Impact of tree choice in metagenomics differential abundance studies

Mahendra Mariadassou

November 24, 2021 - CRI

Work and slides done by Antoine Bichat during his Ph.D., co-supervised by C. Ambroise and J. Plassais

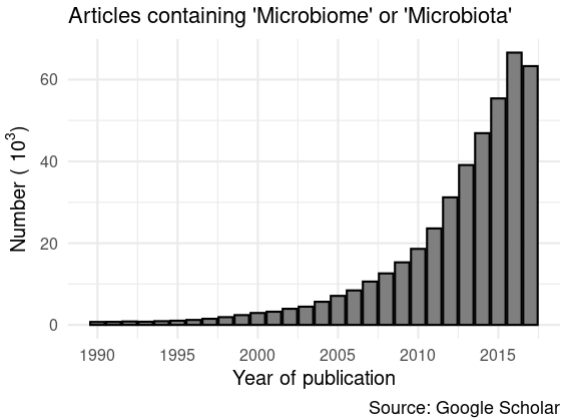
Context

Microbiota

Ecological community of microorganisms that reside in an environmental niche

Some figures for human gut

- 10^{14} bacterial cells in one gut...
- ... weighing 2 kg
- More than 1 500 different species
- More than 10 millions unique genes



Proven associations

- Immune system
- Crohn's disease
- Vaginosis
- Diabete
- Tobacco
- Diet
- Antibiotics
- Birth mode

📘 Gut: The Inside Story of Our Body's Most Underrated Organ (Giulia Enders)

Data - abundances of taxa

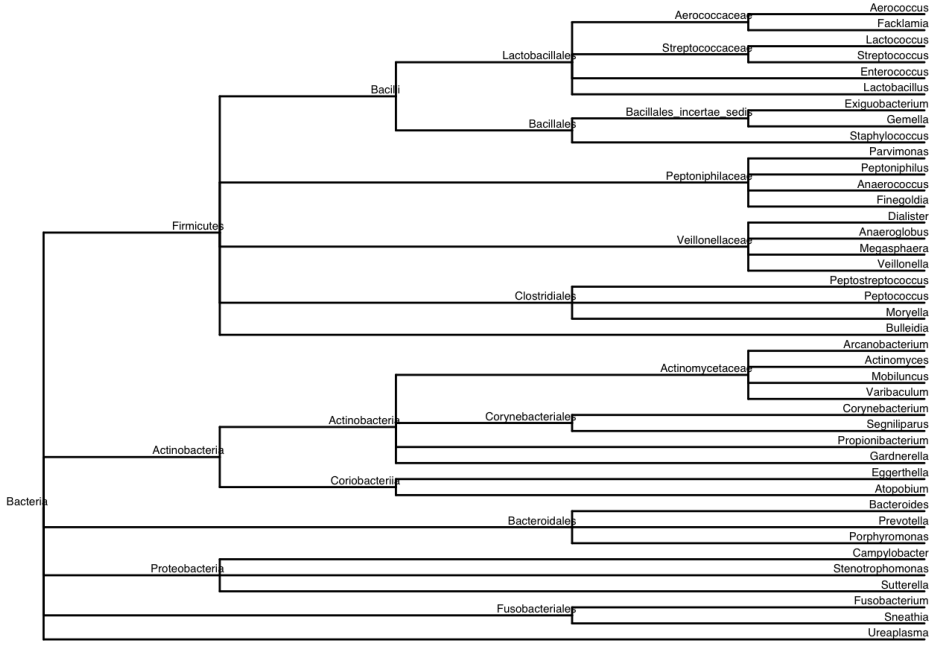

```
Warning: The `n_extra` argument of `print()` is deprecated as of pillar 1.6.2.
Please use the `max_extra_cols` argument instead.

# A tibble: 122 × 395
   Taxa     S001  S002  S003  S004  S005  S006  S007  S008  S009  S010
   <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1 Lactob… 2318  1388  1361  2256    88  1770  1490   119  2136  1790
 2 Prevot…    0     1     1     0   525     7   134   753     0     0
 3 Megasp…    0     1     0     0   402     0     4   102     0     0
 4 Sneath…    0     0     0     0   302     0    35   272     0     0
 5 Atopob…    0     1     0     0    84     0    12    54     0     0
 6 Strept…    0     0     3     0     0     0   138     4     0     2
 7 Dialis…    0     1     0     0   152     4     2   192     0     0
 8 Anaero…    0     1     3     2     0     9    12    13     0     0
 9 Pepton…    0     1     0     0     7     2     6    50     0     0
10 Eggert…    0     0     0     0     2     0     0     7     0     0
# … with 112 more rows
```

- Count data (or compositional) data
- Zero-inflated data
- Correlation between species
- Counts spanning several orders of magnitude: $1 \rightarrow 10^8$

Ravel et al. (2011)

Data - taxonomy


```
# A tibble: 129 × 5
  Phylum         Class          Order           Family            Genus
  <chr>          <chr>          <chr>           <chr>             <chr>
1 Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinobaculum
2 Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Actinomyces
3 Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Arcanobacteriu
4 Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Mobiluncus
5 Actinobacteria Actinobacteria Actinomycetales Actinomycetaceae Varibaculum
# … with 124 more rows
```



Differential abundance studies

Evaluation of binary classifiers

The true condition is usually unknown for real dataset.

The prediction is ususally determined by comparing the *p*-value to $\alpha = 0.05$.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN}) \wedge 1}$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

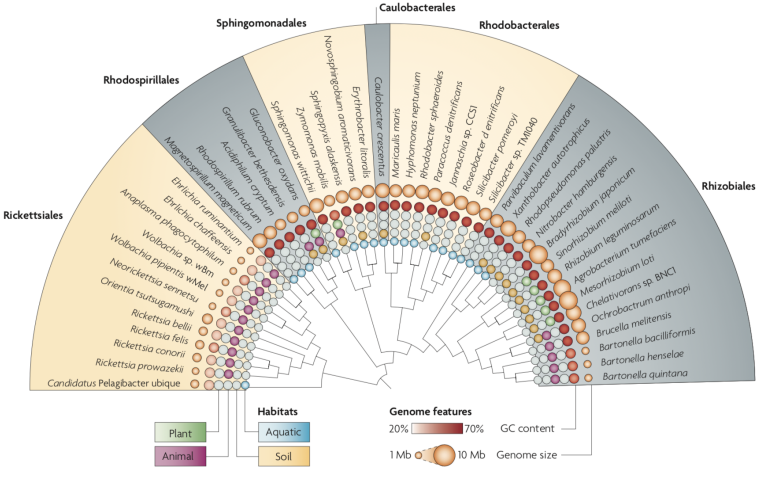
$$\mathrm{FDR} = \frac{\mathrm{FP}}{(\mathrm{TP} + \mathrm{FP}) \wedge 1}$$

Statistical issue

Univariate tests on hundred of taxa

Need for a multiple testing controling procedure!

A hierarchy is available



Can we use it to do it better?

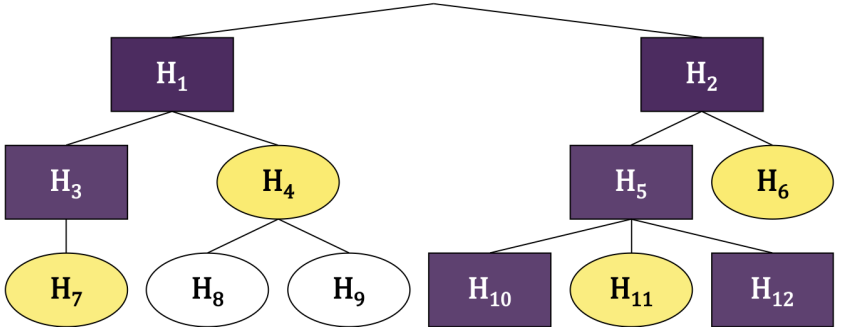
- Hierarchical FDR
- *z*-scores smoothing

📄 Philippot et al. (2010)

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 q



This procedure controls the FDR at level

$$1.44 \times q \times \frac{\#\text{discoveries} + \#\text{families tested}}{\#\text{discoveries} + 1}$$

📄 Yekutieli (2008)

z-scores smoothing

Denote by $\mathbf{z}$ the vector of observed z-scores and μ the vector of "true" z-scores

Assume that $\mathbf{z}|\mu \sim \mathcal{N}_n\left(\mu, \sigma^2\mathbf{I}_m\right)$ and $\mu \sim \mathcal{N}_m\left(\gamma\mathbf{1}, \tau^2\mathbf{C}_\rho\right)$

then

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

and Bayes formula gives

$$\mu^* = \left(\mathbf{I}_m + \frac{\sigma_0^2}{\tau_0^2}\mathbf{C}_{\rho_0}^{-1}\right)^{-1}\left(\frac{\sigma_0^2}{\tau_0^2}\mathbf{C}_{\rho_0}^{-1}\gamma_0\mathbf{1} + \mathbf{z}\right)$$

with σ_0, τ_0, ρ_0 and γ_0 hypermarameters

After smoothing, a multiple testing correction could be done on smoothed values

📄 Xiao et al. (2017)

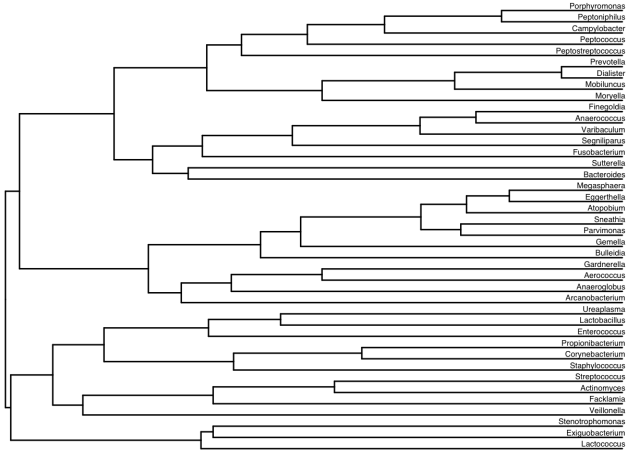
Which tree?

Taxonomy? Phylogeny?

- Proxy for correlations at high-level niches
- Not so much for low-level niches?
- Not available everytime

Correlation tree?

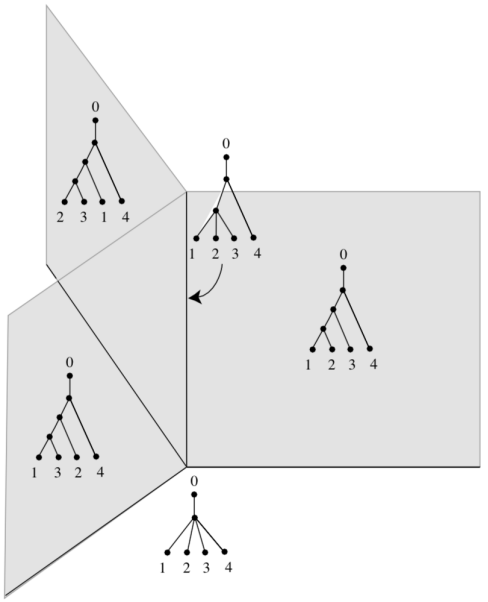
- Actual correlation between taxa
- Computed from data using pairwise correlation

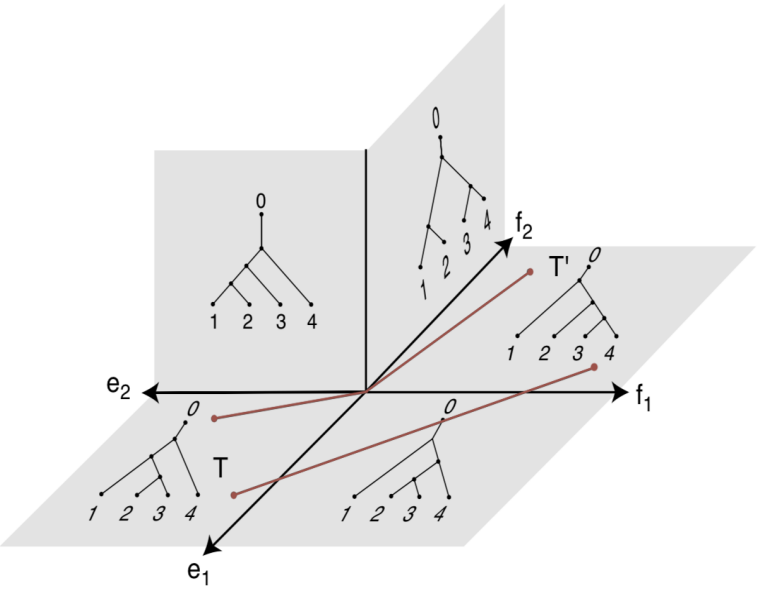


Comparison of trees

Billera-Holmes-Vogtamnn distance

The BHV distance is the length of the unique shortest path between the trees on treespace





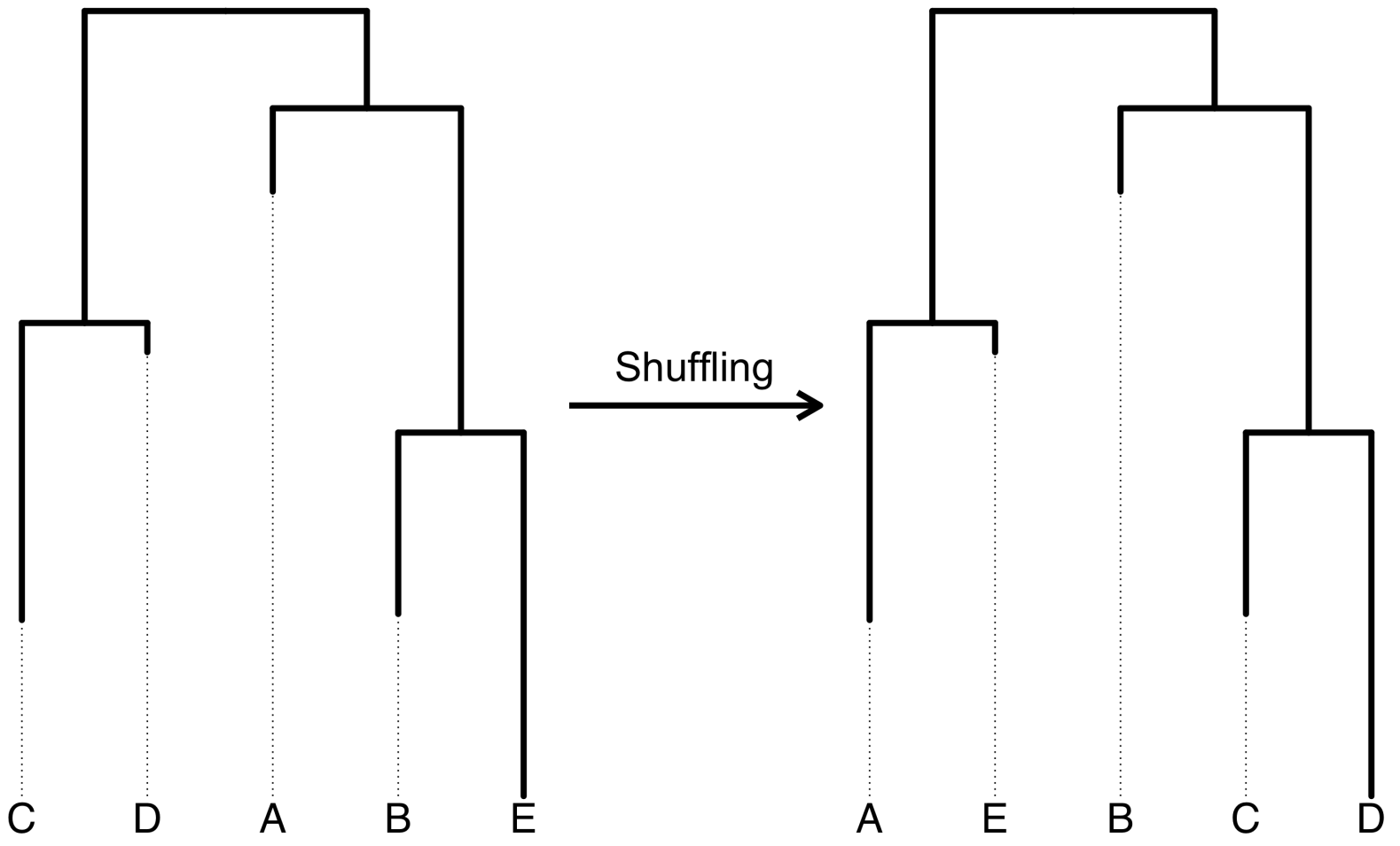
Billera, Holmes & Vogtmann (2001)

Quantifying distance between trees

- **trees of primary interest**
  - correlation tree on original data
  - taxonomy
- **what is the confident 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 taxonomy tip labels

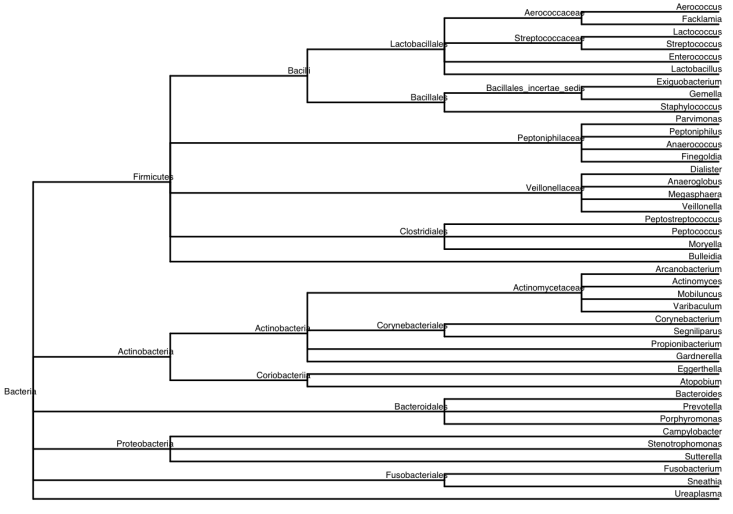
We compute all pairwise distances between these trees

Random shuffling

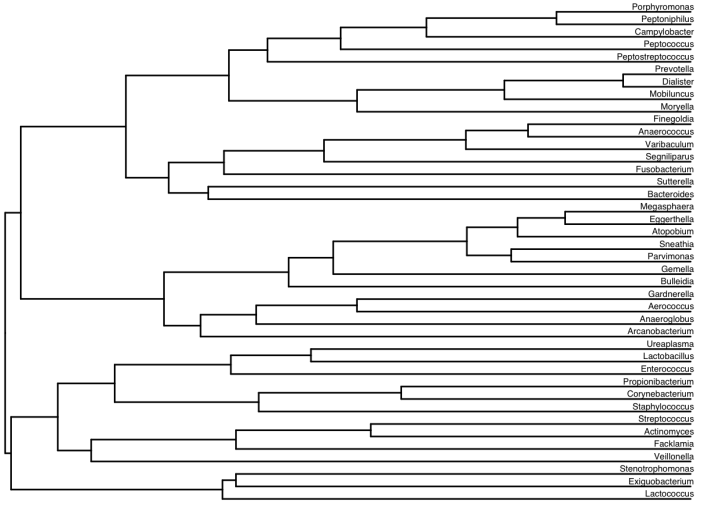


Dataset

- Vaginal microbiome of non pregnant women sequenced by 16S
- 40 different genera after filtering (~ 30 %)



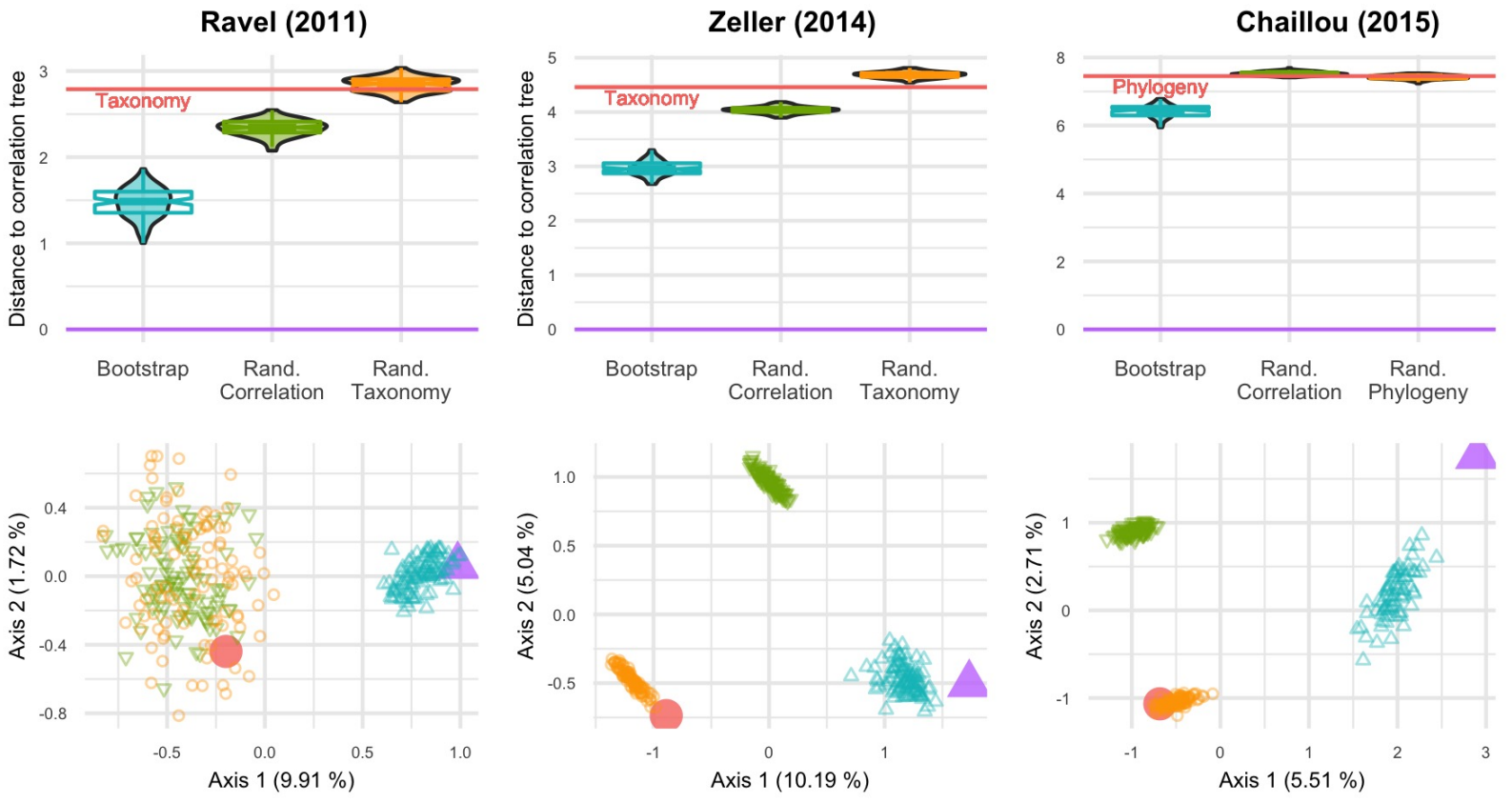
Taxonomy



Correlation tree

📄 Ravel et al. (2011)

Pairwise distances



😀 The correlation tree is different from the taxonomy

Evaluation of hFDR

Chlamydiae dataset

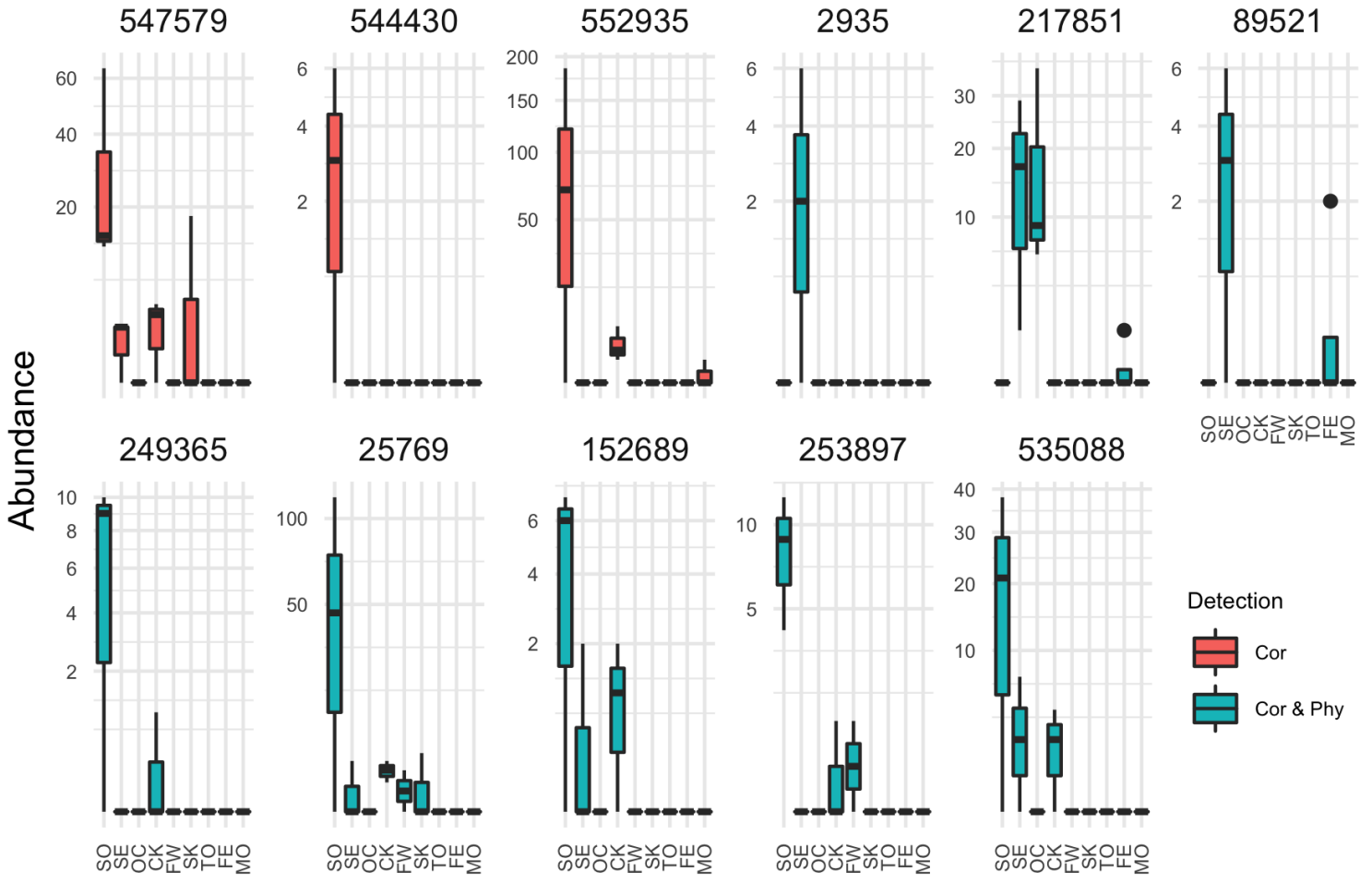
- Small subset of the `GlobalPatterns` dataset narrowed to Chlamydiae phylum
- 21 different OTUs
- 26 samples representing 9 very different environments: soil, ocean, feces, skin...

Method

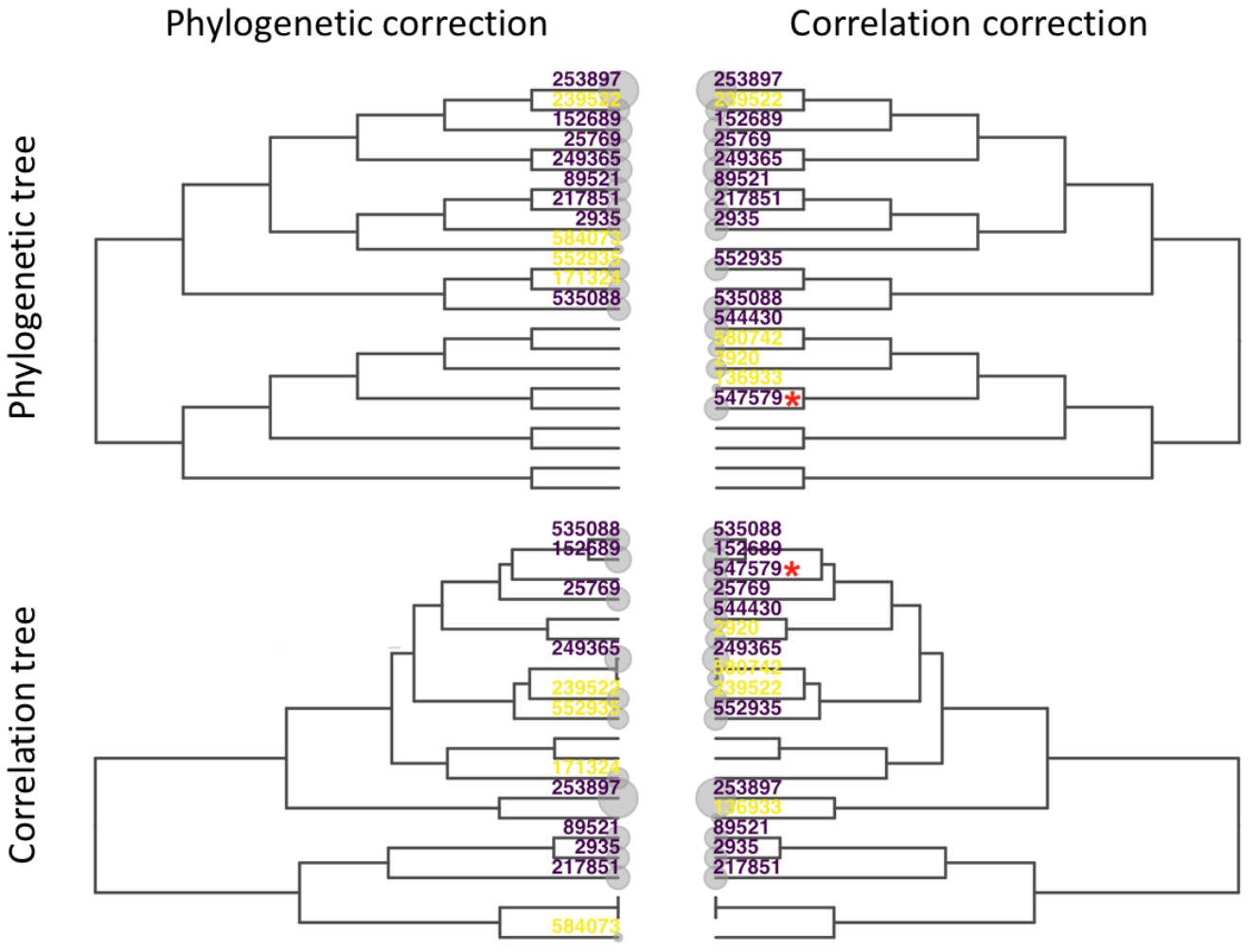
- Find which bacteria are differentially abundant between environments
- Association using Fisher statistic (ANOVA)
- Correction with hierarchical FDR

📄 Caporaso et al. (2011), Sankaran & Holmes (2014)

Abundances of detected species



Representation of evidences on trees



But...

$\alpha = 0.10$ is only the family-level FDR.

The *a posteriori* global FDR is:

- $\alpha' = 0.32$ for phylogenetic correction
- $\alpha' = 0.324$ for correlation correction

A BH procedure at the same global FDR level leads to 15 discoveries (+5)

😀 Using correlation tree instead of taxonomy yields more results

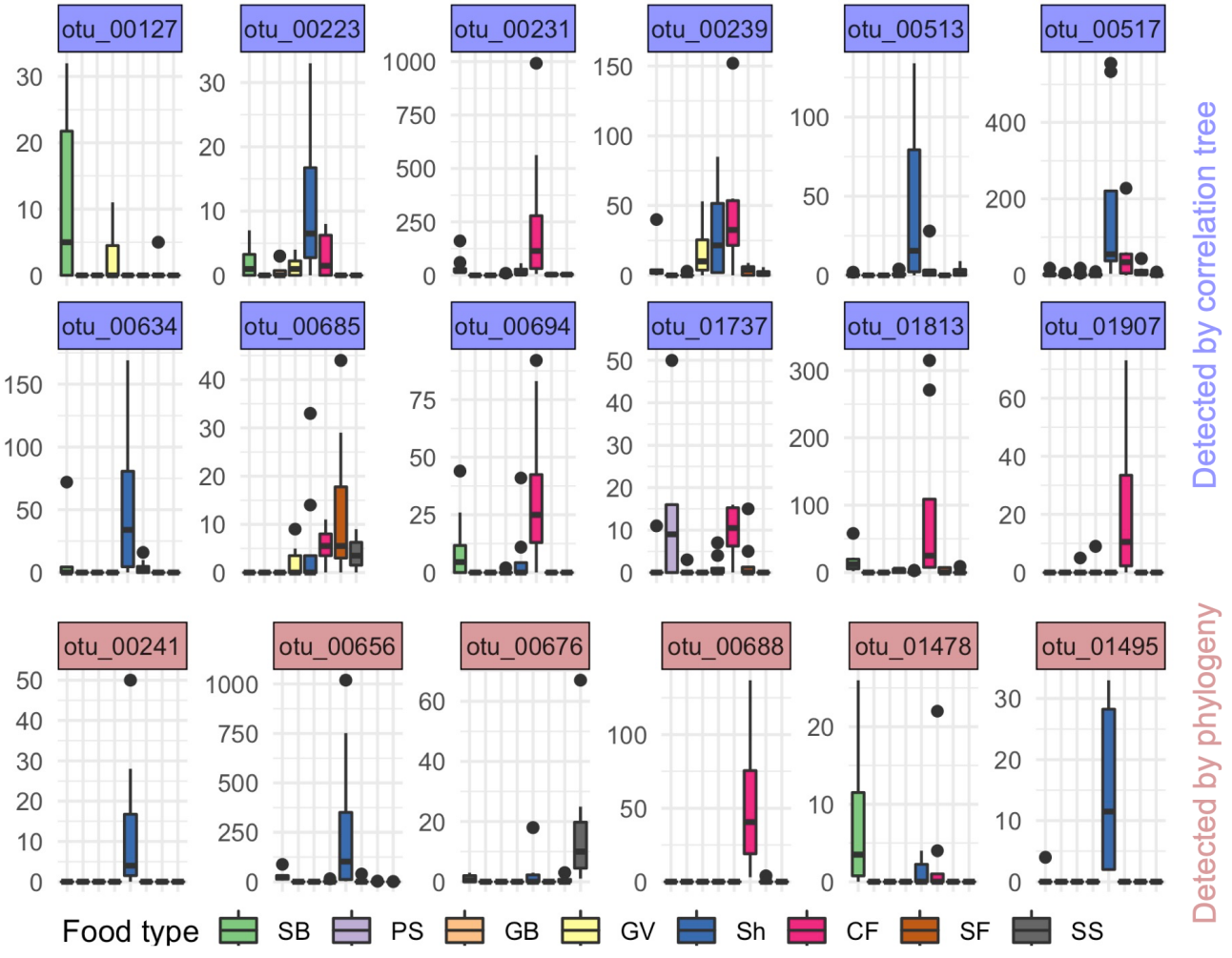
☹ Vanilla BH beats hFDR for a given level

Chaillou Dataset

- Food-associated microbiota of processed meat and seafood products
- 97 different OTUs
- 80 samples across 8 different food type: beef, veal, salmon, shrimp...

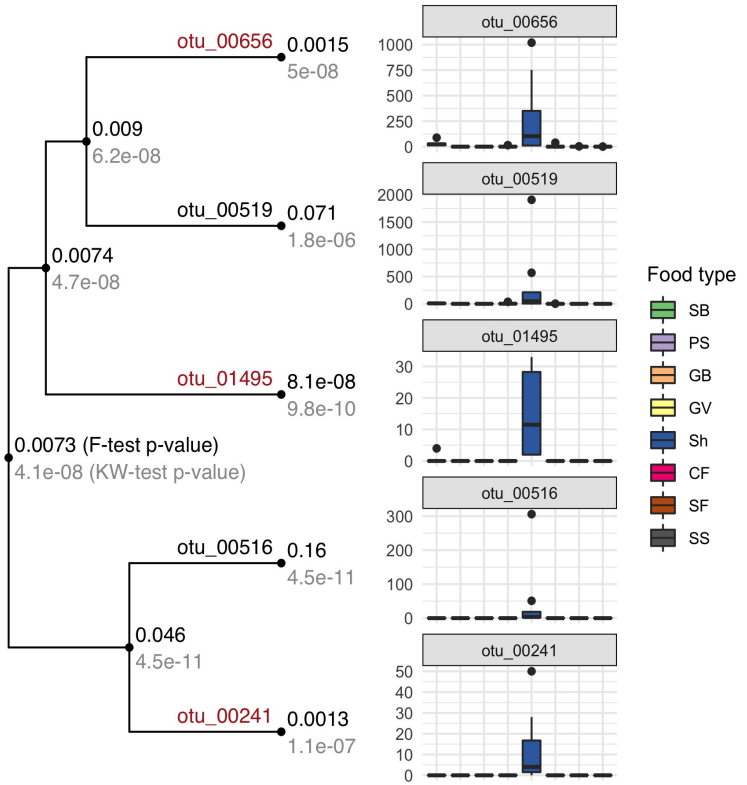
📄 Chaillou et al. (2015)

Abundances of detected species



Impact of the test

Position of selected OTUs on the correlation tree



OTUs in red are only detected by phylogeny

All OTU in the clade are differentially abundant

Obvious unequal variances between groups

KW test is more robust than F-test in this setting

☹ Implemented tests are not appropriate to metagenomic data

Evaluation of z-scores smoothing

Zeller dataset

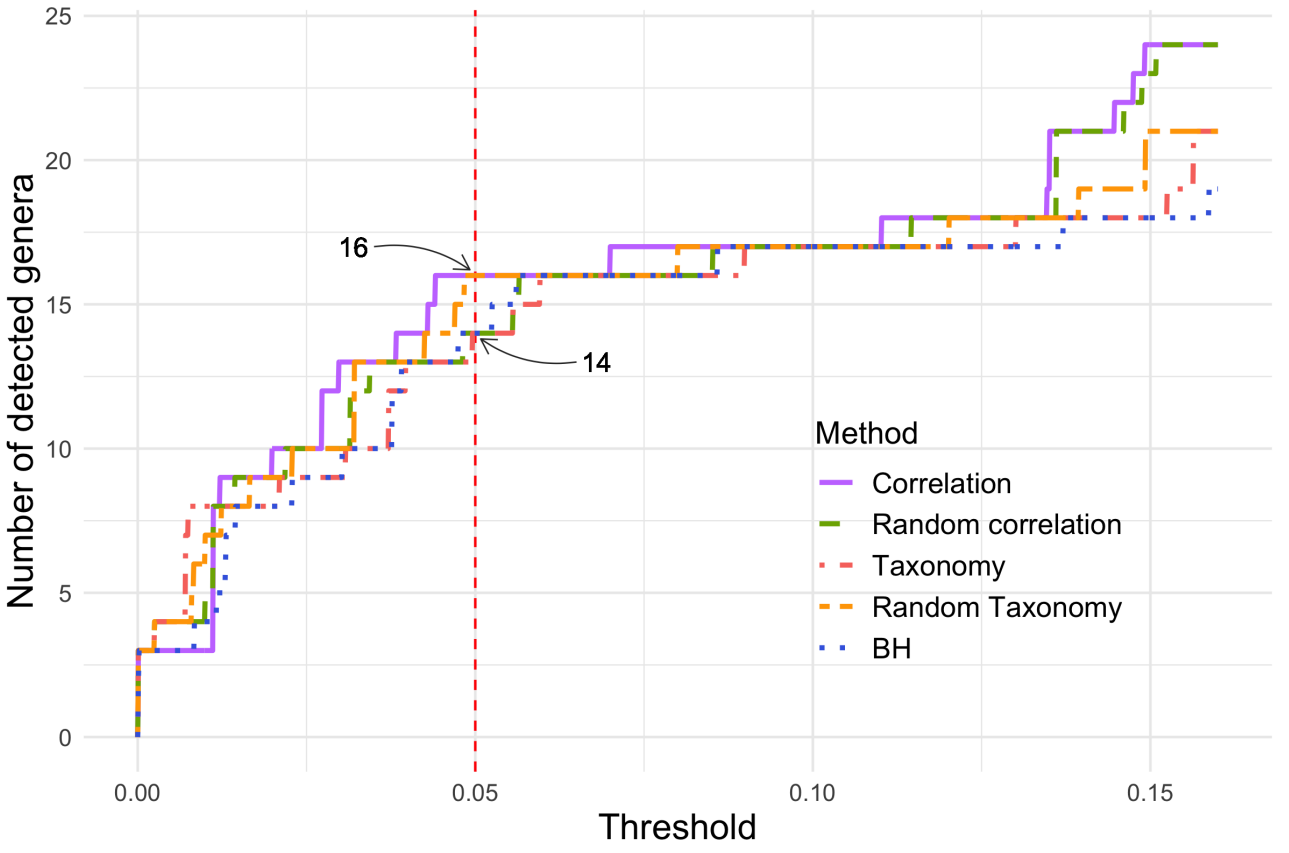
- Dataset from cancer study
- 119 different genera (after filtering)
- 199 samples: 42 adenoma, 91 carcinoma and 66 control

Method

- Find which bacteria are differentially abundant between diseases
- Association using Kruskal-Wallis test
- Correction with hierarchical *p*-value smoothing

📄 Zeller et al. (2014)

Impact of the tree



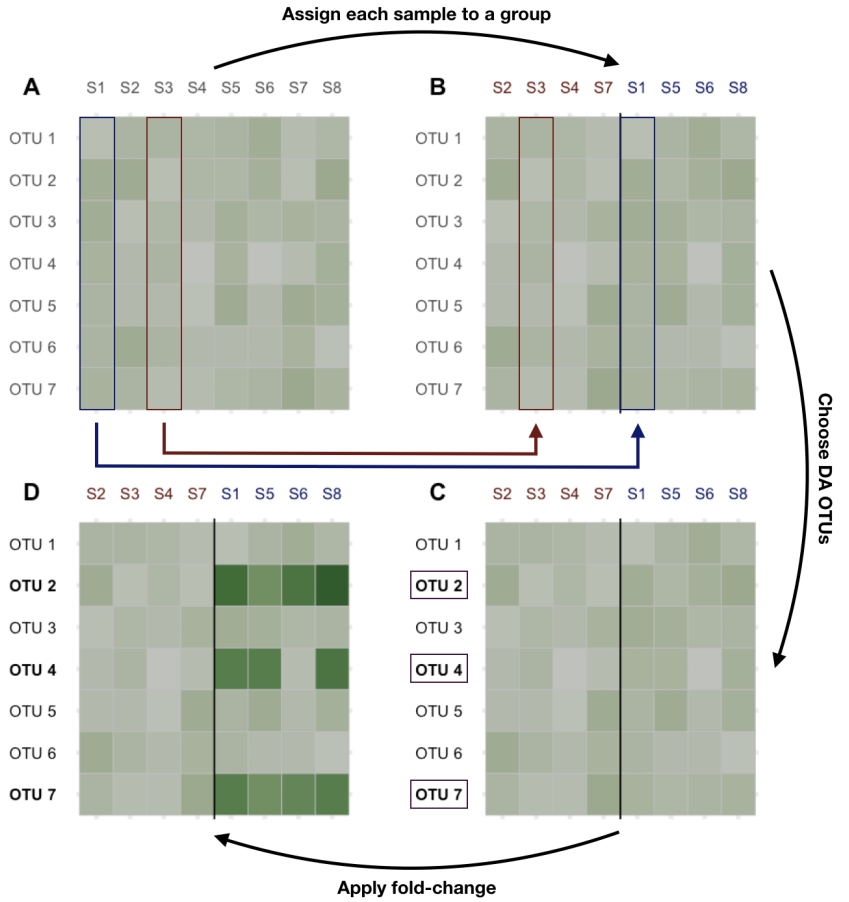
😊 *z*-scores smoothing is slightly better than vanilla BH

☹ All hierachies give highly similar results

Simulations

Workflow

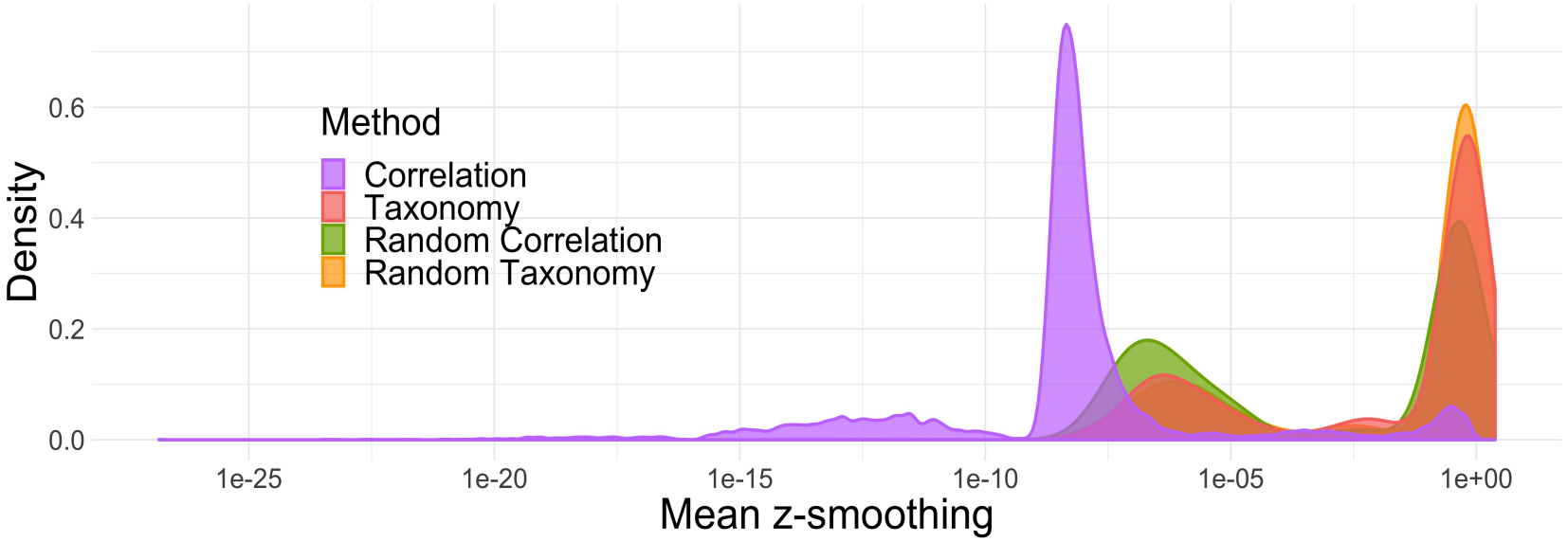
- Simulate DA taxa starting from an homogeneous dataset
- Correction with BH and hierarchical *p*-value smoothing



📄 Brito et al. (2016)

Change in z-scores

$$\text{Mean z-smoothing} = |z_{\text{raw}} - z_{\text{adjusted}}|$$

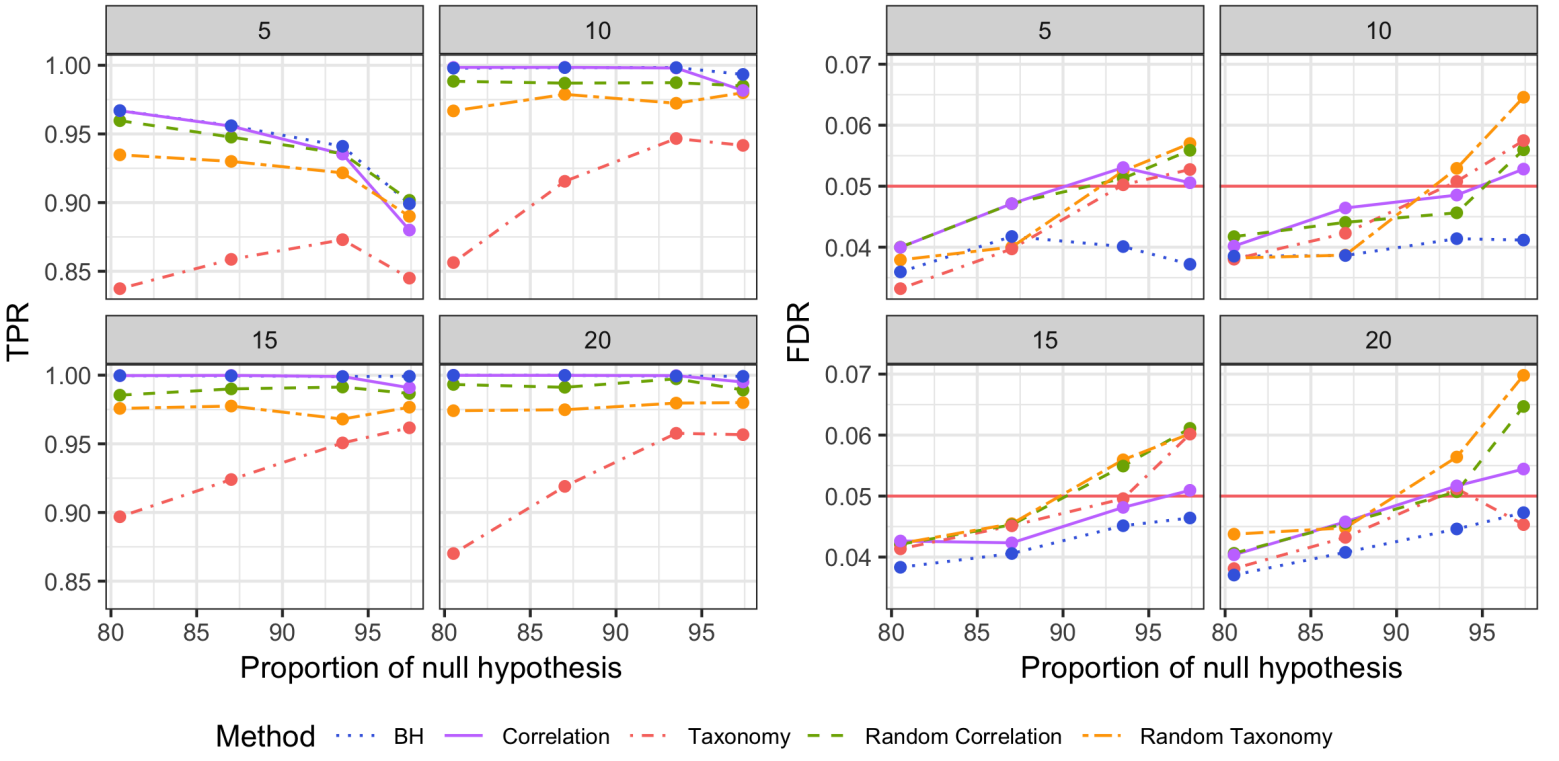


☹️ Taxonomy behaves like random tree and has little impact

😩 In most cases, smoothing has absolutely no impact on the result

enterome bioscience
LaMME MaIAGE

Evaluation



😀 Using correlation tree instead of taxonomy yields more results

😕 Vanilla BH is better

🤔 Taxonomy is worse than random trees

enterome bioscience
LaMME MaIAGE

Conclusions

Conclusions

😀 Correlation tree and taxonomy are very different

😀 Replacing taxonomy tree with correlation tree increases the TPR

☹️ Vanilla BH is more powerful than hFDR

☹️ Bayesian smoothing does not really depend on the tree for z-scores smoothing

☹️ Overall incorporating phylogenetic is not tremendously helpful...

📦 `correlationtree`

📦 `yatah` + CRAN 0.1.0

📦 `evabic`

References

- Bichat, A. et al. "Incorporating Phylogenetic Information in Microbiome Differential Abundance Studies Has No Effect on Detection Power and FDR Control" Frontiers in Microbiology 11:649 (2020). doi: 10.3389/fmicb.2020.00649
- Blander, J. Magarian, et al. "Regulation of inflammation by microbiota interactions with the host." Nature immunology 18.8 (2017): 851.
- Morgan, Xochitl C., et al. “Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment.” Genome biology 13.9 (2012): R79.
- Ravel, Jacques, et al. “Vaginal microbiome of reproductive-age women.” Proceedings of the National Academy of Sciences 108.Supplement 1 (2011): 4680-4687.
- Qin, Junjie, et al. “A metagenome-wide association study of gut microbiota in type 2 diabetes.” Nature 490.7418 (2012): 55.
- Opstelten, Jorrit L., et al. "Gut microbial diversity is reduced in smokers with Crohn's disease." Inflammatory bowel diseases 22.9 (2016): 2070-2077.
- Bokulich, Nicholas A., et al. “Antibiotics, birth mode, and diet shape microbiome maturation during early life.” Science translational medicine 8.343 (2016): 343ra82-343ra82.

Thanks for your attention!

Questions?