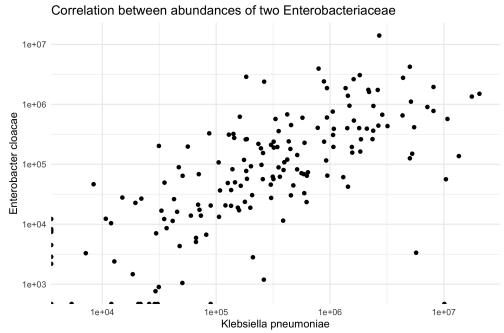
Correction with Benjamini-Hochberg procedure to respect an *a priori* FDR : q^{bh}.

Benjamini and Hochberg [BH95]

Multiple testing problem correction

Correction with Benjamini-Hochberg procedure to respect an *a priori* FDR : q^{bh}.

But it assumes independence between taxa and it is not respected.

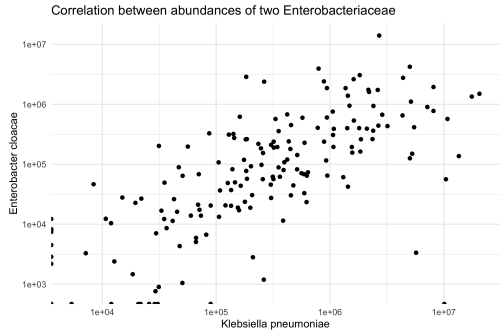


Benjamini and Hochberg [BH95]

Multiple testing problem correction

Correction with Benjamini-Hochberg procedure to respect an *a priori* FDR : q^{bh}.

But it assumes independence between taxa and it is not respected.



One can use Benjamini-Yekutieli correction which does not make any assumption about dependence between taxa but

- it's too conservative,
- we want to correct explicitly for correlation between taxa.

Benjamini and Hochberg [BH95]; Benjamini and Yekutieli [BY01]

Incorporation of hierarchical information

Goal

- Correct explicitly for correlation between taxa.

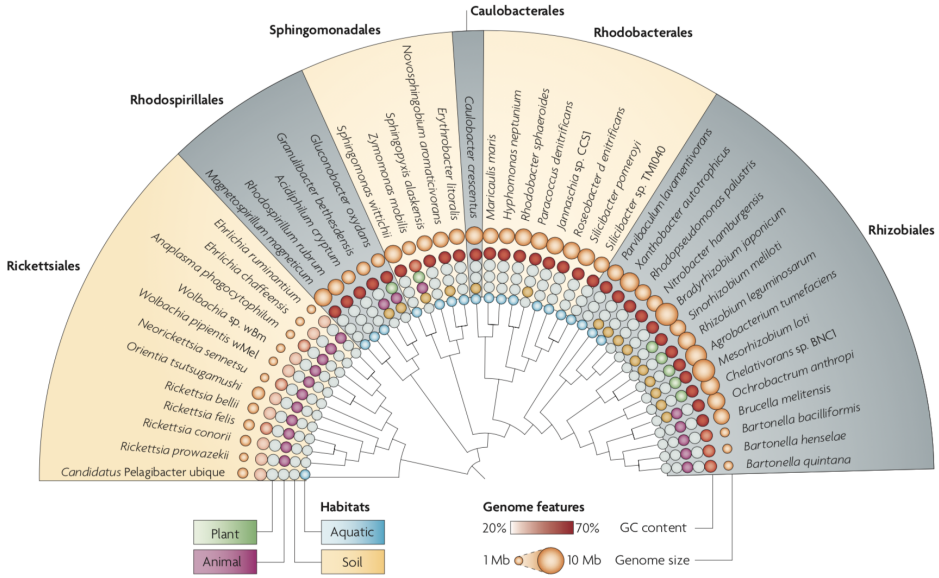
Goal

- Correct explicitly for correlation between taxa.
- Increase power.

Goal

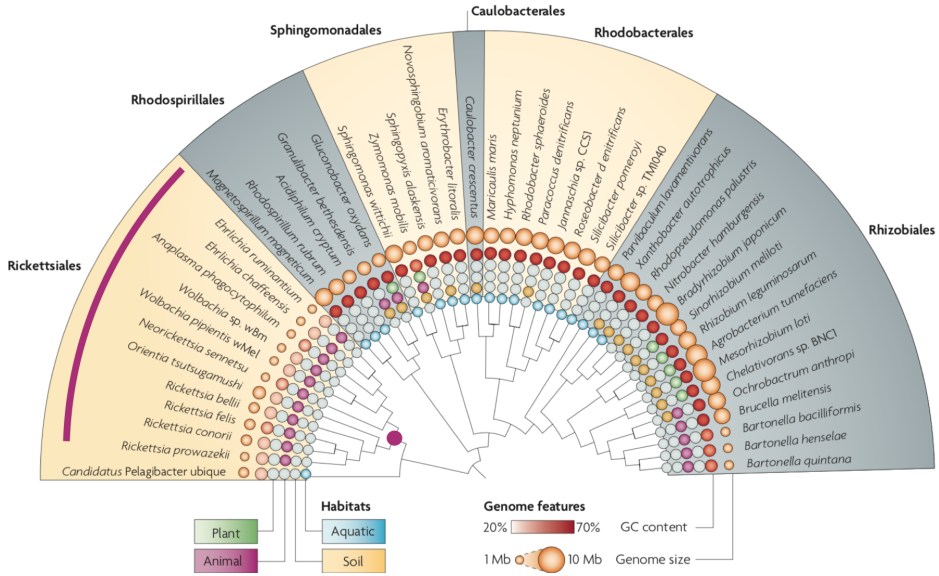
- Correct explicitly for correlation between taxa.
- Increase power.
- Keep FDR under a desired level.

Rationale



Philippot Andersson, et al. [Phi+10]

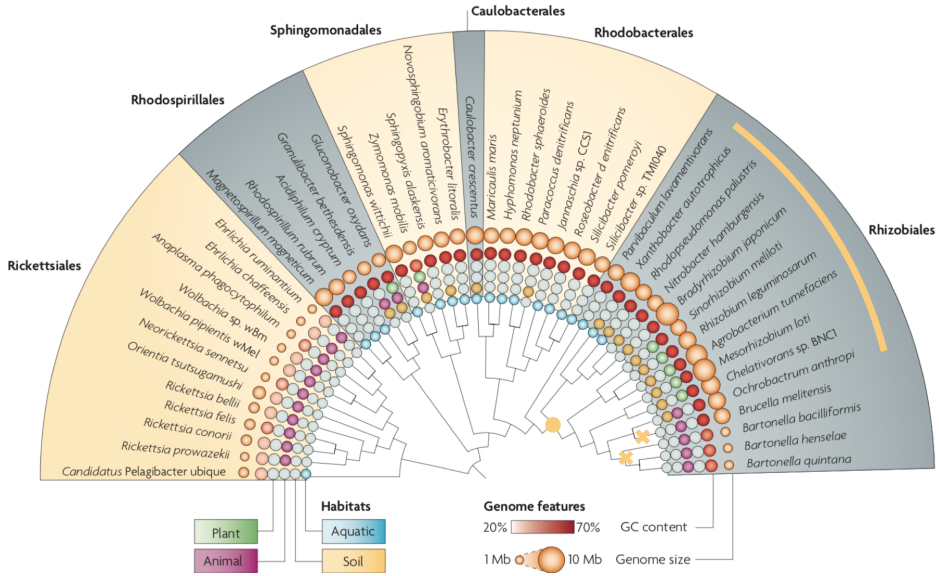
Rationale



These species are associated with animals.

Philippot Andersson, et al. [Phi+10]

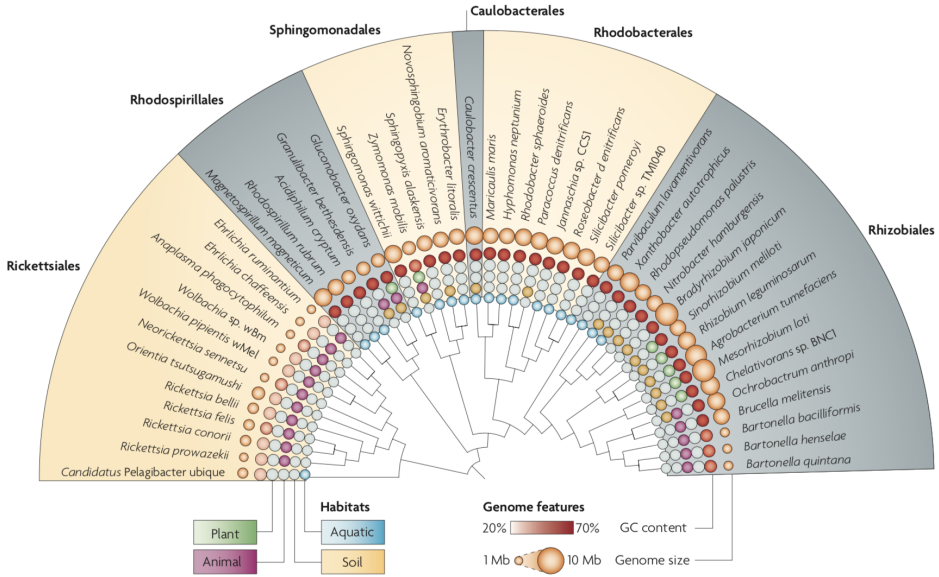
Rationale



These species are associated with soil.

Philippot Andersson, et al. [Phi+10]

Rationale



Already used in

- Hierarchical FDR [Yek08; SH14],
- **TreeFDR [XCC17]**.

Philippot Andersson, et al. [Phi+10]

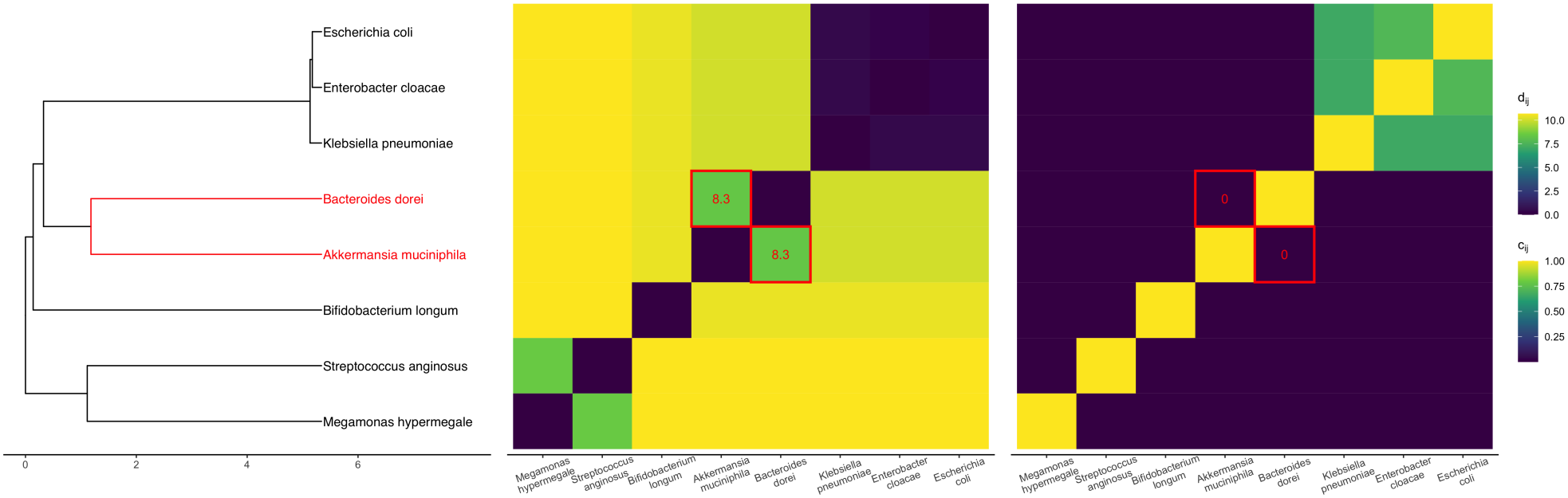
z-scores smoothing with TreeFDR

$\mathfrak{z} = \Phi^{-1}(\mathfrak{p})$ is the vector of observed z-scores and μ the vector of "true" z-scores.

z-scores smoothing with TreeFDR

$\mathfrak{z} = \Phi^{-1}(\mathfrak{p})$ is the vector of observed z-scores and μ the vector of "true" z-scores.

Assume a hierarchical model: $\mathfrak{z} \mid \mu \sim \mathcal{N}_m(\mu, \sigma^2 \mathbf{I}_m)$ and $\mu \sim \mathcal{N}_m(\gamma \mathbf{1}_m, \tau^2 C_\rho)$ with $C_\rho = \left(\exp\left(-2\rho d_{i,j}\right)\right)_{i,j}$.



z-scores smoothing with TreeFDR

Then $\mathcal{Z} \sim \mathcal{N}_m\left(\gamma \mathbf{1}_m, \tau^2 C_\rho + \sigma^2 \mathbf{I}_m\right)$ by Bayes formula and the maximum a posteriori gives

$$\mu^* = \left(\mathbf{I}_m + k^2 C_\rho^{-1}\right)\left(k^2 C_\rho^{-1} \gamma \mathbf{1}_m + \mathcal{Z}\right),$$

with $k = \frac{\sigma}{\tau}$ and ρ_0 hyperparameters to optimize.

z-scores smoothing with TreeFDR

Then $\mathcal{Z} \sim \mathcal{N}_m\left(\gamma \mathbf{1}_m, \tau^2 C_\rho + \sigma^2 \mathbf{I}_m\right)$ by Bayes formula and the maximum a posteriori gives

$$\mu^* = \left(\mathbf{I}_m + k^2 C_\rho^{-1}\right)\left(k^2 C_\rho^{-1}\gamma \mathbf{1}_m + \mathcal{Z}\right),$$

with $k = \frac{\sigma}{\tau}$ and ρ_0 hyperparameters to optimize.

At the end, a multiple testing correction by resampling is done on smoothed values.

Which tree?

- Phylogeny? Taxonomy?

😀 Proxy for correlations at high-level niches.

☹ Not so much for low-level niches?

☹ Not available every time.

Which tree?

- Phylogeny? Taxonomy?

😀 Proxy for correlations at high-level niches.

☹ Not so much for low-level niches?

☹ Not available every time.

- Correlation tree?

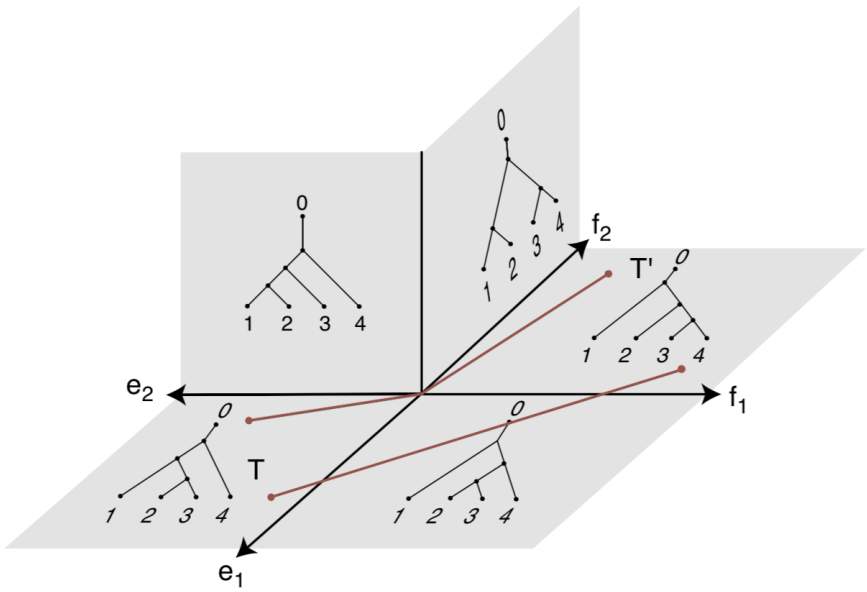
😀 Actual correlation between taxa.

😀 Always available.

☹ Risk of overfitting.

Bichat Plassais, et al. [Bic+20]

Billera-Holmes-Vogtmann distance



- Each tree is mapped into a space composed by merged orthants.
- An orthant corresponds to a topology.
- The BHV distance is the length of the unique shortest path between the trees on treespace.
- It can be computed with a $O\left(m^4\right)$ algorithm.

Billera Holmes, et al. [BHV01]

Quantifying distance between trees

- Trees of primary interest:
  - correlation tree on original data,
  - phylogeny.

Quantifying distance between trees

- Trees of primary interest:
  - correlation tree on original data,
  - phylogeny.
- What is the confidence region for the correlation tree?
  - correlation trees on boostrapped data (resampling on samples).

Quantifying distance between trees

- Trees of primary interest:
  - correlation tree on original data,
  - phylogeny.
- What is the confidence region for the correlation tree?
  - correlation trees on boostrapped data (resampling on samples).
- Are trees significantly closer than two random trees?
  - trees created by random shuffling of correlation tree tip labels,
  - trees created by random shuffling of phylogeny tip labels.

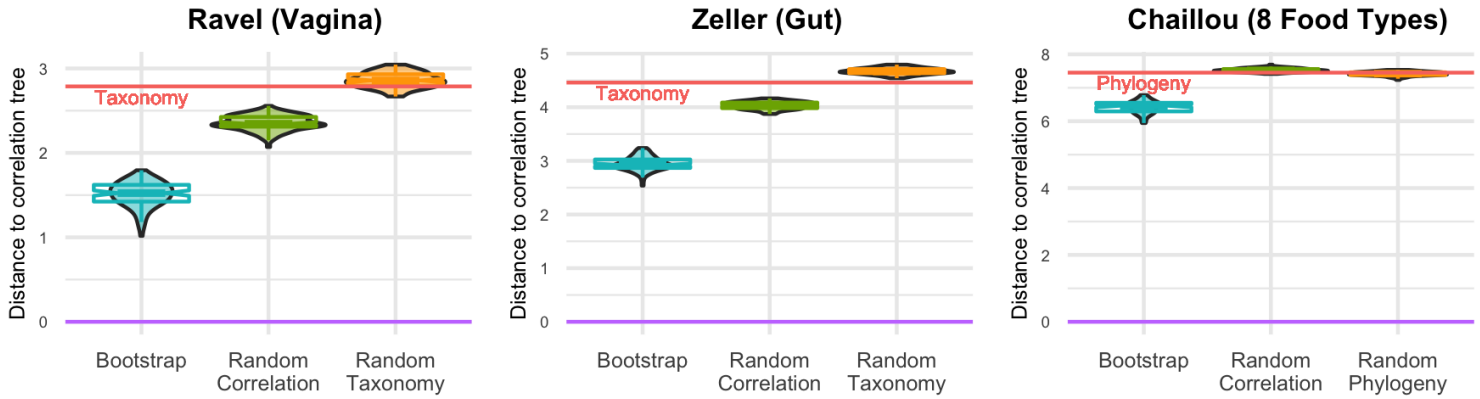
Quantifying distance between trees

- Trees of primary interest:
  - correlation tree on original data,
  - phylogeny.
- What is the confidence region for the correlation tree?
  - correlation trees on boostrapped data (resampling on samples).
- Are trees significantly closer than two random trees?
  - trees created by random shuffling of correlation tree tip labels,
  - trees created by random shuffling of phylogeny tip labels.

We compute all pairwise distances between these trees.

Comparisons between trees

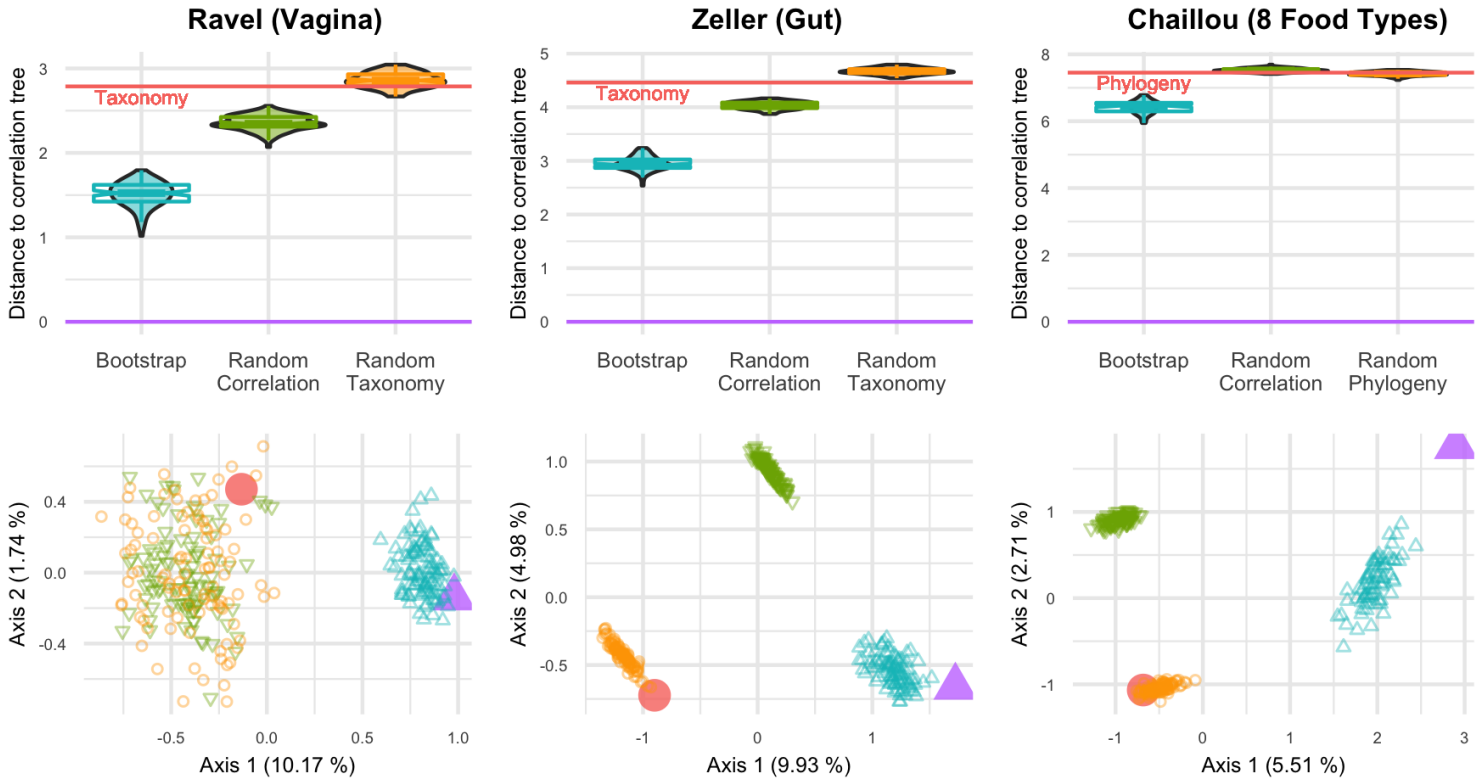
Neither phylogeny nor taxonomy is in the confident region of the correlation tree.



[Bic+20; Rav+11; Zel+14; Cha+15]

Comparisons between trees

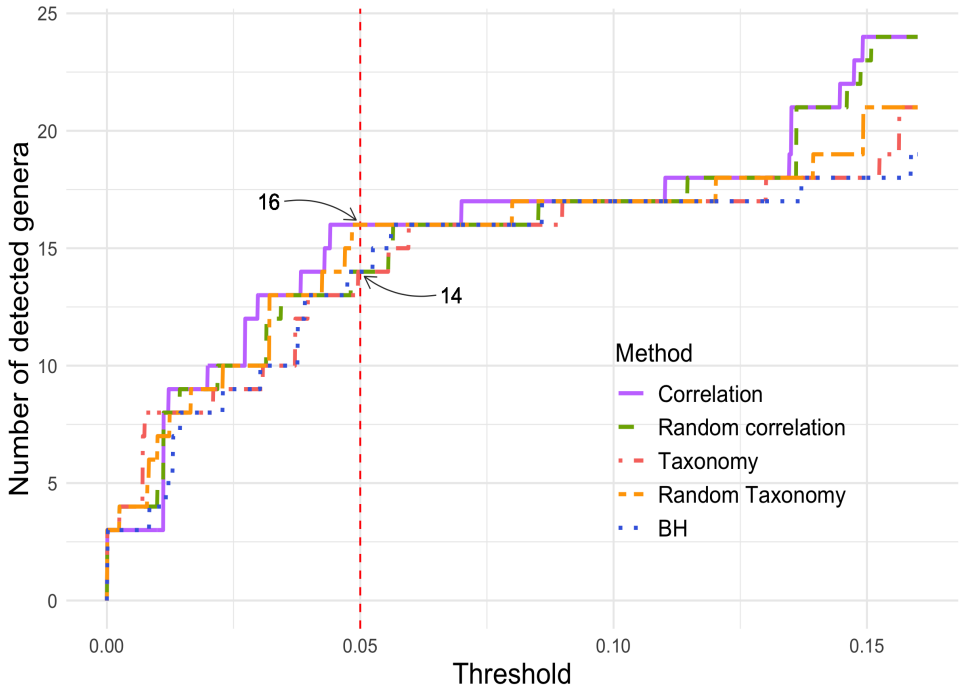
Neither phylogeny nor taxonomy is in the confident region of the correlation tree.



[Bic+20; Rav+11; Zel+14; Cha+15]

Impact of the tree with Zeller dataset

All hierarchies give highly similar results.



- 119 genera
- 199 patients
  - 66 healthy
  - 42 adenoma
  - 91 colorectal cancer

Bichat Plassais, et al. [Bic+20]; Zeller Tap, et al. [Zel+14]

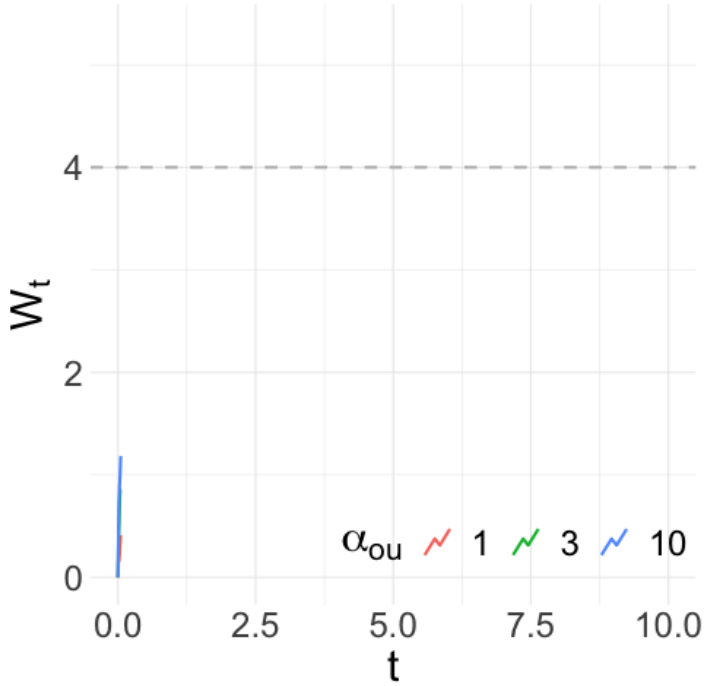
zazou

Z-scores AZ Ornstein-Uhlenbeck

Ornstein-Uhlenbeck process

An Ornstein-Uhlenbeck (OU) process with an optimal value of β_{ou} and a strength of selection $\alpha_{\text{ou}} > 0$ is a Gaussian process that satisfies the SDE

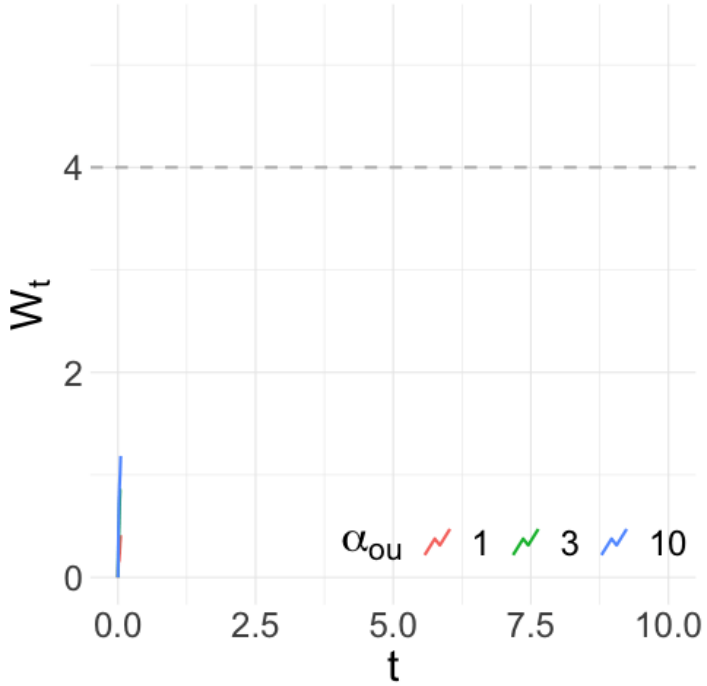
$$\mathrm{d}W_t = -\alpha_{\text{ou}}(W_t - \beta_{\text{ou}})\mathrm{d}t + \sigma_{\text{ou}}\mathrm{d}B_t.$$



Ornstein-Uhlenbeck process

An Ornstein-Uhlenbeck (OU) process with an optimal value of β_{ou} and a strength of selection $\alpha_{\text{ou}} > 0$ is a Gaussian process that satisfies the SDE

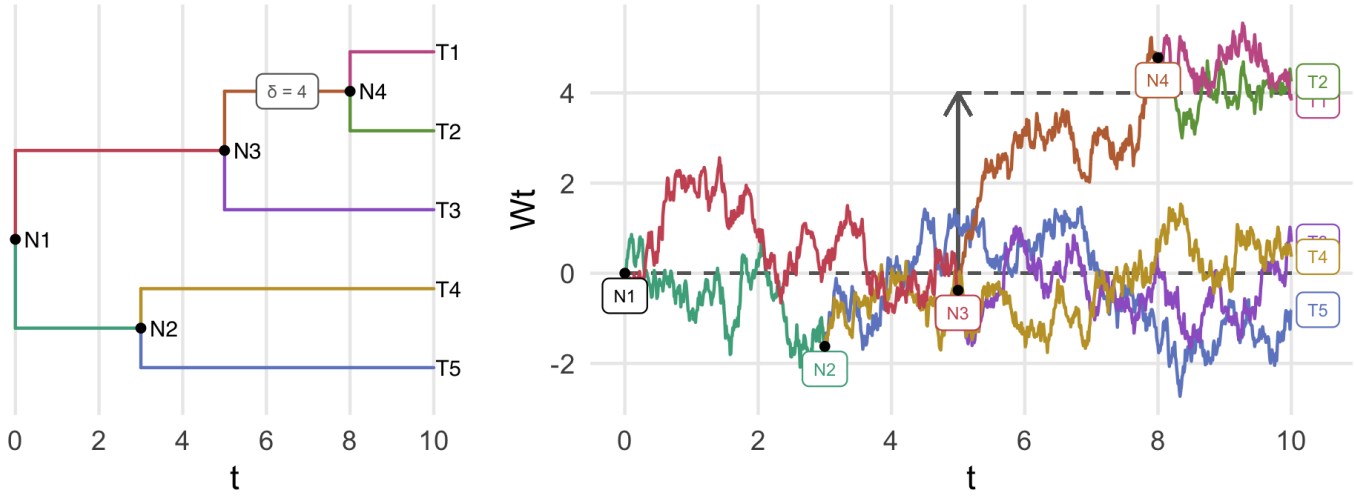
$$\mathrm{d}W_t = -\alpha_{\text{ou}}(W_t - \beta_{\text{ou}})\mathrm{d}t + \sigma_{\text{ou}}\mathrm{d}B_t.$$



W_t is Gaussian with bounded variance and

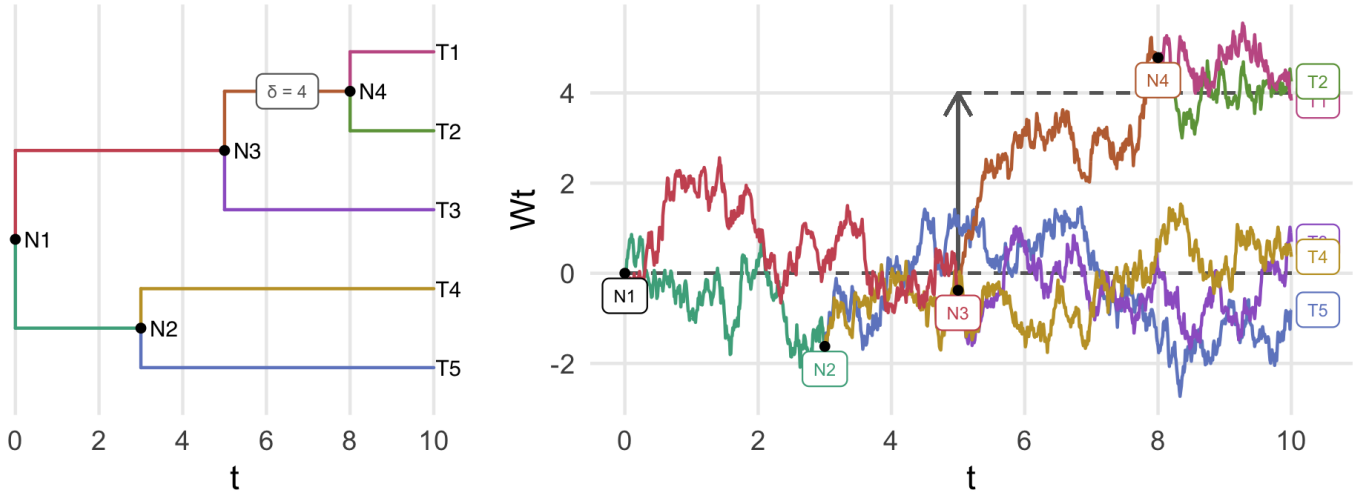
$$W_t \xrightarrow[t\to\infty]{} \mathcal{N}\left(\beta_{\text{ou}}, \frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}}\right).$$

OU process on a tree with shifts δ



Bastide Mariadassou, et al. [BMR17]

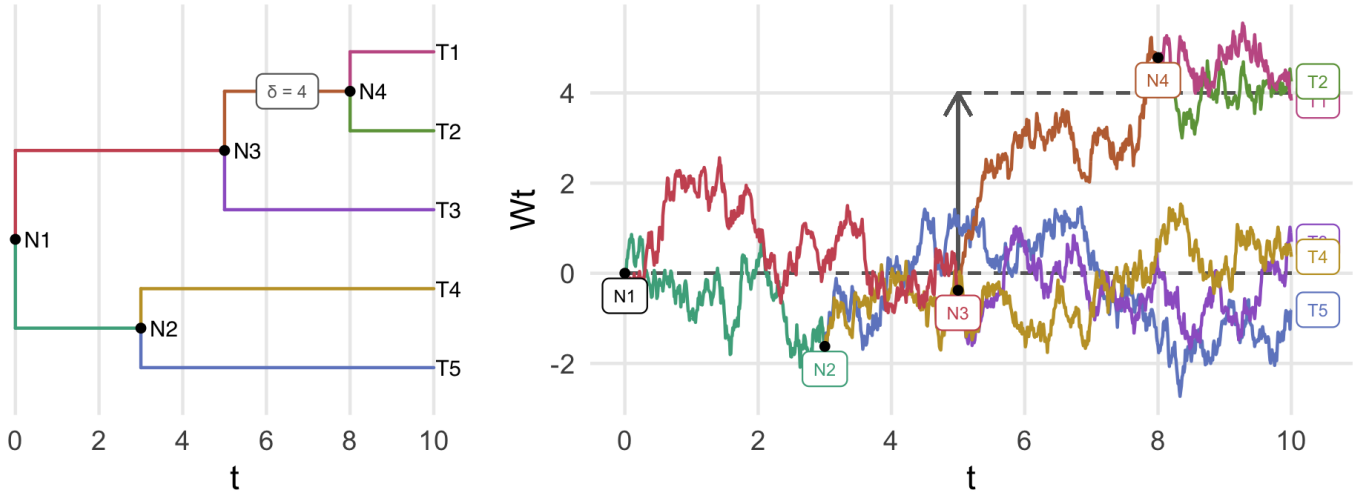
OU process on a tree with shifts δ



- On a branch, the process behaves like on $\mathbb{R}_+$.

Bastide Mariadassou, et al. [BMR17]

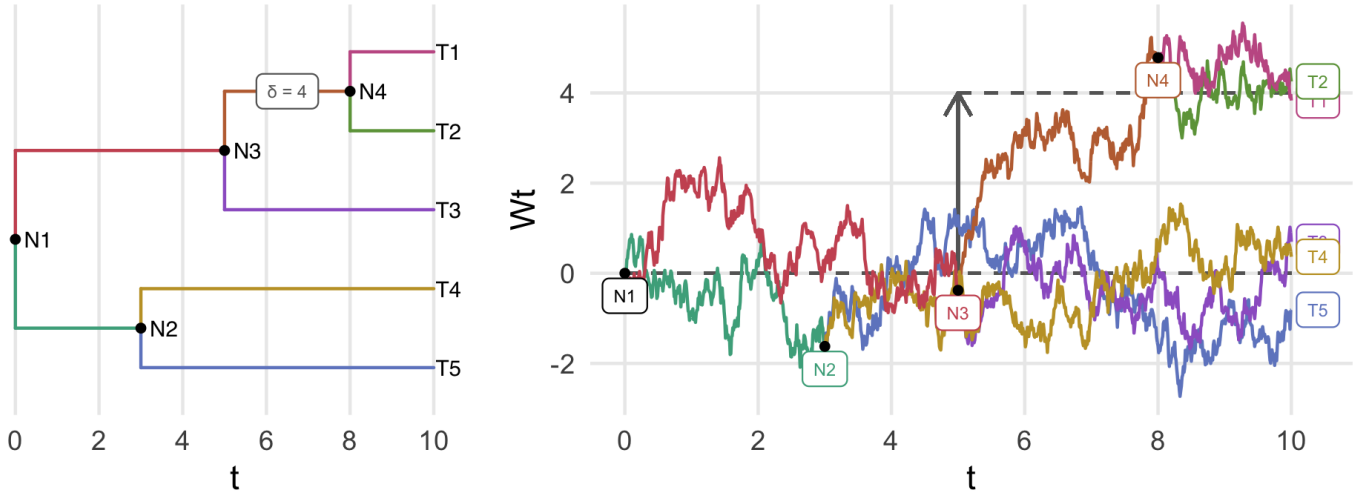
OU process on a tree with shifts δ



- On a branch, the process behaves like on $\mathbb{R}_+$.
- At each node, the process splits into two independent processes with the same initial value.

Bastide Mariadassou, et al. [BMR17]

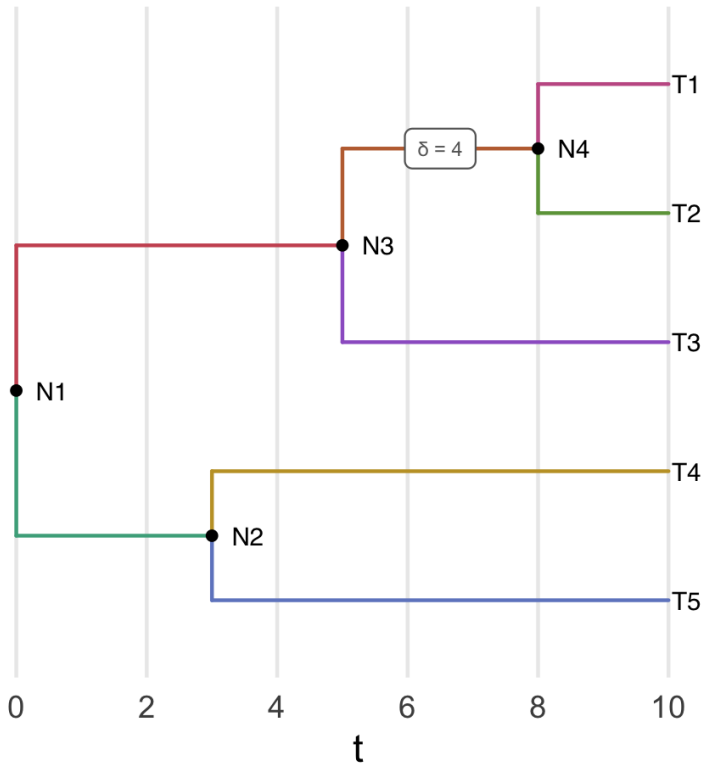
OU process on a tree with shifts δ



- On a branch, the process behaves like on $\mathbb{R}_+$.
- At each node, the process splits into two independent processes with the same initial value.
- The optimal value can shift at a node.

Bastide Mariadassou, et al. [BMR17]

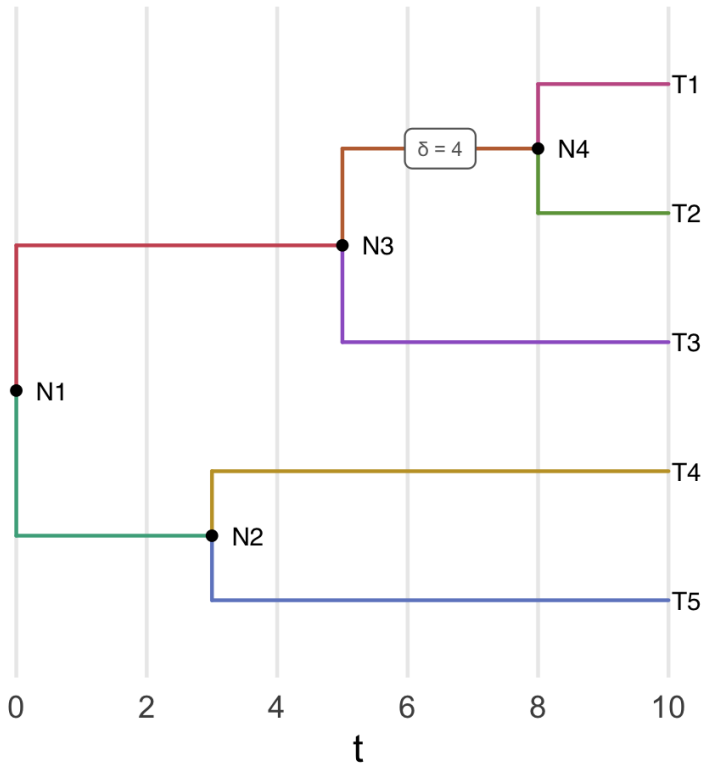
Incidence matrix and vector of shifts



$T = (\mathbb{1}_{\{i \in \mathrm{desc}(j)\}})_{ij} \in \{0,1\}^{m \times n}$ is the incidence matrix of the tree:

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Incidence matrix and vector of shifts



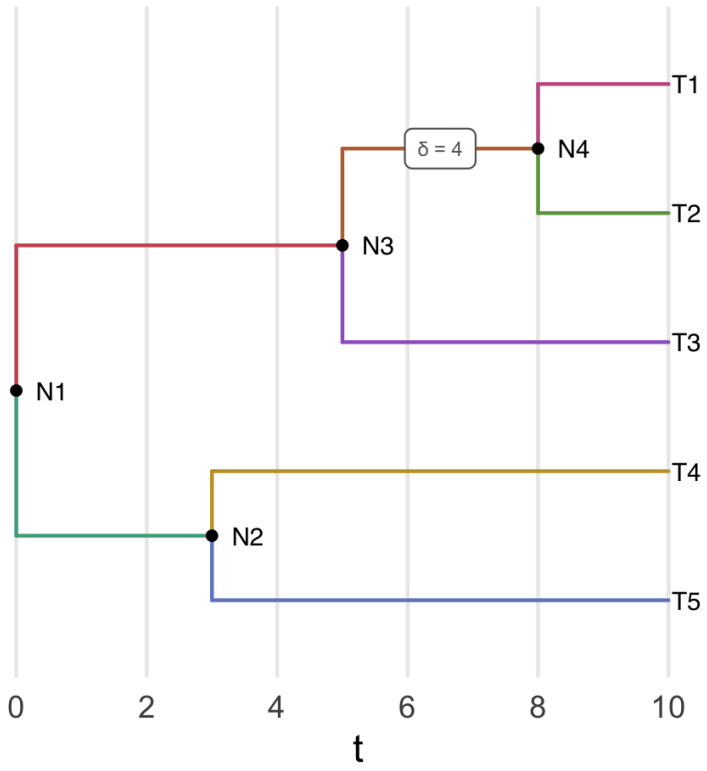
$T = (\mathbb{1}_{\{i \in \mathrm{desc}(j)\}})_{ij} \in \{0,1\}^{m \times n}$ is the incidence matrix of the tree:

$$\begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and δ the vector of shifts:

$$\begin{bmatrix} 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T.$$

Incidence matrix and vector of shifts



$T = (\mathbb{1}_{\{i \in \mathrm{desc}(j)\}})_{ij} \in \{0,1\}^{m \times n}$ is the incidence matrix of the tree:

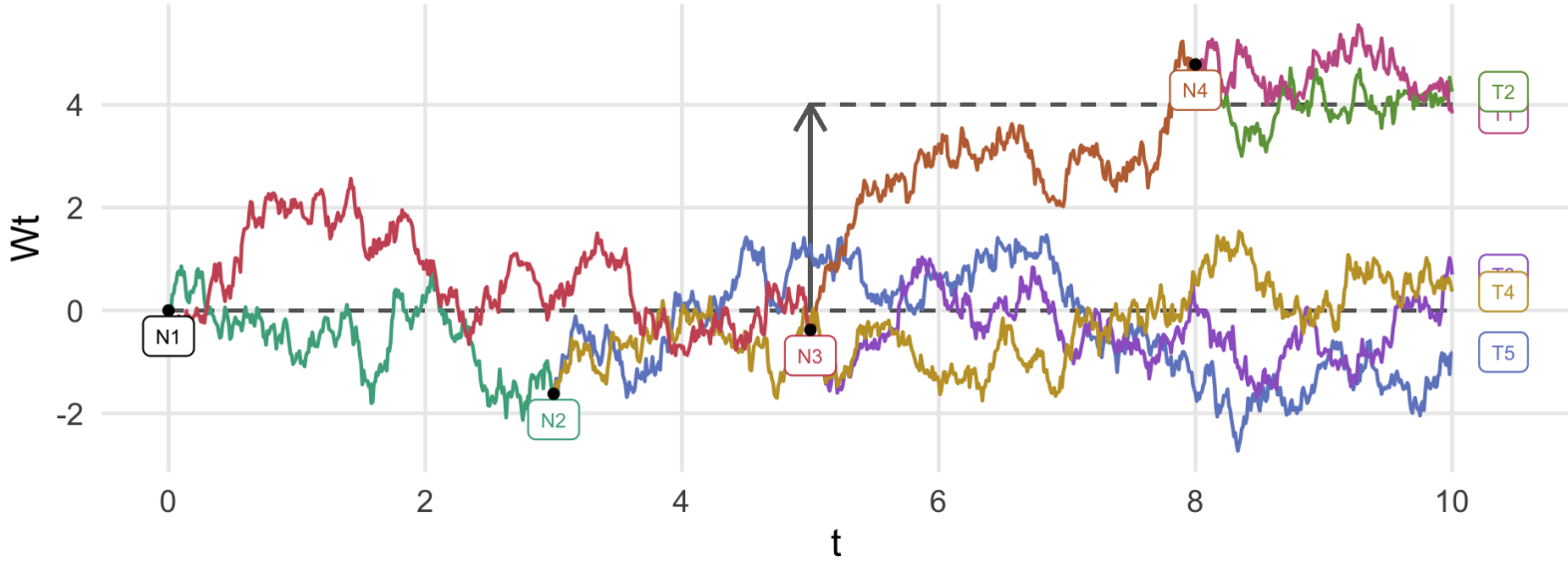
$$\begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and δ the vector of shifts:

$$\begin{bmatrix} 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T.$$

The product $T\delta$ is the vector of optimal values on leaves.

Random variables on leaves



The random variables on leaves are jointly Gaussian $\mathcal{N}_m(T\delta, \Sigma)$ with

$$\Sigma_{i,j} = \frac{\sigma^2_{\text{ou}}}{2\alpha_{\text{ou}}}\left(1 - e^{-2\alpha_{\text{ou}} t_{i,j}}\right) \times e^{-\alpha_{\text{ou}} d_{i,j}}.$$

First assumption

z is the realization of an OU on a tree with shifts δ.

First assumption

$\mathfrak{z}$ is the realization of an OU on a tree with shifts δ.

Then,

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

with $\mu = T\delta$ and Σ depends on α_{ou} and σ_{ou} by

$$\Sigma_{i,j} = \frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}}\left(e^{-\alpha_{\text{ou}} d_{i,j}} - e^{-2\alpha_{\text{ou}} h}\right)$$

for a tree with total height h.

Second assumption

For a taxa i,

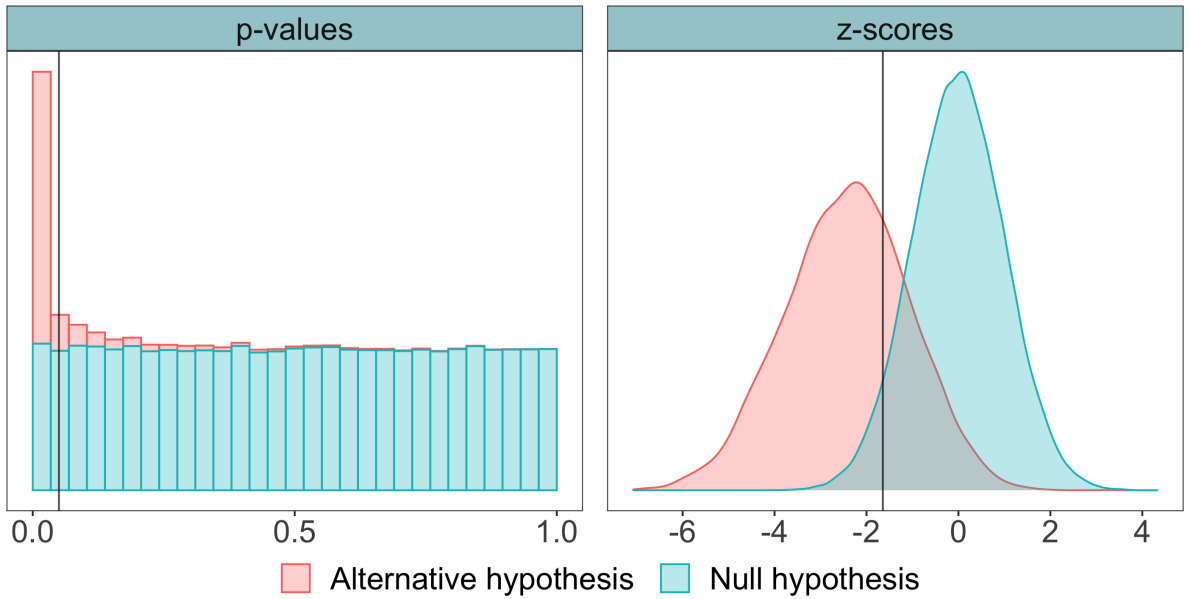
- if $\mathcal{H}_i \in \mathbb{H}_0$, $\mathfrak{p}_i \sim \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(0,1)$,
- if $\mathcal{H}_i \notin \mathbb{H}_0$, $\mathfrak{p}_i \preccurlyeq \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$ with $\mu_i < 0$.

McLachlan and Peel [MP04]

Second assumption

For a taxa i,

- if $\mathcal{H}_i \in \mathbb{H}_0$, $\mathfrak{p}_i \sim \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(0,1)$,
- if $\mathcal{H}_i \notin \mathbb{H}_0$, $\mathfrak{p}_i \preccurlyeq \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$ with $\mu_i < 0$.



McLachlan and Peel [MP04]

Second assumption

For a taxa i,

- if $\mathcal{H}_i \in \mathbb{H}_0$, $\mathfrak{p}_i \sim \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(0,1)$,
- if $\mathcal{H}_i \notin \mathbb{H}_0$, $\mathfrak{p}_i \preccurlyeq \mathcal{U}([0,1])$ so $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$ with $\mu_i < 0$.

Then,

$$\mathfrak{z} \sim \mathcal{N}_m\left(\mu \in \mathbb{R}_-^m, \Sigma\right).$$

One will find **differentially abundant** taxa by finding the **non-zero elements** of μ.

This impose $\Sigma_{i,i} = 1$ so $\sigma_{\text{ou}} = \frac{2\alpha_{\text{ou}}}{1-e^{-2\alpha_{\text{ou}}h}}$.

McLachlan and Peel [MP04]

Estimation of μ

With Σ known, a naive ML estimator gives

$$\hat{\mu} = \underset{\mu \in \mathbb{R}^m_-}{\operatorname{argmin}} \|\mathfrak{z} - \mu\|^2_{\Sigma^{-1},2}.$$

Estimation of μ

With Σ known, a naive ML estimator gives

$$\hat{\mu} = \underset{\mu \in \mathbb{R}_-^m}{\operatorname{argmin}} \|\mathfrak{z} - \mu\|^2_{\Sigma^{-1},2}.$$

To take the tree into account, $\hat{\mu} = T\hat{\delta}$ with

$$\hat{\delta} = \underset{\delta \in \mathbb{R}^n / T\delta \in \mathbb{R}_-^m}{\operatorname{argmin}} \|\mathfrak{z} - T\delta\|^2_{\Sigma^{-1},2}.$$

Estimation of μ

With Σ known, a naive ML estimator gives

$$\hat{\mu} = \underset{\mu \in \mathbb{R}^m_-}{\operatorname{argmin}} \|\mathfrak{z} - \mu\|^2_{\Sigma^{-1},2}.$$

To take the tree into account, $\hat{\mu} = T\hat{\delta}$ with

$$\hat{\delta} = \underset{\delta \in \mathbb{R}^n / T\delta \in \mathbb{R}^m_-}{\operatorname{argmin}} \|\mathfrak{z} - T\delta\|^2_{\Sigma^{-1},2}.$$

To add hierarchically coherent sparsity in our estimate

$$\hat{\delta} = \underset{\delta \in \mathbb{R}^n / T\delta \in \mathbb{R}^m_-}{\operatorname{argmin}} \|\mathfrak{z} - T\delta\|^2_{\Sigma^{-1},2} + \lambda\|\delta\|_1.$$

Estimation of μ (bis)

By Cholesky decomposition, $\Sigma^{-1} = R^T R$

$$\begin{aligned}
\|\mathfrak{z} - T\delta\|^2_{\Sigma^{-1},2} &= (\mathfrak{z} - T\delta)^T \Sigma^{-1} (\mathfrak{z} - T\delta) \\
&= (\mathfrak{z} - T\delta)^T R^T R (\mathfrak{z} - T\delta) \\
&= (R\mathfrak{z} - RT\delta)^T (R\mathfrak{z} - RT\delta) \\
&= (y - X\delta)^T (y - X\delta) = \|y - X\delta\|_2^2
\end{aligned}$$

with $y = R\mathfrak{z}$ and $X = RT$.

Estimation of μ (bis)

By Cholesky decomposition, $\Sigma^{-1} = R^T R$

$$\begin{aligned}
\|\mathfrak{z} - T\delta\|^2_{\Sigma^{-1},2} &= (\mathfrak{z} - T\delta)^T \Sigma^{-1} (\mathfrak{z} - T\delta) \\
&= (\mathfrak{z} - T\delta)^T R^T R (\mathfrak{z} - T\delta) \\
&= (R\mathfrak{z} - RT\delta)^T (R\mathfrak{z} - RT\delta) \\
&= (y - X\delta)^T (y - X\delta) = \|y - X\delta\|_2^2
\end{aligned}$$

with $y = R\mathfrak{z}$ and $X = RT$.

Finally, we fall back to a constrained lasso problem:

$$\hat{\delta} = \underset{\delta \in \mathbb{R}^n / T\delta \in \mathbb{R}^m_-}{\operatorname{argmin}} \|y - X\delta\|_2^2 + \lambda \|\delta\|_1.$$

Numerical resolution

The previous problem could be numerically solved by the shooting algorithm, **iterating unidirectional updates** form the associated problem:

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

Fu [98]

Numerical resolution

The previous problem could be numerically solved by the shooting algorithm, **iterating unidirectional updates** form the associated problem:

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

But Σ and λ are not yet known.

Fu [98]

Estimation of Σ and choice of λ

$\widehat{\Sigma} = \left(\frac{e^{-\hat{\alpha}_{\text{ou}} d_{ij}} - e^{-2\hat{\alpha}_{\text{ou}} h}}{1 - e^{-2\hat{\alpha}_{\text{ou}} h}} \right)_{i,j}$ is determined by $\hat{\alpha}_{\text{ou}}$.

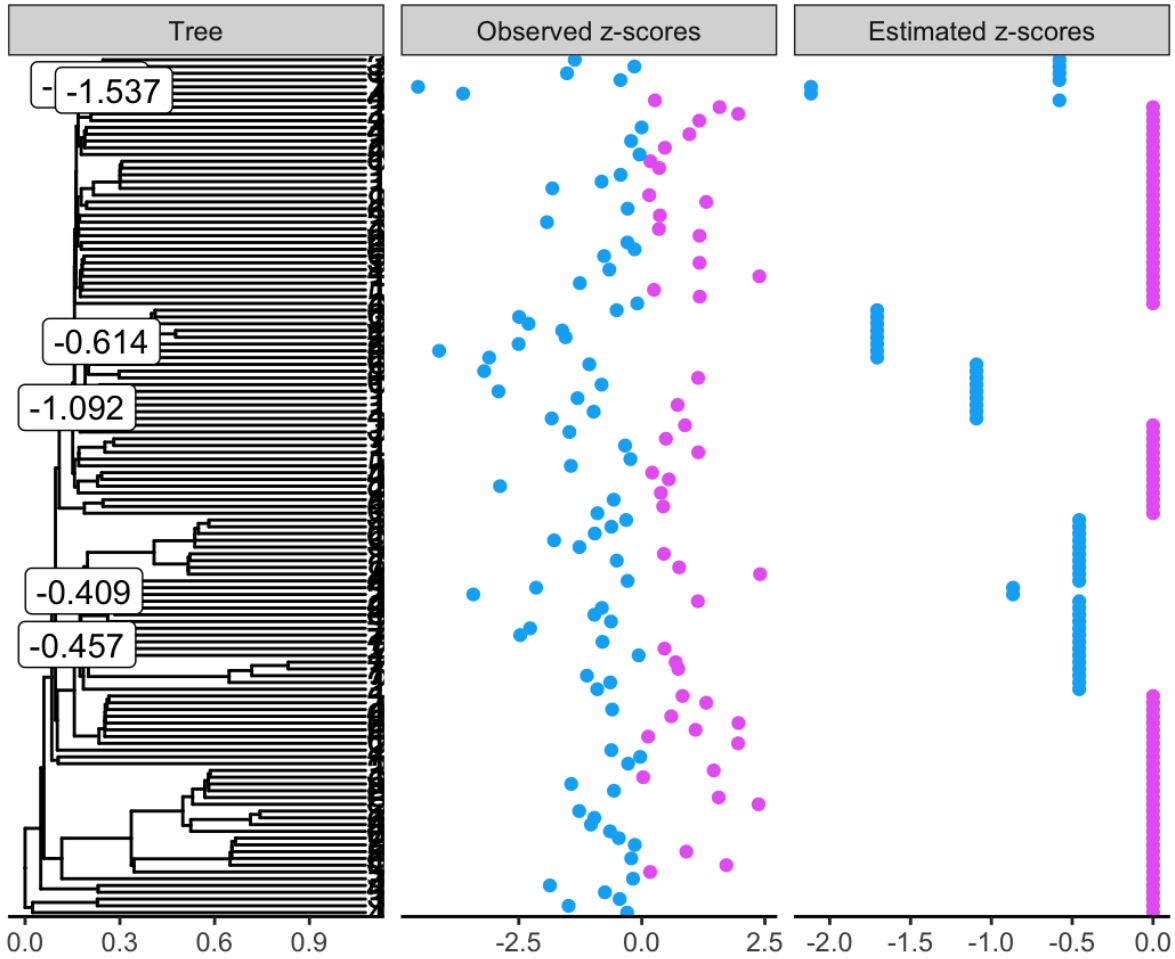
Estimation of Σ and choice of λ

$\widehat{\Sigma} = \left(\frac{e^{-\hat{\alpha}_{\text{ou}} d_{ij}} - e^{-2\hat{\alpha}_{\text{ou}} h}}{1 - e^{-2\hat{\alpha}_{\text{ou}} h}} \right)_{i,j}$ is determined by $\hat{\alpha}_{\text{ou}}$.

The optimal $(\alpha_{\text{ou}}, \lambda)$, is chosen on a bidimensional grid as the argmin of the BIC

$$\|\mathfrak{z} - T\delta_{\alpha_{\text{ou}},\lambda}\|^2_{\Sigma(\alpha_{\text{ou}})^{-1},2} + \log|\Sigma(\alpha_{\text{ou}})| + \|\delta_{\alpha_{\text{ou}},\lambda}\|_0 \log m.$$

Effect of hierarchical smoothing



Find non zero values

Estimation from lasso provides biased estimators without confidence intervals.

We need confidence intervals on $\hat{\delta}$ and $\hat{\mu}$.

Find non zero values

Estimation from lasso provides biased estimators without confidence intervals.

We need confidence intervals on $\hat{\delta}$ and $\hat{\mu}$.

Use of a debiasing procedure

- **score system (ss) [ZZ14],**
- column-wise inverse (ci) [JM13; JM14].

Scaled lasso

The debiasing procedure requires a **initial joint estimator** of $\delta^{(\text{init})}$ and its associated standard error σ.

This can be done with a scaled lasso

$$\left(\hat{\delta}^{(\text{init})}, \hat{\sigma}\right) = \underset{\delta,\sigma}{\operatorname{argmin}} \frac{\|y - X\delta\|_2^2}{2\sigma n} + \frac{\sigma}{2} + \lambda\|\delta\|_1.$$

Sun and Zhang [SZ12]

Score system

The **score system** $S \in \mathbb{R}^{n \times p}$ associated with X and y where s_j is the residual of the (classical) lasso regression of y against X_{-j}:

$$s_j = y - \delta_{\text{lasso}}^{-j} X_{-j}.$$

S is as **weak orthogonalisation** of X.

Debiasing procedure

From the initial estimator $\hat{\delta}_j^{(\text{init})}$ of the scaled lasso, one can do a **one-step correction**

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

Debiasing procedure

From the initial estimator $\hat{\delta}_j^{(\text{init})}$ of the scaled lasso, one can do a **one-step correction**

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

Asymptotically, $\hat{\delta} \sim \mathcal{N}_n(\delta, V)$ with

$$v_{i,j} = \hat{\sigma} \frac{\langle s_i, s_j \rangle}{\langle s_i, x_i \rangle \langle s_j, x_j \rangle}.$$

Debiasing procedure

From the initial estimator $\hat{\delta}_j^{(\text{init})}$ of the scaled lasso, one can do a **one-step correction**

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

Asymptotically, $\hat{\delta} \sim \mathcal{N}_n(\delta, V)$ with

$$v_{i,j} = \hat{\sigma} \frac{\langle s_i, s_j \rangle}{\langle s_i, x_i \rangle \langle s_j, x_j \rangle}.$$

Then the **bilateral confidence interval** for a shift $\hat{\delta}_j$ is

$$\left[\hat{\delta}_j \pm \phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{v_{j,j}} \right].$$

Smoothed p-values

To have the **unilateral hierarchically smoothed p-values** $\mathfrak{p}^{\mathrm{h}}$, we need to propagate the shifts with the incidence matrix T

$$\mathfrak{p}_i^{\mathrm{h}} = \Phi\left(\frac{t_{i.}^T \hat{\delta}}{\left(t_{i.}^T V t_{i.}\right)^{1/2}}\right)$$

with $t_{i.}$ the i^{th} row of T.

Multiple testing correction

Let $\mathrm{t}_i = \frac{t_{i.}^T \hat{\delta}}{\left(t_{i.}^T V t_{i.}\right)^{1/2}}$ be the t-scores, $t_{\max} = \sqrt{2\log m - 2\log\log m}$ and

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

with $R(t) = \sum_{i=1}^m \mathbb{1}_{\{\mathrm{t}_i \leq -t\}}$.

Javanmard Javadi, et al. [JJ19]

Multiple testing correction

Let $\mathrm{t}_i = \frac{t_{i.}^T \hat{\delta}}{\left(t_{i.}^T V t_{i.}\right)^{1/2}}$ be the t-scores, $t_{\max} = \sqrt{2\log m - 2\log\log m}$ and

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

with $R(t) = \sum_{i=1}^m \mathbb{1}_{\{\mathrm{t}_i \leq -t\}}$.

One reject if $\mathrm{t}_i \leq -t^\star$ and the associated **hierarchical q-values** are

$$\mathrm{q}_i^{\mathrm{h}} = \frac{\mathrm{p}_i^{\mathrm{h}} \alpha}{\Phi(-t^\star)}.$$

Javanmard Javadi, et al. [JJ19]

zazou overview

zazou is a recipe with four steps:

- a modeling of z-scores by an Ornstein-Uhlenbeck process on a tree with shifts,

zazou overview

zazou is a recipe with four steps:

- a modeling of z-scores by an Ornstein-Uhlenbeck process on a tree with shifts,
- a constrained scaled lasso to estimate a sparse distribution of the shifts,

zazou overview

zazou is a recipe with four steps:

- a modeling of z-scores by an Ornstein-Uhlenbeck process on a tree with shifts,
- a constrained scaled lasso to estimate a sparse distribution of the shifts,
- a debiasing procedure for the scaled lasso,

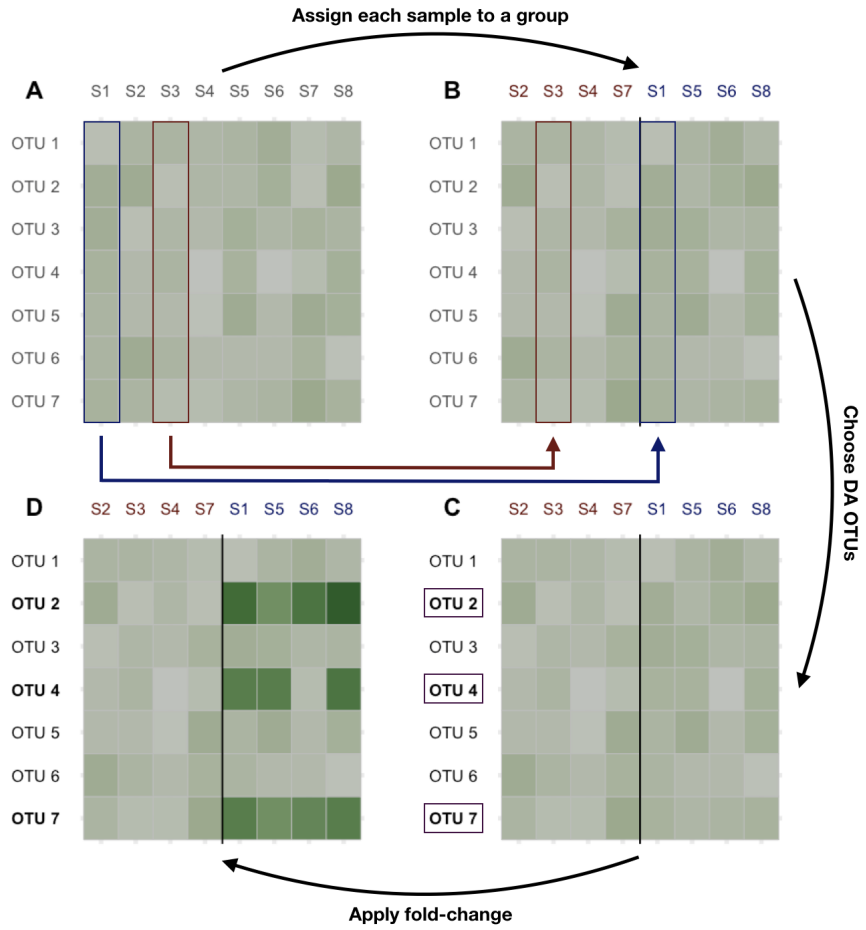
zazou overview

zazou is a recipe with four steps:

- a modeling of z-scores by an Ornstein-Uhlenbeck process on a tree with shifts,
- a constrained scaled lasso to estimate a sparse distribution of the shifts,
- a debiasing procedure for the scaled lasso,
- a debiased lasso designed multiple testing procedure.

Evaluation of zazou

Simulations



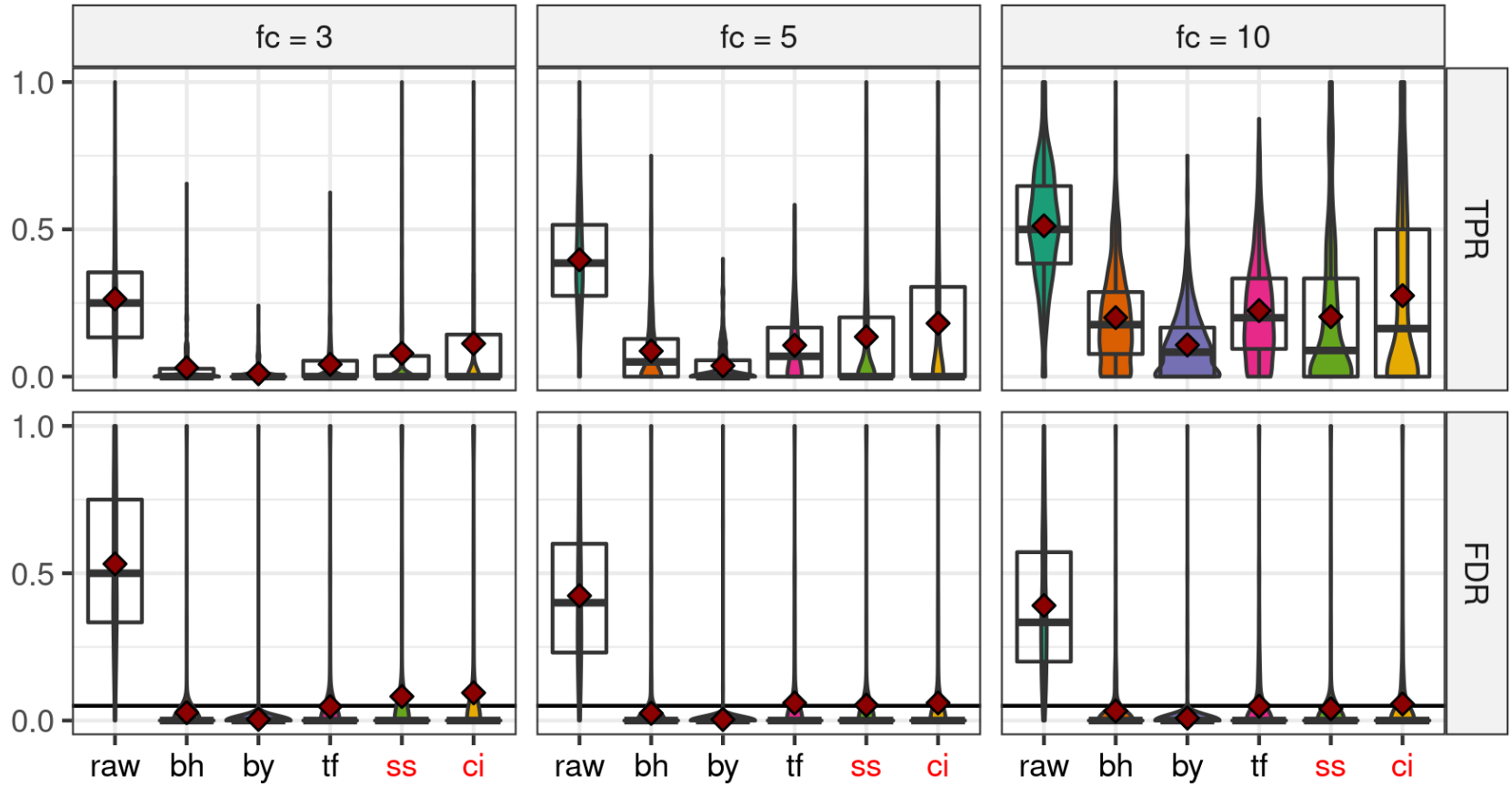
The choice of differentially abundant taxa is done in a hierarchically consistent manner.

- $m = 127, p = 49$ split into two groups.
- Fold-changes of 3, 5 and 10.
- 5 estimated proportions of π_1.
- 100 replicates for each parameters.

Bichat Plassais, et al. [Bic+20]

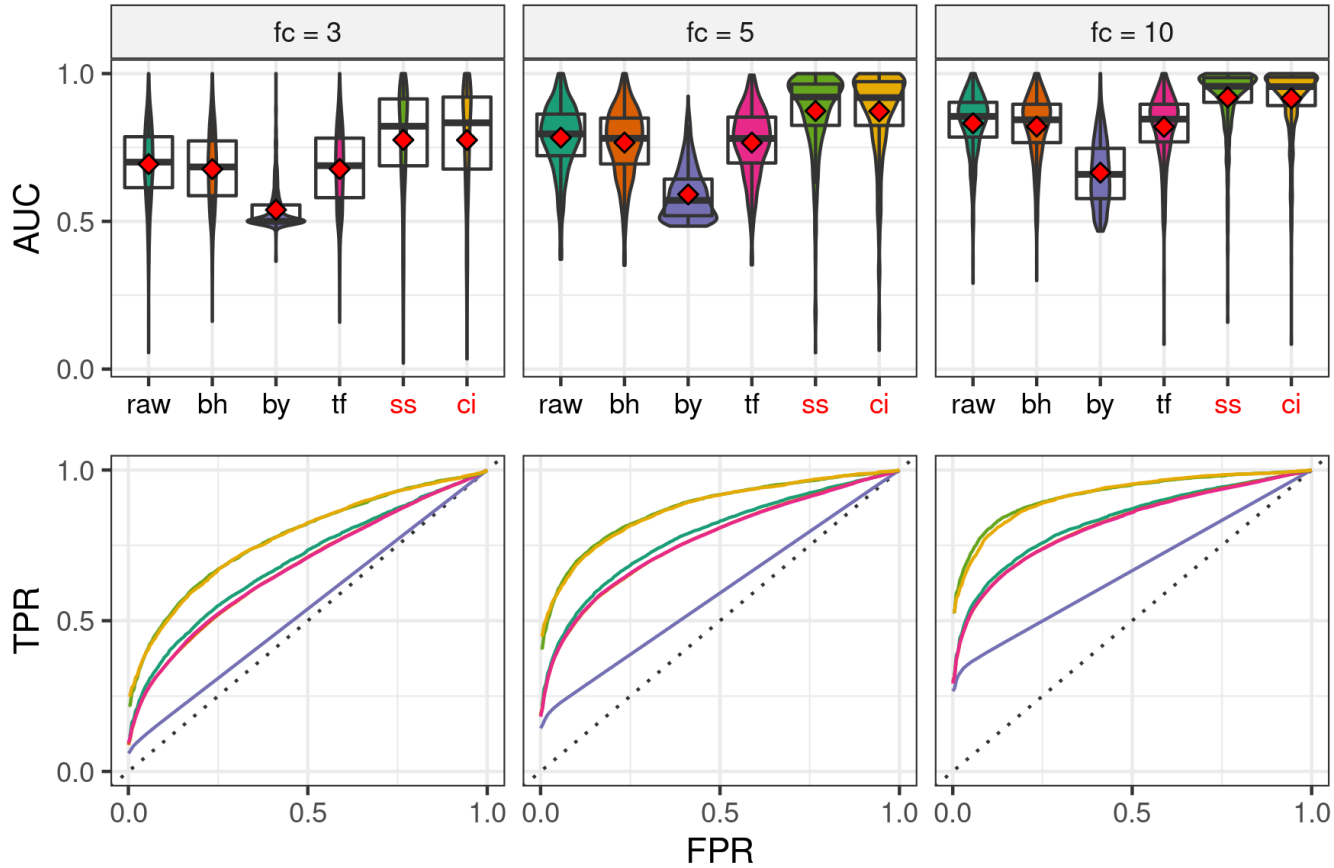
Results

zazou increases the TPR but does not always control the FDR.



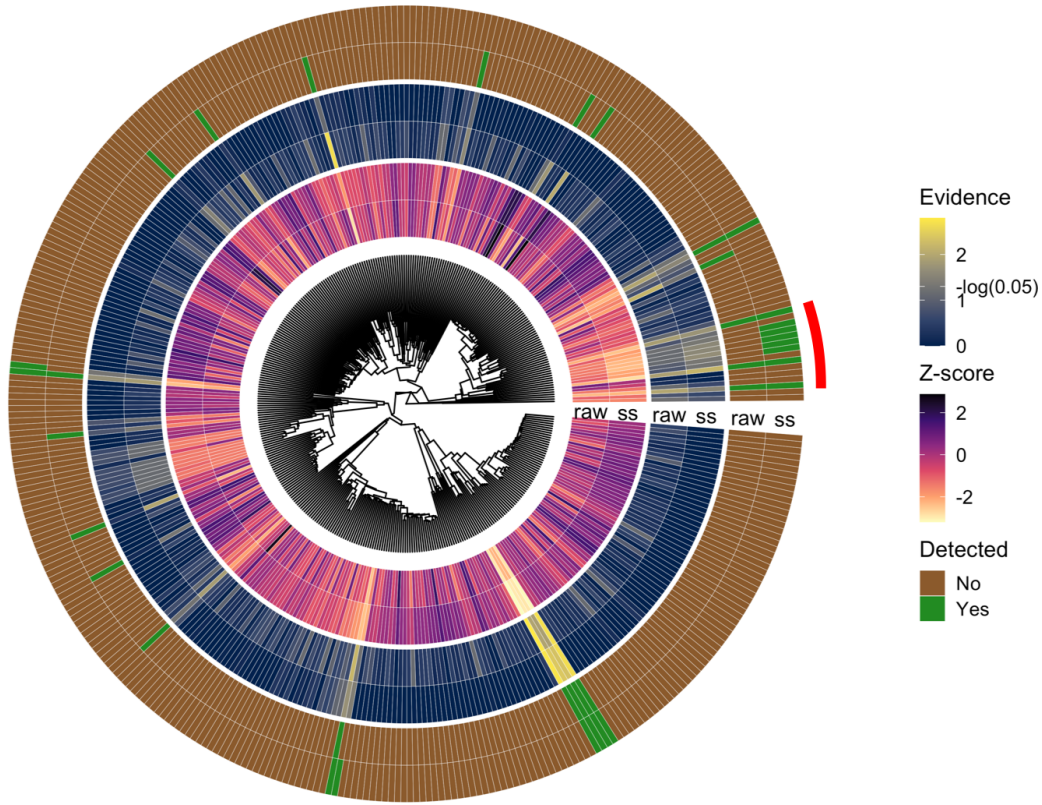
Results

zazou has better AUC and ROC curves.



Association with age

zazou identifies phylogenetically coherent taxa.



Brito Yilmaz, et al. [Bri+16]

Conclusions

Conclusions

- Phylogeny does not capture the true structure of the data.

Conclusions

- Phylogeny does not capture the true structure of the data.
- Previous methods (hFDR, TreeFDR) does not work in practice.

Conclusions

- Phylogeny does not capture the true structure of the data.
- Previous methods (hFDR, TreeFDR) does not work in practice.
- zazou algorithm improves performances a bit ...

Conclusions

- Phylogeny does not capture the true structure of the data.
- Previous methods (hFDR, TreeFDR) does not work in practice.
- zazou algorithm improves performances a bit ...
- ... but the choice of the rejection threshold could be improved.

Conclusions

- Phylogeny does not capture the true structure of the data.
- Previous methods (hFDR, TreeFDR) does not work in practice.
- zazou algorithm improves performances a bit ...
- ... but the choice of the rejection threshold could be improved.
- R packages `correlationtree` and `zazou` are available on GitHub.

Conclusions

- Phylogeny does not capture the true structure of the data.
- Previous methods (hFDR, TreeFDR) does not work in practice.
- zazou algorithm improves performances a bit ...
- ... but the choice of the rejection threshold could be improved.
- R packages `correlationtree` and `zazou` are available on GitHub.
- Articles published in Frontiers in Microbiology and submitted in Statistics and Computing.

Outlooks

- More theoretical work on zazou is required.

Outlooks

- More theoretical work on zazou is required.
- `zazou` could be speed up.

Outlooks

- More theoretical work on zazou is required.
- `zazou` could be speed up.
- Use hierarchical information during testing step, and not only correction step.

Outlooks

- More theoretical work on zazou is required.
- `zazou` could be speed up.
- Use hierarchical information during testing step, and not only correction step.
- Use the framework of prediction instead of association.

Thanks!

Manuscript

abichat.github.io

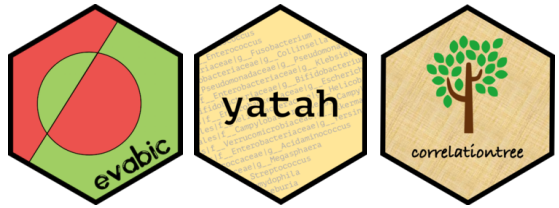
antoinebichat

@_abichat

@abichat

université
PARIS-SACLAY

R packages



evabic evaluates the performance of binary classifiers. It can compute compute 19 different measures with tidy outputs.

yatah provides functions to manage taxonomy when lineages are described like `k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria`.

correlationtree computes correlation trees.

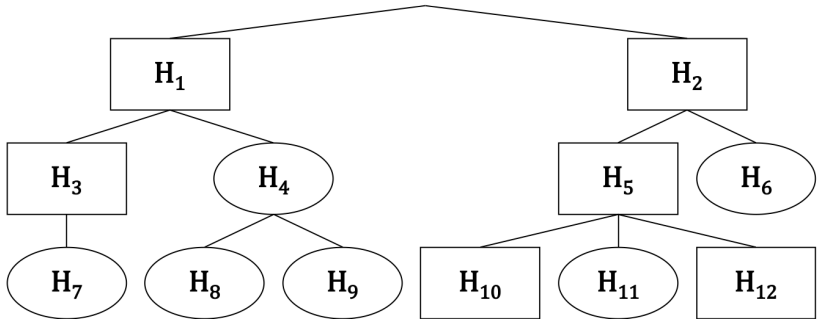
zazou implements the zazou procedure.

hadamardown contains the Université Paris-Saclay PhD manuscript template for bookdown.

Hierarchical FDR

This procedure inscreases statistical power by lessening the number of test to do with a descending method:

- test the family $\mathcal{T}_0$,
- if node t is rejected, test $\mathcal{T}_t = \{H_i \mid \mathrm{Par}(i) = t\}$ with a BH procedure at level α.

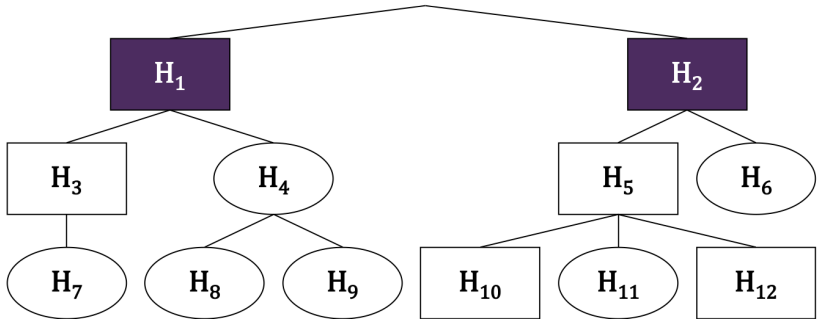


Yekutieli [Yek08]; Sankaran and Holmes [SH14]

Hierarchical FDR

This procedure inscreases statistical power by lessening the number of test to do with a descending method:

- test the family $\mathcal{T}_0$,
- if node t is rejected, test $\mathcal{T}_t = \{H_i \mid \mathrm{Par}(i) = t\}$ with a BH procedure at level α.

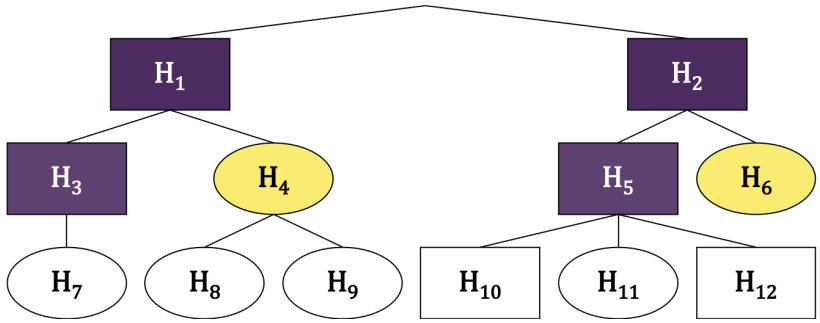


Yekutieli [Yek08]; Sankaran and Holmes [SH14]

Hierarchical FDR

This procedure inscreases statistical power by lessening the number of test to do with a descending method:

- test the family $\mathcal{T}_0$,
- if node t is rejected, test $\mathcal{T}_t = \{H_i \mid \mathrm{Par}(i) = t\}$ with a BH procedure at level α.

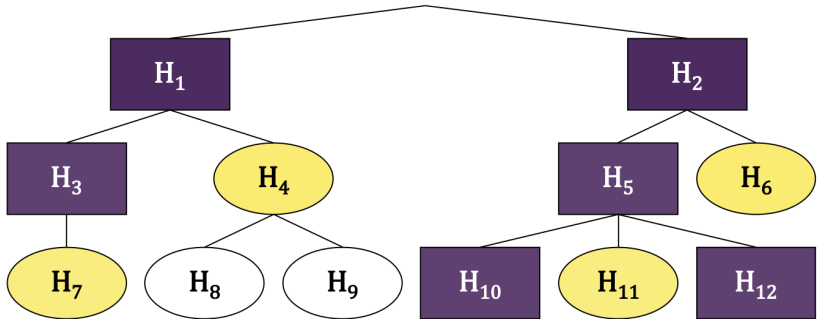


Yekutieli [Yek08]; Sankaran and Holmes [SH14]

Hierarchical FDR

This procedure inscreases statistical power by lessening the number of test to do with a descending method:

- test the family $\mathcal{T}_0$,
- if node t is rejected, test $\mathcal{T}_t = \{H_i \mid \mathrm{Par}(i) = t\}$ with a BH procedure at level α.

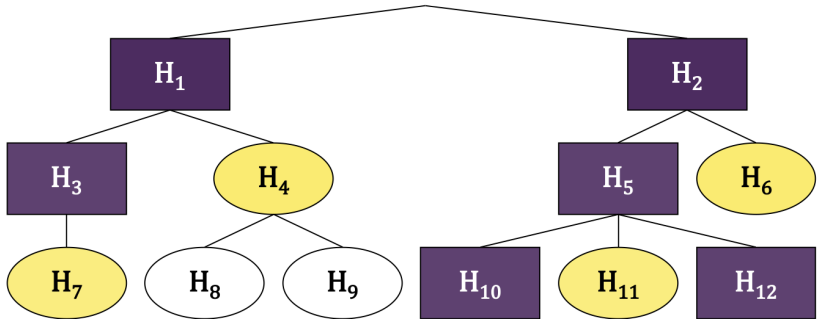


Yekutieli [Yek08]; Sankaran and Holmes [SH14]

Hierarchical FDR

This procedure inscreases statistical power by lessening the number of test to do with a descending method:

- test the family $\mathcal{T}_0$,
- if node t is rejected, test $\mathcal{T}_t = \{H_i \mid \mathrm{Par}(i) = t\}$ with a BH procedure at level α.

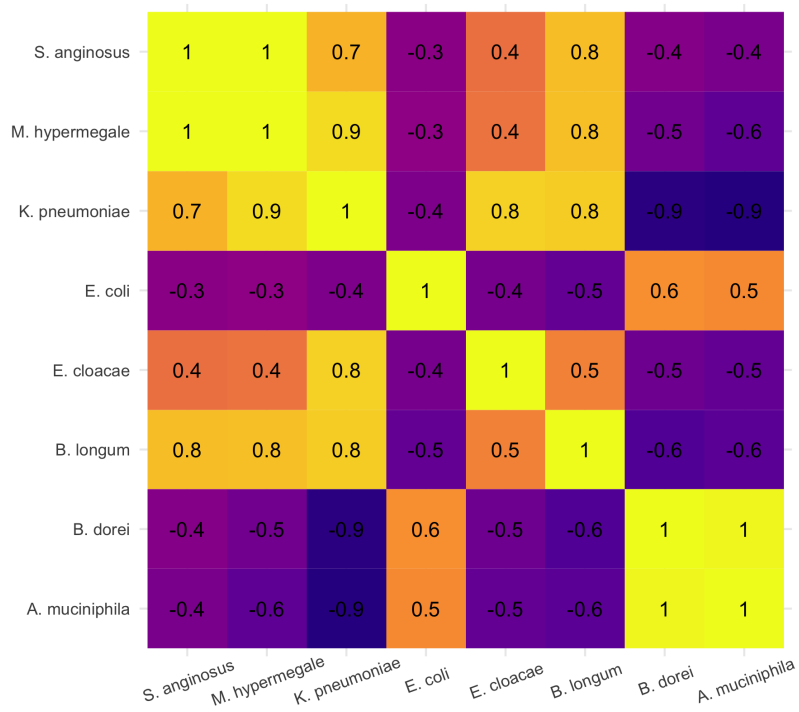


The FDR is controlled at level $\alpha' = 1.44 \times \alpha \times \frac{\#\text{discoveries} + \#\text{families tested}}{\#\text{discoveries} + 1}$.

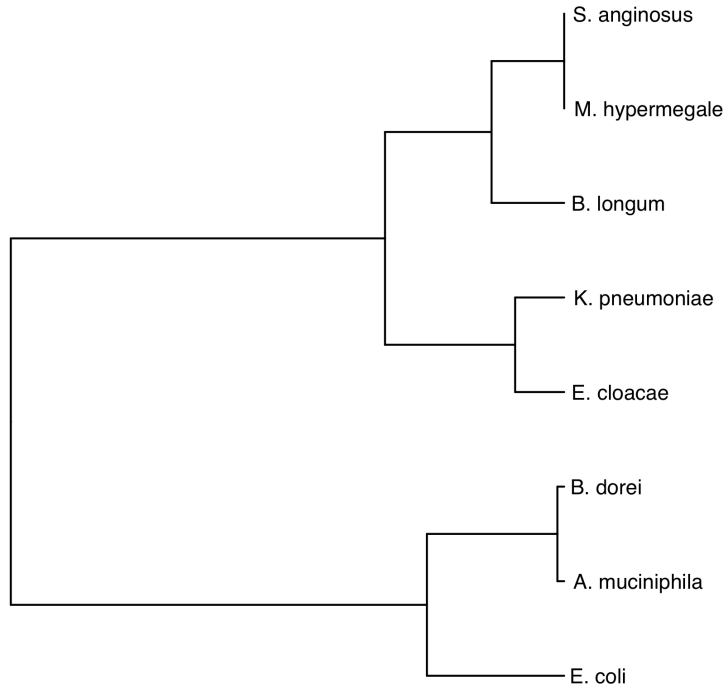
Yekutieli [Yek08]; Sankaran and Holmes [SH14]

Correlation tree

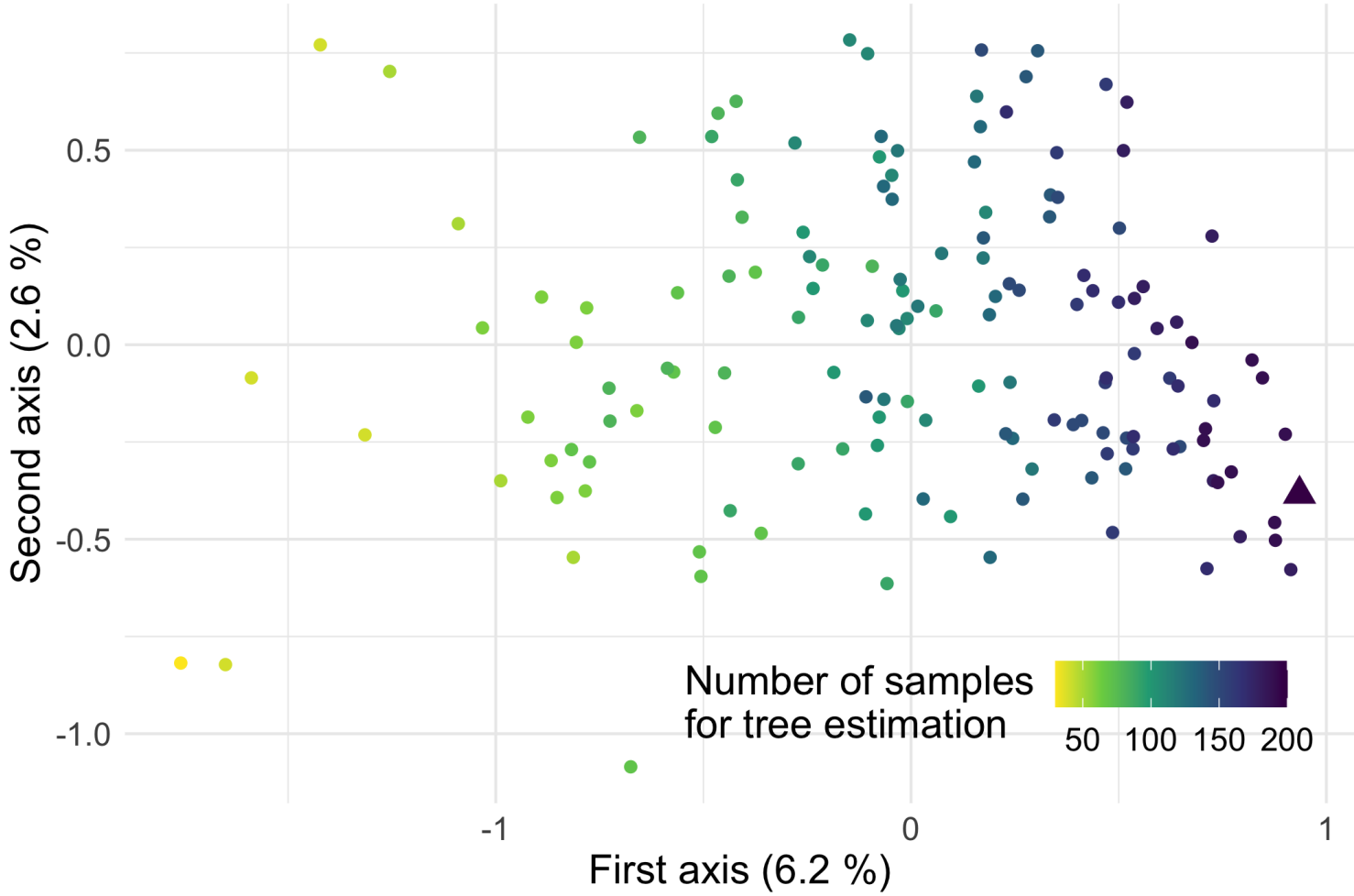
Matrix of pairwise correlation C



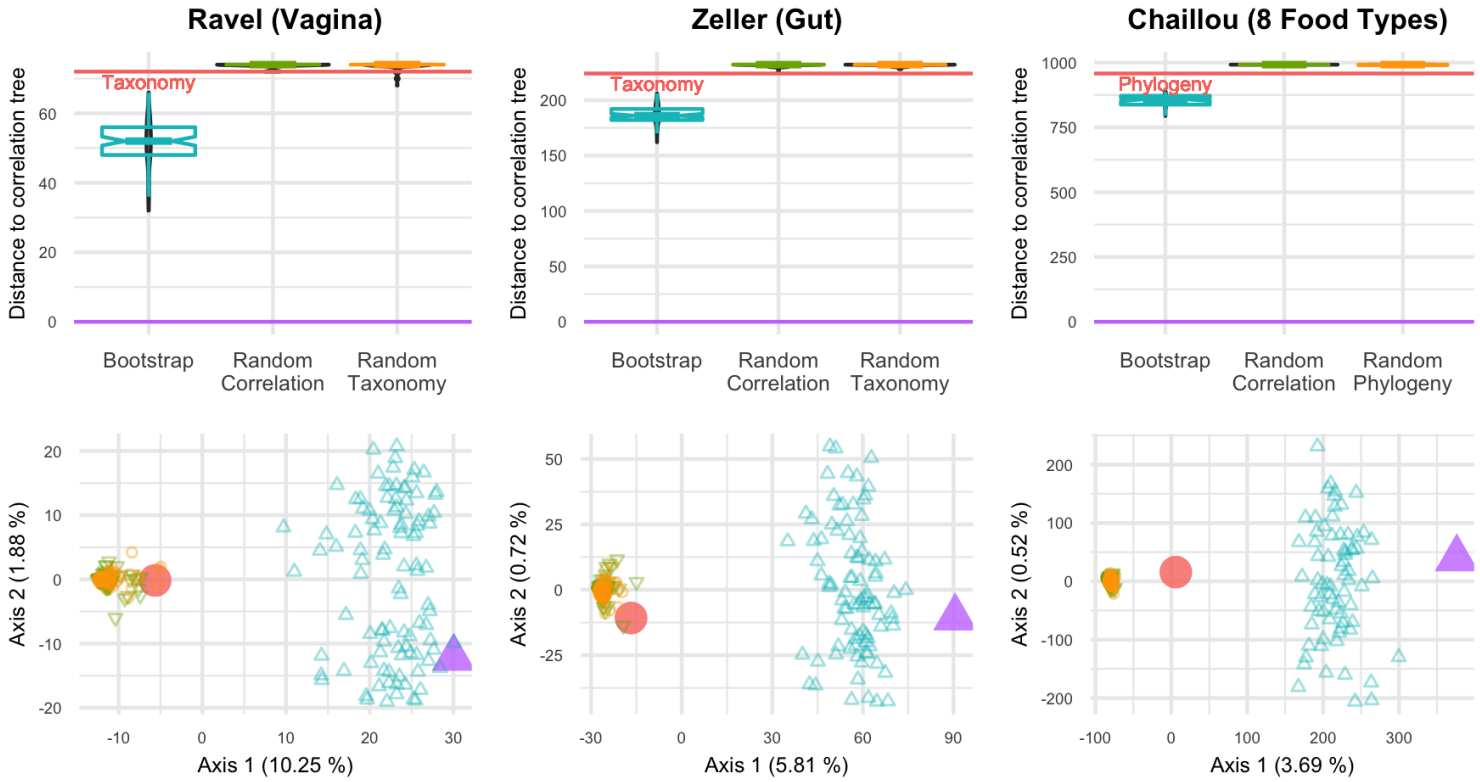
Hierarchical clustering on $1 - C$



Convergence to the correlation tree



With Robinson-Foulds distance



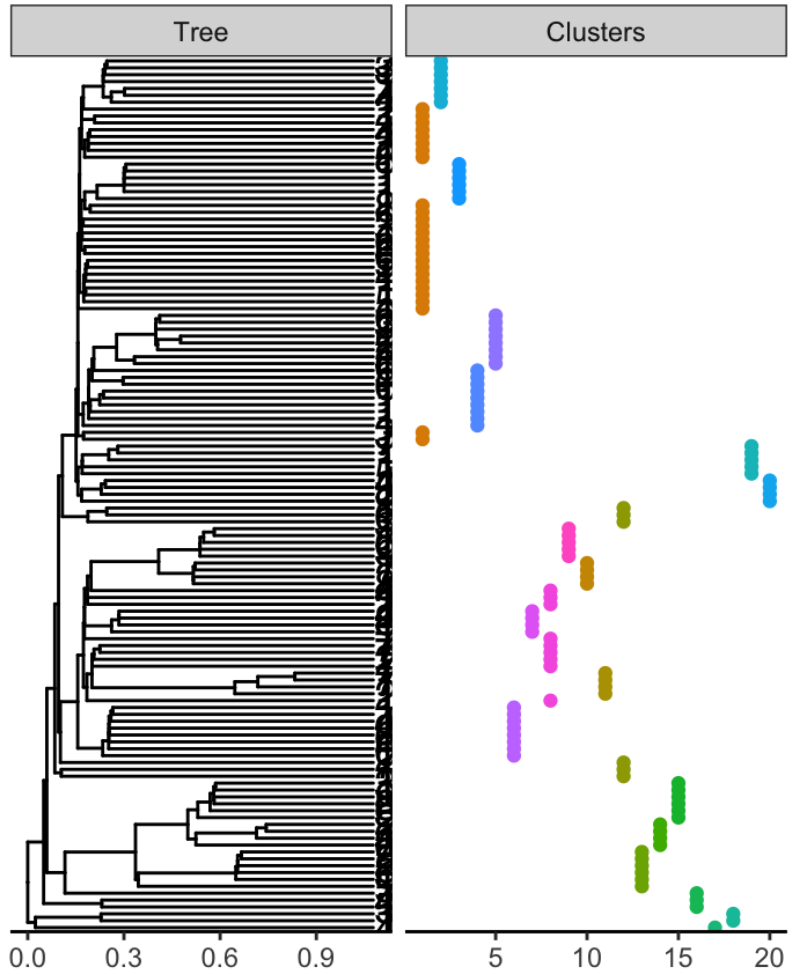
Robinson and Foulds [RF81]

Hierarchically consistent taxa

> The choice of differentially abundant taxa is done in a hierarchically consistent manner.

A partitioning around medoids (PAM) algorithm on the patristic distances matrix $\left(d_{i,j}\right)_{i,j}$ is use for this purpose.

1 to 5 clusters among 20 are selected to be differentially abundant.

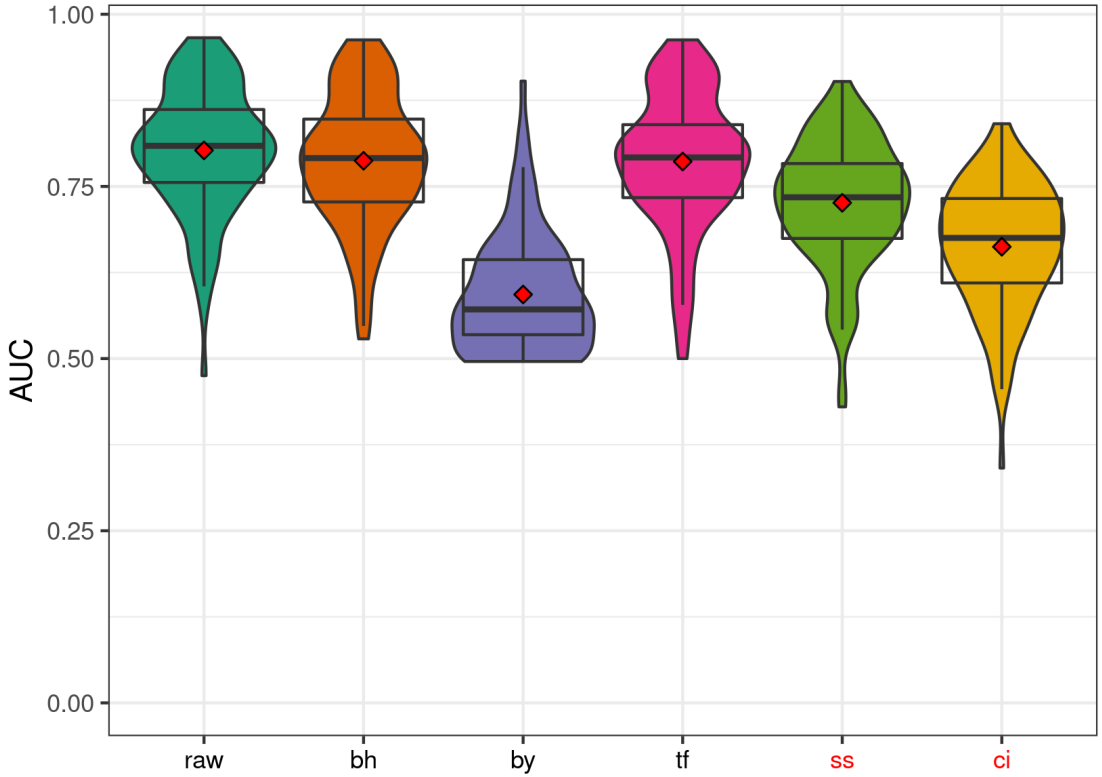


Quantitative results about FDP

(in %)	FDR			P(FDP > 5%)		
method	fc = 3	fc = 5	fc = 10	fc = 3	fc = 5	fc = 10
bh	2.69	2.46	3.23	4.40	7.20	12.20
by	0.40	0.36	0.85	0.40	0.60	2.20
tf	4.75	5.96	4.90	8.72	14.58	18.06
ss	8.17	5.26	3.99	12.40	8.40	8.80
ci	9.41	5.98	5.63	14.80	12.60	18.60

Negative simulations

When differentially abundant taxa are chosen uniformly.



References

Bastide, P., M. Mariadassou, and S. Robin. "Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree Series B Statistical methodology". (2017).

Benjamini, Y. and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289-300.

Benjamini, Y. and D. Yekutieli. "The control of the false discovery rate in multiple testing under dependency". In: *Annals of statistics* (2001), pp. 1165-1188.

Bichat, A., J. Plassais, C. Ambroise, and M. Mariadassou. "Incorporating Phylogenetic Information in Microbiome Differential Abundance Studies Has No Effect on Detection Power and FDR Control". In: *Frontiers in Microbiology* 11 (2020), p. 649. ISSN: 1664-302X.

References

Billera, L. J., S. P. Holmes, and K. Vogtmann. "Geometry of the space of phylogenetic trees". In: *Advances in Applied Mathematics* 27.4 (2001), pp. 733-767.

Bokulich, N. A., J. Chung, T. Battaglia, N. Henderson, M. Jay, H. Li, A. D. Lieber, F. Wu, G. I. Perez-Perez, Y. Chen, and others. "Antibiotics, birth mode, and diet shape microbiome maturation during early life". In: *Science translational medicine* 8.343 (2016), pp. 343ra82-343ra82.

Brito, I. L., S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, M. Tamminen, C. Smillie, J. R. Wortman, and others. "Mobile genes in the human microbiome are structured from global to individual scales". In: *Nature* 535.7612 (2016), pp. 435-439.

Canani, R. B., M. Di Costanzo, L. Leone, M. Pedata, R. Meli, and A. Calignano. "Potential beneficial effects of butyrate in intestinal and extraintestinal diseases". In: *World journal of gastroenterology: WJG* 17.12 (2011), p. 1519.

References

Chaillou, S., A. Chaulot-Talmon, H. Caekebeke, M. Cardinal, S. Christieans, C. Denis, M. H. Desmonts, X. Dousset, C. Feurer, E. Hamon, and others. "Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage". In: *The ISME journal* 9.5 (2015), pp. 1105-1118.

Cryan, J. F., K. J. O'Riordan, C. S. Cowan, K. V. Sandhu, T. F. Bastiaanssen, M. Boehme, M. G. Codagnone, S. Cussotto, C. Fulling, A. V. Golubeva, and others. "The microbiota-gut-brain axis". In: *Physiological reviews* 99.4 (2019), pp. 1877-2013.

David, L. A., C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, and others. "Diet rapidly and reproducibly alters the human gut microbiome". In: *Nature* 505.7484 (2014), pp. 559-563.

Flint, H. J., K. P. Scott, S. H. Duncan, P. Louis, and E. Forano. "Microbial degradation of complex carbohydrates in the gut". In: *Gut microbes* 3.4 (2012), pp. 289-306.

References

Fu, W. J. "Penalized regressions: the bridge versus the lasso". In: *Journal of computational and graphical statistics* 7.3 (1998), pp. 397-416.

Javanmard, A., H. Javadi, and others. "False discovery rate control via debiased lasso". In: *Electronic Journal of Statistics* 13.1 (2019), pp. 1212-1253.

Javanmard, A. and A. Montanari. "Confidence intervals and hypothesis testing for high-dimensional regression". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 2869-2909.

--- "Confidence intervals and hypothesis testing for high-dimensional statistical models". In: *Advances in Neural Information Processing Systems*. 2013, pp. 1187-1195.

References

Kates, A. E., O. Jarrett, J. H. Skarlupka, A. Sethi, M. Duster, L. Watson, G. Suen, K. Poulsen, and N. Safdar. "Household Pet Ownership and the Microbial Diversity of the Human Gut Microbiota". In: *Frontiers in Cellular and Infection Microbiology* 10 (2020), p. 73.

Ley, R. E., D. A. Peterson, and J. I. Gordon. "Ecological and evolutionary forces shaping microbial diversity in the human intestine". In: *Cell* 124.4 (2006), pp. 837-848.

McLachlan, G. J. and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Morgan, X. C., T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, and others. "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment". In: *Genome biology* 13.9 (2012), p. R79.

References

Opstelten, J. L., J. Plassais, S. W. van Mil, E. Achouri, M. Pichaud, P. D. Siersema, B. Oldenburg, and A. C. Cervino. "Gut microbial diversity is reduced in smokers with Crohn's disease". In: *Inflammatory bowel diseases* 22.9 (2016), pp. 2070-2077.

Palleja, A., K. H. Mikkelsen, S. K. Forslund, A. Kashani, K. H. Allin, T. Nielsen, T. H. Hansen, S. Liang, Q. Feng, C. Zhang, and others. "Recovery of gut microbiota of healthy adults following antibiotic exposure". In: *Nature microbiology* 3.11 (2018), pp. 1255-1265.

Philippot, L., S. G. Andersson, T. J. Battin, J. I. Prosser, J. P. Schimel, W. B. Whitman, and S. Hallin. "The ecological coherence of high bacterial taxonomic ranks". In: *Nature Reviews Microbiology* 8.7 (2010), pp. 523-529.

Qin, N., F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. Le Chatelier, J. Yao, L. Wu, and others. "Alterations of the human gut microbiome in liver cirrhosis". In: *Nature* 513.7516 (2014), pp. 59-64.

References

Ravel, J., P. Gajer, Z. Abdo, G. M. Schneider, S. S. Koenig, S. L. McCulle, S. Karlebach, R. Gorle, J. Russell, C. O. Tacket, and others. "Vaginal microbiome of reproductive-age women". In: *Proceedings of the National Academy of Sciences* 108.Supplement 1 (2011), pp. 4680-4687.

Robinson, D. F. and L. R. Foulds. "Comparison of phylogenetic trees". In: *Mathematical biosciences* 53.1-2 (1981), pp. 131-147.

Sankaran, K. and S. Holmes. "structSSI: simultaneous and selective inference for grouped or hierarchically structured data". In: *Journal of statistical software* 59.13 (2014), p. 1.

Sun, T. and C. Zhang. "Scaled sparse linear regression". In: *Biometrika* 99.4 (2012), pp. 879-898.

References

Xiao, J., H. Cao, and J. Chen. "False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing". In: *Bioinformatics* 33.18 (2017), pp. 2873-2881.

Yatsunenko, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, and others. "Human gut microbiome viewed across age and geography". In: *nature* 486.7402 (2012), pp. 222-227.

Yekutieli, D. "Hierarchical false discovery rate-controlling methodology". In: *Journal of the American Statistical Association* 103.481 (2008), pp. 309-316.

Zeller, G., J. Tap, A. Y. Voigt, S. Sunagawa, J. R. Kultima, P. I. Costea, A. Amiot, J. Böhm, F. Brunetti, N. Habermann, and others. "Potential of fecal microbiota for early-stage detection of colorectal cancer". In: *Molecular systems biology* 10.11 (2014), p. 766.

References

Zhang, C. and S. S. Zhang. "Confidence intervals for low dimensional parameters in high dimensional linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014), pp. 217-242.

Zheng, P., B. Zeng, M. Liu, J. Chen, J. Pan, Y. Han, Y. Liu, K. Cheng, C. Zhou, H. Wang, and others. "The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-GABA cycle and schizophrenia-relevant behaviors in mice". In: *Science advances* 5.2 (2019), p. eaau8317.