

# Thèse de doctorat

NNT : 2020SACLAY000



## Prise en compte de l'organisation hiérarchique des espèces pour la découverte de signatures métagénomiques multi-échelles

### Thèse de doctorat de l'Université Paris-Saclay

École Doctorale de Mathématique Hadamard (EDMH) n°574

Spécialité de doctorat : Mathématiques appliquées

Unité de recherche : Laboratoire de mathématiques et modélisation d'Évry

(UEVE), UMR 8071 CNRS-INRA

Référent : Université d'Évry

Thèse présentée et soutenue à Paris, le 9 décembre 2020, par

**Antoine BICHAT**

### Au vu des rapports de :

<b>Étienne Roquain</b> Maître de conférences, HDR – Sorbonne Université	Rapporteur
<b>Joseph Salmon</b> Professeur des universités – Université de Montpellier	Rapporteur

### Composition du jury :

<b>Christophe Ambroise</b> Professeur des universités – Université d'Évry	Directeur
<b>Julie Aubert</b> Ingénierie de recherche – INRAE Paris	Examinateuse
<b>Agathe Guilloux</b> Professeure des universités – Université d'Évry	Examinateuse
<b>Mahendra Mariadassou</b> Chargé de recherche – INRAE Jouy-en-Josas	Co-encadrant
<b>Matthieu Pichaud</b> Danone	Examinateur
<b>Jonathan Plassais</b> Stat-Alliance	Invité
<b>Étienne Roquain</b> Maître de conférences, HDR – Sorbonne Université	Rapporteur
<b>Joseph Salmon</b> Professeur des universités – Université de Montpellier	Rapporteur
<b>Nathalie Vialaneix</b> Directrice de recherche – INRAE Toulouse	Examinateuse



# Table des matières

<b>Introduction</b>	<b>1</b>
<b>Chapitre 1 : Métagénomique</b>	<b>3</b>
1.1 Le microbiote intestinal humain	3
1.1.1 Description	3
1.1.2 Rôle	4
1.1.3 Dysbioses	5
1.1.4 Utilisations	6
1.2 Caractérisation du microbiote	9
1.2.1 Collecte des échantillons	9
1.2.2 Séquençage par gène marqueur	9
1.2.3 Séquençage non ciblé	11
1.2.4 Données métагénomiques	13
1.3 Jeux de données	18
<b>Chapitre 2 : Études d'analyse différentielle</b>	<b>21</b>
2.1 Tests statistiques	21
2.1.1 Analyse de la variance à un facteur	21
2.1.2 Test de Wilcoxon	22
2.1.3 Test de Kruskall-Wallis	22
2.1.4 Autres méthodes	23
2.2 Problématique des tests multiples	24
2.2.1 Évaluation des performances	26
2.2.2 Correction de Bonferroni	27
2.2.3 Correction de Benjamini-Hochberg	28
2.2.4 Correction de Benjamini-Yekutieli	29
2.3 Procédures hiérarchiques pour tests multiples	30
2.3.1 <i>TreeFDR</i>	30
2.3.2 FDR hiérarchique	31
2.3.3 <i>treeclimbR</i>	32

<b>Chapitre 3 : Arbres</b>	<b>35</b>
3.1 Définitions	35
3.2 Distances	36
3.2.1 Distance de Robinson-Foulds	37
3.2.2 Distance cophénétique	37
3.2.3 Distance de Billera-Holmes-Vogtmann (BHV)	37
3.3 Arbres d'intérêt	39
3.3.1 Phylogénie	40
3.3.2 Taxonomie	40
3.3.3 Arbre des corrélations	40
3.4 Comparaison entre les arbres	41
3.4.1 Forêt d'arbres	41
3.4.2 Distance entre les arbres	42
3.4.3 Choix de l'arbre et lissage de z-score	44
3.4.4 Choix de l'arbre et FDR hiérarchique	49
<b>Chapitre 4 : zazou : une nouvelle approche</b>	<b>55</b>
4.1 Processus d'Ornstein-Uhlenbeck	55
4.2 Zazou	57
4.2.1 Modèle	57
4.2.2 Estimation ponctuelle	58
4.2.3 Débiaisage et intervalles de confiance	59
4.2.4 Correction pour tests multiples	60
4.3 Évaluation de la méthode	61
4.3.1 Données simulées	61
4.3.2 Influence de l'âge	63
<b>Chapitre 5 : Problèmes d'analyse numérique</b>	<b>67</b>
5.1 Algorithme du <i>shooting</i>	67
5.2 Minimisation sous contrainte de $x^T Ax$	68
5.3 Projection sur un ensemble de faisabilité	70
<b>Conclusion et perspectives</b>	<b>73</b>
Détection versus prédiction	74
Autres procédures hiérarchiques	75
Amélioration de zazou	75
<b>Digest</b>	<b>77</b>
Chapter I	77
Chapter II	78
Chapter III	78

Chapter IV	79
Chapter V	80
Conclusions and outlooks	80
<b>Annexe A : Notations</b>	<b>83</b>
<b>Annexe B : Productions scientifiques</b>	<b>85</b>
Quantifying the impact of tree choice in metagenomics differential abundance studies with R	85
Incorporating Phylogenetic Information in Microbiome Differential Abundance Studies Has No Effect on Detection Power and FDR Control	87
Hierarchical correction of p-values via a tree running Ornstein-Uhlenbeck process	94
Packages R	99
yatah	99
evabic	99
correlationtree	99
zazou	99
<b>Annexe C : Vignette de zazou</b>	<b>101</b>
<b>Références</b>	<b>109</b>



# Liste des tableaux

1.1	Quelques exemples de maladies associées à une dysbiose.	6
1.2	Jeux de données utilisés dans ce manuscrit.	19



# Table des figures

1.1	Évolution du coût de séquençage (en dollar) d'une mégabase d'ADN, en échelle logarithmique, tiré de <a href="http://www.genome.gov">www.genome.gov</a> . . . . .	4
1.2	Cotation de Seres Therapeutics à partir de l'été 2016. . . . .	8
1.3	Résumé des étapes du séquençage par gène marqueur, tiré de Regier et al. (2019). . . . .	12
1.4	Résumé des étapes du séquençage non ciblé, tiré de Quince, Walker, Simpson, Loman, & Segata (2017). . . . .	13
2.1	Dessin humoristique illustrant la problématique des tests multiples, par XKCD. . . . .	25
2.2	Matrice de confusion. TP est le nombre de vrais positifs, TN le nombre de vrais négatifs, FP le nombre de faux positifs et FN le nombre de faux négatifs. . . . .	26
2.3	Exemple d'une procédure de FDR hiérarchique. Les hypothèses à tester sont notées $H_1$ à $H_{12}$ . L'algorithme commence par tester et rejeter (après correction) $H_1$ et $H_2$ . Puis il teste la famille $(H_3, H_4)$ , car ce sont des enfants de $H_1$ , et rejette $H_4$ mais pas $H_3$ . La famille $(H_7, H_8, H_9)$ n'est pas testée car $H_3$ n'a pas été rejeté. $H_{10}$ est testé et rejeté. L'algorithme procède de même sur les descendants de $H_2$ . En définitive, il y a trois découvertes aux feuilles $(H_{10}, H_{11}$ et $H_{12})$ pour 5 familles testées. Le FDR a <i>posteriori</i> pour les feuilles est alors de $1.44 \times \alpha \times 2$ . . . . .	31
3.1	Exemple d'un arbre enraciné binaire non ultramétrique à $m = 5$ feuilles et $n = 8$ branches. . . . .	36
3.2	Une partie de $\mathcal{T}_4$ , où cinq orthants se rejoignent, tiré de Billera et al. (2001). . . . .	38
3.3	Des chemins traversant plusieurs orthants dans $\mathcal{T}_4$ , tiré de Billera et al. (2001). . . . .	39
3.4	Distances BHV au sein de la forêt d'arbres pour trois jeux de données. . . . .	43
3.5	Distances BHV au sein de la forêt d'arbres pour le jeu de données Chlamydiae. . . . .	44

3.6	Distances RF au sein de la forêt d'arbres pour trois jeux de données.	44
3.7	Création de taxons différentiellement abondants au sein d'un jeu de données. . . . .	46
3.8	Distribution des moyennes des valeurs absolues des différences entre les <i>z</i> -scores avant et après lissage, pour les simulations non paramétriques. . . . .	46
3.9	Moyennes et écart-types de la moyenne des TPR et FDR pour les simulations non paramétriques avec différents fold-changes et proportions d'hypothèses nulles. . . . .	47
3.10	Moyennes et écart-types de la moyenne des TPR et FDR pour les simulations paramétriques avec différents fold-changes et proportions d'hypothèses nulles. . . . .	48
3.11	Moyennes et écart-types de la moyenne des TPR et FDR pour les simulations non paramétriques avec différents fold-changes et proportions d'hypothèses nulles. . . . .	48
3.12	Les évidences brutes sont représentées aux feuilles de la phylogénie (A et C) ou de l'arbre des corrélations (B et D). Les OTUs considérées différentiellement abondantes pour la phylogénie (A et B) ou pour l'arbre des corrélations (C et D) sont en violet. Les OTUs testées mais non rejetées sont en jaune. . . . .	50
3.13	Abondances des OTUs détectées uniquement par l'arbre des corrélations (en bleu) ou par la phylogénie (en rouge). . . . .	51
3.14	Évidences des OTUs détectées avec l'arbre des corrélations (à droite) ou la phylogénie (à gauche). . . . .	52
3.15	Focus sur les cinq OTUs du cadre vert de la figure 3.14. . . . .	53
4.1	Exemple d'un processus d'Ornstein-Uhlenbeck sur un arbre à 5 feuilles. À chaque branchement, le processus se scinde en deux processus indépendants ayant la même valeur initiale. Les paramètres sont conservés sauf lors d'un saut dans la valeur optimale, comme sur la branche conduisant à $N_4$ . . . . .	56
4.2	TPR (haut) et FDR (bas) pour les différentes procédures et différents <i>fold-changes</i> (en colonnes) dans le cadre des simulations positives. . . . .	62
4.3	Distribution des AUC (haut) et courbes ROC (bas) pour les différentes procédures et <i>fold-changes</i> (en colonnes) dans le cadre des simulations positives. . . . .	63
4.4	Distribution des AUC pour les différentes procédures lorsque les taxons différentiellement abondants sont sélectionnés uniformément. . . . .	64





# Introduction

En 1982, lorsque J. Robin Warren et Barry J. Marshall mettent en évidence la relation entre ulcère de l'estomac et présence de la bactérie *Helicobacter pylori*, la communauté scientifique ne les prend pas au sérieux, estimant impossible la survie de micro-organismes dans l'estomac à cause de son acidité. Des études ultérieures leurs donnent cependant raison et Warren et Marshall reçoivent finalement le prix Nobel de physiologie ou médecine en 2005 « pour la découverte de la bactérie *Helicobacter pylori* et son rôle dans les problèmes gastriques et les ulcères de l'estomac ».

Depuis, les connaissances et les données sur ce que l'on appelle désormais le microbiote –et en particulier sur le microbiote intestinal humain– s'accumulent à un rythme effréné. Celles-ci ouvrent la voie à de nouvelles opportunités thérapeutiques, mais soulèvent également de nouvelles questions. Y répondre nécessite des méthodes statistiques adaptées et de plus en plus puissantes.

Nous nous intéressons ici aux méthodes dites « d'abondance différentielle », dont le but est de détecter les espèces dont la présence ou l'abondance peuvent être affectées par un environnement, le statut de l'hôte et plus généralement un facteur d'intérêt. Plus particulièrement, nous considérerons les approches hiérarchiques, où une information de similarité entre espèces, disponible sous la forme d'un arbre, peut être utilisée pour augmenter la puissance statistique du test.

Ce manuscrit commence par un chapitre de contextualisation biologique. Nous y introduisons les concepts de microbiote et de métagénomique à travers l'exemple du microbiote intestinal et les enjeux qui lui sont associés. Puis nous détaillons les différentes étapes du traitement classique des données métagénomiques –collecte, séquençage, prétraitement. Enfin, ce chapitre se termine en présentant les jeux de données qui seront utilisés dans ce manuscrit.

Le deuxième chapitre est un chapitre de contextualisation statistique autour des analyses d'abondances différentielles. Nous y rappelons les tests statistiques dédiés à cette question puis introduisons la problématique des tests multiples et les corrections habituelles. Enfin, nous terminons ce chapitre en présentant trois méthodes d'analyse d'abondance différentielle qui utilisent une information hiérarchique pour augmenter leur puissance statistique.

Après avoir présenté différentes distances entre les arbres, le chapitre 3 présente une évaluation de l'efficacité des méthodes d'abondances différentielles hiérarchiques et l'impact du choix de l'arbre sur celles-ci.

Le chapitre 4 présente la nouvelle approche que nous avons mise au point pour inclure une information hiérarchique dans les études d'abondance différentielle. Il se termine par une évaluation de cette nouvelle méthode sur des jeux de données synthétiques et réelles.

Enfin, le chapitre 5 présente la résolution de trois problèmes d'analyse numériques que nous avons rencontrés au cours de nos recherches.

# Chapitre 1

## Métagénomique

Quel est le point commun entre une poignée de terre, du camembert et votre intestin ? Tous hébergent une communauté de micro-organismes —bactéries, virus, champignons...— collectivement appelés le microbiote. Aussi extrêmes soient-ils, tous les environnements hébergent des communautés microbiennes. Comme l'on pourrait s'y attendre, les bactéries présentes dans le tube digestif d'escargots sous-marins (Aronson, Zellmer, & Goffredi, 2017) ne sont pas les mêmes que celles présentes dans le désert d'Atacama (Araya, González, Cardinale, Schnell, & Stoll, 2020), mais même au sein d'environnements comparables, il existe une grande variabilité tant au niveau des espèces présentes que de leurs abondances.

Depuis la fin des années 2000, le développement des techniques de séquençage haut-débit et la baisse de leurs coûts (voir figure 1.1) ont rendu accessible l'ensemble des génomes du microbiote, aussi appelé microbiome ou métagénome.

### 1.1 Le microbiote intestinal humain

Même si les développements méthodologiques présentés dans cette thèse ont une portée générale et ne sont pas limités à un type de microbiote en particulier, ils ont effectués avec le microbiote intestinal humain en ligne de mire. Nous présentons ici quelques propriétés de ce microbiote afin que le lecteur prenne conscience de son importance et ai un exemple concret auquel se raccrocher par la suite.

#### 1.1.1 Description

Le tractus gastro-intestinal, ou tube digestif, abrite une communauté microbienne composée d'environ cent mille milliards de micro-organismes pour un poids d'environ 2 kg (Ley, Peterson, & Gordon, 2006). Par abus de langage, on l'appelle le microbiote intestinal. Chez les adultes au mode de vie occidental et en bonne santé,

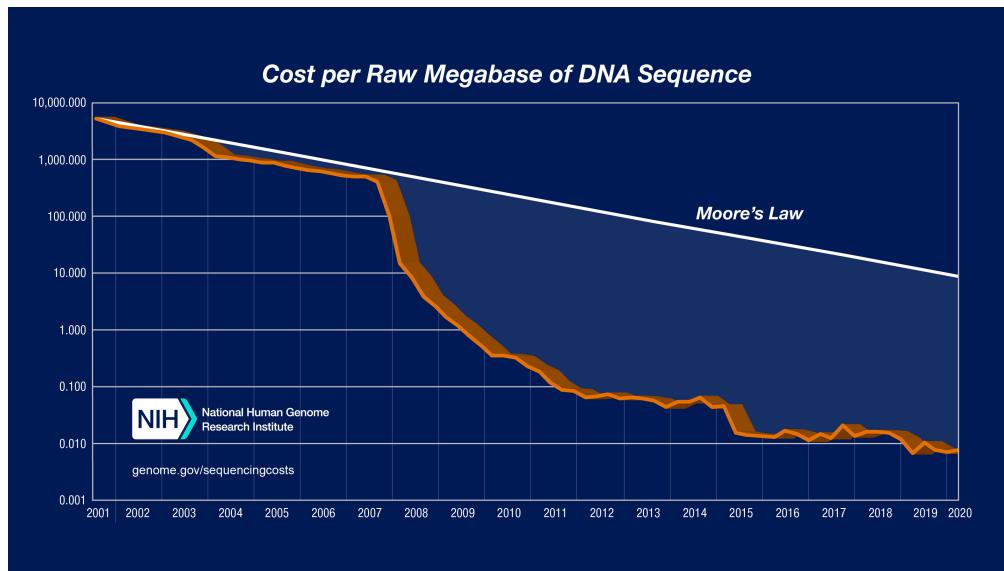


FIGURE 1.1: Évolution du coût de séquençage (en dollar) d'une méga-base d'ADN, en échelle logarithmique, tiré de [www.genome.gov](http://www.genome.gov).

les bactéries appartenant à l'embranchement (*phylum*) des *Firmicutes* représentent plus de 60 % de la composante bactérienne du microbiote. Si l'on y ajoute celles appartenant aux embranchements des *Actinobacteria* et des *Bacteroidetes*, cette fraction monte à 90 % (Zhernakova et al., 2016).

La variabilité inter-individus reflète principalement des facteurs environnementaux ou comportementaux, qui structurent fortement les flores microbiennes, notamment : le régime alimentaire (David et al., 2014), l'âge (O'Toole & Claesson, 2010 ; Yatsunenko et al., 2012) mais aussi la prise d'antibiotiques (Bokulich et al., 2016 ; Palleja et al., 2018), la présence d'un animal de compagnie (Kates et al., 2020), etc.

### 1.1.2 Rôle

Les micro-organismes qui colonisent le tractus gastro-intestinal dégradent les glucides qui n'ont pas été préalablement absorbés par l'hôte (Flint, Scott, Duncan, Louis, & Forano, 2012 ; Rowland et al., 2018). Ces réactions produisent des acides gras à chaîne courte (AGCC) qui sont des sources d'énergies importantes pour l'humain. Plusieurs voies métaboliques, présentes chez différentes bactéries, permettent de produire des AGCC et il n'existe donc un profil unique pour tous les humains. Des études à grande échelles (Arumugam et al., 2011) ont cependant permis de dégager des « profils types », appelés entérotypes, caractérisés entre autres par l'alimentation.

Certaines dégradations de sucres effectuées par le microbiote sont inaccessible aux seules voies métaboliques humaines : le microbiote constitue donc un compagnon indispensable pour assimiler pleinement les nutriments que nous consommons. Citons l'exemple de la population japonaise, qui consomme en moyenne 14.2 g de nori (un type d'algue) par jour et dans laquelle les enzymes porphyranases et agarases responsables de la dégradation des algues sont produites par la bactéries *Bacteroides plebeius*. Cette capacité a été obtenue à la suite d'un transfert horizontal de gène de la part *Zobellia galactanivorans*, une bactéries marine (Hehemann et al., 2010).

Le système immunitaire doit également beaucoup au microbiote intestinal (Blander, Longman, Iliev, Sonnenberg, & Artis, 2017). Parmi les AGCC précédemment cités, on retrouve le butyrate, un métabolite ayant des propriétés anti-inflammatoires et favorisant la prolifération cellulaire au sein de la muqueuse intestinale, ce qui participe à la prévention du cancer colorectal (Canani et al., 2011). La composition du microbiote intestinal a également un effet sur l'efficacité des vaccins (Valdez, Brown, & Finlay, 2014). Eloe-Fadrosh et al. (2013) ont mis en évidence que les patients les plus répondeurs pour un vaccin antityphoïdique étaient ceux présentant une proportion plus importante de *Clostridiales*. Mentionnons enfin dans ce domaine, la bactéries *Bacteroides thetaiotamicron* qui induit une production du peptide antimicrobien LL-37, lequel protège à son tour son hôte contre les infections par *Candida albicans* (Fan et al., 2015).

La signalisation biochimique bidirectionnelle entre le tractus gastro-intestinal et le système nerveux central, communément appelée axe intestin-cerveau, est grandement affectée par le microbiote intestinal. Les résultats les plus spectaculaires sont obtenues chez la drosophile, dans laquelle le microbiote intestinal participe à la modulation des comportements locomoteur (Schretter et al., 2018) et sexuel (Sharon et al., 2010), et chez le rat, dans lequel la production d'un métabolite (l'indole) par certaines bactéries du microbiote induit des troubles du comportement et des troubles de l'anxiété (Jaglin et al., 2018).

### 1.1.3 Dysbioses

Le terme dysbiose désigne un déséquilibre du microbiote, qui se traduit généralement par une perte de diversité ou par la surreprésentation d'une espèce. La table 1.1 présente un ensemble de maladies qui seraient causées par ou associées à une dysbiose.

TABLE 1.1: Quelques exemples de maladies associ es   une dysbiose.

Type de maladie	Maladie
Maladies m�taboliques	Ob�sit� (Turnbaugh et al., 2009) Diab�te de type 2 (Qin et al., 2012) Cirrhose (Qin et al., 2014)
Maladies immunitaires	Maladie de Crohn (Morgan et al., 2012) Syndrome de l'intestin irritable (Chong et al., 2019) Scl�rose en plaques (Cekanaviciute et al., 2017)
Maladies psychiatrique	Asthme (Stokholm et al., 2018) D�pression (Foster & Neufeld, 2013)
Maladies neurologiques	Schizophr�nie (Zheng et al., 2019) Maladie d'Alzheimer (Pistollato et al., 2016)
Autres maladies	Maladie de Parkinson (Bedarf et al., 2017) Syndrome de Gilles de La Tourette (Ding et al., 2019) Eczema (Abrahamsson et al., 2012) Maladies cardiovasculaires (Kelly et al., 2016) Cancer colorectal (Zeller et al., 2014) Ent�rocolite n�crosante (Mai et al., 2011)

#### 1.1.4 Utilisations

Les liens entre microbiote et sant   tant tr s nombreux, la recherche acad mique ainsi que les d partements de recherche et innovation des industries pharmaceutiques et agroalimentaires n'ont pas attendu pour se lancer dans la recherche d'applications et de traitements tirant parti de ces micro-organismes.

Si certaines bact ries sont b n fiques pour l'organisme, pourquoi ne pas augmenter volontairement leur quantit  dans le microbiote ? C'est ce que proposent les approches probiotiques –aussi appel s bioaugmentation– qui consistent   administrer (le plus souvent par voie orale) des bact ries vivantes et non-pathog nes (Gibson et al., 2017). Les probiotiques sont majoritairement consid r s comme des compl ments alimentaires, qui n'ont pas besoin de montrer leur efficacit  pour  tre commercialis s. Des ´tudes cliniques prouvent cependant leur efficacit  dans certains cas, comme l'utilisation d'un probiotique   base d'une souche de *Bifidobacterium longum* pour diminuer la d pression chez les patients souffrant du

syndrome de l'intestin irritable (Pinto-Sánchez et al., 2017).

Plutôt que de fournir directement des souches vivantes, l'approche par prébiotiques fournit des nutriments non digestibles par l'hôte mais stimulant la croissance de certaines bactéries (Gibson et al., 2017). Si les aliments « riches en fibres » peuvent être considérés comme des prébiotiques, des prébiotiques de synthèse font leur apparition sur le marché. Nestlé a par exemple déposé un brevet sur des prébiotiques à base d'oligosaccharides qui réduisent la présence de *Streptococcus* chez l'enfant dans le but de diminuer le risque d'obésité une fois adulte (Sakwinska, Berger, Zolezzi, & Holbrook, 2017). Dans le même registre, la baguette « Amibioté » (contraction entre ami et microbiote), issue d'un partenariat entre l'INRAE et Bridor, contient 11 g de fibres pour 100 g de pain (contre 2.9 g pour une baguette normale) et favorise la croissance de trois bactéries probiotiques.

Pour modifier la composition du microbiote intestinal, la méthode la plus efficace reste la transplantation fécale. Déjà pratiquée en Chine au IV<sup>e</sup> siècle (Zhang, Luo, Shi, Fan, & Ji, 2012), la repopulation de l'intestin d'un sujet malade avec le microbiote d'un sujet sain a fait un retour en force en montrant des résultats spectaculaire pour le traitement des infections à *Clostridium difficile* comparé aux thérapies habituelles (prise d'antibiotiques avec ou sans lavement) (Van Nood et al., 2013). L'infection à *Clostridium difficile* provoque des diarrhées potentiellement mortelles et se produit chez des patients dont le microbiote intestinal a déjà subi une perte de diversité, laissant la place au pathogène pour se développer.

Si la transplantation fécale pourrait servir de traitement à d'autres maladies, comme le syndrome de Gilles de La Tourette (Ding et al., 2019), ce n'est pas le seul dessein dans lequel l'utilisation de cette technique est possible. Des transplantations fécales autologues peuvent être envisagées dans le cas où une altération du microbiote serait à prévoir. Cette technique a été testée avec succès par la société française MaaT Pharma pour des patients souffrant de leucémie aiguë myéloïde. Leurs selles sont collectées avant la chimiothérapie et une transplantation fécale autologue permet de restaurer un microbiote diversifié comparable à celui présent dans le tractus avant le traitement (Mohty et al., 2018). Une utilisation plus originale en est faite chez les koalas : la transplantation fécale leur permet de diversifier les espèces d'eucalyptus qu'ils sont capable de digérer, ce qui augmente leurs chances de survie alors que leur environnement est menacé (Reardon, 2018).

La médiatisation de ces bons résultats a conduit à l'apparition sur les réseaux sociaux de protocoles pour des transplantations fécales « à faire soi-même » (Ekekezie et al., 2020), y compris pour des indications pour lesquelles elles ne sont (pour l'instant) pas recommandées comme le syndrome de l'intestin irritable ou les troubles de l'autisme. En fonction des pays, la thérapie fécale est considérée comme un médicament ou une transplantation au même titre que les organes.

Plutôt que de modifier le microbiote et d'espérer que la nouvelle composition

sera favorable, il est possible d'administrer directement des composés issus des bactéries désirées. C'est sur ce terrain que se place la biotech Enterome, qui a montré que l'administration de peptides issus du microbiote déclencheait une réponse immunitaire capable de s'attaquer à des tumeurs spécifiques (Chene et al., 2019).

Une autre façon de tirer parti du microbiome est de s'en servir comme d'un biomarqueur : regarder sa composition peut être un moyen de déceler une maladie sans avoir recours à des tests invasifs ou plus coûteux. Zeller et al. (2014) ont par exemple proposé un modèle prédictif basé principalement sur les abondances de souches de *Fusobacterium nucleatum*, *Porphyromonas asaccharolytica* et *Peptostreptococcus stomatis* pour détecter de façon précoce un cancer colorectal. Chez les nouveaux-nés, l'entérocolite nécrosante est précédée d'une dysbiose caractéristique (Mai et al., 2011) permettant de l'identifier et de proposer un traitement adapté avant que la maladie ne leur soit fatale.

Toutes ces opportunités ont incité de nombreuses sociétés à se lancer sur le marché du microbiote dans l'espoir de croissances rapides. Mais le réveil peut être douloureux lorsque les espoirs se fracassent sur le mur de la réalité, comme Seres Therapeutics en a fait les frais. Le 29 juillet 2016, l'annonce de l'échec de la phase II de son probiotique SER-109 contre les infections à *Clostridium difficile* a entraîné une chute du cours de son action près de 70 % (figure 1.2). L'entreprise a publié un communiqué en janvier 2017 indiquant que cet échec était dû à une erreur de protocole et qu'elle allait par conséquent entamer un essai de phase III, mais cela n'a pas suffi à rassurer les investisseurs.

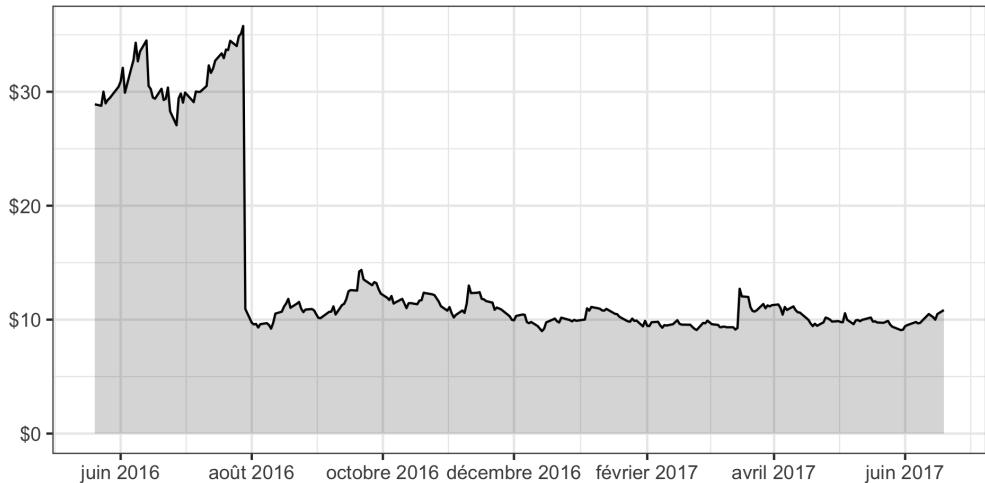


FIGURE 1.2: Cotation de Seres Therapeutics à partir de l'été 2016.

## 1.2 Caractérisation du microbiote

Nous décrivons ici brièvement les différentes méthodes et techniques utilisées pour caractériser le microbiote.

### 1.2.1 Collecte des échantillons

La collecte des échantillons et le mode de conservation des échantillons sont des étapes cruciales pour la reproductibilité des analyses et la fiabilité des résultats. En effet, le temps de stockage avant une congélation à  $-80^{\circ}\text{C}$  (Cuthbertson et al., 2014) ou le nombre de dégels successifs (Sergeant, Constantimou, Cogan, Penn, & Pallen, 2012) peuvent avoir un impact important sur la composition microbienne des échantillons, en affectant certains genres plutôt que d'autres.

Il est donc souhaitable d'avoir les conditions les plus homogènes possibles. Cependant, il peut être relativement compliqué d'obtenir des conditions homogènes de collecte et de conservation, notamment quand il s'agit d'études longitudinales où la récolte d'échantillons s'étale sur plusieurs pays ou plusieurs années, avec des équipes différentes, ou bien quand celle-ci est effectuée au domicile du donneur dans le cadre d'études participatives, telles que l'*American Gut Project* (McDonald et al., 2018).

### 1.2.2 Séquençage par gène marqueur

Le séquençage par gène marqueur, en général la sous-unité 16S de l'ARN ribosomique, permet de faire un inventaire taxonomique des espèces présentes dans le microbiote et de répondre à la question « qui est là ? ». Fait notable, le gène codant pour le 16S est présent chez toutes les espèces de bactéries et archées (Kembel, Wu, Eisen, & Green, 2012). Sa séquence présente une succession de régions conservées, idéales pour l'extraire et l'amplifier par PCR, et des régions hypervariables, qui permettent de déterminer à quel genre (et dans une certaine mesure à quelle espèce) il appartient et de reconstruire une phylogénie. Il bénéficie également de bases de données taxonomiques extrêmement riches (par exemple, SILVA (Quast et al., 2012)). Pour toutes ces raisons, le gène 16S est un marqueur idéal pour identifier les différents micro-organismes présents dans un échantillon et quantifier leur abondance (Morgan & Huttenhower, 2012).

Les différentes étapes d'analyse s'enchaînent comme suit. Après extraction de l'ADN, on amplifie le gène 16S par PCR avec des amorces universelles, calibrées à partir des régions conservées, puis on séquence une ou plusieurs régions hypervariables (pour une longueur totale d'environ 550 paires de base) par séquençage haut débit. Une fois ces portions d'ADN séquencés, on a accès aux séquences brutes des nucléotides qui les composent, appelé lectures ou *reads*.

Ces lectures doivent ensuite subir un contrôle qualité. En effet, les technologies de séquençage produisent des lectures trop courtes ou de mauvaise qualité (Moldolo & Lerat, 2015) qui doivent être éliminées. De plus, lors de l'amplification par PCR, il peut y avoir des évènements d'hybridation entre séquences d'ADN (Meyers, Vartanian, & Wain-Hobson, 1990) qui créent des séquences chimériques et augmentent artificiellement la richesse microbienne. Des algorithmes ont été proposés pour identifier et filtrer les séquences chimériques avant la suite des analyses (Edgar, 2016 ; Wright, Yilmaz, & Noguera, 2012).

Une première approche, dite *affiliation first*, pour identifier à quelles espèces correspondent les lectures séquencées et de trouver une correspondance entre celles-ci et des séquences de références dans des bases de données de gènes 16S comme RDP (Maidak et al., 2000, 1997), SILVA (Quast et al., 2012) ou Greengenes (De-Santis et al., 2006). Cette méthode est rapide et facilement parallélisable mais rend impossible le regroupement et l'analyse des lectures qui n'ont pas d'homologues dans les bases de référence.

L'approche la plus utilisée, dite *clustering first*, consiste à rassembler les lectures au sein de groupes appelés OTUs, pour *Operational Taxonomic Units*. Ce partitionnement se fait en agglomérant toutes les séquences qui ont au moins 97 % de similarité de séquence. Des outils bioinformatiques comme Mothur (Schloss et al., 2009) ou QIIME (Caporaso et al., 2010) utilisent des algorithmes de classification ascendante hiérarchique pour effectuer ce partitionnement, ce qui est coûteux en calcul et en mémoire. Des algorithmes gloutons (Edgar, 2010) permettent d'accélérer et de réduire l'empreinte mémoire de cette étape de partitionnement. À l'issue de cette étape, un représentant est ensuite choisi pour chaque groupe afin de lui assigner une affiliation taxonomique en comparant ce représentant à des bases de références. Les séquences des représentants permettent également de déterminer un arbre phylogénétique des OTUs. Comparée à l'approche *affiliation first*, l'identification par OTU a l'avantage de pouvoir gérer des espèces non présentes dans les bases de données de référence en les considérant simplement comme mal affiliées. Le microbiote est finalement résumée par une table de comptage  $X = (x_{ij})$  où  $x_{ij}$  correspond au nombre de lectures de l'OTU  $i$  dans l'échantillon  $j$ .

Au delà du partitionnement par similarité de séquences, il existe d'autres méthodes pour partitionner l'ensemble des séquences. Mentionnons par exemple le regroupement aux sein d'ASVs, pour *Amplicon Sequence Variants*, qui a vocation à reconstruire les séquences exactes des représentants à l'aide d'un modèle probabiliste des erreurs de séquençage (Callahan, McMurdie, & Holmes, 2017 ; Callahan et al., 2016), ou au sein d'oligotypes, qui se focalisent sur les sites nucléotidiques de grande variabilité (Eren, Borisy, Huse, & Welch, 2014 ; Eren et al., 2013). Le point de différentiation majeur de ces méthodes par rapport aux OTUs est de ne pas donner le même poids à toutes les positions lors de la construction des groupes

de lectures.

La caractérisation par gène marqueur, en particulier le 16S, est bon marché, rapide et s'appuie sur des outils matures. Elle souffre néanmoins de quelques inconvénients :

1. Lors de l'amplification par PCR, outre la possible création de chimères, les taxons très abondants vont avoir plus de chances de voir leurs deux brins s'apparier entre eux plutôt qu'avec une amorce, ce qui brise la chaîne de réPLICATION (Mathieu-Daudé, Welsh, Vogt, & McClelland, 1996). Les taxons peu présents vont au contraire avoir plus de chances d'aller jusqu'au bout de la chaîne de réPLICATION et leur abondance sera surestimée.
2. Le nombre de copies du gène 16S varie entre espèces, dans un rapport de 1 à 21 (Stoddard, Smith, Hein, Roller, & Schmidt, 2015). Utiliser cette information pour corriger la quantification des OTU améliore l'estimation de la composition microbienne, mais le nombre de copies n'est pas toujours disponible (Kembel et al., 2012), en particulier pour les groupes microbiens peu étudiés.
3. Le 16S est limité à la fraction bactérienne du microbiote. D'autres marqueurs comme l'ITS1 et l'ITS2 doivent être utilisées pour la fraction fongique du microbiote. Ces derniers souffrent également du biais du nombre de copies (mais dans un rapport de 1 à 1 000) et de bases de référence nettement moins riches.
4. Le séquençage par gène marqueur ne permet d'obtenir une résolution taxonomique qu'au niveau du genre, éventuellement de l'espèce dans certains cas favorables. Il ne permet pas non plus de déterminer les fonctions ou voies métaboliques présentes dans le microbiote. Ces dernières peuvent en effet être spécifiques aux souches au sein d'une espèce et nécessitent d'adopter une stratégie non ciblée.

### 1.2.3 Séquençage non ciblé

Comme son nom l'indique, le séquençage non ciblé, ou *whole genome shotgun* souvent abrégé en *shotgun*, cible l'ensemble du matériel génétique présent dans les échantillons, et non un unique gène marqueur. Il permet d'étudier le potentiel fonctionnel du microbiote et de répondre à la question « qui peut faire quoi ? ».

On commence par extraire l'ADN contenu dans les cellules puis on le brise par ultrason –ou sonification– afin d'obtenir de courts fragments, de l'ordre d'une centaine de nucléotides. On effectue ensuite une PCR non ciblée avec des amores

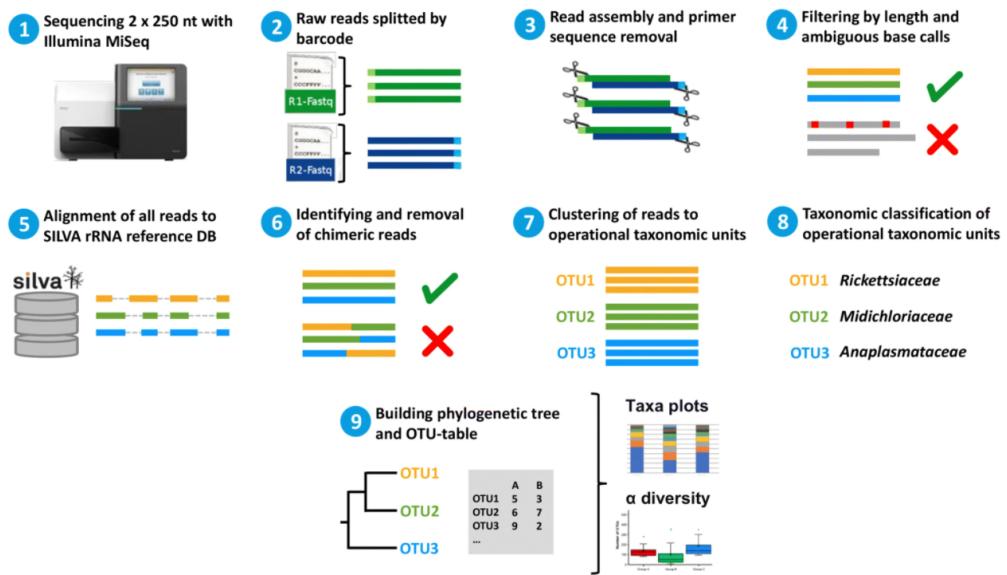


FIGURE 1.3: Résumé des étapes du séquençage par gène marqueur, tiré de Regier et al. (2019).

aléatoires pour amplifier les fragments avant de les séquencer pour obtenir des lectures couvrant l'ensemble du matériel génétique.

La construction d'une table de comptage à partir des lectures est sensiblement plus compliquée que dans l'approche par gène marqueur. Il existe en effet plusieurs manières d'obtenir une table de comptage à partir des lectures : par comparaison de lectures (Maillet, Collet, Vannier, Lavenier, & Peterlongo, 2014 ; Maillet, Lemaitre, Chikhi, Lavenier, & Peterlongo, 2012), par comparaison des profils en  $k$ -mers (Benoit et al., 2016 ; Deorowicz, Kokot, Grabowski, & Debudaj-Grabysz, 2015), par classification exhaustive des lectures (Brady & Salzberg, 2009 ; Kim, Song, Breitwieser, & Salzberg, 2016 ; Ounit, Wanamaker, Close, & Lonardi, 2015 ; Wood & Salzberg, 2014), par recensement de gènes marqueurs (Liu, Gibbons, Ghodsi, Treangen, & Pop, 2011 ; Segata et al., 2012 ; Truong et al., 2015) ou encore par utilisation d'un catalogue de gènes (Coelho et al., 2019 ; Kultima et al., 2012 ; Pons et al., 2010). Nous allons détailler cette dernière méthode, qui nécessite la construction préalable d'un catalogue ou l'utilisation d'un catalogue public (Almeida et al., 2020).

Une fois le catalogue obtenu, chaque lecture est alignée contre celui-ci pour déterminer le gène auquel elle correspond le plus, sur la base de la similarité de séquence. Dans le meilleur des cas, une lecture ne s'aligne que sur un seul gène, mais il arrive dans environ 10 % des cas qu'elle s'aligne sur plusieurs gènes distincts, par exemple parce qu'elle correspond à un domaine protéique partagé par plusieurs séquences. Dans ce dernier cas, plusieurs procédures sont possibles :

- On ne prend pas en compte cette lecture dans le comptage.
- Le compte de cette lecture est réparti uniformément entre les gènes (*i.e.* elle augmente le comptage de chacun des  $n$  gènes sur lesquels elle s'aligne de  $1/n$ ).
- Le compte de cette lecture est réparti entre les gènes, au prorata de leurs comptages obtenus à partir des lectures non ambiguës (*i.e.* si elle s'aligne sur les gènes  $G_1, \dots, G_n$  de comptages respectifs  $A_1, \dots, A_n$  dans les lectures non ambiguës, l'abondance du gène  $i$  est augmentée de  $A_i / \sum_{j=1}^n A_j$ ).

Contrairement à l'approche par gène marqueur, l'approche non-ciblée sur catalogue de gène dresse un inventaire fonctionnel du microbiote, où le gène remplace l'OTU comme descripteur de base du microbiote.

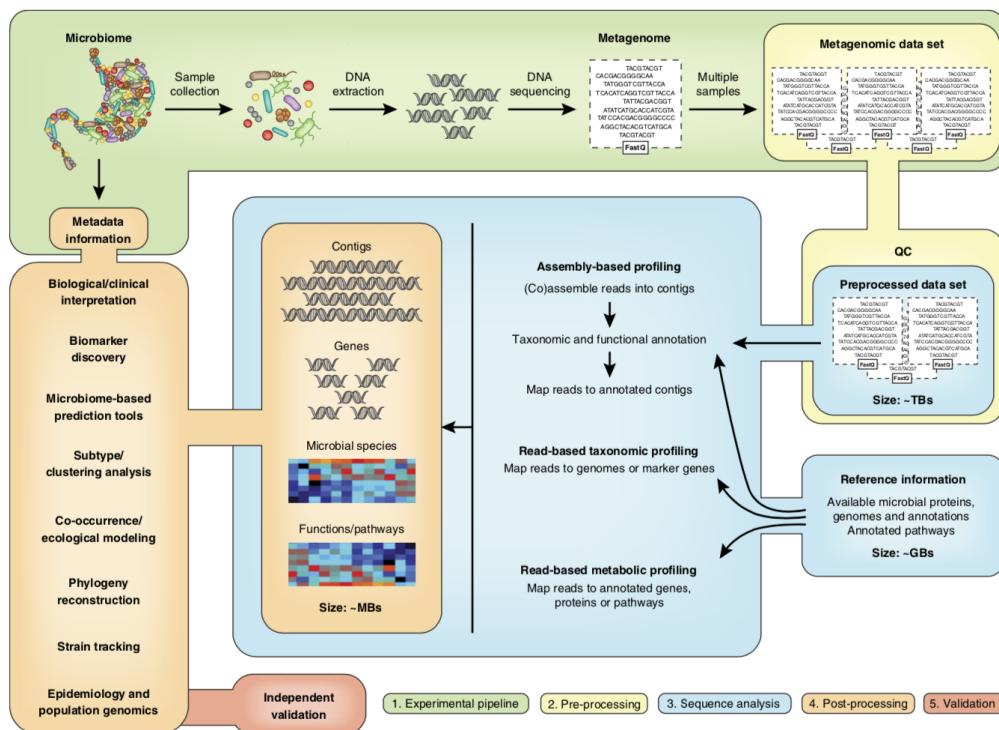


FIGURE 1.4: Résumé des étapes du séquençage non ciblé, tiré de Quince, Walker, Simpson, Loman, & Segata (2017).

#### 1.2.4 Données métagénomiques

Une fois l'annotation effectuée, on dispose d'une table de comptage, c'est à dire du compte des différentes entités considérées dans les échantillons. Dans la suite,

on désignera ces entités sous le terme générique *taxon*, qu'il s'agisse d'espèces, de genres bactériens, d'OTUs, de gènes...

Plusieurs cadres conceptuels existent pour analyser ces données mais les deux plus populaires sont ceux (i) des données de comptage et (ii) des données compositionnelles. Chacun de ces cadres possède des avantages et inconvénients qui lui sont propres et malgré des discussions intenses dans la littérature, (Gloor, Macklaim, Pawlowsky-Glahn, & Egoscue, 2017; Gloor et al., 2016; McMurdie & Holmes, 2014; Vandepitte et al., 2017), il n'existe pas à l'heure actuelle de consensus sur celui à privilégier.

### Données de comptage

Par la façon dont elles sont construites, à savoir en comptant le nombre de lectures de chaque taxon dans chaque échantillon, les tables d'abondance sont composées de données de comptage, à valeurs dans  $\mathbb{N}$ .

Une approche naturelle serait de considérer ces comptes comme issus d'une loi de Poisson, dont les masses sont telles que  $\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$  pour  $k \in \mathbb{N}$ . Cependant, l'espérance et la variance d'une loi de Poisson sont égales, alors que les données observées sont surdispersées : elles possèdent une plus grande variabilité qu'attendue pour une distribution de Poisson (Anders & Huber, 2010 ; Robinson & Smyth, 2007). Afin de prendre en compte une telle surdispersion dans les données, il est préférable d'utiliser une loi binomiale négative (Zhang et al., 2017), qui peut être vu comme un mélange de loi de Poisson, avec une densité  $\Gamma$  sur le paramètre de la loi de Poisson.

Dans sa définition classique, la loi binomiale négative de paramètres  $n$  et  $p$  compte le nombre d'échecs avant d'obtenir  $n$  succès pour un événement binaire, avec  $p$  comme probabilité de succès. Sa fonction de répartition est alors telle que

$$\mathbb{P}(X = k) = \binom{k + n - 1}{n - 1} p^n (1 - p)^k.$$

On préférera utiliser ici une autre paramétrisation, celle de moyenne  $m$  et de paramètre de dispersion  $\alpha \geq 0$ . Lorsque  $\alpha = 0$ , on retombe sur une loi de Poisson. Dans le cas contraire,

$$\mathbb{P}(X = k) = \frac{\Gamma(k + r)}{k! \Gamma(r)} \left( \frac{m}{r + m} \right)^r \left( \frac{r}{r + m} \right)^k$$

où  $r = \frac{1}{\alpha}$  et  $\Gamma : z \mapsto \int_0^{+\infty} t^{z-1} e^{-t} dt$  est la fonction gamma.

Une autre particularité des données métagénomiques est leur grande proportion de zéros : les comptages nuls sont surreprésentés et représentent fréquemment 90 % ou plus des coefficients de la table. Ceux-ci peuvent être des zéros structurels –de

nature biologique – où le taxon n'est pas présent dans le jeu de données, ou des zéros d'échantillonnage – de nature probabiliste – où le taxon n'a pas été détecté alors qu'il est présent (mais en faible quantité) dans le jeu de données. Pour modéliser cette fréquence élevée de comptages nuls, on utilise alors des lois avec excès de zéros. Si  $X$  suit une loi sur  $\mathbb{N}$ , on peut créer  $Y$  avec excès de zéros comme suit :

$$\begin{cases} \mathbb{P}(Y = 0) = p_0 + (1 - p_0)\mathbb{P}(X = 0) \\ \mathbb{P}(Y = k) = (1 - p_0)\mathbb{P}(X = k) \end{cases} \quad \text{pour } k \in \mathbb{N}^*,$$

où  $p_0$  est la proportion de zéros structurels. Si  $X$  suit une loi binomiale négative,  $Y$  suivra alors une loi binomiale négative avec excès de zéros, ou ZINB (pour *Zero-Inflated Negative Binomial*) (Xinyan, Himel, & Nengjun, 2016).

La somme des comptages des taxons dans un échantillon est appelé *profondeur de séquençage* et correspond au nombre de lectures produites par le séquenceur pour l'échantillon. La profondeur *cible* (typiquement 100 000 lectures par échantillons) est imposée par le scientifique mais la profondeur *effective* peut varier d'un facteur 4 pour des échantillons avec la même cible (50 000 lectures pour le premier, 200 000 pour le deuxième). Il est d'usage d'utiliser un facteur de normalisation pour prendre en compte ces différences et rendre les comptages comparables entre échantillons.

## Données compositionnelles

L'autre grand point de vue considère que les comptages sont contraints par la profondeur de séquençage et ne nous permettent donc d'étudier que les *abondances relatives* (par opposition aux *abondances absolues*) des différents taxons dans l'échantillon. En pratique, les comptages des  $m$  taxons de chaque échantillon sont divisés par la profondeur de séquençage de l'échantillon pour reconstruire des vecteurs d'abondance relative, à valeurs dans le simplexe  $\mathcal{S}^m = \{x \in \mathbb{R}_+^{*,m}, \sum_{i=1}^m x_i = 1\}$  (Gloor & Reid, 2016).

Ces vecteurs de proportion peuvent ensuite être modélisées comme des tirages de lois de Dirichlet. La loi de Dirichlet  $\mathcal{D}(\alpha)$  sur  $\mathcal{S}^m$  de paramètre  $\alpha = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}_+^{*,m}$  a pour densité

$$f(x_1, \dots, x_m) = \frac{1}{B(\alpha)} \prod_{i=1}^m x_i^{\alpha_i - 1},$$

où  $B : \alpha \mapsto \frac{\prod_{i=1}^m \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^m \alpha_i)}$  est la fonction beta.

Des approches plus sophistiquées, basées sur des lois hiérarchiques multinomiales Dirichlet, permettent de modéliser en plus le fait que les comptages sont

une version bruitée du vecteur de proportion (Holmes, Harris, & Quince, 2012).

$$p \sim \mathcal{D}(\alpha),$$

$$\mathbb{P}(X_1 = n_1, \dots, X_m = n_m \mid p) = \frac{n!}{\prod_{i=1}^m n_i!} \prod_{i=1}^m p_i^{n_i}.$$

La transformation en proportions préserve certaines propriétés, par exemple le rang d'un taxon au sein d'un échantillon, mais requiert une attention particulière pour d'autres opérations élémentaires, comme le calcul d'une composition moyenne. Aitchison (1982) est le premier à proposer une géométrie *compositionnelle* du simplexe, qui diffère de la géométrie euclidienne standard. Dans cette géométrie, le simplexe possède ses propres opérations internes de perturbation  $\oplus$  ou de composition  $\odot$  ainsi que sa propre distance  $d_a$  définie par

$$d_a(x, y) = \sqrt{\frac{1}{2m} \sum_{i=1}^m \sum_{j=1}^m \left( \ln \left( \frac{x_i}{x_j} \right) - \ln \left( \frac{y_i}{y_j} \right) \right)^2} \text{ pour } x, y \in \mathcal{S}^m.$$

Un point régulièrement évoqué, que la géométrie d'Aitchison corrige, est la présence de corrélations négatives fallacieuses dans le jeu de données normalisé par la somme. Pour s'en convaincre, on peut considérer 2 variables indépendantes  $X, Y$  et constater que  $p_1 = \frac{X}{X+Y}$  et  $p_2 = \frac{Y}{X+Y} = 1 - p_1$  sont corrélées négativement. Aitchison propose plusieurs transformations, collectivement appelées *xlr*, pour plonger le simplexe dans l'espace euclidien standard et pouvoir ainsi appliquer les méthodes d'analyse multivariée standards aux données compositionnelles (Pawlowsky-Glahn, Egozcue, & Tolosana Delgado, 2007).

- **Le ratio logarithmique additif** (*additive log-ratio*) (Aitchison, 1986)

$$\text{alr} : x \in \mathcal{S}^m \mapsto \left( \ln \left( \frac{x_1}{x_m} \right), \dots, \ln \left( \frac{x_{m-1}}{x_m} \right) \right) \in \mathbb{R}^{m-1}.$$

Cette transformation souffre de deux problèmes : elle ne conserve pas les distances entre le simplexe et  $\mathbb{R}^{m-1}$  et nécessite d'utiliser un taxon de référence (Albarède, 1996 ; Pawlowsky-Glahn et al., 2007).

- **Le ratio logarithmique centré** (*centered log-ratio*) (Aitchison, 1986)

$$\text{clr} : x \in \mathcal{S}^m \mapsto \left( \ln \left( \frac{x_1}{g_m(x)} \right), \dots, \ln \left( \frac{x_m}{g_m(x)} \right) \right) \in \mathbb{R}^m$$

où  $g : x \mapsto \sqrt[m]{\prod_{i=1}^m x_i}$  est la fonction de moyenne géométrique. Bien que cette transformation conserve les distances et ne nécessite plus d'utiliser un taxon de

référence, le simplexe est plongé dans un sous-espace vectoriel de dimension  $m - 1$  de  $\mathbb{R}^m$  défini par  $\{y \in \mathbb{R}^m : y^T 1_m = 0\}$ . Autrement dit, la somme des coordonnées du projeté doit être nulle.

- **Le ratio logarithmique isométrique (*isometric log-ratio*)** (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, & Barcelo-Vidal, 2003).

$$\text{ilr} : x \in S^m \mapsto \Psi^T \text{clr}(x) = (y_1, \dots, y_{m-1}) \in \mathbb{R}^{m-1}$$

où  $\Psi$  est une base orthonormée quelconque du sous-espace vectoriel  $\text{clr}(S^m) = \{y \in \mathbb{R}^m : y^T 1_m = 0\}$ . Cette transformation conserve les distances et nécessite une base adaptée. Un choix classique de base est donné par la matrice de Helmert privée de sa première ligne (et représentée ici pour  $m = 4$ )

$$\Psi^T = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} & 0 \\ 1/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} & -3/\sqrt{12} \end{pmatrix},$$

pour laquelle on obtient

$$y_{i-1} = \frac{1}{\sqrt{i(i+1)}} \ln \left( \frac{x_i}{\left( \prod_{j=1}^{i-1} x_j \right)^{\frac{1}{i-1}}} \right) \text{ pour } i \in \llbracket 2, m \rrbracket.$$

Cette base produit des coordonnées  $y_i$  facilement interprétable :  $y_i$  mesure la balance entre  $x_i$  et la moyenne géométrique de  $x_1$  à  $x_{i-1}$ .

Si l'on dispose de l'arbre phylogénétique des taxons, d'autres contrastes interprétables peuvent également être utilisés, par exemple la balance entre le sous-arbre gauche et le sous-arbre droit d'un nœud de l'arbre (Silverman, Washburne, Mukherjee, & David, 2017).

Enfin, Xia, Sun, & Chen (2018) proposent une autre base, qui donne les coordonnées suivantes.

$$y_i = \frac{1}{\sqrt{i(i+1)}} \ln \left( \frac{\prod_{j=1}^i x_j}{(x_i + 1)^i} \right) \text{ pour } i \in \llbracket 1, m-1 \rrbracket.$$

Les transformations xlr ne tolèrent pas des proportions nulles. La solution généralement adoptée consiste à ajouter un pseudo-compte de 1 (ou 1/2) à tous les taxons avant de calculer les proportions. Cette solution, si elle permet en pratique de s'abstraire des zéros lors des calculs, présente néanmoins l'inconvénient majeur d'induire des pics dans la densité des coordonnées transformées sans pour autant permettre de gérer explicitement les zéros structurels.

## 1.3 Jeux de données

Nous présentons ici les jeux de données utilisés dans ce manuscrit. Chaque jeu de données porte le nom du premier auteur de l'étude dont il est extrait. Tous sont disponibles dans le matériel supplémentaire de l'article d'origine ou dans le *package R {curatedMetagenomicData}* qui met à disposition de manière homogène des jeux de données de métagénomique (Pasolli et al., 2017; R Core Team, 2020).

### Brito

Brito et al. (2016) comparent une cohorte de 81 Nord-Américains urbains avec une cohorte de 171 Fidjiens ruraux en utilisant à la fois du séquençage 16S (que nous utiliserons) et du séquençage non ciblé. Leurs travaux ont montré que la variation du régime alimentaire se reflète dans la variation du potentiel fonctionnel des gènes du microbiote, les Fidjiens ayant par exemple plus de gènes spécialisés dans la dégradation de l'amidon. En ne gardant que les 112 adultes fidjiens de cette cohorte pour former un groupe d'échantillons homogènes, il reste 77 OTUs.

### Chaillou

L'étude de Chaillou et al. (2015) s'intéresse au microbiote alimentaire de produit carnés (bœuf haché, veau haché, merguez de volaille et dés de lardons) et de produits issus de la mer (filet de cabillaud, crevette, filet saumon et saumon fumé) pour étudier le rôle du microbiote dans l'altération de l'aliment. Les 64 échantillons, répartis uniformément entre les différents aliments, ont mis en évidence une perte de diversité microbienne concomitante à l'altération et ont permis d'identifier des espèces associées à une altération précoce. En ne conservant que les taxons ayant une prévalence supérieure à 5 % de prévalence, on conserve 499 OTUs dont 97 assignées à l'embranchement des *Bacteroidetes*.

### Chlamydiae

Seul exception à la règle de nommage, le jeu de donnée Chlamydiae constitue une sous-partie du jeu de données construit par Caporaso et al. (2011). Ce dernier contient 26 échantillons répartis au sein de 8 environnements très différents (selles, bouche, eau, sol, sédiments, océan, eau douce calme et eau douce vive) pour étudier la diversité microbienne globale et calibrer l'effort de séquençage nécessaire à une bonne caractérisation. Le sous-jeu de données est restreint aux 21 OTUs assignées à l'ordre des *Chlamydiales* et a servi d'exemple pour *{StructSSI}* dans Sankaran & Holmes (2014).

### Ravel

Le jeu de données présenté dans Ravel et al. (2011) concerne le microbiote vaginal de 396 femmes nord-américaines, n'ayant pas atteint la ménopause, issues de différents groupes ethniques et sujettes ou non à des vaginoses. Le séquençage par gène marqueur 16S a permis d'identifier cinq *archétypes* de communautés microbiennes : quatre d'entre eux sont dominés par une espèce de *Lactobacillus* qui acidifie le milieu et empêche le développement de bactéries responsables de vaginose, le dernier correspond à une diversité bactérienne élevée et est associé à des risques accrus de vaginose. 40 genres différents, de prévalence supérieure à 5 % sont présents dans ce jeu de données.

### Wu

Wu et al. (2011) se sont intéressé aux relations entre microbiote et régime alimentaire, parmi lesquels l'importance de la consommation d'alcool. Bien qu'un changement de régime alimentaire pendant une courte durée (10 jours) a un impact significatif sur le microbiote mais son ampleur reste modeste. Ce jeu de données, qui comprend 98 échantillons répartis à égalité entre sujets à faible et forte consommation d'alcool, contient 400 OTUs.

### Zeller

Le jeu de données issu de Zeller et al. (2014) contient 42 patients ayant un adénome, 91 patients ayant un cancer colorectal et 66 volontaires sains. Le but de l'étude est d'étudier les associations entre la composition du microbiote et le statut du patient, notamment pour trouver des biomarqueurs de la maladie. Tous les échantillons ont été caractérisés par l'approche gène marqueur (à l'aide du gène 16S) et par l'approche non-ciblée. En ne conservant que les taxons ayant une prévalence supérieure à 5%, l'approche gène marqueur a permis d'identifier 119 genres différents tandis que l'approche non-ciblée a identifié 878 MSPs –une autre entité métagénomique (Plaza Oñate et al., 2018).

TABLE 1.2: Jeux de données utilisés dans ce manuscrit.

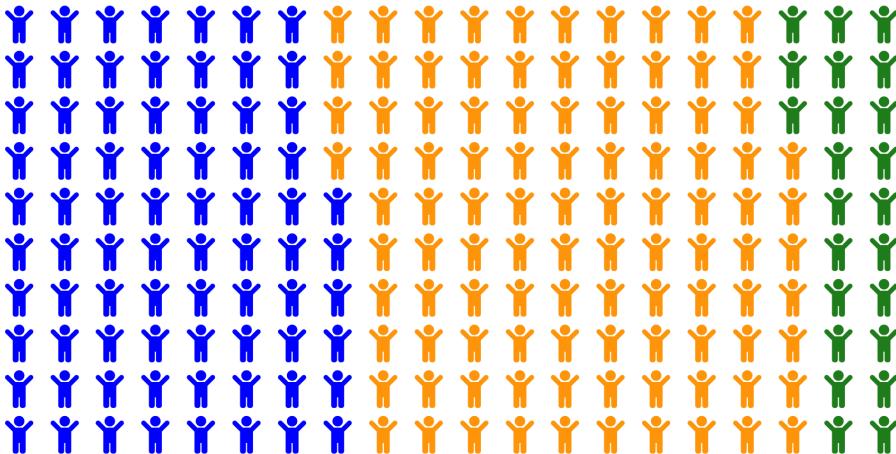
Jeu de données	Microbiote	Rang	Taxons	Échantillons
Brito et al. (2016)	Intestinal	OTU	77	112
Chaillou et al. (2015)	Alimentaire	OTU	499/97	64
Chlamydiae	Varié	OTU	21	26
(Caporaso et al., 2011)				
Ravel et al. (2011)	Vaginal	Genre	40	396
Wu et al. (2011)	Intestinal	OTU	400	98

Jeu de données	Microbiote	Rang	Taxons	Échantillons
Zeller et al. (2014)	Intestinal	Genre	119	199
Zeller et al. (2014)	Intestinal	MSP	878	199

# Chapitre 2

## Études d'analyse différentielle

Ce chapitre est un chapitre bibliographique sur les méthodes d'analyses différentes existantes.



### 2.1 Tests statistiques

#### 2.1.1 Analyse de la variance à un facteur

Une première approche pour tester s'il y a un effet du groupe sur l'abondance est d'utiliser un modèle d'ANOVA à un facteur.

Avec  $k$  groupes, on renomérote les abondances des  $N$  individus tels que  $y_{i,j}$  soit l'abondance du  $j^{\text{ème}}$  individu au sein  $i^{\text{ème}}$  groupe. Avec  $n_i$  individus dans le groupe  $i$ , la moyenne des abondances au sein du groupe est  $y_{i,\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{i,j}$  et la moyenne générale est  $y_{\bullet,\bullet} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{i,j}$ .

Sous les hypothèses de normalité et d'homoscédasticité des résidus, la statistique de test  $F$ , définie par

$$F = \frac{\sum_{i=1}^k n_i (y_{i,\bullet} - \bar{y}_{\bullet,\bullet})^2 / (k-1)}{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_{i,\bullet})^2 / (N-k)},$$

suit une loi de Fisher à  $k-1$  et  $N-k$  degrés de liberté.

$F$  correspond au ratio entre la variance des moyennes de chaque groupe et la variance totale. On rejette alors  $\mathcal{H}_0 = \{\text{tous les groupes ont la même moyenne}\}$  si  $F$  est suffisamment grande et on considère que le groupe a une influence sur l'abondance.

### 2.1.2 Test de Wilcoxon

Le test de Wilcoxon-Mann-Whitney (Mann & Whitney, 1947; Wilcoxon, 1992) permet de tester si, pour deux échantillons, la probabilité qu'une valeur tirée au hasard dans un échantillon soit plus petite qu'une valeur tiré au hasard dans le second est égale à la probabilité qu'elle soit plus grande. L'hypothèse nulle est alors  $\mathcal{H}_0 = \{\mathbb{P}(X < Y) = \mathbb{P}(Y < X)\}$  pour  $X$  et  $Y$  tirés dans chacune des populations. En pratique, on utilise ce test non-paramétrique pour vérifier si les deux populations suivent la même distribution ou non.

Formellement, soit  $X = (x_1, \dots, x_{n_1})$  et  $Y = (y_1, \dots, y_{n_2})$  les deux échantillons à comparer pour lesquels il n'y a pas d'égalité dans leurs valeurs. Il est possible de définir sans ambiguïté le rang de chaque individu au sein de l'échantillon concaténé de taille  $n_1 + n_2$ , puis  $R_1$  la somme des rangs des individus du premier groupe. Sous  $\mathcal{H}_0$ , la statistique de test

$$U = R_1 - \frac{n_1(n_1+1)}{2}$$

suit asymptotiquement une loi normale de moyenne  $\frac{n_1 n_2}{2}$  et de variance  $\frac{n_1 n_2 (n_1+n_2+1)}{12}$ . On rejette  $\mathcal{H}_0$  lorsque  $|U| > \Phi^{-1}(1-\alpha)$ . Il est possible de généraliser lorsqu'il y a des égalités dans les valeurs au sein d'un échantillon ou entre les échantillons. Dans ce cas là, la variance de la loi asymptotique doit être ajustée en conséquence.

### 2.1.3 Test de Kruskall-Wallis

Le test de Kruskall-Wallis (Kruskal & Wallis, 1952) est une généralisation à  $k$  groupes du test de Wilcoxon-Mann-Whitney. Il est utilisé pour déterminer si les  $k$  échantillons proviennent d'une même population ou si au moins un des échantillons a une distribution différente des autres.

Comme pour Wilcoxon-Mann-Whitney, on commence par concaténer les  $N = \sum_{i=1}^k n_i$  observations des  $k$  échantillons puis calculer la somme des rangs au sein de chaque groupe :  $R_1, \dots, R_k$ . Si les observations sont indépendantes et qu'il n'y a pas d'égalité dans leurs valeurs, la statistique de test

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) \quad (2.1)$$

suit une loi du  $\chi^2$  à  $k - 1$  degrés de libertés.  $H$  peut être réécrite comme

$$\frac{N-1}{N} \frac{\sum_{i=1}^k n_i \left( \frac{R_i}{n_i} - \frac{N+1}{2} \right)^2}{\frac{N^2-1}{12}},$$

qui, à un facteur multiplicatif près, correspond à la statistique de test d'une ANOVA à un facteur sur les rangs et où la variance au dénominateur n'a pas besoin d'être estimée (il s'agit de la variance de la loi uniforme sur  $\llbracket 1, N \rrbracket$ ).

Lorsque des égalités sont présentes au sein des échantillons,  $H$  est calculée en divisant l'expression (2.1) par

$$1 - \frac{\sum_{i=1}^k t_i^3 - t_i}{N^3 - N},$$

où  $t_i$  est le nombre d'égalités au sein du  $i^{\text{ème}}$  échantillon. Après cette correction, si les  $n_i$  ne sont pas trop petits, cette nouvelle statistique  $H$  corrigée suit toujours une loi qui  $\chi^2$  à  $k - 1$  degrés de liberté.

On rejette l'hypothèse nulle sur l'égalité des rangs moyens  $\mathcal{H}_0 = \left\{ \frac{R_1}{n_1} = \dots = \frac{R_k}{n_k} \right\}$  lorsque  $H$  est plus grande que la valeur seuil déterminée par  $\alpha$ .

#### 2.1.4 Autres méthodes

Les données métagénomiques ayant les spécificités précédemment présentées dans la section 1.2.4 (non-normalité, surabondance de zéros, somme contrainte...), des méthodes spécifiques ont été développées pour les appréhender. Nous pouvons par exemple citer *edgeR* (Robinson, McCarthy, & Smyth, 2010), *DESeq2* (Love, Huber, & Anders, 2014), *metagenomeSeq* (Paulson, Stine, Bravo, & Pop, 2013) ou *mbzinb* (Chen et al., 2018). Celles-ci peuvent utiliser des modèles linéaires généralisés ou des modèles avec excès de zéros pour essayer d'appréhender au mieux les données.

Nous n'avons pas testé ces méthodes dans nos analyses. Nous nous intéressons à l'apport de l'information hiérarchique et avons préféré utiliser le même test classique (Wilcoxon ou Kruskall-Wallis) dans toutes nos comparaison. La nouvelle procédure que nous proposons dans le chapitre 4 peut cependant être appliquée sur les sorties de méthodes quelconques, pourvu qu'elles fournissent un vecteur de  $p$ -valeurs.

### ***edgeR* et *DESeq2***

Proposés initialement pour de l'analyse de données de transcriptomique, *edgeR* (Robinson et al., 2010) et *DESeq2* (Love et al., 2014) ajustent un modèle de régression linéaire généralisé avec une structure d'erreur binomiale négative, afin de prendre en compte la surdispersion des données.

Le compte du taxon  $i$  dans l'échantillon  $j$  appartenant au groupe  $g$  est modélisé par

$$Y_{i,j} \sim \text{NB}(s_j \mu_{i,g}, \alpha_{i,g})$$

où  $s_j$  est un facteur de normalisation qui modélise le nombre total de lectures dans l'échantillon  $j$ ,  $\mu_{i,g}$  est l'abondance du taxon  $i$  dans le groupe  $g$  et  $\alpha_{i,g}$  sa dispersion. La comparaison entre les groupes  $g$  et  $g'$  se formule alors  $\mathcal{H}_0 : \{\mu_{i,g} = \mu_{i,g'}\}$  contre  $\mathcal{H}_1 : \{\mu_{i,g} \neq \mu_{i,g'}\}$ .

### ***mbzinb***

*mbzinb* (Chen et al., 2018) modélise également les comptes de lectures. En notant de plus  $p_{i,g}$  la proportion de zéros pour le taxon  $i$  dans les échantillons du groupe  $g$  (et donc  $1 - p_{i,g}$  sa prévalence), on a un nouveau modèle avec excès de zéros :

$$Y_{i,j} \sim p_{i,g} \delta_0 + (1 - p_{i,g}) \text{NB}(s_j \mu_{i,g}, \alpha_{i,g}),$$

où  $\delta_0$  est la masse de Dirac en 0.

### ***ALDEx2***

Contrairement aux méthodes précédentes, *ALDEx2* (Fernandes et al., 2014) considère les données métagénomiques comme des données compositionnelles. Ce test commence par normaliser les données pour les plonger dans le simplexe, puis ajuste une distribution de Dirichlet sur celles-ci. Elle génère alors des réalisations conformément à la loi de Dirichlet apprise via une méthode de Monte-Carlo afin d'augmenter artificiellement la taille du jeu de données. Enfin, les jeux de données artificiels sont projetés dans  $\mathbb{R}^p$  via la transformation `clr` avant d'effectuer des tests d'abondance différentiels classiques pour obtenir des  $p$ -valeurs.

## 2.2 Problématique des tests multiples

À chaque fois qu'un test est réalisé, il y a une probabilité  $\alpha$  que celui-ci rejette à tort une hypothèse nulle. Si  $m$  tests indépendants sont réalisés, la probabilité qu'au moins une hypothèse rejetée le soit à tort est alors de  $1 - (1 - \alpha)^m$ , qui dépasse 0.9 pour  $m = 50$  et  $\alpha = 0.05$ . En métagénomique, il est courant d'effectuer plusieurs

centaines de tests simultanés, et la nécessité de corriger pour la multiplicité des tests est d'autant plus importante.

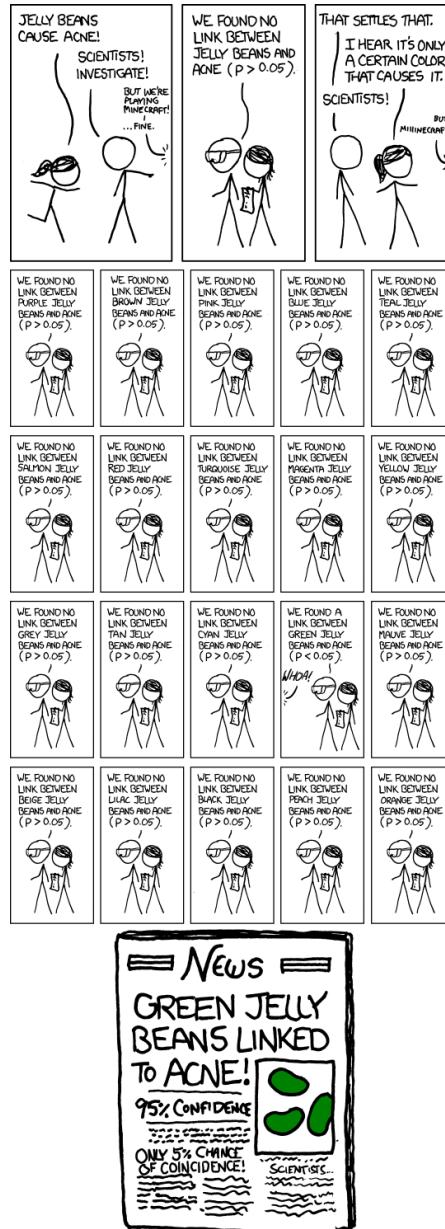


FIGURE 2.1: Dessin humoristique illustrant la problématique des tests multiples, par XKCD.

### 2.2.1 Évaluation des performances

Lorsqu'une série de tests est réalisée et que la vérité est connue, il est possible de comparer le résultat avec l'attendu et de compter les effectifs dans les quatre configurations possibles.

Lorsqu'un test rejette l'hypothèse nulle à raison, on parlera de vrai positif et s'il la rejette à tort, il s'agira d'un faux positif. Si un test ne parvient pas à rejeter l'hypothèse nulle alors que celle-ci est vraie, il s'agit d'un vrai négatif alors que s'il aurait dû la rejeter, il s'agit d'un faux négatif. Sur l'ensemble des tests, le nombre de vrais positifs, vrais négatifs, faux positifs et faux négatifs sont notés respectivement TP, TN, FP et FN. Ces quantités sont agrégées dans la *matrice de confusion*, présentée dans la figure 2.2.

		Vraie condition	
		Condition positive	Condition négative
Détection	Déetecté positif	TP	FP
	DéTECTÉ negatif	FN	TN

FIGURE 2.2: Matrice de confusion. TP est le nombre de vrais positifs, TN le nombre de vrais négatifs, FP le nombre de faux positifs et FN le nombre de faux négatifs.

À partir de ces mesures primaires, on peut définir le taux de vrais positifs (*true positive rate*), appelé également puissance, sensibilité ou rappel :

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

qui représente la proportion d'espèces vraiment différentiellement abondantes que la méthode arrive à détecter. C'est une quantité que l'on souhaite maximiser. Cependant, un test qui détecterait toutes les espèces comme différentiellement abondantes aurait un TPR au maximum, à 1. Il faut donc aussi pouvoir contrôler que l'on ne rejette pas trop à tort, en contrôlant soit directement le nombre de faux positifs FP, soit la proportion de faux positifs parmi les positifs (*false discovery proportion*) :

$$\text{FDP} = \frac{\text{FP}}{\text{TP} + \text{FP}}.$$

Le FDP est lié à la précision PPV (*positive predictive value*) via  $\text{PPV} = 1 - \text{FDP}$ . Celle-ci mesure la crédence que l'on peut accorder au test sachant qu'il a rejeté l'hypothèse nulle.

Une fois que l'on maximise la puissance d'un test tout en contrôlant le nombre ou la proportion de fausses découvertes, il est intéressant d'avoir des métriques qui combinent ces métriques secondaires, comme l'exactitude (*accuracy*), qui est la proportion de fois où le test a correctement assigné les bactéries :

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

ou du score  $F_1$  qui est la moyenne harmonique entre la précision et le rappel :

$$F_1 = \frac{2}{\frac{1}{\text{TPR}} + \frac{1}{\text{PPV}}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}.$$

Ces deux quantités sont à valeurs dans  $[0, 1]$  et plus elles sont proches de 1, plus le modèle est performant.

Les mesures présentées auparavant sont valables quelque soit le seuil à partir duquel on rejette l'hypothèse nulle, couramment égal à 0.05. Il peut alors être intéressant de regarder ces quantités comme une fonction du seuil de rejet  $t$  :  $\text{TPR}(t)$ ,  $\text{FDP}(t)$ ,  $\text{ACC}(t)\dots$

On définit alors la courbe ROC (*receiver operating characteristic*) comme étant la sensibilité en fonction de la spécificité pour les différents seuil, soit l'ensemble des points  $(\text{FDP}(t), \text{TPR}(t))_{t \in [0,1]}$ .

Lorsque  $t = 0$ , on ne rejette jamais, il n'y a ni faux positif ni vrai positif et on est en  $(0, 0)$ . Lorsque  $t = 1$ , on rejette tout le temps,  $\text{FN} = \text{TN} = 0$  et on est en  $(1, 1)$ . FDP et TPR sont des fonctions croissantes en  $t$ . En  $(0, 1)$ , il n'y a ni faux négatif ni faux positif et le test est toujours correct. La courbe ROC relie donc  $(0, 0)$  à  $(1, 1)$  et plus elle s'approche de  $(0, 1)$ , mieux c'est. L'aire sous la courbe ROC, dite AUC pour *area under the curve*, permet de quantifier la qualité d'un classificateur indépendamment du seuil choisi. Plus l'AUC est proche de 1, plus le classificateur est performant.

### 2.2.2 Correction de Bonferroni

La correction de Bonferroni a pour objectif de contrôler le *family-wise error rate*, la probabilité de faire au moins une erreur dans les découvertes en changeant le seuil auquel on rejette l'hypothèse nulle. Si au lieu de rejeter en dessous du seuil

$\alpha$ , on rejette en dessous de  $\frac{\alpha}{m}$ , on contrôle le FWER au niveau  $\alpha$  :

$$\begin{aligned} \text{FWER} &= \mathbb{P}(\text{FP} \geq 1) \\ &= \mathbb{P}\left(\bigcup_{i \in \mathbb{H}_0} \left\{ \mathbf{p}_i \leq \frac{\alpha}{m} \right\}\right) \\ &\leq \sum_{i \in \mathbb{H}_0} \mathbb{P}\left(\left\{ \mathbf{p}_i \leq \frac{\alpha}{m} \right\}\right) \\ &= m_0 \frac{\alpha}{m} \\ &\leq \alpha. \end{aligned}$$

Pour plus de praticité, on travaille avec les  $q$ -valeurs associées à la correction de Bonferroni définies par

$$\mathbf{q}^{\text{bonf}} = m * \mathbf{p},$$

qui sont, elles, comparées au seuil  $\alpha$ .

Ainsi, avec une probabilité de  $1 - \alpha$ , il n'y a aucun faux positif pour l'ensemble des tests réalisés.

Cependant, les procédures qui contrôlent le FWER comme les corrections de Bonferroni ou de Holm, une alternative plus puissante (Holm, 1979), sont très conservatrices et peu utilisées en métagénomique.

### 2.2.3 Correction de Benjamini-Hochberg

Au lieu de n'autoriser aucun faux positif à un risque  $\alpha$ , Benjamini & Hochberg (1995) proposent de contrôler la proportion de fausses découvertes parmi les découvertes. Il s'agit alors de garder la quantité  $\text{FDP} = \frac{\text{FP}}{\text{R} \vee 1}$  en dessous d'un seuil  $\alpha$ , où  $\text{R} = \text{TP} + \text{FP}$  est le nombre d'hypothèses rejetées.

Sous réserve d'une indépendance entre les tests, la procédure de Benjamini-Hochberg (BH) contrôle l'espérance du FDP, appelée FDR (pour *false discovery rate*), au seuil  $\alpha$ . Celle-ci se fait en trois étapes :

- ordonner les  $p$ -valeurs  $\mathbf{p}_{(1)}, \dots, \mathbf{p}_{(m)}$  et poser  $\mathbf{p}_{(0)} = 0$ ;
- trouver le rang  $\hat{\ell} = \max \left\{ \ell \in \llbracket 0, m \rrbracket \mid \mathbf{p}_{(\ell)} \leq \frac{\alpha \ell}{m} \right\}$ ;
- rejeter les  $\hat{\ell}$  hypothèses correspondant aux plus petites  $p$ -valeurs.

En effet, notons  $V_i = \mathbb{1}_{\{H_i \text{ est rejetée}\}}$  pour  $i \in \mathbb{H}_0$ . On a alors  $\text{FDP} = \frac{\text{FP}}{\text{R} \vee 1} = \sum_{i \in \mathbb{H}_0} \frac{V_i}{\text{R} \vee 1}$ .

Fixons  $i \in \mathbb{H}_0 \neq \emptyset$ . S'il y a  $k$  hypothèses rejetées, alors  $H_i$  est rejetée si et seulement si  $\mathbf{p}_i \leq \frac{\alpha k}{m}$  et alors  $V_i = \mathbb{1}_{\{\mathbf{p}_i \leq \frac{\alpha k}{m}\}}$ . Définissons  $\text{R}(\mathbf{p}_i \rightarrow 0)$  le nombre

d'hypothèses rejetées lorsqu'on fixe  $\mathbf{p}_i$  à 0. S'il y a  $k$  hypothèses rejetées et que  $\mathbf{p}_i \leq \frac{\alpha k}{m}$ ,  $H_i$  est rejetée et fixer  $\mathbf{p}_i$  à 0 ne change pas le nombre d'hypothèses rejetées :  $R = R(\mathbf{p}_i \rightarrow 0)$ . À l'inverse, si  $\mathbf{p}_i > \frac{\alpha k}{m}$ ,  $H_i$  n'est pas rejetée et  $V_i = 0$ . La concaténation des deux résultats précédents donne  $V_i \mathbb{1}_{\{R=k\}} = V_i \mathbb{1}_{\{R(\mathbf{p}_i \rightarrow 0)=k\}}$ .

En notant  $\mathcal{F}_i = \{\mathbf{p}_1, \dots, \mathbf{p}_{i-1}, \mathbf{p}_{i+1}, \dots, \mathbf{p}_m\}$  la tribu engendrée par les  $p$ -valeurs sauf  $\mathbf{p}_i$ ,

$$\begin{aligned} \mathbb{E} \left[ \frac{V_i}{R \vee 1} \middle| \mathcal{F}_i \right] &= \mathbb{E} \left[ \sum_{k=1}^m \frac{V_i \mathbb{1}_{\{R=k\}}}{k} + \frac{0}{0 \vee 1} \middle| \mathcal{F}_i \right] \\ &= \sum_{k=1}^m \mathbb{E} \left[ \frac{V_i \mathbb{1}_{\{R(\mathbf{p}_i \rightarrow 0)=k\}}}{k} \middle| \mathcal{F}_i \right] \\ &= \sum_{k=1}^m \mathbb{1}_{\{R(\mathbf{p}_i \rightarrow 0)=k\}} \mathbb{E} \left[ \frac{\mathbb{1}_{\{\mathbf{p}_i \leq \frac{\alpha k}{m}\}}}{k} \middle| \mathcal{F}_i \right] \\ &= \sum_{k=1}^m \mathbb{1}_{\{R(\mathbf{p}_i \rightarrow 0)=k\}} \frac{\alpha}{m} \\ &= \frac{\alpha}{m}. \end{aligned}$$

La troisième égalité provient du fait que  $R(\mathbf{p}_i \rightarrow 0)$  est connue conditionnellement à  $\mathcal{F}_i$ , la quatrième découle du fait que  $\mathbf{p}_i \sim \mathcal{U}([0, 1])$  et la dernière résulte du fait qu'en fixant une  $p$ -valeur à 0, on va rejeter au moins une fois et  $R(\mathbf{p}_i \rightarrow 0)$  est compris entre 1 et  $m$ .

Finalement,

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{E} \left[ \sum_{i \in \mathbb{H}_0} \frac{V_i}{R \vee 1} \right] = \sum_{i \in \mathbb{H}_0} \frac{\alpha}{m} \leq \alpha.$$

En terme de  $q$ -valeurs,

$$q_{(i)}^{\text{bh}} = \min \left\{ \min_{j \geq i} \left\{ \frac{m \mathbf{p}_{(j)}}{j} \right\}, 1 \right\},$$

où  $\mathbf{p}_{(1)}, \dots, \mathbf{p}_{(m)}$  sont les  $p$ -valeurs réordonnées et  $q_{(i)}^{\text{bh}}$  est la  $q$ -valeur associée à  $\mathbf{p}_{(i)}$ .

#### 2.2.4 Correction de Benjamini-Yekutieli

Si les tests ne sont pas indépendants, il est possible d'appliquer la procédure de Benjamini-Yekutieli (BY) (Benjamini & Yekutieli, 2001) qui ne requiert aucune hypothèse d'indépendance entre les tests. Il s'agit d'une modification dans la procédure de Benjamini-Hochberg, où l'on ne compare plus à  $\frac{\alpha \ell}{m}$  mais à  $\frac{\alpha \ell}{m \sum_{i=1}^m \frac{1}{\ell}}$ .

Les  $q$ -valeurs sont alors

$$\mathfrak{q}_{(i)}^{\text{by}} = \min \left\{ \min_{j \geq i} \left\{ \sum_{i=1}^m \frac{1}{i} \frac{m \mathfrak{p}_{(j)}}{j} \right\}, 1 \right\}.$$

Cette procédure très générique présente l'inconvénient d'être extrêmement conservatrice et de faire peu de découvertes.

## 2.3 Procédures hiérarchiques pour tests multiples

Jusqu'alors, les corrections proposées font l'hypothèse (*a priori* fausse) que les tests sont indépendants ou vérifient une hypothèse technique de dépendance positive, comme BH, ou fonctionnent quelque soit la relation de dépendance entre les tests, au prix d'un fort conservatisme, comme BY.

Il serait plus intéressant d'utiliser explicitement la relation de dépendance entre les tests afin d'augmenter la puissance statistique. C'est ce que proposent les méthodes présentées dans cette section.

Cette section ne nécessite qu'une définition intuitive de ce qu'est un arbre : un graphe acyclique connexe orienté dont les noeuds terminaux (feuilles) sont étiquetés et ayant ou non des longueurs de branches. Une définition plus rigoureuse en sera donnée dans le chapitre 3.

### 2.3.1 *TreeFDR*

*TreeFDR* (Xiao, Cao, & Chen, 2017) est une procédure de lissage des  $z$ -scores suivie d'une procédure de correction par permutation implémentée dans le package `{StructFDR}`.

Dans ce modèle hiérarchique, les  $z$ -scores  $\mathfrak{z} = \Phi^{-1}(\mathfrak{p})$  sont vus comme la réalisation un vecteur gaussien multivarié de moyenne  $\mu$  :

$$\mathfrak{z} | \mu \sim \mathcal{N}_m(\mu, \sigma^2 \mathbf{I}_m).$$

À partir de l'arbre, on calcule la matrice des distances patristiques entre feuilles  $D = (d_{i,j})_{i,j}$  que l'on converti en une matrice de corrélation  $C_\rho = (\exp(-2\rho D_{i,j}))_{i,j}$ . Cette matrice de corrélation est ensuite utilisée pour décrire les corrélations entre les composantes de  $\mu$  :

$$\mu \sim \mathcal{N}_m(\gamma \mathbf{1}_m, \tau^2 C_\rho).$$

L'estimateur du maximum *a posteriori* de  $\mu$  est alors

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

avec  $k = \frac{\sigma}{\tau}$ .

Les hyperparamètres  $\rho$  et  $k$  contrôlent le niveau de lissage du modèle : les  $z$ -scores issus d'un même clade vont être regroupés vers une valeur commune. De hautes valeurs de  $k$  ou de faibles valeurs de  $\rho$  entraînent un fort lissage.

### 2.3.2 FDR hiérarchique

Le FDR hiérarchique (hFDR) est une procédure proposée par Yekutieli (2008) et implémentée dans le package R `{structSSI}` (Sankaran & Holmes, 2014).

Contrairement à *TreeFDR*, le FDR hiérarchique a besoin d'avoir une  $p$ -valeur à chaque noeud interne. Dans le cas de données métagénomiques, celles-ci peuvent facilement être obtenues en effectuant un test d'abondance différentielle sur la somme des bactéries qui descendent du noeud considéré.

C'est un algorithme descendant qui teste les hypothèses par *familles* –c'est à dire tous les enfants d'un même noeud– au niveau  $\alpha$  en corrigeant avec BH à chaque fois. Plus précisément, on commence par tester la famille de la racine (avec une correction de BH). Puis, pour chaque noeud rejeté au sein de cette famille, on va tester ses enfants (en corrigeant toujours avec BH). À l'inverse, si un noeud n'est pas rejeté, aucun de ses enfants directs et de ses descendants n'est testé. L'algorithme se termine une fois arrivé aux feuilles ou lorsqu'il ne reste que des noeuds qui n'ont pas pu être rejettés. La figure 2.3 illustre cette procédure sur un arbre à 6 feuilles.

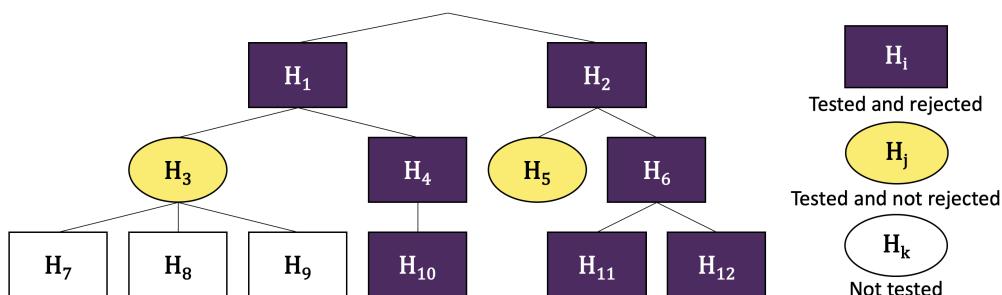


FIGURE 2.3: Exemple d'une procédure de FDR hiérarchique. Les hypothèses à tester sont notées  $H_1$  à  $H_{12}$ . L'algorithme commence par tester et rejeter (après correction)  $H_1$  et  $H_2$ . Puis il teste la famille ( $H_3, H_4$ ), car ce sont des enfants de  $H_1$ , et rejette  $H_4$  mais pas  $H_3$ . La famille ( $H_7, H_8, H_9$ ) n'est pas testée car  $H_3$  n'a pas été rejeté.  $H_{10}$  est testé et rejeté. L'algorithme procède de même sur les descendants de  $H_2$ . En définitive, il y a trois découvertes aux feuilles ( $H_{10}, H_{11}$  et  $H_{12}$ ) pour 5 familles testées. Le FDR a *posteriori* pour les feuilles est alors de  $1.44 \times \alpha \times 2$ .

Alors que  $\alpha$  est le niveau de contrôle *a priori* intra-familles, si l'on ne considère que les découvertes au niveau des feuilles, hFDR garanti un contrôle *a posteriori* au niveau

$$\alpha' = 1.44 \times \alpha \times \frac{\#\text{découvertes} + \#\text{familles testées}}{\#\text{découvertes} + 1}.$$

Cette méthode souffre de plusieurs problèmes majeurs. Tout d'abord, le contrôle du FDR est fait seulement *a posteriori*, et plusieurs essais sont nécessaires pour contrôler le FDR au niveau souhaité, sans avoir la certitude que cela soit possible. De plus, pour que la procédure descende jusqu'aux feuilles, il faut que le signal d'abondance différentielle soit détectable dès la famille de la racine, sinon la procédure s'arrête dès le début, et cela arrive souvent en pratique (Huang et al., 2020).

Enfin, dans son implémentation par Sankaran & Holmes (2014), les  $p$ -valeurs d'entrées ne sont pas données en argument par l'utilisateur mais calculées au sein la procédure via une ANOVA à un facteur. Ce type de test n'est malheureusement pas adapté aux données métagénomiques, ce qui fait perdre de la puissance à la procédure (Bichat, Plassais, Ambroise, & Mariadassou, 2020). Pour ces raisons, la procédure hFDR ne sera pas comparée aux autres méthodes dans la section 4.

### 2.3.3 *treeclimbR*

Huang et al. (2020) ont récemment proposé *treeclimbR*, une procédure ascendante qui permet de sélectionner directement des clades différentiellement abondants.

Tout d'abord, pour chaque nœud interne, on agrège les abondances en les sommant sur ses descendants. Puis, pour chaque noeud  $i$  (interne ou feuille), on effectue un test qui renvoie une  $p$ -valeur  $\mathbf{p}_i$  ainsi que le signe  $\mathbf{s}_i \in \{-1, 1\}$  associé à la direction du changement d'abondance.

On calcule ensuite un score à chaque nœud défini par

$$U_i(t) = \left| \frac{\sum_{k \in \text{desc}(i)} \mathbf{s}_k \mathbb{1}_{\{\mathbf{p}_k < t\}}}{\#\text{desc}(i)} \right|$$

où  $t \in [0, 1]$  est un hyperparamètre que l'on va estimer dans la suite et  $\text{desc}(i)$  est l'ensemble des descendants de  $i$ . Lorsque  $U_i(t)$  est proche de 1, cela signifie que les descendants de  $i$  sont différentiellement abondants et dans le même sens. À l'inverse, quand  $U_i(t)$  s'approche de 0, soit les espèces ne sont pas différentiellement abondantes, soit elles le sont dans des directions opposées.

Pour chaque  $t$  candidat sur une grille, on parcours l'arbre depuis la racine et on arrête la descente d'une branche lorsque l'on rencontre un nœud  $i$  tel que

$U_i(t) = 1$  : il s'agit d'un nœud terminal pour la procédure, qui représente tous ses descendants.

On effectue ensuite une correction pour test multiple (Benjamini-Hochberg) sur les nœuds terminaux et les feuilles qui ne font pas partie de la descendance d'un nœud terminal.

Il reste à sélectionner le meilleur  $t$  candidat. On utilise trois critères éliminatoires. Le premier est un choix *a posteriori* d'une borne maximale  $t_{\max}$  pour  $t$  (dépendante des résultats) et on exclut les candidats tels que  $t \in [0, t_{\max}]$ , ceci permettant de contrôler le FDR au niveau des feuilles parmi les candidats restants. Puis on ne conserve que les candidats ayant le plus grand nombre de feuille rejetées, ce qui maximise la puissance. Et enfin, parmi les candidats restants, on ne garde que ceux qui minimisent le nombre de nœuds rejetés, pour retenir le niveau de résolution le plus adapté.



# Chapitre 3

## Arbres

### 3.1 Définitions

Un arbre est un graphe connexe acyclique. Les noeuds de degré 1 sont appelés feuilles, par oppositions aux noeuds internes.

Un arbre enraciné est un graphe connexe acyclique dirigé possédant une unique racine. Chaque noeud  $i$ , à l'exception de la racine, a un unique parent noté  $\text{pa}(i)$ . Les feuilles sont les noeuds n'ayant pas de fils. Il est possible d'enraciner un arbre non-raciné en choisissant un noeud quelconque et en orientant les arêtes de la racine vers les feuilles. La branche menant au noeud  $i$  est dénotée  $b_i$  et a pour longueur  $\ell_i$ . On parle de topologie d'un arbre lorsqu'on fait abstraction des ses longueurs de branche.

On note  $\text{mrca}(i, j)$  l'ancêtre commun le plus récent (*most recent common ancestor*) aux noeuds  $i$  et  $j$  :  $\text{mrca}(i, j) = \text{pa}^k(i)$  avec  $k = \text{argmin}_{p \in \mathbb{N}} \{\exists q \in \mathbb{N} : \text{pa}^p(i) = \text{pa}^q(j)\}$ , et  $\text{desc}(i)$  les descendants de  $i$ , soit l'ensemble des noeuds ayant  $i$  pour ancêtre,  $i$  inclus :  $\text{desc}(i) = \{j : \exists k \in \mathbb{N} : \text{pa}^k(j) = i\}$ .

Le noeud  $i$  est à distance  $t_i$  de la racine et on note  $t_{i,j} = t_{\text{mrca}(i,j)}$ .

La distance, dite patristique, entre les noeuds  $i$  et  $j$  est notée  $d_{i,j} = t_i + t_j - 2t_{i,j}$ .

$\tilde{\ell}_{i,j} = t_i - t_{i,j}$  est la distance entre le noeud  $i$  et l'ancêtre commun le plus récent à  $i$  et  $j$ .

Un arbre enraciné est **binnaire** si tous ses noeuds internes ont exactement deux fils. En particulier, un arbre binaire à  $m$  feuilles possède  $n = 2m - 2$  branches.

Un noeud est **polytomique** s'il a plus de deux fils.

Un arbre est **ultramétrique** s'il est raciné et si toutes les feuilles sont à la même distance de la racine.

Pour un arbre enraciné, la **matrice d'incidence**  $T = (t_{i,j}) \in \{0, 1\}^{m \times n}$  encode les relations de descendance entre noeuds au sein de l'arbre. Pour une feuille  $i$  et un noeud quelconque  $j$ , le coefficient  $t_{i,j}$  vaut 1 si  $i \in \text{desc}(j)$ . En particulier, la

colonne  $t_j$  est l'indicatrice des feuilles issues du noeud  $j$  tandis que la ligne  $t_i$  est l'indicatrice des ancêtres de la feuille  $i$ .

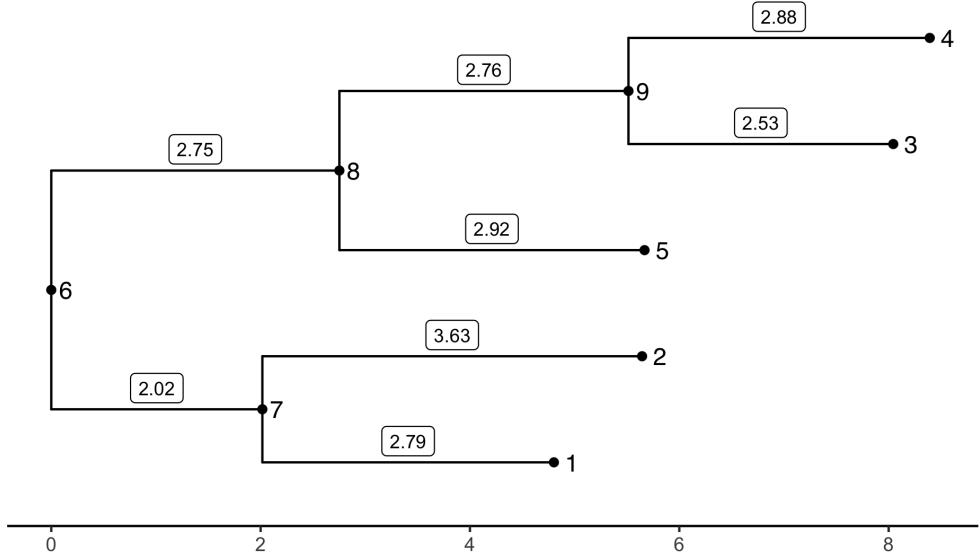


FIGURE 3.1: Exemple d'un arbre enraciné binaire non ultramétrique à  $m = 5$  feuilles et  $n = 8$  branches.

Illustrons les notations de cette section à l'aide de la figure 3.1. On a alors  $\text{pa}(1) = \text{pa}(2) = 7$ ,  $\text{desc}(8) = \{3, 4, 5, 8, 9\}$ ,  $\text{mrca}(5, 3) = 8$ ,  $\ell_1 = 2.79$ ,  $t_{3,4} = t_9 = \ell_8 + \ell_9 = 5.51$ ,  $t_{1,2} = t_7 = 2.02$ ,  $d_{3,5} = t_3 + t_5 - t_{3,5} = \ell_5 + \ell_9 + \ell_3 = 8.21$ ,  $\ell_{3,5} = t_3 - t_{3,5} = \ell_9 + \ell_3 = 5.29$  et

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}.$$

## 3.2 Distances

Les arbres sont des objets compliqués et plusieurs distances entre arbres ont été proposées dans la littérature. Nous en décrivons brièvement trois, qui mettent l'accent sur des aspects différents de l'arbre : topologie, longueurs de branche, combinaison des deux.

### 3.2.1 Distance de Robinson-Foulds

La distance de Robinson-Foulds (Robinson & Foulds, 1981) entre deux arbres racinés  $T$  et  $T'$  correspond intuitivement au nombre de branches spécifiques à un des deux arbres. Formellement, chaque nœud interne  $j$  définit une bipartition des feuilles et peut être associé à l'ensemble  $C_j \subset \llbracket 1, m \rrbracket$  des feuilles issues du nœud  $j$ . La colonne  $j$  de la matrice d'incidence précédemment introduite correspond en particulier à un codage binaire de  $C_j$ . L'arbre  $T$  est ensuite recodé comme l'ensemble  $\mathcal{C}^T = \{C_j : j \in \text{nœud interne de } T\}$  de ses bipartitions. La distance de Robinson-Foulds entre  $T$  et  $T'$  est définie comme le cardinal de la différence symétrique entre  $\mathcal{C}^T$  et  $\mathcal{C}^{T'}$  :

$$d(T, T') = |\mathcal{C}^T \setminus \mathcal{C}^{T'}| + |\mathcal{C}^{T'} \setminus \mathcal{C}^T|.$$

### 3.2.2 Distance cophénétique

La distance cophénétique entre deux arbres enracinés, introduite par Sokal & Rohlf (1962), est basée sur les matrice des distances patristiques  $D^T = (d_{i,j}^T)_{i,j}$  entre paires de feuilles dans l'arbre  $T$ , en particulier sur la vectorisation  $v^T$  de la partie triangulaire supérieure de  $D^T$ , de longueur  $\frac{m(m-1)}{2}$ . La distance cophénétique  $d(T, T')$  entre arbres  $T$  et  $T'$  est définie comme le complémentaire du coefficient de corrélation de Pearson entre  $v^{T_1}$  et  $v^{T_2}$  :

$$d(T, T') = 1 - \frac{\frac{2 \sum_{i < j} (d_{i,j}^T - \bar{d}^T)(d_{i,j}^{T'} - \bar{d}^{T'})}{m(m-1)}}{\sqrt{\frac{2 \sum_{i < j} (d_{i,j}^T - \bar{d}^T)^2}{m(m-1)}} \sqrt{\frac{2 \sum_{i < j} (d_{i,j}^{T'} - \bar{d}^{T'})^2}{m(m-1)}}} = 1 - \text{cor}(v^T, v^{T'})$$

où  $\bar{d}^T = \frac{2}{m(m-1)} \sum_{1 \leq i < j \leq m} d_{i,j}^T$ . Cette distance ne prend en compte que la topologie de l'arbre.

### 3.2.3 Distance de Billera-Holmes-Vogtmann (BHV)

La distance BHV (Billera, Holmes, & Vogtmann, 2001) est construite en plongeant l'espace  $\mathcal{T}_m$  des arbres à  $m$  feuilles dans  $\mathbb{R}^m \times \mathbb{R}^{2^m-m-1}$  puis en considérant la distance des plus courts chemins dans le sous-espace induit.

Formellement, un arbre binaire est définie par (i) sa topologie et (ii) ses longueurs de branches. Il existe  $m$  branches terminales (menant à une feuille) et  $2^m - m - 1$  branches internes distinctes, caractérisées chacune par l'ensemble des feuilles issues de cette feuille. Toutes ces branches ont une longueur positive. L'espace  $\mathcal{T}_m$  est donc naturellement plongé dans  $\mathbb{R}_+^m \times \mathbb{R}_+^{2^m-m-1}$ . Le produit cartésien

distingue les longueurs des branches externes (celles qui mènent aux feuilles, partagées par tous les arbres) de celle des branches internes, caractéristique d'une topologie donnée. Mais il n'en constitue qu'une sous-variété. En effet, à topologie  $T$  fixée, chaque longueur de branche peut varier dans  $\mathbb{R}_+$ . L'ensemble des arbres binaires de topologie  $T$ , branches terminales omises, est donc identifiable à l'orthant ouvert  $\mathbb{R}_+^{*,m-2}$ . Si on fait tendre la longueur d'une branche interne vers 0, on se ramène à une topologie non-binaire  $\tilde{T}$ , dans laquelle un unique nœud a exactement trois fils. Cette topologie peut également s'obtenir par contraction de branche interne à partir de deux autres topologies binaires  $T'$  et  $T''$ . La topologie non-binaire  $\tilde{T}$  correspond à une *frontière commune*, un sous-orthant de dimension  $m-3$  inclus dans chacun des trois orthants de dimension  $m-2$  caractérisant  $T$ ,  $T'$  et  $T''$ . De façon générale, les sous-orthants de dimension  $m-2-k$  peuvent être identifiés à des topologies dégénérées, obtenues en contractant  $k$  branches internes d'un arbre binaire.

Il existe  $(2m-3)!! = (2m-3) \times (2m-5) \times \cdots \times 5 \times 3 \times 1 = \prod_{i=0}^{m-2} 2i + 1$  topologies binaires (Cavalli-Sforza & Edwards, 1967) qui correspondent à autant d'orthants de dimension  $m-2$ .  $\mathcal{T}_m$  est donc constitué de  $(2m-3)!!$  orthants (un par topologie) de dimension  $m-2$ , collés entre eux par des sous-orthants de dimension inférieure et correspondant à des topologies non-binaires. La figure 3.2 montre une jonction de tels orthants dans le cas où  $m=4$ .

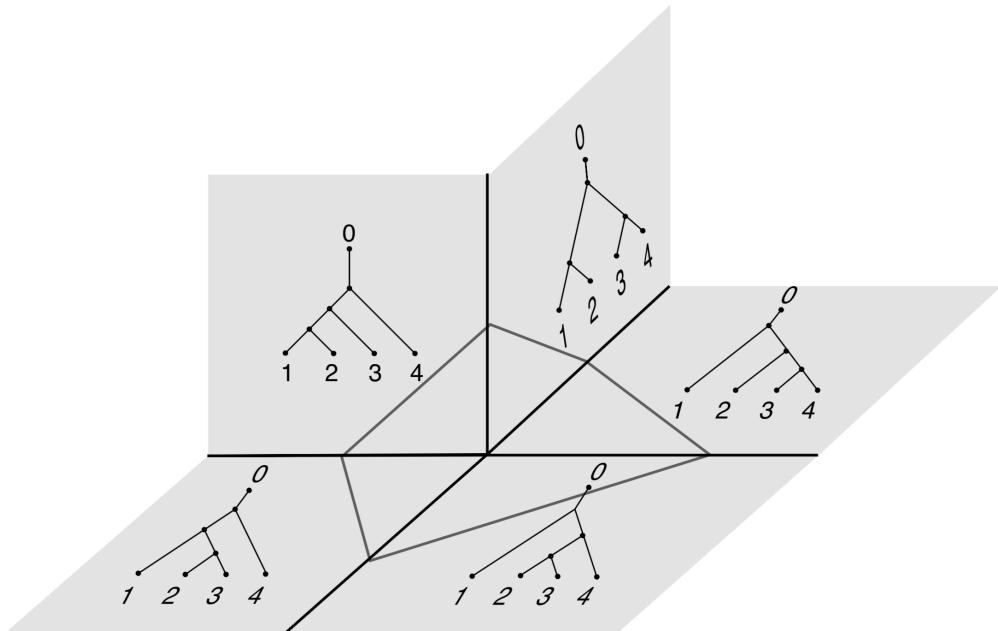


FIGURE 3.2: Une partie de  $\mathcal{T}_4$ , où cinq orthants se rejoignent, tiré de Billera et al. (2001).

Billera et al. (2001) montrent que  $\mathcal{T}_m$  est connexe et possède une courbure négative. En particulier, pour toute paire d'arbre  $T$  et  $T'$ , il existe un plus court chemin dans  $\mathcal{T}_m$  entre  $T$  et  $T'$ . La distance BHV est la distance géodésique dans  $\mathcal{T}_m$ , c'est à dire la distance du plus court chemin. Lorsque  $T$  et  $T'$  appartiennent au même orthant (c'est à dire partagent la même topologie), la distance géodésique coïncide avec la distance euclidienne dans l'orthant.

La figure 3.3 illustre un exemple plus complexe dans lequel le plus court chemin traverse plusieurs orthants. Dans cet exemple, la distance géodésique est la somme des distances euclidiennes du chemin parcouru dans chaque orthant. La distance BHV prend en compte la topologie, les longueurs de branche et la géométrie de  $\mathcal{T}_m$  mais son calcul effectif est nettement plus coûteux que celui des distances cophénétiques et de Robinson-Foulds (Owen & Provan, 2010).

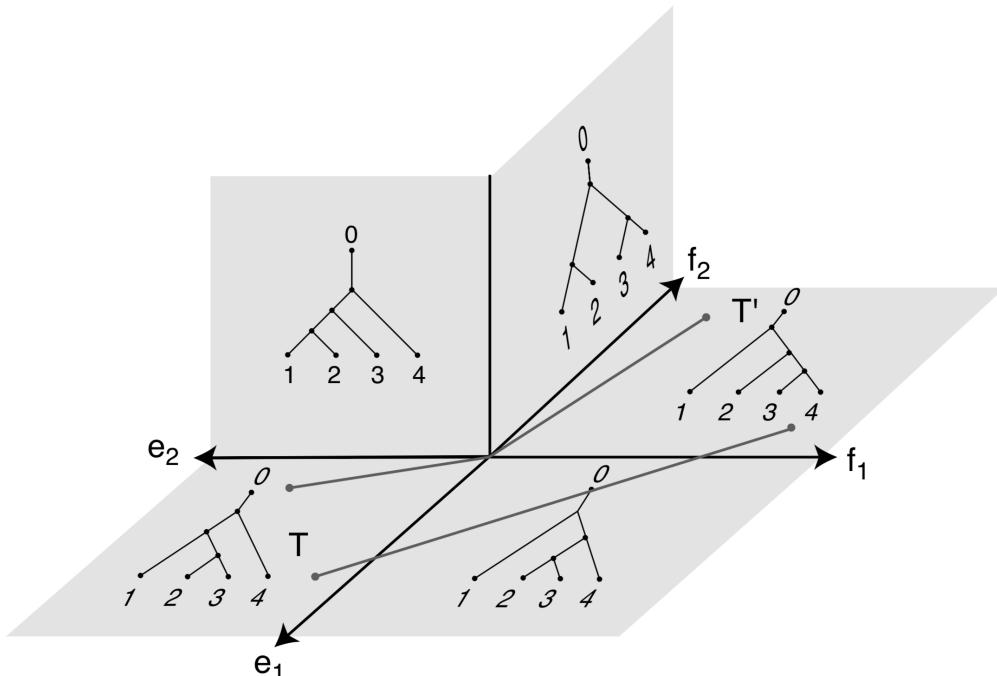


FIGURE 3.3: Des chemins traversant plusieurs orthants dans  $\mathcal{T}_4$ , tiré de Billera et al. (2001).

### 3.3 Arbres d'intérêt

Dans nos analyses, nous considérons trois types d'arbre : taxonomique, phylogénétique ou de corrélation. Les deux premiers représentent l'histoire évolutive des taxons et sont reconstruits à partir de données phénotypiques et moléculaires. Le

dernier est reconstruit à partir des tableaux d'abondance et représente les similités en terme de profils d'abondance.

### 3.3.1 Phylogénie

L'arbre phylogénétique, ou phylogénie, traduit les relations de parenté entre organismes. Pour un séquençage avec gène marqueur, l'arbre est reconstruit à partir des séquences représentatives (Price, Dehal, & Arkin, 2010). Intuitivement, la longueur d'une branche reflète la distance évolutive entre une séquence et sa séquence parentale, mesuré par le nombre moyen de substitutions par paire de bases sur la branche.

### 3.3.2 Taxonomie

La classification linnéenne du vivant regroupe les espèces au sein de groupes cohérents de plus en plus larges : genres, familles, ordres, classes, embranchements et règnes. Cette hiérarchie peut être représentée sous la forme d'un arbre, dit taxonomie, et est disponible dans des bases de données comme celle du NCBI (Geer et al., 2010). Cette taxonomie est construite à partir de données phénotypiques ou moléculaires mais est indépendante des données d'abondance. De plus, l'étiquetage des espèces et des rangs supérieurs permet d'avoir un cadre conceptuel unifié entre différentes études. En revanche, l'arbre n'est pas binaire mais au contraire fortement polytomique et les branches n'ont pas de longueur naturelle. Elles seront arbitrairement fixées à 1 dans la suite des analyses.

### 3.3.3 Arbre des corrélations

Afin de rendre compte de la structure observée des données, nous avons défini un arbre, dit *arbre des corrélations*, construit à partir des tables d'abondances. Dans un premier temps, la table d'abondance est utilisée pour construire une matrice de dissimilarité  $D = (d_{i,j})$  entre espèces, définie par :

$$d_{i,j} = 1 - \frac{\text{cov}(\text{rg}_{\tilde{X}_i}, \text{rg}_{\tilde{X}_j})}{\sigma_{\text{rg}_{\tilde{X}_i}} \sigma_{\text{rg}_{\tilde{X}_j}}}$$

où  $\text{rg}_X$  désigne les variables de rang du vecteur  $X$  et  $\tilde{X}_i = \{X_i : X_i + X_j > 0\}$ ,  $\tilde{X}_j = \{X_j : X_i + X_j > 0\}$  sont les profils d'abondances de  $X_i$  et  $X_j$  privés des coordonnées simultanément nulles. Le deuxième terme du membre de droite correspond à la correlation de Spearman entre  $\tilde{X}_i$  et  $\tilde{X}_j$ . La matrice  $D$  est utilisée pour construire une classification ascendante hiérarchique avec la méthode de Ward. L'arbre des corrélations correspond au dendrogramme de la classification. Par construction,

les taxons au profil d'abondance similaire sont regroupés ensemble et la longueur des branches correspond au coût de fusion entre deux sous-arbres.

Contrairement aux arbres taxonomiques et phylogénétiques, l'arbre des corrélations est construit à partir des abondances et il est toujours possible de le calculer, y compris lorsque les « taxons » ne sont pas des espèces mais d'autres entités (gènes, MSPs, etc) pour lesquelles ni la phylogénie, ni la taxonomie ne sont définies.

Cependant, l'arbre des corrélations étant estimé à partir des données, il est sensible à la variabilité de celles-ci. C'est d'autant plus le cas que pour les couples d'espèces rares où il y a beaucoup de zéros partagés et l'estimation de leur corrélation est très imprécise. Ce problème peut être résolu en filtrant les espèces à faible abondance ou prévalence mais ce sont souvent ces espèces rares qui jouent un rôle crucial dans le fonctionnement des écosystèmes (Jousset et al., 2017).

Enfin, si les données servent à la fois à construire l'arbre des corrélations et à effectuer les tests, il faut faire attention à ne pas surapprendre de celles-ci.

## 3.4 Comparaison entre les arbres

Notre objectif est de comparer l'arbre des corrélations à l'arbre taxonomique (ou phylogénétique). Nous nous intéressons plus précisément aux trois questions suivantes :

- L'arbre des corrélations est-il significativement différent de l'arbre taxonomique ?
- L'arbre des corrélations est-il plus proche de l'arbre taxonomique qu'un arbre aléatoire ?
- Quel est l'impact de l'arbre sur les procédures d'abondance différentielle hiérarchiques ?

Pour répondre aux deux premières questions, nous utilisons une forêt d'arbres pour construire une *région de confiance* autour de l'arbre des corrélations et une *distance typique* entre arbres aléatoires. Une réponse à la dernière question sera apportée en testant différents choix d'arbre pour les procédures hiérarchiques sur des jeux de données simulées et réelles.

### 3.4.1 Forêt d'arbres

L'arbre des corrélations étant estimé à partir des données et assez variable, nous déterminons une région de confiance autour de celui-ci à l'aide de la méthode du *bootstrap* (Felsenstein, 1985 ; Wilgenbusch, Huang, & Gallivan, 2017). Pour ce faire,

nous créons  $N_B$  nouvelles tables d’abondance par un ré-échantillonnage avec remise sur les échantillons (*i.e.* les colonnes) puis, pour chaque table ainsi construite, nous calculons un nouvel arbre de corrélation.

Nous générerons également des arbres aléatoires, sous l’hypothèse nulle, en permutant les labels des feuilles d’un arbre de référence. Cette procédure conserve le nombre de branches et la distribution des polytomies. Elle est de ce fait plus adaptée qu’un tirage uniforme dans l’espace des arbres, qui favorise les topologies symétriques binaires et ne génère pas de nœuds polytomiques.  $N_T$  arbres aléatoires sont générés de cette manière en partant de la taxonomie et  $N_C$  en partant de l’arbre des corrélations.

Au total, nous avons donc une forêt comprenant  $2 + N_B + N_C + N_T$  arbres ayant les mêmes feuilles.

### 3.4.2 Distance entre les arbres

Nous calculons la matrice des distances RF et BHV entre toutes les paires d’arbres de notre forêt, que nous exploitons de deux façons différentes.

Nous regardons d’abord la distance entre chaque arbre et l’arbre des corrélations. La partie supérieure de la figure 3.4 représente les boîtes à moustaches ainsi que les diagrammes en violon de ces distances pour trois jeux de données. L’arbre des corrélations est significativement plus proche de ses répliquats *bootstrapés* que de la taxonomie ou des arbres aléatoires ( $p < 10^{-16}$  avec un test des étendues de Tukey). De plus, la taxonomie est aussi loin de l’arbre des corrélations que peut l’être une taxonomie aléatoire ( $p > 0.05$ ).

Ensuite, ayant accès à toutes les distances deux à deux nous pouvons également effectuer une analyse en coordonnées principales (PCoA, *Principal Component Analysis*) (Gower, 1966) sur la forêt d’arbres. La PCoA est une méthode de réduction de dimension des données. Comme l’analyse en composantes principales (ACP), elle cherche à construire des axes décorrélés entre eux qui vont maximiser l’inertie du nuage de points. Mais contrairement à l’ACP, la PCoA s’appuie sur une matrice de distances et non sur un tableau de descripteurs individus  $\times$  variables. Dans  $\mathbb{R}^n$ , effectuer une PCoA avec les distances euclidiennes est équivalent à faire une ACP.

Dans la partie inférieure de la figure 3.4, nous avons les deux premiers axes de la PCoA avec la distance BHV pour différents jeux de données, ce que Jombart, Kendall, Almagro-Garcia, & Colijn (2017) et Wilgenbusch et al. (2017) appellent des paysages d’arbres (*tree landscapes*). Nous apercevons deux ou trois îlots (Jombart et al., 2017) : un pour l’arbre de corrélations et ses répliquats *bootstrapés*, qui matérialisent sa région de confiance, un pour la taxonomie et les taxonomies aléatoires et un dernier pour les arbres aléatoires construits à partir de l’arbre des corrélations –les deux derniers étant éventuellement confondus, comme c’est le cas

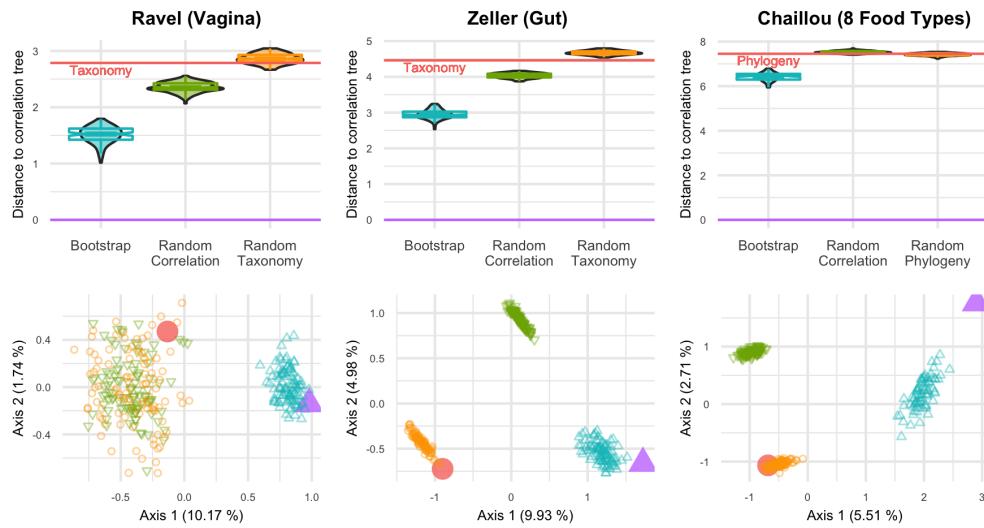


FIGURE 3.4: Distances BHV au sein de la forêt d'arbres pour trois jeux de données.

pour les données de Ravel et al. (2011). Bien que le premier axe ne représente que 5 à 10 % de l'inertie totale, il exclut systématiquement la taxonomie de la région de confiance de l'arbre des corrélations.

Ces résultats doivent être mis en regard de ceux obtenus avec le jeu de données *Chlamydiae* de Caporaso et al. (2011), où la phylogénie se positionne au sein de la région de confiance de l'arbre des corrélations (figure 3.5). Ce jeu de données se caractérise par sa faible taille : seulement 26 échantillons et beaucoup de taxons peu abondants ou peu prévalents, ce qui produit une région de confiance très large. De plus, il contient des échantillons provenant de 8 environnements très différents (océan, selles, sol...), et ces niches écologiques se retrouvent à la fois dans la phylogénie (Philippot et al., 2010) et dans l'arbre des corrélations.

À l'aide de ces résultats, nous pouvons affirmer que l'arbre des corrélations ne capte pas le même signal que la taxonomie ou la phylogénie, en particulier lorsque l'on se concentre sur un seul biome. Les taxons avec un profil d'abondance similaire sont regroupés ensemble dans l'arbre des corrélations, par construction, mais pas dans la taxonomie ou la phylogénie, ce qui ne fait daucun de ces deux derniers arbres un bon candidat pour trouver des groupes de taxons différentiellement abondants.

Nous obtenons les mêmes conclusions en utilisant la distance RF, comme le montre la figure 3.6.

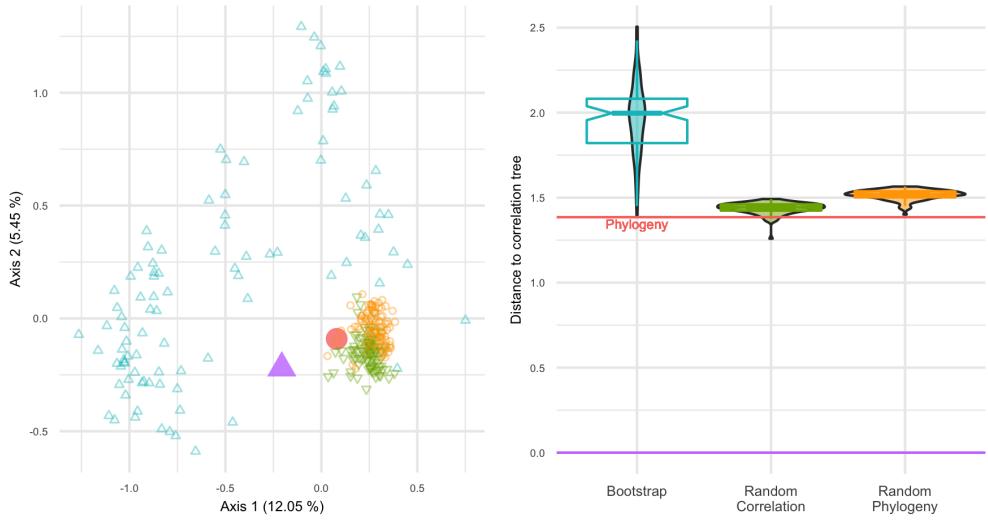


FIGURE 3.5: Distances BHV au sein de la forêt d’arbres pour le jeu de données Chlamydiae.

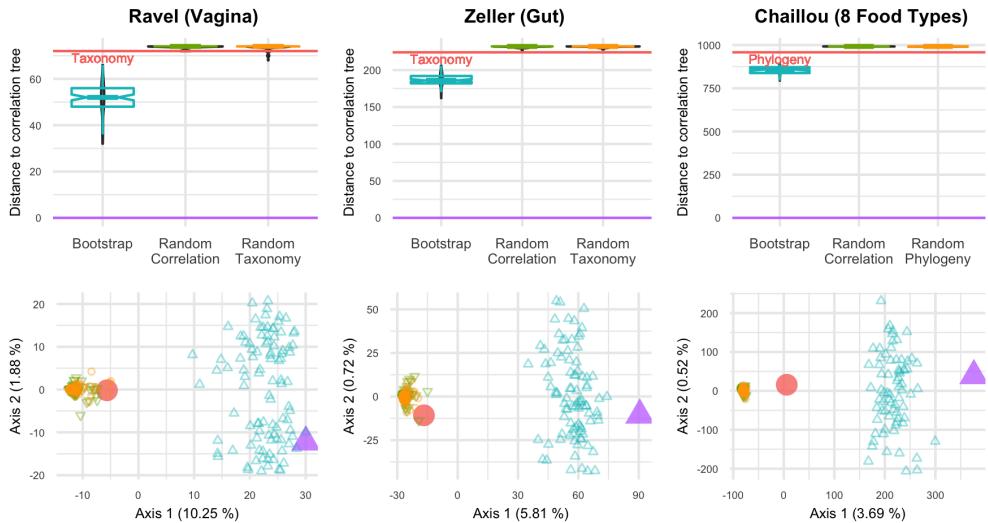


FIGURE 3.6: Distances RF au sein de la forêt d’arbres pour trois jeux de données.

### 3.4.3 Choix de l’arbre et lissage de z-score

L’arbre des corrélations étant différent de l’arbre taxonomique, nous avons regardé l’impact du choix de l’arbre sur la procédure *TreeFDR* (Xiao et al., 2017), à la fois sur des données simulées et sur un jeu de données réel.

Afin de simuler un jeu de données métagénomique, nous avons d'abord appris du jeu de données de Wu et al. (2011) les paramètres d'une loi Dirichlet-multinomiale  $\mathcal{D}(\gamma)$ . Nous créons ensuite un jeu de données échantillon par échantillon comme suit : (i) la profondeur de séquençage  $N_i$  est tirée une loi binomiale négative de moyenne 10000 et de dispersion 25, (ii) le vecteur de proportion  $\alpha_i$  est tirée dans une distribution de Dirichlet de paramètre  $\gamma$  puis (iii) les comptages de l'échantillon  $i$  sont tirés dans une loi multinomiale de paramètre  $N_i$  et  $\alpha_i$ .

Une fois le jeu de données simulées obtenu, il reste à créer des taxons différentiellement abondants. La procédure est illustrée dans la figure 3.7. Tout d'abord, les échantillons sont aléatoirement assignés à un groupe  $A$  ou  $B$  (quadrant B). Ensuite, les taxons différentiellement abondants sont sélectionnés uniformément dans l'arbre (quadrant C). Enfin, l'abondance de ces taxons au sein du groupe  $B$  est multipliée par un *fold-change* (quadrant D).

On qualifiera ce schéma de simulation de *paramétrique*.

Il est également possible d'utiliser directement les comptages d'un jeu de données réel homogène, comme celui des individus sains de Brito et al. (2016) –au lieu d'en simuler suivant une Dirichlet-multinomiale– puis d'appliquer la procédure de la figure 3.7 pour générer de l'abondance différentielle. On parlera alors d'un schéma *non-paramétrique* –par opposition à la simulation paramétrique, précédemment décrite.

Nous avons généré des jeux de données d'abondance différentielle puis évalué la performance des corrections Benjamin-Hochberg et de *TreeFDR* avec l'arbre des corrélations, la taxonomie, un arbre des corrélations aléatoire et une taxonomie aléatoire.

Regardons dans un premier temps les simulations non paramétriques (issues d'un jeu de données réelles). Tout d'abord, notons que la procédure échoue dans 4 % des simulations pour l'ensemble des arbres, et jusqu'à 8 % quand on se limite à l'arbre des corrélations. Notons ensuite que les hyperparamètres  $k$  et  $\rho$  qui contrôlent le niveau de lissage des  $z$ -scores sont assez éloignés de 1 (très bas pour  $k$ , très haut pour  $\rho$ ), ce qui réduit considérablement l'impact du lissage. Les distributions de la figure 3.8 montrent que dans plus de la moitié des simulations, le lissage déplace le  $z$ -score de moins de  $10^{-2}$ . Dans le cas où l'arbre utilisé est celui des corrélations, un déplacement des  $z$ -scores d'amplitude supérieure  $10^{-2}$  ne se produit que dans 5 % des simulations.

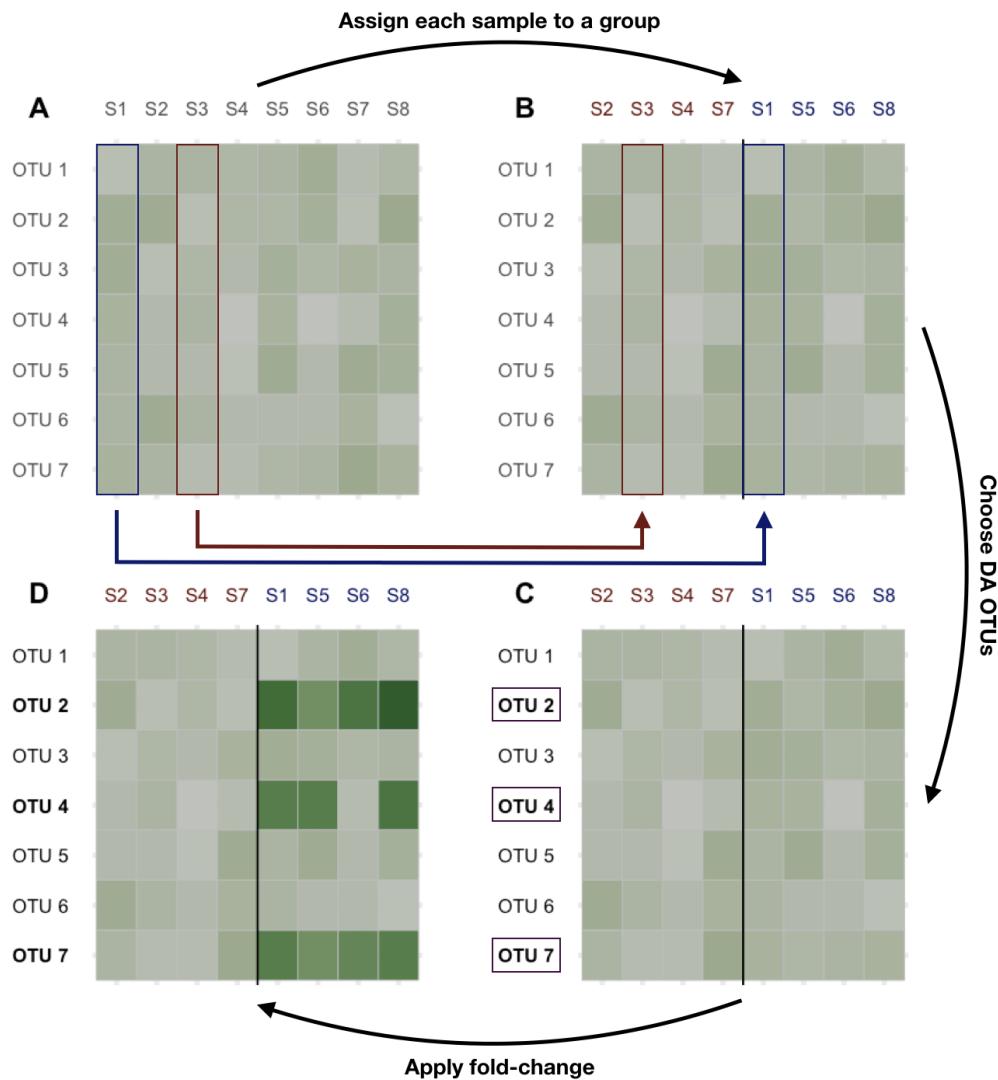


FIGURE 3.7: Création de taxons différentiellement abondants au sein d'un jeu de données.

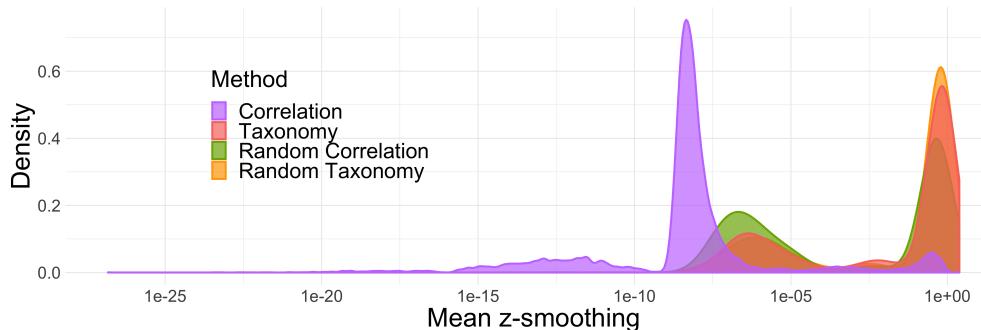


FIGURE 3.8: Distribution des moyennes des valeurs absolues des différences entre les  $z$ -scores avant et après lissage, pour les simulations non paramétriques.

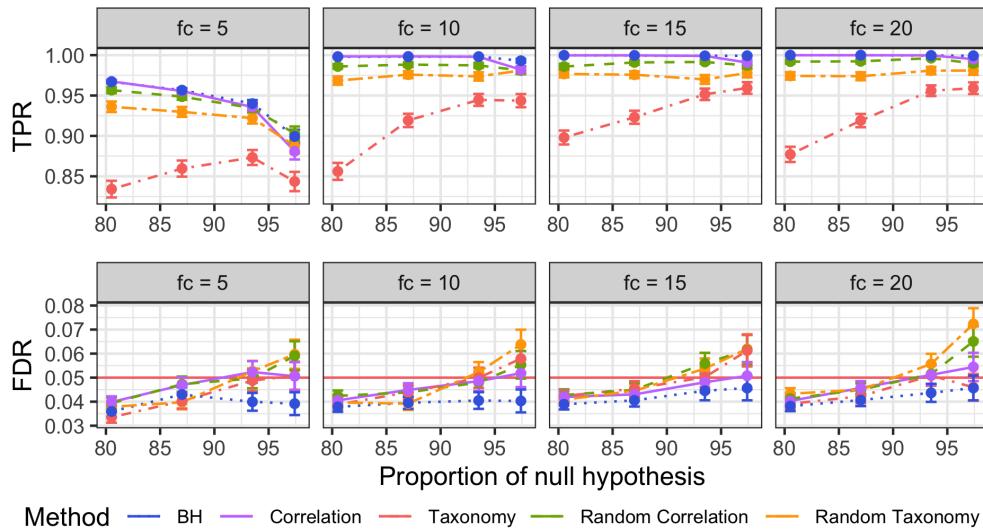


FIGURE 3.9: Moyennes et écart-types de la moyenne des TPR et FDR pour les simulations non paramétriques avec différents fold-changes et proportions d’hypothèses nulles.

Au niveau du contrôle du FDR, seul BH contrôle systématiquement le taux de faux positifs en deçà de 5 % (figure 3.9 bas). Les procédures hiérarchiques peuvent dépasser ce seuil, allant jusqu’à 7 % lorsque la proportion réelle d’espèces différemment abondantes est faible ( $\leq 10\%$ ). Enfin, BH est la méthode qui a le plus important TPR, quelque soit le *fold-change* ou la proportion d’hypothèses nulles (figure 3.9 haut). Devant ces résultats, nous pouvons affirmer que l’utilisation d’un arbre dans la procédure de lissage de *TreeFDR*, même adapté aux données d’abondances comme l’arbre des corrélations, ne permet pas d’avoir des procédures plus puissantes que BH.

Les simulations paramétriques donnent des résultats semblables aux simulations non-paramétriques, comme le montre la figure 3.10. On note cependant un effondrement du TPR avec ce schéma de simulation.

Nous avons également testé l’influence de l’arbre sur la procédure de lissage avec le jeu de données de Zeller et al. (2014). Celui-ci a été analysé à deux niveaux de granularité différents : genre et MSP. La figure montre le nombre de genres (à gauche) ou de MSPs (à droite) détectés pour différents seuils jusqu’à  $\alpha = 0.15$ .

Dans les deux cas, l’arbre des corrélations détecte pour la majorité des seuils plus de taxons que BH mais la différence est très faible, y compris en comparaison avec les arbres aléatoires. Si l’on regarde au niveau des MSPs à  $\alpha = 0.05$ , l’arbre des corrélations permet de détecter 5 taxons de plus que la procédure BH classique. Ces taxons ne sont cependant pas regroupés dans des clades différentiellement abondants. Il semblerait alors que cela soit plutôt la correction de permutation

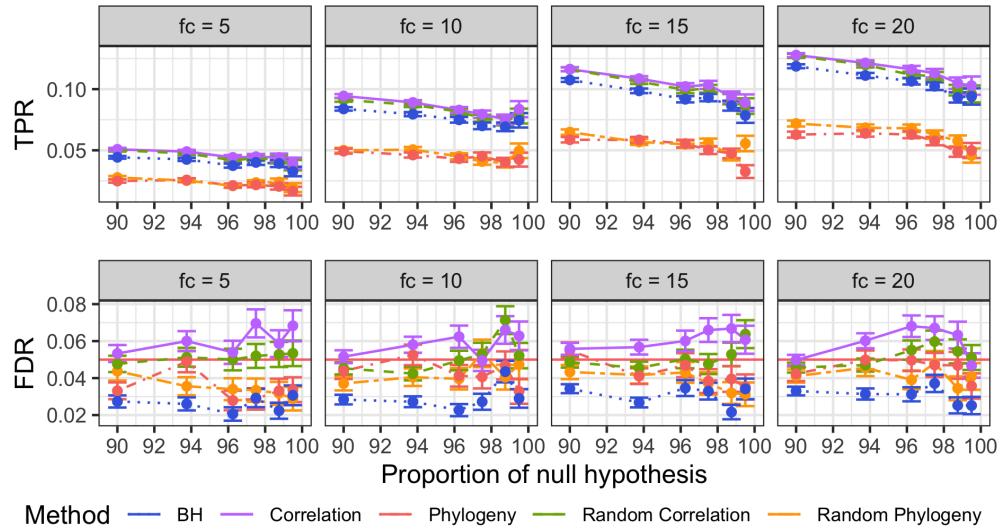


FIGURE 3.10: Moyennes et écart-types de la moyenne des TPR et FDR pour les simulations paramétriques avec différents fold-changes et proportions d'hypothèses nulles.

post-lissage, qui a pour objectif de contrôler le *FDR discret*, plutôt que le lissage lui-même qui produise cet effet. Le FDR discret prend en compte des distributions non continues des *p*-valeurs sous l'hypothèse nulle et est adapté aux données de comptage. Cette méthode de correction a été déclarée plus performante pour détecter des taxons différemment abondants que la procédure BH classique (Jiang et al., 2017).

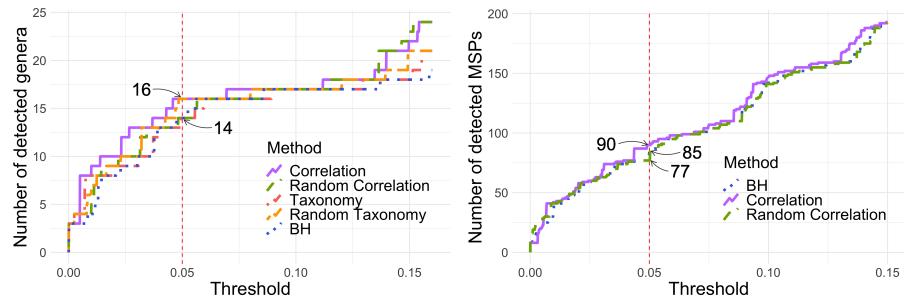


FIGURE 3.11: Moyennes et écart-types de la moyenne des TPR et FDR pour les simulations non paramétriques avec différents fold-changes et proportions d'hypothèses nulles.

### 3.4.4 Choix de l'arbre et FDR hiérarchique

Le fait que le FDR hiérarchique ne permette pas de spécifier *a priori* un niveau de FDR cible mais uniquement de le calculer *a posteriori* pour une liste d'espèces différentiellement abondantes le rend inadapté à des simulations comme celles de la section 3.4.3. C'est pourquoi nous ne regarderons l'impact du choix de l'arbre sur cette procédure qu'au travers de jeux de données réelles.

Nous avons d'abord réanalysé le jeu de données Chlamydiae (Sankaran & Holmes, 2014), comme dans l'article original, en utilisant un seuil pour la correction au niveau des familles  $\alpha = 0.1$ . Avec la phylogénie, 8 OTUs ont été détectés comme différentiellement abondantes et la procédure garantit un FDR *a posteriori* de  $\alpha' = 0.32$ . Si l'on utilise l'arbre des corrélations à la place, on trouve 3 OTUs supplémentaires pour un FDR *a posteriori* comparable de  $\alpha' = 0.324$ . Les boîtes à moustaches des abondances de ces OTUs (figure 3.12 E, F) montrent qu'elles sont bien plus abondantes dans les échantillons de sol que dans les autres biomes, ce qui confirme leur statut d'OTU différentiellement abondante.

La figure 3.12 résume cette analyse en indiquant quelles OTUs sont détectées avec l'une ou l'autre des méthodes et en précisant les évidences brutes ( $\epsilon = -\log p$ ) correspondantes aux feuilles des arbres. Intéressons nous plus particulièrement à l'OTU 547 579 (marquée d'une étoile rouge) qui a été détectée par l'arbre des corrélations (figure 3.12 D) mais pas par la phylogénie (figure 3.12 A). Elle n'a pas été testée par la phylogénie car elle est entourée d'espèces non différentiellement abondantes qui masquent le signal (figure 3.12 B). À l'inverse, dans l'arbre des corrélations (figure 3.12 D), elle se situe dans un clade où toutes les OTUs sont différentiellement abondantes, ce qui permet à la procédure de rejeter tous les sous-arbres contenant 547 579 jusqu'à la feuille correspondant à 547 579, sans s'arrêter au niveau des branches internes.

On peut remarquer que le FDR *a posteriori* est assez élevé, à 0.324. En effectuant une correction par BH à ce même niveau, on détecte 15 OTUs différentiellement abondantes, soit quatre de plus que la correction avec l'arbre des corrélations. Ceci peut s'expliquer par le fait que hFDR contrôle le FDR dans le pire des cas tandis que le FDR effectif pourrait être bien plus bas que cette borne pessimiste (Yekutieli, 2008).

Avec cette approche descendante, l'arbre des corrélations est plus adapté que la phylogénie. En regroupant les espèces corrélées et donc potentiellement différentiellement abondantes au sein d'un même sous-arbre, il permet de concentrer les différents signaux dans une portion d'arbre et d'éviter qu'ils ne soient dilués dans tout l'arbre, comme c'est le cas avec la phylogénie.

Nous avons ensuite analysé le jeu de données Chaillou (Chaillou et al., 2015) restreint aux *Bacteroidetes*. L'utilisation d'un seuil  $\alpha = 0.01$  au niveau des familles a conduit à un contrôle du FDR *a posteriori* à  $\alpha' = 0.04$  à la fois pour la phylogénie

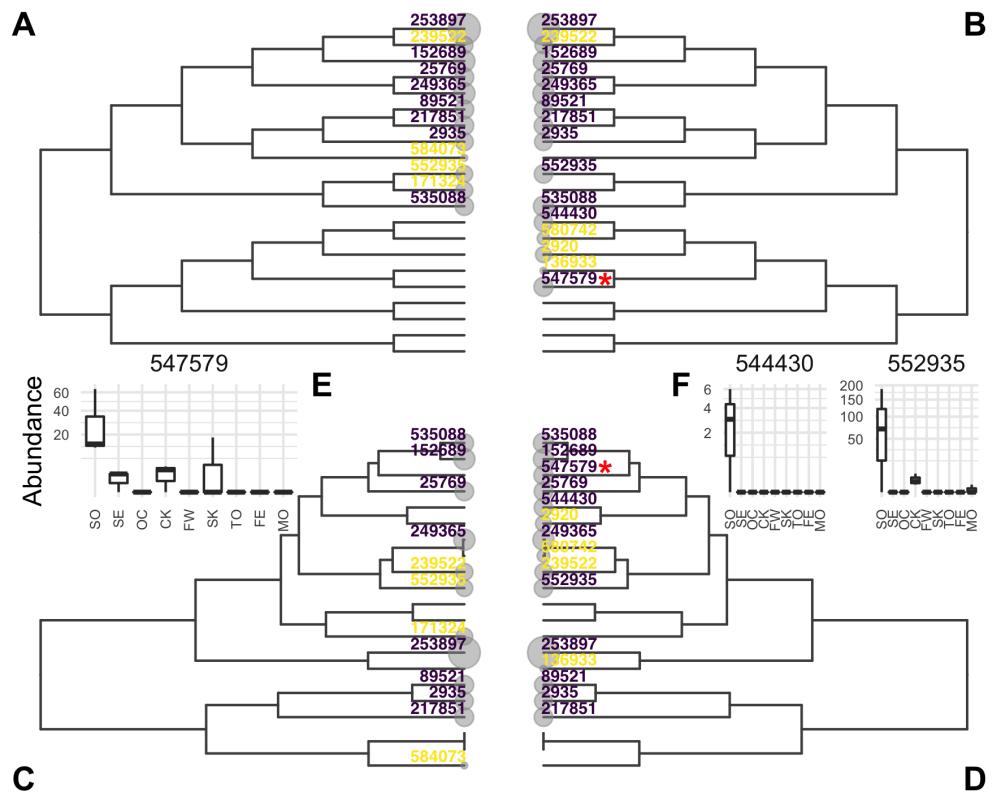


FIGURE 3.12: Les évidences brutes sont représentées aux feuilles de la phylogénie (A et C) ou de l'arbre des corrélations (B et D). Les OTUs considérées différemment abondantes pour la phylogénie (A et B) ou pour l'arbre des corrélations (C et D) sont en violet. Les OTUs testées mais non rejetées sont en jaune.

et l'arbre des corrélations. Le premier arbre a détecté 28 OTUs différemment abondantes contre 34 pour la phylogénie.

En observant l'abondance des 22 OTUs détectées simultanément par les deux procédures ou des 18 détectées par une seule des deux (figure 3.13), on remarque que chacune d'entre elles (i) est absente ou en dessous du seuil de détection dans au moins un des aliments et (ii) a de fortes prévalences et abondances dans au moins un autre aliment. Ceci valide leur caractère différemment abondant et permet de les considérer comme des vrais positifs.

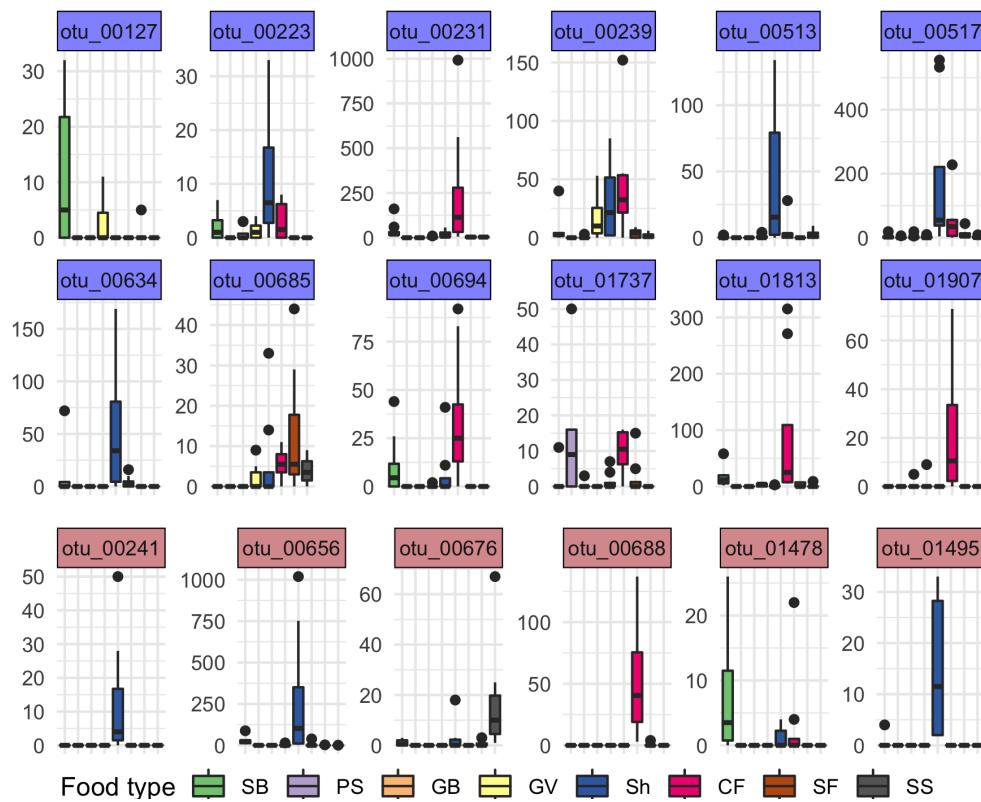


FIGURE 3.13: Abondances des OTUs détectées uniquement par l'arbre des corrélations (en bleu) ou par la phylogénie (en rouge).

Si l'on regarde les OTUs détectées sur leurs arbres respectifs, comme représentées sur la figure 3.14, les OTUs détectées uniquement par l'arbre des corrélations sont éparpillées dans la phylogénie, de manière similaire à ce qu'on a pu observer pour le jeu de données Chlamydiae. À l'inverse, les OTUs différemment abondantes uniquement dans la phylogénie sont proches d'OTUs détectées pour l'arbre des corrélations mais ne sont pas détectées à cause de la faible puissance du test de Fisher, pratiqué par défaut dans StructSSI et peu adapté aux données métagénomiques.

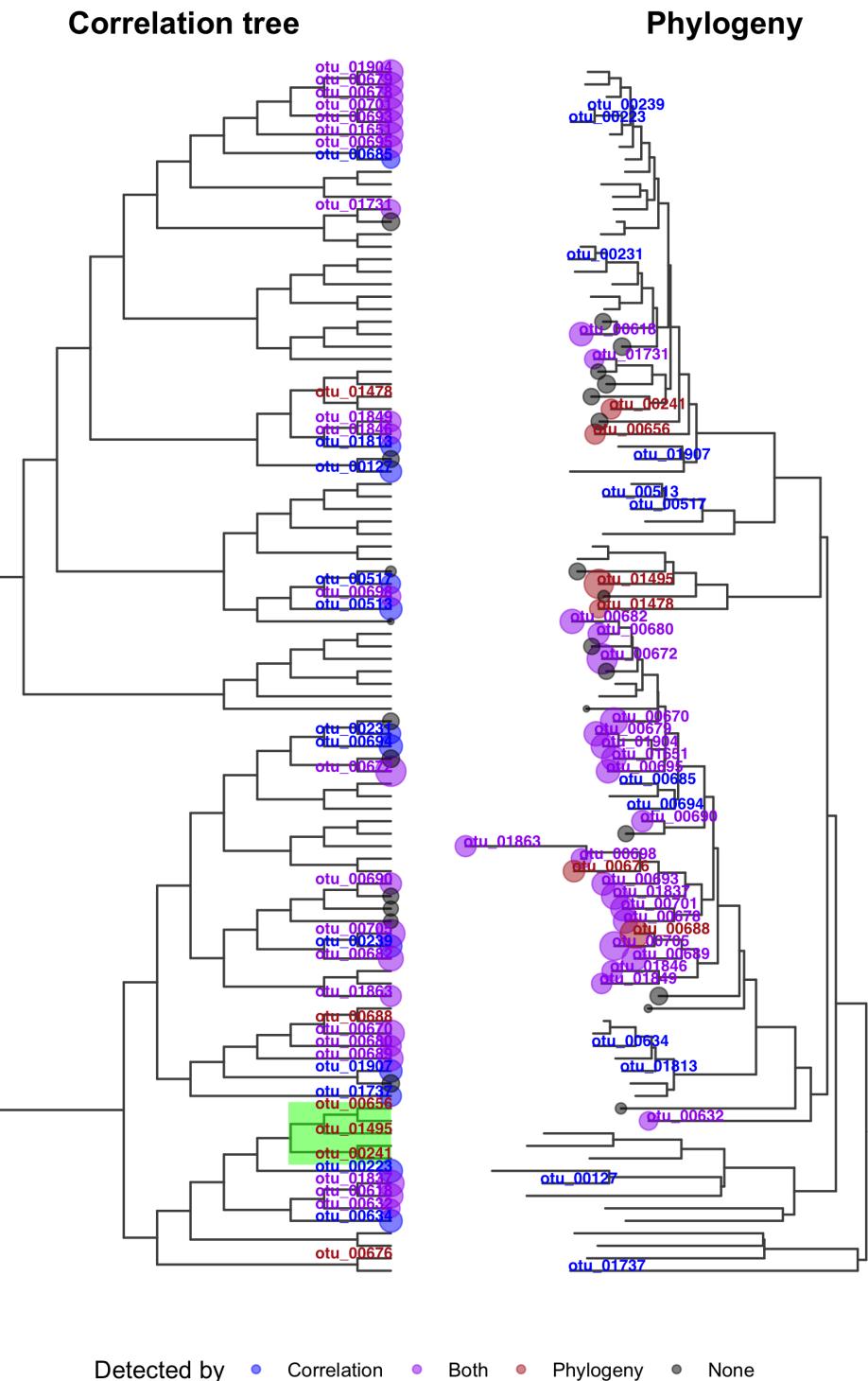


FIGURE 3.14: Évidences des OTUs détectées avec l'arbre des corrélations (à droite) ou la phylogénie (à gauche).

Si l'on regarde plus attentivement les cinq OTUs du cadre vert de la figure 3.14, on a trois OTUs qui ont été détectées uniquement par la phylogénie et non par l'arbre des corrélations alors qu'elles appartiennent au même clade de cinq individus dans celui-ci. La figure 3.15 met l'accent sur ces cinq OTUs, qui ne sont présentes quasiment que dans les crevettes. Le test de Fisher couplé au bruit causé par l'agrégation des données fait qu'on ne descend pas dans le sous-arbre en question. L'inaptitude de l'arbre des corrélations à identifier ces taxons est donc imputable à l'utilisation d'un test inadapté : remplacer le test de Fisher par un test non paramétrique de Kruskall-Wallis aurait permis à la procédure hiérarchique sur l'arbre des corrélations d'identifier l'ensemble des OTUs du clade.

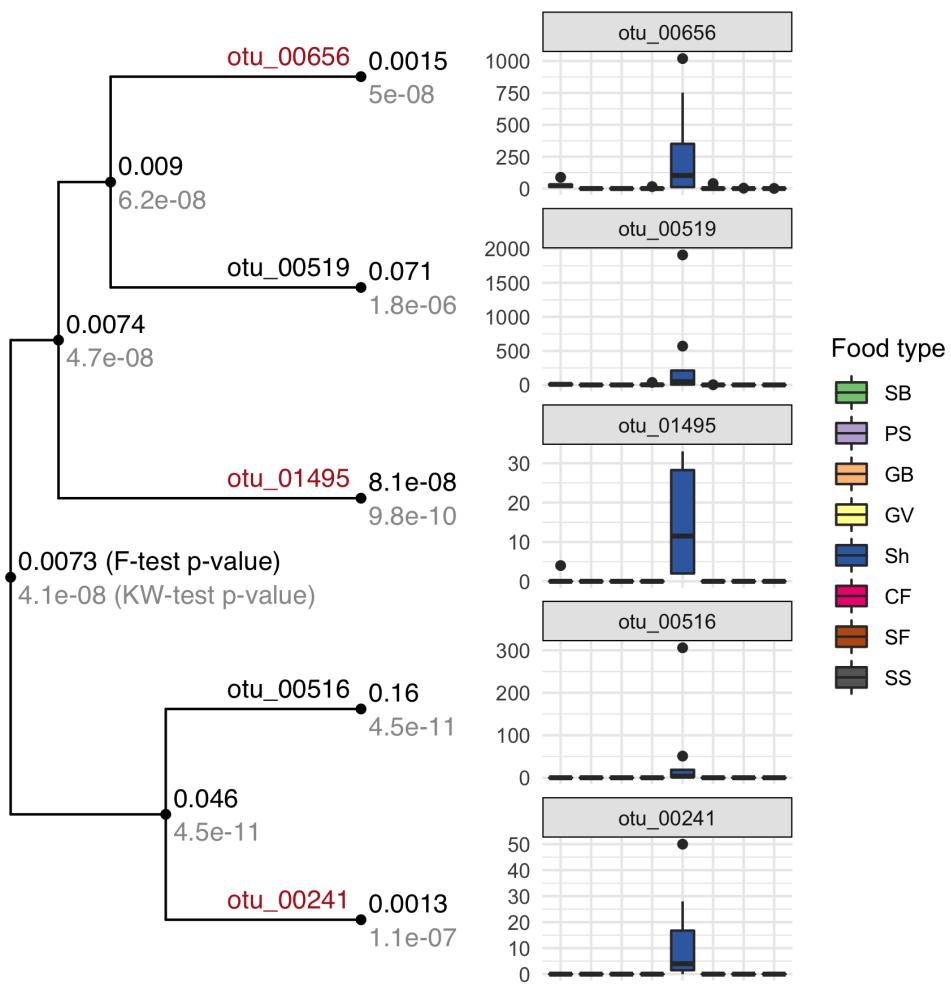


FIGURE 3.15: Focus sur les cinq OTUs du cadre vert de la figure 3.14.



# Chapitre 4

## ***zazou* : une nouvelle approche**

### **4.1 Processus d'Ornstein-Uhlenbeck**

Un processus d'Ornstein-Uhlenbeck (OU) de force de rappel  $\alpha_{\text{ou}}$ , de valeur optimale  $\beta_{\text{ou}}$ , et d'écart-type  $\sigma_{\text{ou}}$  est un processus gaussien qui satisfait l'équation différentielle stochastique suivante :

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

Si  $W_0$  est connu et fixé, l'espérance du processus vaut  $\mathbb{E}[W_t] = W_0 e^{-\alpha_{\text{ou}}t} + \beta_{\text{ou}}(1 - e^{-\alpha_{\text{ou}}t})$  et la covariance du processus est donnée par

$$\text{Cov}[W_t, W_s] = \frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}} \left( e^{-\alpha_{\text{ou}}|t-s|} - e^{-\alpha_{\text{ou}}(t+s)} \right).$$

Le processus est gaussien et il admet pour loi limite  $\mathcal{N}\left(\beta_{\text{ou}}, \frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}}\right)$ , dont la variance est finie.

De par leurs propriétés, les processus d'Ornstein-Uhlenbeck sont devenus populaires pour modéliser l'évolution de traits biologiques continus, comme la masse corporelle des mammifères (Freckleton, Harvey, & Pagel, 2003).

Il est également possible de faire évoluer un processus d'Ornstein-Uhlenbeck sur un arbre (Bastide, Mariadassou, & Robin, 2017). Le long d'une branche, les paramètres du processus sont fixes. À chaque nœud, une branche se divise (en deux dans le cas d'un arbre binaire) et le processus donne naissance à deux copies indépendantes ayant la même valeur initiale au point de branchement. Cela induit notamment une dépendance statistique entre tous les descendants d'un même ancêtre. Cette dépendance est d'autant plus forte que l'ancêtre est récent. Sur la figure 4.1, le processus vert partant de  $N_1$  jusqu'à  $N_2$  donne naissance à deux processus  $T_4$  et  $T_5$ , jaune et bleu, lorsqu'il arrive au nœud  $N_2$ .

De plus, à chaque branchement, un changement dans les paramètres du processus est susceptible de se produire. Dans ce cas, le processus garde la même valeur

au noeud mais continue sa trajectoire avec les nouveaux paramètres. C'est le cas en  $N_3$  dans la figure 4.1 où le processus orange a subi un changement dans sa valeur optimale  $\beta_{\text{ou}}$  par rapport au processus rouge : la trajectoire est continue et le processus dérive vers la nouvelle valeur optimale.

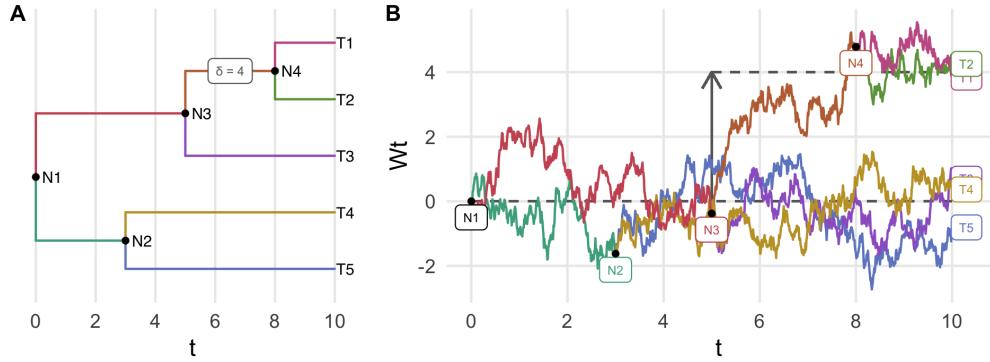


FIGURE 4.1: Exemple d'un processus d'Ornstein-Uhlenbeck sur un arbre à 5 feuilles. À chaque branchemet, le processus se scinde en deux processus indépendants ayant la même valeur initiale. Les paramètres sont conservés sauf lors d'un saut dans la valeur optimale, comme sur la branche conduisant à  $N_4$ .

Le produit matriciel de la matrice d'incidence  $T$  par le vecteur de sauts  $\delta$  permet d'effectuer la somme cumulée des sauts le long des branches pour obtenir la valeur optimale du processus aux feuilles. Dans l'exemple de la figure 4.1,

$$\beta_{\text{ou},\{\text{feuilles}\}} = T\delta = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

En notant  $\beta_{\text{ou},i}$  la valeur optimale du processus sur la branche descendant au noeud  $i$ , la loi à ce noeud conditionnellement à son parent est

$$X_i | X_{\text{pa}(i)} \sim \mathcal{N} \left( X_{\text{pa}(i)} e^{-\alpha_{\text{ou}} \ell_i} + \beta_{\text{ou},i} (1 - e^{-\alpha_{\text{ou}} \ell_i}), \frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}} (1 - e^{-2\alpha_{\text{ou}} \ell_i}) \right).$$

Ceci permet d'avoir une expression pour la covariance entre les nœuds  $i$  et  $j$ . Dans le cas où  $P = \text{pa}(i) = \text{pa}(j)$ , on a

$$\begin{aligned}\text{Cov}[X_i, X_j] &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \mathbb{E}[X_j] \\ &= \mathbb{E}[\mathbb{E}[X_i X_j | P]] - \mathbb{E}[\mathbb{E}[X_i | P]] \mathbb{E}[\mathbb{E}[X_j | P]] \\ &= \mathbb{E}\left[\left(Pe^{-\alpha_{\text{ou}}\ell_i} + \beta_{\text{ou},i}(1 - e^{-\alpha_{\text{ou}}\ell_i})\right)\left(Pe^{-\alpha_{\text{ou}}\ell_j} + \beta_{\text{ou},j}(1 - e^{-\alpha_{\text{ou}}\ell_j})\right)\right] \\ &\quad - \mathbb{E}\left[Pe^{-\alpha_{\text{ou}}\ell_i} + \beta_{\text{ou},i}(1 - e^{-\alpha_{\text{ou}}\ell_i})\right] \mathbb{E}\left[Pe^{-\alpha_{\text{ou}}\ell_j} + \beta_{\text{ou},j}(1 - e^{-\alpha_{\text{ou}}\ell_j})\right] \\ &= e^{-\alpha_{\text{ou}}(\ell_i + \ell_j)} \mathbb{E}[P^2] - e^{-\alpha_{\text{ou}}(\ell_i + \ell_j)} \mathbb{E}[P]^2 \\ &= e^{-\alpha_{\text{ou}}d_{i,j}} \text{Var}[P] \\ &= \frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}} (1 - e^{-2\alpha_{\text{ou}}t_{i,j}}) \times e^{-\alpha_{\text{ou}}d_{i,j}}.\end{aligned}$$

Cette expression reste valable quelque soit le niveau de parenté entre  $i$  et  $j$ . Ainsi, pour un arbre ultramétrique de longueur  $h$ , la matrice de variance-covariance du vecteur gaussien des feuilles (numérotées de 1 à  $m$ ) est composée des éléments

$$\frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}} (e^{-\alpha_{\text{ou}}d_{i,j}} - e^{-2\alpha_{\text{ou}}h}). \quad (4.1)$$

En particulier, la variance à toutes les feuilles est  $\frac{\sigma_{\text{ou}}^2}{2\alpha_{\text{ou}}} (1 - e^{-2\alpha_{\text{ou}}h})$ .

## 4.2 Zazou

### 4.2.1 Modèle

L'approche que nous avons développée s'appuie sur les  $z$ -scores aux feuilles et fait deux hypothèses :

1. les  $z$ -scores sont la réalisation d'un processus d'Ornstein-Uhlenbeck avec sauts sur l'arbre phylogénétique,
2. sous  $\mathcal{H}_1$ ,  $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$  avec  $\mu_i < 0$ .

La première hypothèse nous permet de définir la distribution jointe des  $z$ -scores comme

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

où  $\mu$  dépend de  $\delta$ , le vecteur des sauts du processus, via la relation  $\mu = T\delta$  et  $\Sigma$  dépend de  $\alpha_{\text{ou}}$  et  $\sigma_{\text{ou}}$  via l'équation (4.1).

La seconde hypothèse est classique lorsqu'on travaille sur les  $z$ -scores (McLachlan & Peel, 2004) et suppose l'équivalence entre trouver les hypothèses alternatives et déterminer les composantes non-nulles du vecteur  $\mu$ . Sous  $\mathcal{H}_0$  et  $\mathcal{H}_1$ , la

variance aux feuilles doit valoir 1, ce qui impose

$$\hat{\sigma}_{\text{ou}} = \frac{2\hat{\alpha}_{\text{ou}}}{1 - e^{-2\hat{\alpha}_{\text{ou}}h}},$$

de sorte que  $\hat{\Sigma}$  dépend uniquement de  $\hat{\alpha}_{\text{ou}}$ .

### 4.2.2 Estimation ponctuelle

En supposant  $\hat{\alpha}_{\text{ou}}$  (et donc  $\hat{\Sigma}$ ) connue, une application naïve du maximum de vraisemblance donnerait

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

comme estimateur de  $\mu$ . En réalité, c'est la position des sauts qui nous intéresse et nous souhaitons donc plutôt avoir un estimateur de  $\delta$ . En utilisant la relation  $\mu = T\delta$ , celui-ci est donné par

$$\hat{\delta} = \underset{\delta \in \mathcal{D}}{\operatorname{argmin}} \|\mathbf{z} - T\delta\|_{\Sigma^{-1}, 2}^2,$$

où  $\mathcal{D} = \{\delta \in \mathbb{R}^n / T\delta \in \mathbb{R}_-^m\}$  est l'ensemble de faisabilité pour les sauts qui induisent des composantes négatives aux feuilles. Bien que le problème soit convexe (la fonction objective est convexe, tout comme l'ensemble de faisabilité), la matrice  $T$  n'est pas plein rang et l'estimateur  $\hat{\mu}$  n'est donc pas unique.

Nous lui préférons donc un estimateur parcimonieux, obtenu en rajoutant une contrainte  $\ell_1$  (Tibshirani, 1996) à la fonction objectif :

$$\hat{\delta} = \underset{\delta \in \mathcal{D}}{\operatorname{argmin}} \|\mathbf{z} - T\delta\|_{\Sigma^{-1}, 2}^2 + \lambda \|\delta\|_1.$$

En utilisant la décomposition de Cholesky  $\Sigma^{-1} = R^T R$ , ce nouveau problème peut se ramener au problème bien connu du lasso, avec une contrainte convexe sur  $\delta$  :

$$\hat{\delta} = \underset{\delta \in \mathcal{D}}{\operatorname{argmin}} \|y - X\delta\|_2^2 + \lambda \|\delta\|_1, \quad (4.2)$$

où  $y = R\mathbf{z} \in \mathbb{R}^m$  et  $X = RT \in \mathbb{R}^{m \times n}$ .

Le problème (4.2) est convexe en  $\delta$  et sa résolution est possible avec l'algorithme détaillé dans la section 5.1, qui est une modification de l'algorithme du *shooting* (Fu, 1998).

Il reste maintenant à déterminer  $\alpha$  et  $\lambda$ . Ceux-ci ne pouvant être obtenus directement à partir des données, nous allons sélectionner le couple qui minimise le critère BIC suivant :

$$(\hat{\alpha}_{\text{ou}}, \hat{\lambda}) = \underset{\alpha > 0, \lambda \geq 0}{\operatorname{argmin}} \| \mathbf{z} - T\delta_{\alpha, \lambda} \|_{\Sigma(\alpha)^{-1}, 2}^2 + \log |\Sigma(\alpha)| + \|\delta_{\alpha, \lambda}\|_0 \log m,$$

où  $\delta_{\alpha, \lambda}$  est la solution du problème (4.2) pour  $\alpha$  et  $\lambda$ . En pratique, une grille bidimensionnelle donne les valeurs du couple à tester. Cette approche a été préférée à l'alternative usuelle de la validation croisée pour ne pas avoir à gérer la dépendance entre les  $z$ -scores.

### 4.2.3 Débiaisage et intervalles de confiance

L'estimateur lasso est connu pour être biaisé (Javanmard & Montanari, 2013) et ne produit pas d'intervalles de confiance pour les  $\hat{\delta}_i$ . Nous utilisons donc une procédure de débiaisage, comme celles proposées dans Zhang & Zhang (2014) ou Javanmard & Montanari (2013) et Javanmard & Montanari (2014). Ces deux procédures fonctionnent suivant le même principe. Tout d'abord, au lieu d'avoir un estimateur initial de  $\delta$  comme dans (4.2), nous avons besoin d'un estimateur couplé de  $\delta$  et de sa variance  $\sigma$ , qui peut être obtenu par un *scaled lasso* (Sun & Zhang, 2012) et qui sera notre estimateur initial :

$$(\hat{\delta}^{(\text{init})}, \hat{\sigma}) = \underset{\delta \in \mathcal{D}, \sigma > 0}{\operatorname{argmin}} \frac{\|y - X\delta\|_2^2}{2\sigma m} + \frac{\sigma}{2} + \lambda \|\delta\|_1.$$

L'estimation jointe est faite de façon itérative en alternant des étapes de mise à jour de  $\hat{\delta}^{(\text{init})}$  par lasso et de mise à jour de  $\hat{\sigma}$  par résolution exacte, via l'expression  $\hat{\sigma} = \frac{\|y - X\hat{\delta}^{(\text{init})}\|_2}{\sqrt{m}}$ . Une fois cette estimation initiale obtenue, Zhang & Zhang (2014) proposent de calculer un système de score  $S$ , qu'on peut comprendre comme une orthogonalisation faible de  $X$ , pour corriger  $\hat{\delta}^{(\text{init})}$ . La colonne  $s_j$  de  $S$  s'obtient comme étant le résidu de la régression lasso de  $x_j$  contre  $X_{-j}$ , le reste des colonnes de  $X$ .

Puis, l'estimateur débiaisé s'obtient alors en en corrigeant  $\hat{\delta}_j^{(\text{init})}$  comme suit :

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

Javanmard & Montanari (2013) et Javanmard & Montanari (2014) proposent une correction alternative

$$\hat{\delta} = \hat{\delta}^{(\text{init})} + \frac{1}{m} S X^T (Y - X \hat{\delta}^{(\text{init})}),$$

basée sur un système de score différent. La matrice  $S$  est cette fois-ci un inverse généralisé de  $M = \frac{X^T X}{m}$ , construit colonne par colonne en résolvant les problèmes suivants :

$$\begin{cases} s_j = \underset{s \in \mathbb{R}^n}{\operatorname{argmin}} \quad s^T M s \\ \text{t.q. } \|Ms - e_j\|_\infty \leq \gamma \end{cases}$$

où  $e_j \in \mathbb{R}^n$  est le  $j^{\text{ème}}$  vecteur de la base canonique définie par  $e_{ij} = \delta_{i,j}$ . Dans chacune des deux méthodes, sous des hypothèses standard en régression en grande dimension,  $\hat{\delta} \sim \mathcal{N}_n(\delta, V)$  ce qui permet d'obtenir un intervalle de confiance bila-téral au niveau  $\alpha$  pour  $\hat{\delta}_j$  :

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

Avec le premier système de score,  $V$  se calcule élément par élément :

$$v_{i,j} = \hat{\sigma}^2 \frac{\langle s_i, s_j \rangle}{\langle s_i, x_i \rangle \langle s_j, x_j \rangle},$$

tandis qu'avec le second,  $V = \frac{S M S^T}{m}$ . En exprimant la  $i^{\text{ème}}$  composante de  $\mu$  à l'aide de  $\beta$  comme étant  $\mu_i = t_{i.}^T \delta$  où  $t_{i.}$  est la  $i^{\text{ème}}$  ligne de  $T$ , on obtient également un intervalle de confiance unilatéral au niveau  $\alpha$  pour  $\hat{\mu}_i$  :

$$\left[ -\infty, \hat{\mu}_i + \sqrt{t_{i.}^T V t_{i.}} \phi^{-1} (1 - \alpha) \right],$$

et la  $p$ -valeur associée à cet intervalle, qui teste  $\mathcal{H}_0 = \{\mu_i = 0\}$  contre  $\mathcal{H}_1 = \{\mu_i < 0\}$ , est donc

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

qui est la  $p$ -valeur lissée hiérarchiquement.

#### 4.2.4 Correction pour tests multiples

Une fois ces  $p$ -valeurs lissées obtenues, Javanmard, Javadi, & others (2019) proposent une méthode de correction pour tests multiples conçue spécialement pour le lasso débiaisé et qui repose sur les  $t$ -scores  $t_i = \frac{t_{i.}^T \hat{\delta}}{(t_{i.}^T V t_{i.})^{1/2}}$ .

Définissons  $t_{\max} = \sqrt{2 \log m - 2 \log \log m}$  puis

$$t^* = \inf \left\{ 0 \leq t \leq t_{\max} : \frac{2m(1 - \Phi(t))}{R(t) \vee 1} \leq \alpha \right\}, \quad (4.3)$$

où  $R(t) = \sum_{i=1}^m \mathbb{1}_{\{t_i \leq -t\}}$  est le nombre d'hypothèses nulles rejetées au niveau  $t$ . Si l'infimum en (4.3) est  $+\infty$ , on pose  $t^* = \sqrt{2 \log m}$ .

On rejette  $\mathcal{H}_0$  si  $t_i \leq -t^*$  ce qui amène à définir les  $q$ -valeurs hiérarchiques

$$q_i^h = \frac{\mathfrak{p}_i^h \alpha}{\Phi(-t^*)},$$

et l'hypothèse nulle associée est rejetée si  $q_i^h < \alpha$ . Cette procédure contrôle le FDR au niveau  $\alpha$  (Javanmard et al., 2019).

*Remarque.* Comme  $t^*$  dépend de  $\alpha$ , les  $q$ -valeurs dépendent aussi de  $\alpha$  et doivent être calculées pour chaque niveau de FDR cible, contrairement à la procédure BH( $\alpha$ ) standard.

## 4.3 Évaluation de la méthode

### 4.3.1 Données simulées

Afin d'évaluer la qualité de notre procédure, nous avons simulé des données de manière non-paramétrique, comme présenté dans la section 3.4.3 et dans la figure 3.7 : à partir d'un jeu de données homogènes, on sélectionne des taxons différentiellement abondants, on assigne un groupe  $A$  ou  $B$  à chacun des échantillon et on multiplie les abondances de chaque taxon différentiellement abondant dans le groupe  $B$  par un *fold-change* prédéterminé.

Trois déclinaisons de simulations ont été adoptées. Dans la première, dite positive, les taxons sont sélectionnés pour former des groupes dans la phylogénie et un *fold-change* de 3, 5 ou 10 est ensuite appliqué. Ceci permet de créer des simulations pour lesquelles l'arbre est réellement informatif. Plus précisément, pour sélectionner les taxons différentiellement abondants, on applique un algorithme des  $k$ -médoïdes (Reynolds, Richards, Iglesia, & Rayward-Smith, 2006) à la matrice des distances patristiques (Sneath, Sokal, & others, 1973), ce qui donne des groupes de taxons à faible distance patristique les uns des autres, qui correspondent généralement à un sous-arbre de la phylogénie. Un ou plusieurs groupes sont alors sélectionnés aléatoirement et leurs taxons sont déclarés différentiellement abondants.

Les deux autres déclinaisons de la procédure de simulation de jeux de données sont dites négatives. Pour l'une, les taxons sont sélectionnées uniformément dans

l’arbre puis un *fold-change* de 5 est appliqué, pour créer des simulations dans lesquelles le modèle est mal spécifié. Pour l’autre le *fold-change* appliqué vaut 1, ce qui permet de voir comment réagit l’algorithme lorsqu’aucun taxon n’est différentiellement abondant.

La figure 4.2 présente les résultats dans le cas des simulations positives. Si les procédures de corrections classiques (BH et BY) contrôlent bien le FDR à 5 %, ce n’est pas le cas pour les procédures hiérarchiques. Dans la majorité des cas, le FDR des procédures hiérarchiques reste en dessous de 6 % mais pour la procédure à système de score (ss) avec un fold-change de 5 et pour celle avec l’inverse généralisé par colonne (ci) avec in fold-change de 3, il passe à 8 et 9 % respectivement. Dans toutes les configurations, BY a le TPR le plus faible, BH et *TreeFDR* (tf) se comportent de manière semblable, conformément aux résultats présentés dans Bichat et al. (2020), et les deux variantes de *zazou* obtiennent le meilleur TPR.

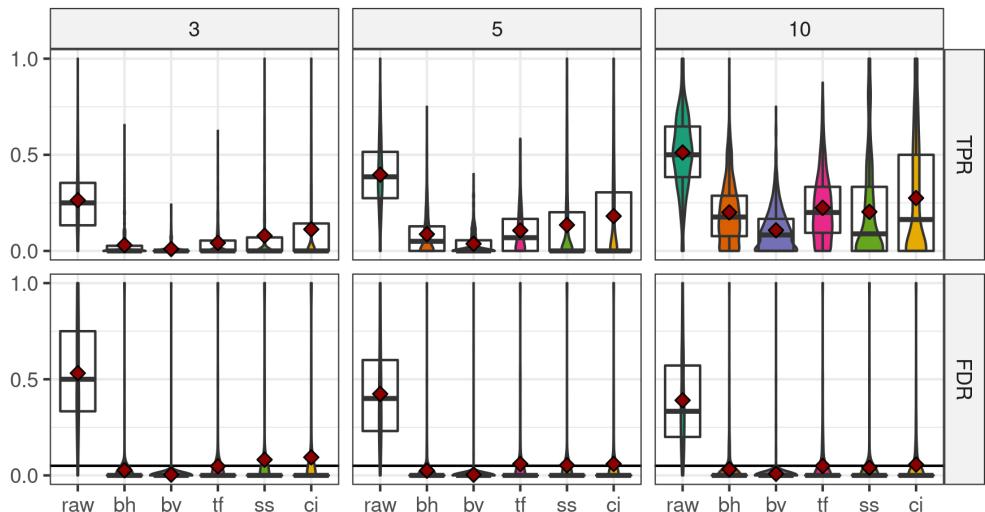


FIGURE 4.2: TPR (haut) et FDR (bas) pour les différentes procédures et différents *fold-changes* (en colonnes) dans le cadre des simulations positives.

Le taux de faux positif plus élevé qu’attendu pour les résultats de *zazou* suggère que le choix du seuil de détection proposé dans Javanmard et al. (2019) n’est pas complètement adapté. Nous comparons donc les performances des différentes méthodes à l’aide de l’AUC, qui est une mesure indépendante du seuil. La figure 4.3 met en évidence les meilleures performances de *zazou*, dans ses deux variantes. En regardant la partie gauche des courbes ROC, on s’aperçoit que *zazou* est plus performant dès les premières découvertes et que cela n’est pas un effet compen-

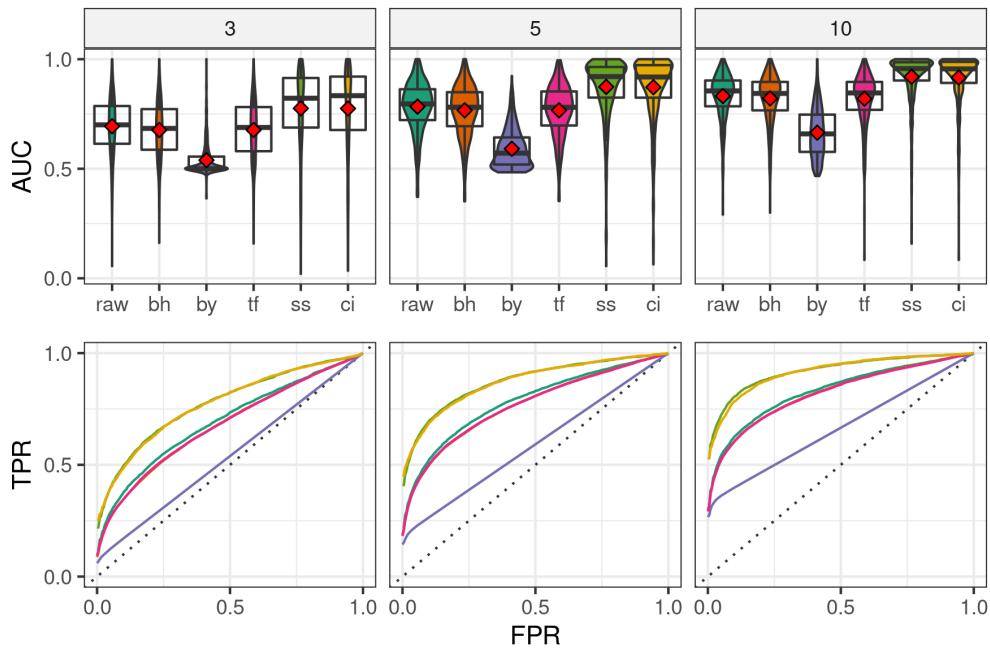


FIGURE 4.3: Distribution des AUC (haut) et courbes ROC (bas) pour les différentes procédures et *fold-changes* (en colonnes) dans le cadre des simulations positives.

satoire des seuils plus élevés. Comme mentionné précédemment, BH et *TreeFDR* obtiennent des performances similaires et moins bonnes que celles de *zazou*. BY est la méthode la moins satisfaisante.

Dans le cas des simulations négatives (figure 4.4), l'imposition d'une contrainte hiérarchique inadaptée fait perdre 15 à 20 points d'AUC à *zazou* par rapport à BH. Ce phénomène ne se retrouve pas dans les résultats de *TreeFDR*, qui est pourtant également une procédure hiérarchique, grâce à une astuce d'implémentation. En effet, *TreeFDR* effectue une correction de BH en parallèle de la procédure de lissage, et si cette dernière détecte bien moins de taxons que BH, les résultats de BH sont renvoyés à la place de ceux obtenus par lissage (Bichat et al., 2020; Xiao et al., 2017).

Enfin, pour les simulations où aucun taxon n'est différentiellement abondant ( $fc = 1$ ), les procédures ne détectent aucun faux positif, comme attendu.

### 4.3.2 Influence de l'âge

Nous avons comparé l'effet des différentes procédures de correction lors d'une analyse d'abondance différentielle entre des 112 adultes et 34 enfants au sein du jeu

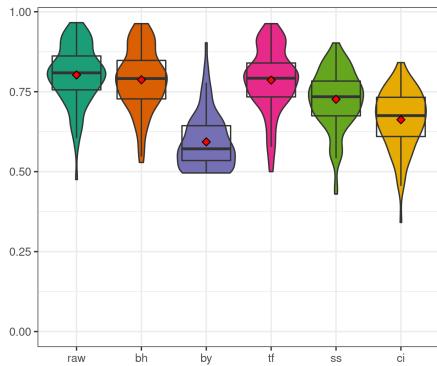


FIGURE 4.4: Distribution des AUC pour les différentes procédures lorsque les taxons différentiellement abondants sont sélectionnés uniformément.

de données de Brito et al. (2016) des populations insulaires. Des tests de Wilcoxon ont été effectués sur les 387 espèces présentes et 21 espèces ont été détectées à 5 % sans correction. Après correction par BH, BY, *TreeFDR* ou *treeclimbR*, aucune espèce n'est détectée. En revanche, *zazou* en détecte certaines avec ses deux variantes : 17 pour ss et 6 pour ci.

La figure 4.5 montre que les espèces détectées par *zazou* ne forment pas un sous-ensemble de celles obtenues sans correction. Au contraire, *zazou* identifie des espèces proches de certaines détectées sans correction, majoritairement dans la zone mise en évidence par le bandeau rouge.

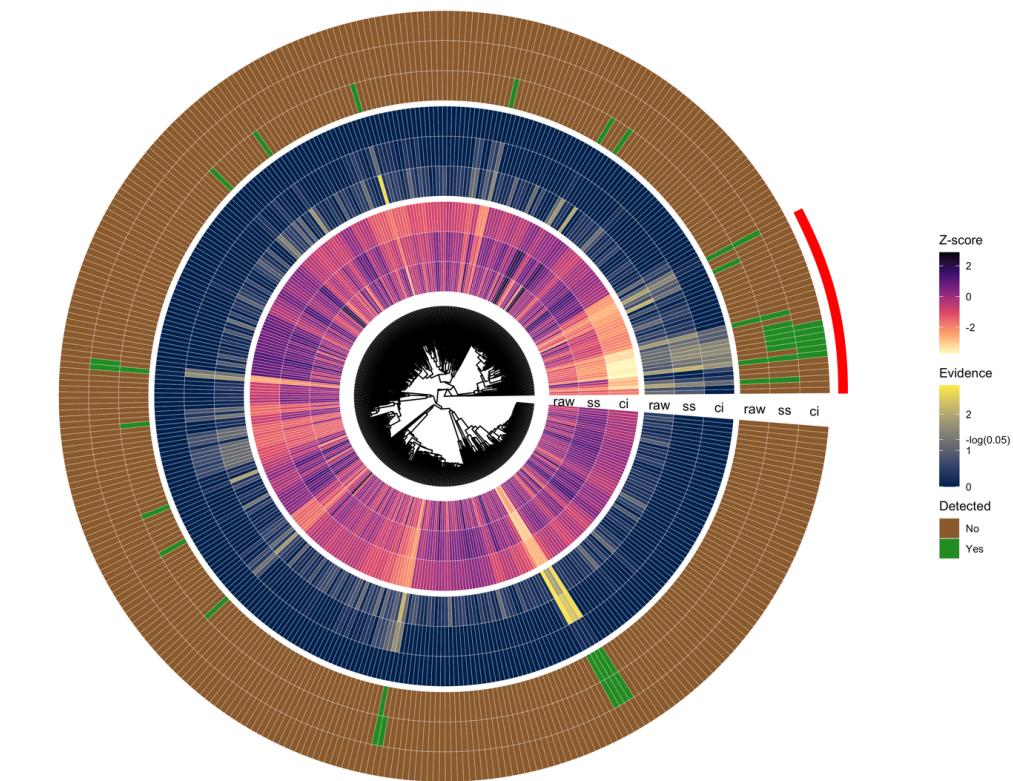


FIGURE 4.5: Phylogénie des 387 espèces du jeu de données des Fidjiens. Les cercles intérieur, central et extérieur représentent respectivement les  $z$ -scores bruts, les évidences corrigées et les statuts détecté ou non associé aux espèces pour différentes procédures (sans correction et les deux variantes de *zazou*).



# Chapitre 5

## Problèmes d'analyse numérique

Ce chapitre présente trois résolutions que nous avons apportées aux problèmes d'analyses numériques rencontrés lors de l'élaboration de l'algorithme *zazou*, présenté dans la section 4.2.

Ce sont des algorithmes itératifs qui permettent de converger jusqu'à une approximation raisonnable de la solution.

### 5.1 Algorithme du *shooting*

L'algorithme du *shooting* est une méthode de résolution numérique du lasso proposée par Fu (1998). Nous allons ici décrire la variante que nous avons développée lorsque l'espace d'existence du paramètre est contraint.

Le problème (4.2) peut-être réécrit dans sa forme générale

$$\theta^* = \underset{\theta \in \mathbb{R}^p, U\theta \in \mathbb{R}^q_-}{\operatorname{argmin}} \|y - X\theta\|_2^2 + \lambda\|\theta\|_1, \quad (5.1)$$

où  $y \in \mathbb{R}^n$ ,  $X \in \mathbb{R}^{n \times p}$  et  $U \in \mathbb{R}^{q \times p}$ . Il s'agit d'un problème convexe et donc nous pouvons facilement adapter l'algorithme initial pour résoudre itérativement un problème unidirectionnel sur chaque coordonnée.

Pour isoler la  $j$ ème coordonnée  $\theta_j$ , nous décomposons les termes du problème en

$$\|y - X\theta\|_2^2 + \lambda\|\theta\|_1 = \|y - X_{-j}\theta_{-j} - \theta_j x_j\|_2^2 + \lambda|\theta_j| + \lambda\|\theta_{-j}\|_1$$

et

$$U\theta = U_{-j}\theta_{-j} + \theta_j u_j.$$

Ainsi, le problème unidirectionnel associé à la  $j$ ème coordonnée du problème (5.1) est

$$\begin{cases} \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} h(\theta) = \frac{1}{2}\|y - z - x\theta\|_2^2 + \lambda|\theta| \\ \text{t.q. } u + v\theta \leq 0 \end{cases} \quad (5.2)$$

Notons  $I_+ = \{i : v_i > 0\}$  et  $I_- = \{i : v_i < 0\}$  puis  $\theta_{\min} = \max_{i \in I_+} -\frac{u_i}{v_i}$  et  $\theta_{\max} = \min_{i \in I_-} -\frac{u_i}{v_i}$ .

Le problème (5.2) est faisable si et seulement si

1.  $\theta_{\min} \leq \theta_{\max}$ ,
2. pour chaque  $i$ ,  $v_i = 0$  entraîne  $u_i \leq 0$ .

Sous ces deux conditions, la région de faisabilité est  $\mathcal{I} = [\theta_{\min}, \theta_{\max}]$ .

Le sous-gradient de  $h$  est

$$\partial h(\theta) = \begin{cases} -(y - z)^T x + x^T x \theta - \lambda & \text{si } \theta < 0, \\ -(y - z)^T x + x^T x \theta + \lambda & \text{si } \theta > 0, \\ -(y - z)^T x + \lambda[-1, 1] & \text{si } \theta = 0, \end{cases}$$

et les potentiels minimiseurs de  $h$ , tels que  $0 \in \partial h(\theta)$ , sont alors

$$\begin{cases} \frac{(y - z)^T x + \lambda}{x^T x} & \text{si } (y - z)^T x < -\lambda, \\ \frac{(y - z)^T x - \lambda}{x^T x} & \text{si } (y - z)^T x > \lambda, \\ 0 & \text{si } |(y - z)^T x| \leq \lambda. \end{cases} \quad (5.3)$$

Si le problème (5.2) est faisable, par convexité de  $h$ , sa solution est obtenue en projetant (5.3) sur l'ensemble de faisabilité  $\mathcal{I}$  :

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

où  $P_{\mathcal{I}} : x \mapsto \max(\theta_{\min}, \min(x, \theta_{\max}))$  est la projection sur  $\mathcal{I}$ .

## 5.2 Minimisation sous contrainte de $x^T A x$

Soit  $A \in \mathbb{R}^{n \times n}$  une matrice définie positive,  $e$  un vecteur de  $\mathbb{R}^n$  et  $\gamma$  un réel strictement positif. Nous avons mis en place un algorithme de résolution au problème de minimisation suivant :

$$\begin{cases} x^* = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \quad x^T A x \\ \text{t.q. } \|Ax - e\|_\infty \leq \gamma. \end{cases} \quad (5.4)$$

D'après le théorème spectral, il existe  $U, D \in \mathbb{R}^{n \times n}$  et  $\lambda \in \mathbb{R}_+^n$  tels que

- $A = UDU^T$ ,
- $U^T U = \mathbf{I}_n$ ,
- $D = \text{Diag}(\lambda_1, \dots, \lambda_n)$ ,
- $\lambda_1 \geq \dots \geq \lambda_r > 0 = \lambda_{r+1} = \dots = \lambda_n$  avec  $r = \text{rang}(A)$ .

La solution  $x^*$  du problème initial (5.4) est liée à la solution  $x^\diamond$  du problème complémentaire

$$\begin{cases} x^\diamond = \underset{x \in \mathbb{R}^n}{\text{argmin}} \quad x^T D x \\ \text{t.q. } \|Bx - e\|_\infty \leq \gamma \end{cases} \quad (5.5)$$

par  $x^* = Ux^\diamond$ , avec  $B = UD$ .

La résolution du problème complémentaire (5.5) se fait en trouvant d'abord un  $x$  dans l'ensemble de faisabilité, par exemple avec la méthode proposée dans la section 5.3, puis en mettant à jour coordonnée par coordonné jusqu'à convergence.

Pour  $j \in \llbracket 1, n \rrbracket$ , le problème de minimisation de  $x^T D x$  en  $x_j$  peut s'écrire

$$\underset{x_j \in \mathbb{R}}{\text{argmin}} \lambda x_j^2 + \sum_{k \neq j} x_k^2 \lambda_k,$$

dont la solution non contrainte évidente est 0, qu'il faudra projeter sur l'ensemble de faisabilité.

Explicitons la contrainte en  $x_j$  donnée dans le problème (5.5). Écrivons d'abord la décomposition de  $Bx$  en  $Bx = B_{-j}x_{-j} + x_j b_j = c + x_j b_j$  avec  $c = B_{-j}x_{-j} \in \mathbb{R}^n$ .

On a alors

$$\begin{aligned} & \|Bx - e\|_\infty \leq \gamma \\ \iff & \|c + x_j b_j - e\|_\infty \leq \gamma \\ \iff & |c_k + x_j b_{k,j} - e_k| \leq \gamma \quad \forall k \\ \iff & -\gamma \leq c_k + x_j b_{k,j} - e_k \leq \gamma \quad \forall k \\ \iff & -\gamma + e_k - c_k \leq x_j b_{k,j} \leq \gamma + e_k - c_k \quad \forall k \\ \iff & x_j \in [l_k, u_k] = \mathcal{I}_k \quad \forall k \\ \iff & x_j \in [\max((l_k)_k), \min((u_k)_k)] = \bigcap_k \mathcal{I}_k = \mathcal{I} \end{aligned}$$

où

$$l_k = \begin{cases} \frac{-\gamma + e_k - c_k}{b_{k,j}} & \text{si } b_{k,j} > 0 \\ \frac{\gamma + e_k - c_k}{-b_{k,j}} & \text{si } b_{k,j} < 0 \\ \begin{cases} -\infty & \text{si } |c_k - e_k| \leq \gamma \\ +\infty & \text{sinon} \end{cases} & \text{si } b_{k,j} = 0 \end{cases}$$

et

$$u_k = \begin{cases} \frac{\gamma + e_k - c_k}{b_{k,j}} & \text{si } b_{k,j} > 0 \\ \frac{-\gamma + e_k - c_k}{-b_{k,j}} & \text{si } b_{k,j} < 0 \\ \begin{cases} +\infty & \text{si } |c_k - e_k| \leq \gamma \\ -\infty & \text{sinon} \end{cases} & \text{si } b_{k,j} = 0 \end{cases}$$

Le problème (5.5) est faisable si et seulement si  $\mathcal{I} \neq \emptyset$  si et seulement si  $\max((l_k)_k) \leq \min((u_k)_k)$ . Dans ce cas, le minimum sous contrainte est atteint en  $P_{\mathcal{I}}(0)$ , la projection de 0 sur  $\mathcal{I}$ .

*Remarque.* S'il existe un  $k$  tel que  $b_{k,j} = 0$  et  $|c_k - e_k| > \gamma$  alors  $\mathcal{I} = \mathcal{I}_k = \emptyset$  et le problème n'est pas faisable. À l'inverse, si pour tout  $k$ ,  $b_{k,j} = 0$  et  $|c_k - e_k| \leq \gamma$  alors  $\mathcal{I} = \mathbb{R}$  et la solution est 0.

L'algorithme de résolution avance en mettant à jour la  $j^{\text{ème}}$  coordonnée de  $x^\diamond$  à  $P_{\mathcal{I}}(0)$  et continue d'itérer sur les coordonnées jusqu'à convergence.

### 5.3 Projection sur un ensemble de faisabilité

Nous appliquons ici le problème de projection sur une intersection d'ensembles convexes au cas rencontré dans le problème (5.5), c'est à dire quand l'ensemble est de la forme  $\|Mx - e\|_\infty < \gamma$ .

Formellement, si  $M \in \mathbb{R}^{m \times n}$ ,  $e \in \mathbb{R}^m$  et  $\gamma > 0$ , nous cherchons à obtenir un point dans  $\mathcal{C} = \{x \in \mathbb{R}^n, \|Mx - e\|_\infty < \gamma\}$ , que nous supposons non vide.

Définissons  $f : x \mapsto \max(f_1(x), \dots, f_m(x), -\varepsilon)$ , où  $f_j : x \mapsto |m_j^T x - e_j| - \gamma$  et  $\varepsilon > 0$ . Comme  $\mathcal{C} \neq \emptyset$ ,  $\operatorname{argmin} f \subset \mathcal{C}$  et  $\min f = -\varepsilon$ .

Nous utilisons une descente du gradient dont les itérations successives sont donnés par  $x^{(0)} \in \mathbb{R}^n$  et

$$x^{(k+1)} = x^{(k)} - \alpha_k g_k$$

où  $\alpha_k$  est le pas et  $g_k = \partial f(x^{(k)})$ .

Comme  $\min f = -\varepsilon$  est connu, nous pouvons utiliser la longueur de pas de Polyak (Polyak, 1987)

$$\alpha_k = \frac{f(x^{(k)}) + \varepsilon}{\|g_k\|_2^2},$$

qui est optimale en un certain sens.

Il nous reste maintenant à expliciter  $g_k$ .

Si pour chaque  $i \in \llbracket 1, m \rrbracket$ ,  $f_i(x^{(k)}) < -\varepsilon$  alors  $x^{(k)} \in \mathcal{C}$  et l'algorithme peut s'arrêter.

Dans le cas contraire, il existe  $j \in \llbracket 1, m \rrbracket$  tel que  $f(x^{(k)}) = f_j(x^{(k)})$  et alors

$$\begin{aligned} g_k &= \partial f_j(x^{(k)}) \\ &= \begin{cases} m_j & \text{si } m_j^T x^{(k)} - e_j > 0, \\ [-m_j, m_j] & \text{si } m_j^T x^{(k)} - e_j = 0, \\ -m_j & \text{si } m_j^T x^{(k)} - e_j < 0, \end{cases} \end{aligned}$$

puis  $\|g_k\|_2^2 = m_j^T m_j$ .



# Conclusion et perspectives

Dans cette thèse, nous avons exploré différentes pistes d'amélioration des méthodes d'analyse d'abondance différentielle en métagénomique tirant parti d'une information hiérarchique sur les taxons.

Après avoir présenté la métagénomique et les méthodes d'abondance différentielle, hiérarchiques ou non, nous avons introduit dans le chapitre 3 l'arbre des corrélations. Par construction, cet arbre regroupe les taxons aux profils d'abondance similaires. Intuitivement, les taxons différemment abondants devraient être regroupés dans l'arbre, c'est à dire y former des sous-arbres. Ce regroupement de signal permet aux procédures hiérarchiques utilisant l'arbre des corrélations d'être plus puissantes que celles utilisant la taxonomie ou la phylogénie. Cependant, les méthodes existantes souffrent de gros désavantages : le lissage des  $z$ -scores de Xiao et al. (2017) a du mal à prendre en compte l'arbre et son avantage comparatif par rapport à BH vient avant tout (i) de sa deuxième étape de calcul des  $p$ -valeurs par permutation et de (ii) son implémentation, qui se rabat sur BH lorsque la procédure hiérarchique ne retrouve pas assez de taxons identifiés par BH. De même, le FDR hiérarchique ne permet pas de contrôler le FDR à un niveau spécifié *ex ante* mais uniquement de calculer un niveau de contrôle *ex post*, à l'issue de la procédure. Dans sa formulation la plus naïve, il est même incapable de descendre de la racine et d'explorer l'arbre lorsqu'on l'applique à certaines types de données, par exemple les données compositionnelles. En conséquence, il est préférable d'utiliser la procédure BH plutôt que ces premières méthodes hiérarchiques.

Devant l'absence de procédures hiérarchiques satisfaisantes, nous avons développé notre propre méthode d'analyse d'abondance différentielle hiérarchique, présentée dans le chapitre 4 et baptisée *zazou*. Il s'agit d'une combinaison de quatre méthodes statistiques connues que nous avons adaptées à notre problème :

- une modélisation des  $z$ -scores par un processus d'Ornstein-Uhlenbeck sur un arbre avec sauts,
- une régression lasso pour estimer de façon parcimonieuse la localisation des sauts dans l'arbre,
- une procédure de débiaisage du lasso,

- une correction pour tests multiples adaptée au lasso débiaisé.

Lorsque les taxons différentiellement abondants respectent la structure de l'arbre, *zazou* est une procédure plus puissante que la procédure BH. Ce n'est en revanche plus le cas lorsque les espèces différentiellement abondantes sont réparties aléatoirement dans l'arbre : la régularisation indue par un arbre non informatif fait alors perdre de la puissance statistique.

Enfin, nous avons proposé des solutions effectives aux problèmes d'analyse numérique rencontrés lors de l'implémentation de cette procédure dans le package *{zazou}*.

## Détection versus prédition

Jusqu'à présent, nous nous sommes seulement intéressés à la détection de taxons différentiellement abondants en répondant à la question : « à partir du statut des échantillons, que peut-on dire des taxons qui les composent ? ». Il est possible de retourner ce problème pour en faire un problème de prédition : « à partir de la composition des échantillons, que peut-on dire sur leur statut ? ». Ce problème n'a pas été abordé dans ce manuscrit et nous allons tenter de donner ici quelques pistes de réflexion.

L'approche la plus classique pour effectuer des prédictions est la régression linéaire. Dans ce cadre mathématique, Park, Hastie, & Tibshirani (2007) ont montré que si des groupes de prédicteurs sont suffisamment corrélés, il est plus intéressant –du point de vue de l'erreur de prédition– d'effectuer une régression simple sur la moyenne de leur composantes plutôt que de déterminer un coefficient par variable.

Illustrons ceci par un exemple où il n'y a qu'un seul bloc de variables corrélées et standardisées  $x_1, \dots, x_m$  telles que

$$X^T X = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}.$$

Au lieu de modéliser la variable réponse  $y$  par  $y_i \sim \sum_j \beta_j x_{i,j} + \varepsilon_i$ , il est préférable d'utiliser  $y_i \sim \beta^* \sum_j \frac{x_{i,j}}{m} + \varepsilon_i$  dès lors que  $\rho$  est suffisamment proche de 1.

Ce résultat est intéressant mais l'analyse est limitée à un design peu réaliste, où la corrélation est identique entre toutes les paires de variables. Des extensions à des designs plus réalistes, avec des groupes disjoints de variables corrélées ou une structure de corrélation connue entre les variables –par exemple de type OU sur un arbre– seraient intéressantes pour des applications à l'étude du microbiote.

## Autres procédures hiérarchiques

Si l'on regarde les procédures hiérarchiques existantes, elles prennent toutes en entrée un vecteur de  $p$ -valeurs ou de  $z$ -scores. Pour tester ces méthodes, nous avons eu recours à des tests non paramétriques sur les rangs, comme Wilcoxon ou Kruskal-Wallis. Il serait alors intéressant de regarder la combinaison de tests conçus spécifiquement pour les données métagénomiques, comme ceux présentés dans la section 2.1.4, avec des procédures hiérarchiques pour évaluer la plus-value de la hiérarchie pour des procédures de tests sophistiquées.

Dans les méthodes actuelles, la hiérarchie intervient toujours dans un deuxième temps, en tant que structure de lissage ou de correction. À notre connaissance, il n'existe pas de test qui prenne directement en compte cette hiérarchie. Des modèles linéaires à matrice de variance-covariance conditionnée par une hiérarchie pourraient convenir.

Les plus proches sont les modèles gaussiens multivariées utilisées en méthodes comparatives phylogénétiques :

$$\text{vec}(Z) \sim \mathcal{N}_{np}(0, \Sigma \otimes R)$$

dans lesquelles  $\Sigma$  capture la dépendances entre taxons (typiquement la matrice de covariance d'un OU sur arbre),  $R$  la dépendance entre observations au sein d'un taxon et où  $\otimes$  est le produit de Kronecker. Ces modèles peuvent être adaptés à des données de comptage en adoptant une approche hiérarchique :  $Z$  est une variable latente et les comptages  $Y$ , conditionnellement à  $Z$ , sont indépendants et tirées dans une loi discrète, typiquement Poisson ou binomiale négative (avec ou sans excès de zéros). S'ils permettent théoriquement sous cette forme de gérer les données de comptages typiquement observées en métagénomique, ils sont plutôt utilisées en pratique pour faire de la détection de ruptures en écologie (Bastide, Ané, Robin, & Mariadassou, 2018) et ne permettent pas à notre connaissance de faire des tests d'abondance différentielle.

## Amélioration de zazou

*zazou* souffre d'un léger problème lors de l'inférence des paramètres du processus d'Ornstein-Uhlenbeck sur l'arbre : il arrive parfois que le modèle choisi par le critère BIC ne contienne aucun saut. Si la procédure de débiaisage fait qu'il est ensuite tout de même possible de détecter des taxons différemment abondants, on aimerait pouvoir être moins strict dans notre sélection de modèle. D'autres procédures de sélection de modèle via un BIC phylogénétique (Khabbazian, Kriebel, Rohe, & Ané, 2016) ou une heuristique de pente (Baudry, Maugis, & Michel, 2012) ont été testées en alternative au BIC traditionnel mais les résultats ne sont pas

convaincants. Dans le contexte de *zazou*, le critère idéal sélectionnerait trop de sauts plutôt que pas assez, l'étape final de tests permettant de filtrer les taxons issus des sauts identifiés à tort.

Une autre piste d'amélioration consiste à adapter l'étape de débiaisage aux hypothèses de notre problème. Lors de l'inférence ponctuelle des sauts, le modèle est en effet conditionné pour que les moyennes aux feuilles soient à valeurs dans  $\mathbb{R}_+$ . Cependant, lors du débiaisage, les valeurs aux feuilles ne sont plus contraintes à vivre dans cette demi-droite. L'imposition de contraintes rend l'analyse plus complexe mais il serait intéressant de vérifier son impact sur les propriétés théorique du débiaisage.

Enfin et dans la lignée du point précédent, bien que l'algorithme se montre efficace, les hypothèses requises pour avoir les résultats théoriques des différentes méthodes employées dans *zazou* ne sont pas complètement respectées. Il serait intéressant de creuser cet aspect théorique.

# Digest

## Chapter I

Microbiome, defined as the collection of microbes that inhabit a given environment, and metagenome, defined as the collection of their genes, have been the focus of growing attention for over a decade. In the human gut, microbial communities are responsible for carbohydrate degradation (Flint et al., 2012 ; Rowland et al., 2018) and help the immune system to regulate inflammation (Blander et al., 2017). However, microbiome deregulation can also lead to moderate to severe disease, like Crohn's disease (Morgan et al., 2012), depression (Foster & Neufeld, 2013), or necrotizing enterocolitis (Mai et al., 2011). Food and pharmaceutical industries saw the tremendous potential of the gut microbiome began developing product to take advantage of it. Among a lot of different methods, probiotics to lower depressive symptoms (Pinto-Sanchez et al., 2017), prebiotics to enhance specific bacteria and lower obesity's risk (Sakwinska et al., 2017) or fecal transplant to fight against lethal forms of diarrhea (Van Nood et al., 2013) have been tested and proved to be highly efficient.

Several techniques are available to establish the composition of the microbiome. The marker gene sequencing approach, also called metabarcoding, aims at targeting a universal but rapidly evolving gene (usually the 16S rRNA gene when studying bacteria) whose sequence acts as a barcode to identify the species it originates from (Morgan & Huttenhower, 2012). However, it suffers from several drawbacks like the restriction to bacteria and archaea or the impossibility to obtain any information on the biological functions present in or expressed by the community. To bypass those limitations, the whole genome shotgun (WGS) sequencing approach amplifies all the genetic material to reconstruct a functional and more complete view of the sample. It comes however with added complexity : the sequences must be mapped to preconstructed reference catalogues (Quince et al., 2017).

By construction, metagenomics data consist of counts (number of reads per species or per genes) and can be modeled as such. Several models try to take into account overdispersion by using negative binomial models (Zhang et al., 2017) or

overrepresentation of zeros by using zero-inflated models (Xinyan et al., 2016). However, several authors also suggest that the counts are misleading and that metagenomics data are compositional in essence and should be analyzed using an appropriate framework (Gloor & Reid, 2016).

## Chapter II

In order to detect taxa that are differentially abundant between groups, several procedures have been developed. Among them, Wilcoxon-Mann-Whitney (Mann & Whitney, 1947; Wilcoxon, 1992) and Kruskall-Wallis (Kruskal & Wallis, 1952) tests are generic rank-based tests whose null hypothesis is a common count distribution in all groups. There also are specific tests tailored to the count or compositional aspect of sequencing-based omics data like *edgeR* (Robinson et al., 2010), *DESeq2* (Love et al., 2014), *mbzinb* (Chen et al., 2018) or *ALDEx2* (Fernandes et al., 2014). The former two were first designed for transcriptomics data and then imported to microbiome data whereas the latter two were designed from the ground up for microbiome data.

All the previously mentioned tests are univariate : they test one taxon at the time for differential abundance and require a subsequent step of multiple testing correction to avoid a high number of false discoveries. The Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) is the best known and most popular of those correction procedures : it control the expected proportion of false discoveries  $\mathbb{E} \left[ \frac{FP}{TP+FP} \right]$  at a certain level, defined *a priori*.

Recently, new procedures have been proposed to take a hierarchical structure (mainly phylogeny) into account during differential abundance analyses. *TreeFDR* (Xiao et al., 2017) is a two-steps procedure based on smoothing of *z*-scores followed by a permutational correction technique. The smoothing is based on a hierarchical model whose hyperparameters control the level of smoothing. Hierarchical FDR, theorized by Yekutieli (2008) and implemented by Sankaran & Holmes (2014), considers instead a top-down approach : it progresses down the tree, from the root to the leaves, and tests increasingly smaller groups of taxons until it can not reject the null hypothesis anymore. Finally, Huang et al. (2020) developed *treeclimbR*, which select differentially abundant clades based on a composite score computed for each internal node of the tree.

## Chapter III

This chapter addresses the questions of which tree to consider in hierarchical approaches. Phylogeny reflects the evolution history between taxa and is created

from nucleotidic sequences from the dataset. Taxonomy is a highly polytomic tree of nested ranks that is available in reference databases (Geer et al., 2010). We introduce the correlation tree, a tree built from the pairwise correlation matrix of abundance data using a hierarchical clustering procedure.

To compare those trees, we consider several distances. Formally, rooted trees are directed acyclic networks : they can be compared using the Robison-Foulds distance (Robinson & Foulds, 1981) which focuses solely on topologies (trees without branches lengths), the cophenetic distance (Sokal & Rohlf, 1962) which focuses on path lengths in the tree, or the Billera-Holmes-Vogtmann distance (Billera et al., 2001) which embeds trees in a geometric space and consider shortest paths in that space.

We first investigate whether the phylogeny (or the taxonomy) is close to the correlation tree. We perform pairwise BHV distance computation on forest of trees covering the phylogeny, the correlation tree, bootstrapped versions of correlation tree and the random trees followed by a PCoA. It appears that the phylogeny is neither in the confidence region of the correlation tree, nor closer to the correlation tree than a random tree.

We then apply hierarchical procedures with the different trees. The correlation tree performs better than the phylogeny, as it naturally captures similarity of abundance profiles and thus groups taxons in a relevant way. However, the classical BH procedure outperforms the hierarchical procedure, no matter what tree is used.

## Chapter IV

The Ornstein-Uhlenbeck process with optimal value  $\beta_{ou}$  is defined as the solution of the stochastic differential equation :

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

It is a well suited framework to model evolution of continuous phylogenetic traits (Freckleton et al., 2003), especially because of its  $\beta_{ou}$  centered Gaussian limit law. It has been used as a basic block to build more complex models, by considering OU processes on a tree and by considering piecewise linear functions for  $\beta_{ou}$  (Bastide et al., 2017).

We proposed a new model, *zazou*, where the *z*-scores arise from an OU process on a tree with shifts on its optimal values. Moreover, we assume that under  $\mathcal{H}_1$ ,  $\mathbf{z}_i \sim \mathcal{N}(\mu_i, 1)$  with  $\mu_i < 0$  (McLachlan & Peel, 2004) : finding the alternative hypotheses is equivalent to finding the non-zero components of  $\mu$ . This can be reframed as a constrained version of the well-known lasso (Tibshirani, 1996) to have a point estimator of the  $\mu_i$ . We then enhance this estimator using two desparsifications procedures from Zhang & Zhang (2014) and Javanmard & Montanari (2014) to

debias the lasso estimate and build confidence intervals and in turn compute  $p$ -values for each of the components of  $\mu$ . Those  $p$ -values act as tree-smoothed  $p$ -values. The last part of *zazou* is the application of a multiple testing correction designed for desparsified lasso (Javanmard et al., 2019) on the computed  $p$ -values.

We evaluate our procedure on both synthetic and real data. When the tree is informative in the simulations, *zazou* outperforms BH and *TreeFDR* in terms of TPR but does not control the FDR at the correct level. Using a threshold independent approach, we show that *zazou* also outperforms competing methods in terms of ROC curves and AUC values. However, when the tree is not informative, forcing an irrelevant constraint leads to a significant loss of AUC.

## Chapter V

In this last chapter, we resolved three numerical analysis problems. The proposed algorithms are iterative and converge to a reasonable approximation of the solution.

The first one is a variation of the lasso with a linear constraint :

$$\begin{cases} \theta^* = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \\ \text{s.t. } U\theta \in \mathbb{R}_+^q \end{cases}$$

where  $y \in \mathbb{R}^n$ ,  $X \in \mathbb{R}^{n \times p}$  and  $U \in \mathbb{R}^{q \times p}$ .

The second one is a minimization problem of a quadratic form subject to an infinite norm constraint :

$$\begin{cases} x^* = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} x^T Ax, \\ \text{s.t. } \|Ax - e\|_\infty \leq \gamma, \end{cases}$$

where  $A \in \mathbb{R}^{n \times n}$  is a positive semidefinite matrix,  $e \in \mathbb{R}^n$  and  $\gamma > 0$ .

The final one is a feasibility problem, where the goal is to find an element  $x$  that satisfies :

$$\{x \in \mathbb{R}^m : \|Mx - e\|_\infty < \gamma\},$$

with  $M \in \mathbb{R}^{m \times n}$ ,  $e \in \mathbb{R}^m$  and  $\gamma > 0$  sufficiently large.

## Conclusions and outlooks

The issue considered in this manuscript is the detection of differentially abundant taxa among several groups. The inverse problem is a prediction problem where one wants to predict the groups from the taxa. Including hierarchical information for prediction has already been done in a weak manner for linear regression by

Park et al. (2007). The hierarchical information was the partition of variables in groups with high covariance among them. They proved that it is more interesting to estimate one coefficient per group than one per variable.

Existing hierarchical procedures take precomputed  $p$ -values or  $z$ -scores as input. Saying that, hierarchical information always come as a second step, for smoothing or correction. To our knowledge, there is no test that use this information.

*zazou* suffers from a minor problem during model selection. Sometimes, the BIC selected model has no shifts. Alternative model selection procedures (Baudry et al., 2012 ; Khabbazian et al., 2016) have been tested but without success. The optimal model selection procedure would select a model with too many shifts, the useless one will be removed by the other steps.

Some theoretical work is still requires. During the desparsification step to constraint shifts on branches such that their sum on the leafs are non-positive, and to check if the hypothesis of the several combined methods are respected.



# Annexe A

## Notations

Nous répertorions ici l'ensemble des notations classiques suffisamment courantes pour ne pas être explicitées au fil de l'eau.

$\mathbb{N}$  : ensemble des entiers naturels (0 inclus)

$\mathbb{R}$  : ensemble des réels

$\llbracket a, b \rrbracket : [a, b] \cap \mathbb{N}$

$E_+, E_-$  ou  $E^*$  :  $E$  restreint à ses éléments positifs, négatifs ou non nuls

$E^{n \times m}$  : ensemble des matrices à  $n$  lignes et  $m$  colonnes à coefficients dans  $E$

$M^T$  : transposée de  $M$

$|M|$  : déterminant de  $M$

$m_j$  :  $j^{\text{ème}}$  colonne de  $M$

$m_{i,j}$  : élément de  $M$  à l'intersection de la  $i^{\text{ème}}$  ligne et  $j^{\text{ème}}$  colonne

$M_{-j}$  : matrice obtenue lorsqu'on retire à  $M$  sa  $j^{\text{ème}}$  colonne

$\mathbf{I}_n$  : matrice identité de dimension  $n$

$\mathbf{1}_n$  : matrice colonne de taille  $n$  composée uniquement de 1

$\text{Diag}(x_1, \dots, x_n)$  : matrice diagonale dont la diagonale est  $x_1, \dots, x_n$  définie par

$(x_i \delta_{i,j})_{i,j \in \llbracket 1, n \rrbracket}$

$\|x\|_q$  : norme  $q \in \mathbb{N}^*$  d'un vecteur défini par  $(\sum_{i=1}^n |x_i|^q)^{\frac{1}{q}}$

$\|x\|_\infty$  : norme infinie d'un vecteur défini par  $\max\{|x_1|, \dots, |x_n|\}$

$\langle x, y \rangle$  : produit scalaire entre  $x$  et  $y$

$x \vee y$  : maximum entre  $x$  et  $y$

$f^{(n)}$  : composée  $n$ -ième de  $f$  avec elle-même

$\nabla^2 f$  : hessienne de  $f$  définie par  $\nabla^2 f : x \mapsto (\partial_i \partial_j f(x))_{i,j} \in \mathbb{R}^{n \times n}$

$\binom{n}{k}$  : coefficient binomial défini par  $\frac{n!}{k!(n-k!)}$

$\mathbb{E}[X]$  : espérance de  $X$

$\text{Var}[X]$  : espérance de  $X$

$\text{Cov}[X, Y]$  : covariance entre  $X$  et  $Y$

$\mathcal{U}(E)$  : loi uniforme sur  $E$

- $\mathcal{N}(\mu, \sigma^2)$  : loi normale de moyenne  $\mu$  et de variance  $\sigma^2$
- $\mathcal{N}_d(\mu, \Sigma)$  : loi normale multidimensionnelle de moyenne  $\mu$  et de matrice de variance-covariance  $\Sigma$  en dimension  $d$
- $\Phi$  : fonction de répartition de la loi normale centrée réduite
- $\mathcal{H}_0$  : hypothèse nulle
- $\mathcal{H}_1$  : hypothèse alternative
- $\mathbb{H}_0$  : indices des hypothèses dans l'ensemble des vraies hypothèses nulles
- $\mathbb{1}$  : indicatrice de l'événement  $A$  définie par  $\mathbb{1}_A = \begin{cases} 1 & \text{si } A \\ 0 & \text{sinon} \end{cases}$
- $\delta_{i,j}$  : symbole de Kronecker entre  $i$  et  $j$  défini par  $\delta_{i,j} = \mathbb{1}_{i=j}$
- $(x_{(1)}, \dots, x_{(n)})$  : réordonnement du vecteur  $x$  défini par  $x_{(1)} = \min_i x_i$  et  $x_{(p)} = \min_{y \in \{x_1, \dots, x_n\} \setminus \{x_{(1)}, \dots, x_{(p-1)}\}} y$  pour  $p \in \llbracket 2, n \rrbracket$

## Annexe B

### Productions scientifiques

#### **Quantifying the impact of tree choice in metagenomics differential abundance studies with R**

Ce poster a été présenté à *UseR! 2019*, et a remporté le premier prix dans la catégorie *Biostatistique*.

# Quantifying the impact of tree choice in metagenomics differential abundance studies with R

Antoine Bichat<sup>1,2</sup>, Mahendra Mariadassou<sup>3</sup>, Jonathan Plassais<sup>2</sup> and Christophe Ambroise<sup>1</sup>

1. LaMME - Université d'Évry-Val-d'Essonne; 2. Enterome; 3. MaIAGE - INRA

Microbiota	
<ul style="list-style-type: none"><li>Ecological community of microorganisms that resides in an environmental niche</li><li><math>10^{14}</math> bacteria in the gut among 1500 species</li><li>Associations with:<ul style="list-style-type: none"><li>metabolism (diet, obesity, drug absorption, ...)</li><li>diseases (IBD, allergies, diabetes...)</li><li>behavior (smokers, antibiotics, C-section...)</li><li>environment (pet, water...)</li></ul></li></ul>	

Objectives	
<ul style="list-style-type: none"><li>Find which bacteria are differentially abundant between two or more groups</li><li>Use a FDR multiple testing correction to prevent false positives (one test per bacteria)</li><li>Incorporate hierarchical information to increase power</li><li>Which tree?</li></ul>	

Hierarchical False Discovery Rate	
The z-scores $\mathbf{z} = \Phi^{-1}(\mathbf{p})$ are smoothed using the following hierarchical model:	
$\mathbf{z}   \mu \sim \mathcal{N}_n(\mu, \sigma^2 \mathbf{I}_m) \quad \mu \sim \mathcal{N}_m(\gamma \mathbf{1}, \tau^2 \mathbf{C}_\rho)$	
where $\mathbf{C}_\rho = (\exp(-2\rho \mathbf{D}_{i,j}))$ with $\mathbf{D}$ the patristic distance matrix between taxa from the tree. By applying Bayes's formula:	
$\mathbf{z} \sim \mathcal{N}_m(\gamma \mathbf{1}, \tau^2 \mathbf{C}_\rho + \sigma^2 \mathbf{I}_m)$	
$\mu^* = \left( \mathbf{I}_m + \frac{\sigma_0^2}{\tau_0^2} \mathbf{C}_{\rho_0}^{-1} \right)^{-1} \left( \frac{\sigma_0^2}{\tau_0^2} \mathbf{C}_{\rho_0}^{-1} \gamma \mathbf{1} + \mathbf{z} \right)$	
Finally, a permutation-based FDR control is applied on $\mu^*$	

## Take-home message

- The tree choice has little impact on detection power
- Benjamini-Hochberg procedure is still the most powerful method and the only one which respects the FDR control
- The ease of creating R packages greatly increases the reproducibility of analysis
- tidyverse** and especially list-columns allow to write elegant and efficient R code when manipulating non-standard structures (trees, statistical model outputs...)

## References

Xiao, Jian, Hongyan Cao, and Jun Chen. **False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing.** Bioinformatics 33.18 (2017): 2873-2881.

Bokulich, Nicholas A., et al. **Antibiotics, birth mode, and diet shape microbiome maturation during early life.** Science translational medicine 8.343 (2016): 343ra82-343ra82.

Opstelten, Jorrit L., et al. **Gut microbial diversity is reduced in smokers with Crohn's disease.** Inflammatory bowel diseases 22.9 (2016): 2070-2077.

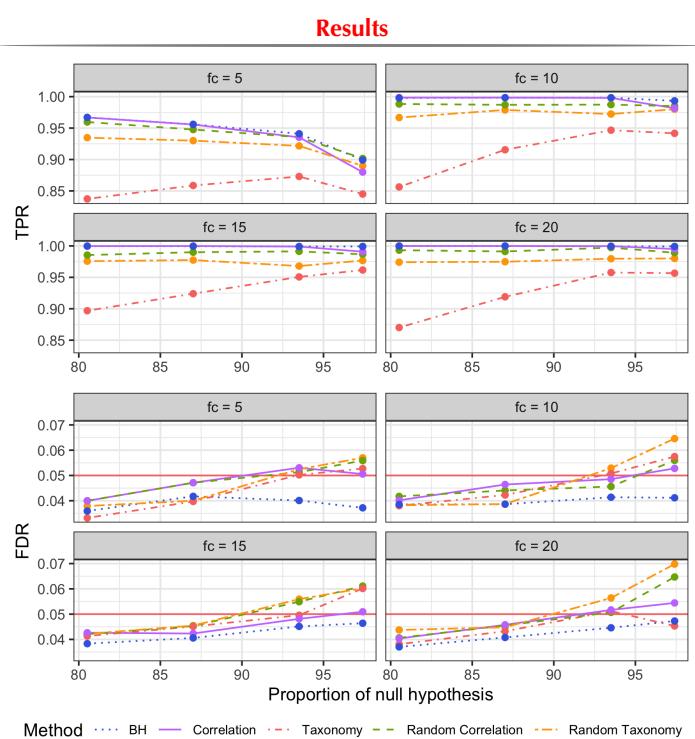
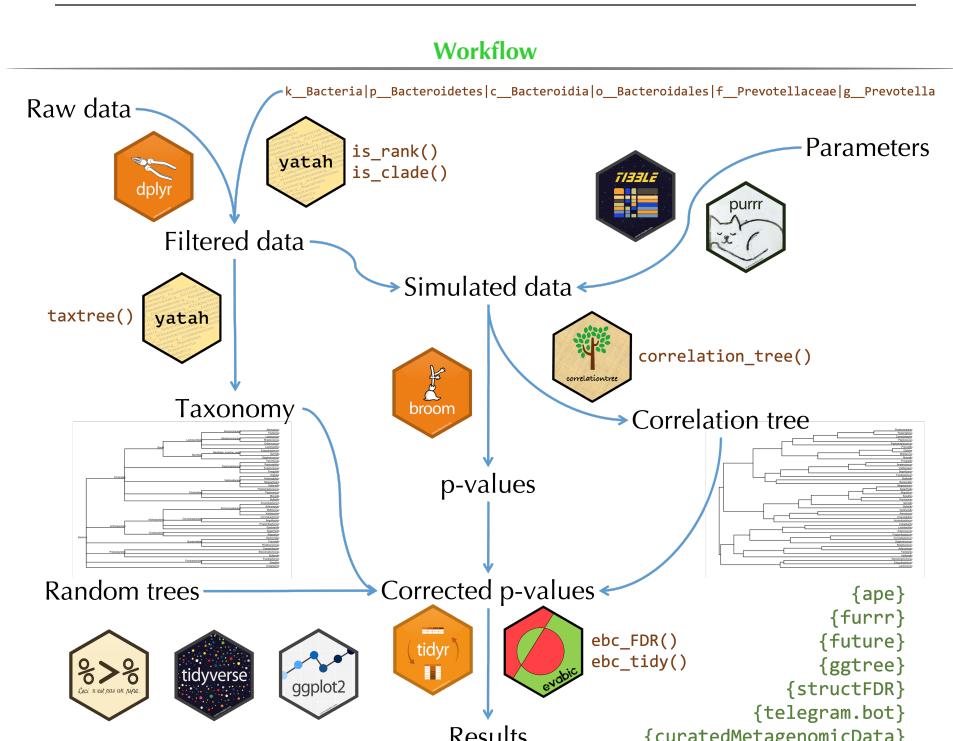
## Contact Information



abichat@enterome.com  
abichat.github.io  
antoinebichat  
@\_abichat  
@abichat



Data: taxonomy and abundance										
Phylum	Class	Order	Family	Genus	S001	S002	S003	S004	S005	...
Actinobacteria	Coriobacteriia	Coriobacteriales	Atopobiaceae	Atopobium	84	0	12	54	0	...
Actinobacteria	Coriobacteriia	Eggerthellales	Eggerthellaceae	Eggerthella	2	0	0	7	0	...
Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	525	7	134	753	0	...
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	88	1770	1490	119	2136	...
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	0	0	138	4	0	...
Firmicutes	Negativicutes	Veillonellales	Veillonellaceae	Dialister	152	4	2	192	0	...
Firmicutes	Negativicutes	Veillonellales	Veillonellaceae	Megasphaera	402	0	4	102	0	...
Fusobacteria	Fusobacteriia	Fusobacteriales	Leptotrichiaceae	Sneathia	302	0	35	272	0	...



## Incorporating Phylogenetic Information in Microbiome Differential Abundance Studies Has No Effect on Detection Power and FDR Control

Cet article a été publié dans *Frontiers in Microbiology*.



# Incorporating Phylogenetic Information in Microbiome Differential Abundance Studies Has No Effect on Detection Power and FDR Control

Antoine Bichat<sup>1,2</sup>, Jonathan Plassais<sup>2</sup>, Christophe Ambroise<sup>1</sup> and Mahendra Mariadassou<sup>3\*</sup>

<sup>1</sup> LaMME, Université Paris-Saclay, CNRS, Université d'Évry val d'Essonne, Évry, France, <sup>2</sup> Enterome, Paris, France, <sup>3</sup> MalAGE, INRAE, Université Paris-Saclay, Jouy-en-Josas, France

## OPEN ACCESS

### Edited by:

Guillermina Hernandez-Raquet,  
Institut National de Recherche pour  
l'Agriculture, l'Alimentation et  
l'Environnement (INRAE), France

### Reviewed by:

Marlis Reich,  
University of Bremen, Germany  
Leo Mikael Lahti,  
University of Turku, Finland

### \*Correspondence:

Mahendra Mariadassou  
[mahendra.mariadassou@inrae.fr](mailto:mahendra.mariadassou@inrae.fr)

### Specialty section:

This article was submitted to  
Systems Microbiology,  
a section of the journal  
*Frontiers in Microbiology*

Received: 02 August 2019

Accepted: 20 March 2020

Published: 15 April 2020

### Citation:

Bichat A, Plassais J, Ambroise C and  
Mariadassou M (2020) Incorporating  
Phylogenetic Information in  
Microbiome Differential Abundance  
Studies Has No Effect on Detection  
Power and FDR Control.  
*Front. Microbiol.* 11:649.  
doi: 10.3389/fmicb.2020.00649

We consider the problem of incorporating evolutionary information (e.g., taxonomic or phylogenetic trees) in the context of metagenomics differential analysis. Recent results published in the literature propose different ways to leverage the tree structure to increase the detection rate of differentially abundant taxa. Here, we propose instead to use a different hierarchical structure, in the form of a correlation-based tree, as it may capture the structure of the data better than the phylogeny. We first show that the correlation tree and the phylogeny are significantly different before turning to the impact of tree choice on detection rates. Using synthetic data, we show that the tree does have an impact: smoothing *p*-values according to the phylogeny leads to equal or inferior rates as smoothing according to the correlation tree. However, both trees are outperformed by the classical, non-hierarchical, Benjamini–Hochberg (BH) procedure in terms of detection rates. Other procedures may use the hierarchical structure with profit but do not control the False Discovery Rate (FDR) *a priori* and remain inferior to a classical Benjamini–Hochberg procedure with the same nominal FDR. On real datasets, no hierarchical procedure had significantly higher detection rate than BH. Intuition advocates that the use of hierarchical structures should increase the detection rate of differentially abundant taxa in microbiome studies. However, our results suggest that current hierarchical procedures are still inferior to standard methods and more effective procedures remain to be invented.

**Keywords:** microbiome, metagenomics, multiple testing, false discovery rate, correlation, phylogeny, taxonomy

## 1. INTRODUCTION

The microbiota, loosely defined as the collection of microbes that inhabit a given environment, has become an increasingly important research topic in the last two decades as it proves to either play an active role or be associated with health conditions (Lynch and Pedersen, 2016; Opstelten et al., 2016). For instance, specific changes in microbiome composition have been associated to Inflammatory Bowel Diseases (IBD) (Morgan et al., 2012) and liver cirrhosis (Qin et al., 2014). The microbiota also influences efficiency of cancer therapy (Routy et al., 2018) and there is a growing interest in finding biomarker microbes that could be used to predict the response to

treatment (Behrouzi et al., 2019). The effect of the microbiota is not limited to human health: works in plant biology show that the root microbiota can improve resistance to stress (Trivedi et al., 2017). Molecules produced by the microbiota can also have a profound impact on stress tolerance (Bernardo et al., 2017), plant health (Mendes et al., 2011), and pathogen control (Bartoli et al., 2018).

There are two main approaches to profile the microbiome using sequence data: amplicon sequencing and whole genome shotgun (WGS) sequencing. In amplicon sequencing, a marker-gene that acts a bacterial taxonomic "barcode" (e.g., the 16S rRNA gene) is first amplified and then sequenced. The resulting sequences are then used to build a taxonomic profile of the sample. By contrast, no prior amplification of a specific region is required for WGS sequencing as it sequences fragments from the whole metagenome. Although WGS sequencing is less affected by technical bias than amplicon sequencing and can profile both taxonomic and functional composition of the microbiome, it suffers from higher costs and requires complex bioinformatics pipelines. We focus in this work on taxonomic profiles.

In the amplicon approach, sequence reads are first clustered into Operational Taxonomic Units (OTUs) using either a 97% sequence similarity threshold (Caporaso et al., 2010), threshold-free agglomerative approaches (Mahé et al., 2015; Escudie et al., 2017) or divisive approaches to produce taxonomic oligotypes (Eren et al., 2015) or Amplicon Sequence Variants (ASVs) (Callahan et al., 2016). Divisive and threshold-free agglomerative approaches achieve finer taxonomic resolutions than the threshold-based similarity approach. Using WGS in the ecosystems where a bacterial gene catalog is available, such as the human gut (Li et al., 2014) or the pig gut (Xiao et al., 2016), the standard approach consists in mapping the reads against the catalog and then clustering the bacterial genes based on their abundance profiles to produce metagenomic species (MGS) (Nielsen et al., 2014) or clusters of co-abundant genes to reconstruct microbial pan-genomes (MSP) (Plaza-Oinate et al., 2018). We will refer to taxa, noting that the term can designate OTUs, ASVs, oligotypes, MGSs, MSPs and generally any feature found in abundance tables (obtained by counting the number of copies of each feature in each sample).

The microbial taxa share a common evolutionary history that can be encoded by a phylogenetic tree. For amplicon sequencing, the phylogenetic tree of taxa can even be reconstructed based on the sequence divergence of taxa (Price et al., 2010). Related taxa are generally thought to perform similar biological functions. For example, Philippot et al. (2010) shows a strong association between taxonomic lineage and ecological niche in soil microbiota. Chaillou et al. (2015) reports similar associations in food microbial ecosystems.

These associations suggest that the biological functions responsible for a given phenotype exhibit a phylogenetic signal and should thus be shared by closely related species. This prompted the development of several tree-based hierarchical methods, built under the assumption that taxa associated with a phenotype of interest are clustered in the tree (Martiny et al., 2017).

Carroll et al. (2014) considers group-based procedures, with groups defined as clades of the tree. Sankaran and Holmes (2014) proposes an implementation of the hierarchical testing procedure of Yekuteli (2008) aimed at leveraging the phylogenetic tree of the taxa to increase statistical power while controlling the False Discovery Rate (FDR). The FDR is unfortunately only known *a posteriori*, and the implemented testing-procedure is limited to one-way ANOVA with no correction for differences in sequencing depths. Matsen and Evans (2013) and Washburne et al. (2017) develop phylogenetic eigenvalues decomposition of species compositions for exploratory data analysis. Finally, Xiao et al. (2017) uses the tree as a regularization structure to shrink the test statistics of close-by taxa toward the same value. They use a permutational procedure to control the FDR and report good empirical control of the FDR but the method lacks theoretical grounding.

Unfortunately for phylogeny-based methods, the association between ecological niche and taxonomy reported in Philippot et al. (2010) holds for high-rank taxa but breaks down for lower-rank taxa. Indeed, phylogeny reflects the global evolutionary relatedness but the genes responsible for a specific phenotype may have a substantially different history, especially if they are transmitted horizontally rather than vertically, as is frequently the case for bacteria. In particular, mobile elements driving adaptation (Kazanjan, 2004) are likely to be spread out in the phylogeny (Brito et al., 2016) and the phylogenetic clades will not reflect their distribution across species. We question in this work the premise that the phylogenetic (or taxonomic) tree is the relevant hierarchical structure to incorporate in differential studies. We advocate instead the use of a correlation-tree: a clustering tree build from co-abundance data taxa, where taxa with highly correlated abundances are very close in the tree. We argue that the correlation tree is a better proxy of biological functions than the phylogeny and can increase the detection with no loss of FDR control.

Using the classical Billera–Holmes–Vogtmann (Billera et al., 2001) and Robinson–Foulds (Robinson and Foulds, 1981) distances on the treespace, we study the distance between the phylogenetic tree and the correlation trees in several previously published datasets. The datasets cover the vaginal microbiome (Ravel et al., 2011), the gut microbiome (Zeller et al., 2014), food-associated microbiomes (Chaillou et al., 2015) and microbiomes from a global survey (Caporaso et al., 2011). The former two have a narrow environmental range, as they encompass only one ecosystem, whereas the latter two have a broader range, as they encompass several ecosystems. We compare those distances to the average distance between (i) a focal tree (phylogeny or correlation) and a random tree and (ii) between two random trees to investigate the relationship between proximity in the tree and correlated abundances. We then assess the impact of tree selection on differential studies using both extensive simulation studies and reanalysis of previously published datasets. We compare the results obtained with the phylogeny, the correlation tree, and the standard Benjamini–Hochberg correction. Finally, we discuss the pros and cons of using one or the other in hierarchical procedures and some limitations of our work.

## 2. MATERIALS AND METHODS

### 2.1. Trees

We consider in this study different hierarchical structures, or trees: the phylogenetic tree, the taxonomic tree and the correlation tree.

#### 2.1.1. Phylogenetic Tree

The phylogeny encodes the common evolutionary history of the taxa. In the amplicon context, it is usually reconstructed based on the sequence divergence of the marker-gene (Price et al., 2010) and branch lengths correspond to the expected number of substitutions per nucleotide.

#### 2.1.2. Taxonomic Tree

When the phylogeny is not available but taxonomic annotations are, we fall back on the taxonomic tree instead. Inner nodes correspond to coarse taxonomic ranks (e.g., phylum, class, order, etc.). The hierarchical structure is reconstructed from lineages extracted from regularly updated databases like the one from NCBI (Geer et al., 2009). Branch lengths correspond to the number of levels in the hierarchy: e.g., a branch between species-level and genus-level nodes has length 1, a branch between species-level and genus-level nodes has length 2. Unlike phylogenetic trees, taxonomic trees are highly polytomic.

#### 2.1.3. Correlation Tree

The correlation tree is based on the abundance profiles of taxon across samples and built in the following way. We first compute the pairwise correlation matrix, using the Spearman correlation and excluding "shared zeros", i.e., samples where both taxa are absent. We then change this correlation matrix into a dissimilarity matrix using the transformation  $x \mapsto 1 - x$ . Finally, we use hierarchical clustering with Ward linkage on this matrix to create the correlation tree. Branch lengths correspond to the dissimilarity cost of merging two subtrees.

### 2.2. Distances Between Trees

We consider two different distances between trees: the Robinson–Foulds distance, or RTF (Robinson and Foulds, 1981), the Billera–Holmes–Vogtmann distance, or BHV (Billera et al., 2001). Those distances are computed using different characteristics of the tree (topology, branch lengths, etc.) and emphasize different features.

The RF distance is defined on topologies, i.e., trees without branch lengths, and based on elementary operations: branch contraction and branch expansion. A branch contraction step creates a polytomy in the tree by shrinking a branch and merging its two ending nodes whereas a branch expansion step resolves a polytomy by adding a branch to the tree. For any pair of trees, it is possible to turn one tree into the other using only elementary operations. The RF distance is the smallest number of operations required to do so. Note that the RF distance gives the same importance to all branches, no matter how short or long.

The BHV distance is defined on trees and accounts for both topology and branch length. All possible trees are embedded into a common treespace with a complex geometry. Trees with the same topology are mapped to the same orthant, and hyperplanes

share a common boundary if and only if they are at RF-distance 2 (one contraction and one expansion step away). For any pair of trees, there is a path in treespace between those two trees. The BHV distance is the length of the shortest of those paths. It can be thought of as the generalization of the RF-distance that upweights long branches and downweights short branches.

### 2.3. Forest of Trees

We generated a forest of bootstrapped trees and a forest of random trees in the following way. For the bootstrapped forest, we generated  $N_B$  bootstrap datasets using resampling with replacement (Felsenstein, 1985; Wilgenbusch et al., 2017). Each bootstrap dataset was used to compute a correlation matrix and a correlation tree as detailed in section 2.1.

Random trees were generated from a seed tree by shuffling the leaves labels. This allowed us to generate a forest of random trees with the same number of branches as the seed tree. This is especially important for RF-distances as the scale with the number of branches and we want to study both non-binary taxonomic trees with a high number of polytomies and low number of branches and binary correlation trees, with a high number of branches. We generated  $N_T$  random trees from the taxonomic tree and  $N_C$  from the correlation tree.

### 2.4. Testing Tree Equality

The correlation tree is reconstructed from abundance profiles rather than molecular sequences and/or lineages and may therefore be poorly estimated. We use the bootstrap forest to compute a confidence region around the correlation tree. The random trees were used to create a null distribution of distances between random trees.

The full set of  $2 + N_B + N_T + N_C$  trees was used to construct RF and BHV distance matrices. The distance matrices were then used to visualize a 2D-projection of all trees via Principal Coordinates Analysis (PCoA) (Gower, 1966; Jombart et al., 2017; Wilgenbusch et al., 2017). Bootstrap trees were used to test whether the taxonomy was in the confidence region of the correlation tree whereas random trees were used to test whether the taxonomic and correlation trees were closer to each other than to random trees.

We also compared the distance from the correlation tree to each group of trees using a one-way ANOVA.

### 2.5. Differential Abundances Studies

The literature abounds in differential analysis methods dedicated to abundance data (Soneson and Delorenzi, 2013). Most of them differ in the normalization and preprocessing steps (Dillies et al., 2013; Chen et al., 2018). Count data coming from metagenomic studies are very similar to those found in RNA-Seq studies. The former one may exhibit more zeros entries but the same types of normalizations and statistical models can be used for both types of data.

As the focus of the paper is not on normalization procedure, we therefore used only a simple and classic normalization (Chen et al., 2018) to assess the impact of taking into account the data hierarchical structure in the differential abundance testing.

We briefly present two methods for differential abundance testing (DAT) that leverage a tree-like structure: z-score smoothing as proposed in Xiao et al. (2017) and hFDR as proposed in Yekuteli (2008).

#### 2.5.1. z-Scores Smoothing

Given any taxa-wise DAT procedure, p-values ( $p_1, \dots, p_n$ ) are first computed for each taxa (leaves of the tree) and then transformed to z-scores using the inverse cumulative distribution function of the standard Gaussian. Similarly, the tree is first transformed into a patristic distance matrix ( $D_{ij}$ ) and then into a correlation matrix  $C_\rho = (\exp(-2\rho D_{ij}))$  between taxa. The z-scores  $z = (z_1, \dots, z_n)$  are then smoothed using the following hierarchical model:

$$z \mid \mu \sim \mathcal{N}_m(\mu, \sigma^2 I_m)$$

$$\mu \sim \mathcal{N}_m(\gamma I_m, \tau^2 C_\rho)$$

where  $\mu$  captures the effect size of each taxa. The maximum a posteriori estimator  $\mu^*$  of  $\mu$  is given by

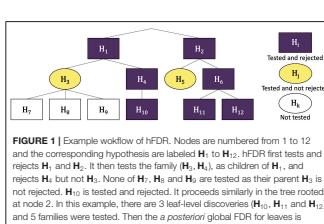
$$\mu^* = (I_m + k C_\rho^{-1})^{-1} (k C_\rho^{-1} \gamma I_m + z) \quad \text{where } k = \sigma^2 / \tau^2$$

and the FDR is controlled using a resampling procedure. This method intuitively pulls effect sizes of taxa close-by in the tree toward the same value. In particular, a differential taxon with large effect size and small p-value but surrounded by non-differential taxa in its phylogenetic neighborhood will be considered a fluke; its smoothed effect size will be shrunk toward zero and its corrected p-value will increase toward non-significance. Likewise, a taxa that is barely differential but phylogenetically close to differential taxa will be rescued toward significance: its effect size will increase and its p-value decrease. Extreme smoothing creates clades where all taxa are simultaneously differential or simultaneously non-differential.  $k$  and  $\rho$  are hyperparameters controlling the level of smoothing. Low ( $\tau^2$ , high) values of  $\rho$  ( $\tau^2$ ,  $k$ ) correspond to high smoothing. Finally,  $k$ ,  $\gamma$ , and  $\rho$  are estimated using generalized least-squares.

#### 2.5.2. Hierarchical FDR

Hierarchical FDR (hFDR) considers a different framework where differential abundance can be tested not only for a single taxa but also for groups of taxa, corresponding to inner nodes or clades of the tree. hFDR uses a top-down approach: tests are performed sequentially and only for nodes whose parent node were previously rejected. Formally, the procedure is described in Algorithm 1.

```
Algorithm 1 Hierarchical FDR
1:  $\mathcal{D} \leftarrow \emptyset$                                 Initialize discoveries
2:  $\mathcal{S} \leftarrow \text{Root}$                             Initialize stack
3: while  $\mathcal{S} \neq \emptyset$  do
4:   choose  $N$  in  $\mathcal{S}$ 
5:    $\mathcal{N} \leftarrow \text{BH}_\alpha(\text{ch}(N))$  Discoveries in children of  $N$ 
6:    $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{N}$                           Update discoveries
7:    $\mathcal{S} \leftarrow (\mathcal{S} \setminus N) \cup (\mathcal{N} \setminus \mathcal{L})$         Update stack
8: end while
9: return  $\mathcal{D}$  for full-tree discoveries or  $\mathcal{D} \cap \mathcal{L}$  for leaves discoveries
```



**FIGURE 1 |** Example workflow of hFDR. Nodes are numbered from 1 to 12 and the corresponding hypothesis tests are  $H_1$  to  $H_{12}$ . hFDR first tests and rejects  $H_1$ ,  $H_2$ , and  $H_3$ . Then it moves to node  $H_5$  and rejects  $H_5$  but not  $H_4$ . None of  $H_6$ ,  $H_7$ , and  $H_8$  are tested as their parent  $H_2$  is not rejected.  $H_9$  is tested and rejected. It proceeds similarly in the tree rooted at node 2; in this example, there are 3 leaf-level discoveries ( $H_{10}$ ,  $H_{11}$ , and  $H_{12}$ ) and 5 families were tested. Then a *a posteriori* global FDR for leaves is  $1.44 \times \alpha \times 2$ .

The hFDR procedure is illustrated in Figure 1.

#### 2.5.3. Implementations

These two algorithms are implemented in R packages (R Core Team, 2018): **structFDR** (Xiao et al., 2017) for the z-scores smoothing and **structSSI** (Sankaran and Holmes, 2014) for hFDR.

The z-scores smoothing algorithm as implemented in **structFDR** includes a fallback to standard, non-hierarchical, independent tests when too few taxa are detected. It was not part of the original algorithm and we therefore used a vanilla implementation, with the fallback (see modified code in **correlationtree** package), to specifically evaluate the impact of the tree in the procedure. **structFDR** requires the user to specify its test. We used non-parametric ones: Wilcoxon rank sum for settings with two groups and Kruskal-Wallis (Hollander and Wolfe, 1973) for settings with three or more groups.

In contrast, the hFDR procedure is only available for one-way ANOVA on the groups, and corresponding f-test, and does not correct for differences in sequencing depths. Moreover, we noticed that the global FDR control was off by the corrective factor of 1.44 in Equation (1). We corrected the output of **structSSI** to use the correct FDR values in our analyses.

### 2.6. Methods Evaluation

We tested the impact of tree choice on the performance of both procedures (z-score smoothing and hFDR) on real data and synthetic data simulated from real dataset in one of two following ways. The code and data used to perform the simulations are available on the github repository [github.com/abichat/correlationtree\\_analysis](https://github.com/abichat/correlationtree_analysis).

#### 2.6.1. Parametric Simulations

The parametric simulations use the following scheme. First, a Dirichlet-multinomial model  $\mathcal{D}(\gamma)$  is fitted to the gut microbiome dataset of healthy patients from Wu et al. (2011). Second, a homogenous dataset is created by sampling count vectors  $S_i$  from the Dirichlet-Multinomial distribution: (i) a proportion vector  $\alpha_i$  is drawn from  $\mathcal{D}(\gamma)$ , (ii) the sequencing depth  $N$  is drawn from a negative binomial distribution  $\mathcal{NB}(10,0.0025)$  with mean 10,000 and size 25 and finally (iii) the counts  $S_i$  of sample  $i$  are sampled from a multinomial distribution  $\mathcal{M}(N, \alpha_i)$ . We acknowledge that Dirichlet-multinomial distributions can only sample negatively correlated species but the goal here is to closely reproduce the simulation scheme from Xiao et al. (2018).

Differential abundances are then produced as follows. First, each sample is randomly assigned to class A or B. Second,  $n_{H_i}$  taxa (representing up to 20% of all taxa) were sampled uniformly among all taxa. Finally, the abundances of those taxa are multiplied by a fold-change (chosen in  $\{5, 10, 15, 20\}$ ) in group B. The process is illustrated in Figure 2.

#### 2.6.2. Non-parametric Simulations

Non-parametric simulations proceeded like the parametric ones detailed in section 2.6.1 with three major differences. First, we used a different dataset with homogeneous samples: the gut microbiome of healthy individuals from North America and Fiji Islands (Brito et al., 2016). Second, we did not fit a Dirichlet-Multinomial to the original dataset but used it as such to preserve the potential complex correlation structure present in the dataset. Finally, differentially abundant taxa were sampled only from highly prevalent taxa (prevalence  $> 90\%$ ) to ensure that DAT procedures were affected by effect size (fold-change) and hierarchical correction, rather than by sparsity.

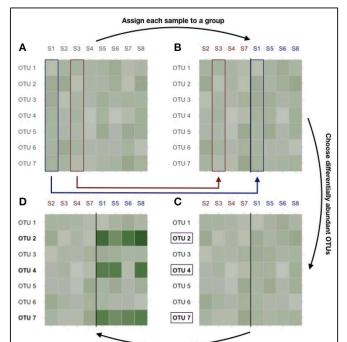
#### 2.6.3. Accuracy Evaluation

We used true positive rate (TPR) and FDR to evaluate the performance of z-scores smoothing used with five different trees: no tree or standard Benjamini–Hochberg (BH), taxonomy, correlation tree, random taxonomy and random correlation tree. BH is our baseline and the random trees are here to evaluate the impact of uninformative trees, with different granularity levels, on the procedure.

We evaluated hFDR by comparing the results obtained using either the taxonomy or the correlation tree in several datasets.

### 2.7. Datasets

We used seven different datasets for the experimental part (see Table 1 for a summary). One was used to study the difference between correlation and phylogenetic trees,



**TABLE 1 |** Summary table of the different datasets used in this study with information on biome type, taxonomic rank used for the analysis, corresponding number of taxa, number of samples and analyses performed on the dataset: comparison of the correlation and taxonomic trees (Tree), creation of synthetic datasets (Simulations), or impact of the tree on differential abundance procedures (DA).

Dataset	Biome	Rank	Taxa	Samples	Analyses	Publication	
Chlamydia Varied	OTU	21	26	Tree & DA		Caporaso et al., 2011	
Ravel	Vaginal	Genus	40	996	Tree	Ravel et al., 2011	
Wu	Gut	OTU	400	98	Simulations	Wu et al., 2011	
Zeller	Gut	Genus	119	199	Tree & DA	Zeller et al., 2014	
Zeller	MSP	Gut	MSP	878	199	DA	Zeller et al., 2014
Challou	Food	OTU	499/67	64	Tree & DA	Challou et al., 2015	
Brito	Gut	OTU	77	112	Simulations	Brito et al., 2016	

on to assess the impact of tree choice on difference abundance testing, three for both and the last two to generate synthetic datasets as described previously. All datasets used in this study are available on the github repository [github.com/abichat/correlationtree\\_analysis](https://github.com/abichat/correlationtree_analysis).

Three of the four datasets used for tree comparison (Ravel, Challou, and Zeller) were chosen because they are well-suited for bootstrapping correlation trees: they had enough samples and enough variability in taxa counts to ensure that a

meaningful correlation tree could be computed on bootstrapped datasets. They also represent diverse microbiome with contrasted biodiversity levels: vaginal microbiome for Ravel, food-associated microbiome for Chaillou and gut microbiome for Zeller. Briefly, Ravel et al. (2011) studied a cohort of 396 North-American women from 4 ethnic groups using metabarcoding on the V1-V2 region of 16S rRNA gene. Chaillou et al. (2015) studied food-associated microbiota of 80 processed meat and seafood products using metabarcoding on the V3-V4 region of the 16S rRNA gene. Zeller et al. (2014) considered the gut microbiota of 199 subjects (42 with adenomas, 91 with colorectal cancer and 66 healthy ones), using both shotgun sequencing and metabarcoding on the V4 region of 16S rRNA gene. Zeller refers to the 16S rRNA fraction of the data. Details of bioinformatics treatments used to produce abundance count tables are available in the respective publications. All datasets were aggregated at a given taxonomic level and taxa with a prevalence lower than 5% were filtered out.

The fourth one (Chlamydia) was used in Sankaran and Holmes (2014) to assess the performance of hFDR and is an excerpt from the data collected in Caporaso et al. (2011). It consists of bacteria from the Chlamydia phylum and is distributed with *StructSSI* (Sankaran and Holmes, 2014). Finally, the Zeller MSP data originates from the same study as the Zeller data (Zeller et al., 2014). It was created from the shotgun data by reconstructing Metagenomics Species Pan-Genomes (MSPs) abundance count table, as reported in Plaza-Obreque et al. (2011), where the phylogeny is within the confidence region of the correlation (Figure S1). Note however that this dataset is very small (26 samples) and has many taxa with low abundances, resulting in an extremely large confidence region for the correlation tree. It is also the only one that covers environments ranging from stool to soil and freshwater and thus, for which ecological niche and taxonomy may overlap (Philippot et al., 2010).

In light of these results, we find that the phylogeny is different from the correlation tree, especially when focusing on a single biome. In other words, taxa with similar abundance profiles are not clustered in the phylogeny and the phylogeny may therefore not be a good proxy to find groups of differentially abundant taxa.

Similar results are observed when using RF distance instead of BHV distance (Figure S2).

### 3.2. Pros and Cons of the Different Trees

Although phylogenies (resp. taxonomies) are evolutionary (resp. ecologically) meaningful and increasingly available, they do not capture similarities between taxa in terms of abundance profiles. For example, if abundances are driven by a phenotype regulated by a mobile element (e.g., an antibiotic resistance gene), evolutionary and ecological histories are not informative. Furthermore, when performing differential abundance analyses with genes (metatranscriptomics) or metagenomics-based taxa such as MSPs and metagenome-assembled genomes, many of which are poorly annotated, neither a taxonomy nor a phylogeny is available.

In contrast, the correlation tree is constructed from the abundance data and can thus always be used. By its very definition, it clusters taxa with similar abundance profiles. Unfortunately, it suffers from limitations of its own. First, it

procedures (BH) in terms of detection power and FDR control and assess whether some topologies are better than others and whether z-score smoothing outperforms standard BH. Finally, we analyze the Chlamydia (varied biome), Chaillou (food microbiome), and Zeller (gut microbiome) datasets using the hFDR hierarchical procedure to assess the same points for this procedure.

### 3.1. The Taxonomy Differs From the Correlation Tree

In all studied datasets, the correlation tree is closer to its bootstrap replicates than to either the taxonomy or the randomized trees (Figure 3, top row). The differences are statistically significant ( $p < 10^{-16}$ , one-way ANOVA with Tukey's HSD *post-hoc* test).

Similarly, the PCoA results (Figure 3, bottom row) highlight two or three island-like structures (Jombart et al., 2017); one for the correlation tree and its bootstrap replicates, one for the taxonomy and its randomized replicates and the final one for randomized correlation trees. All random trees can belong to the same island, as seen in the Ravel dataset. The first axis of PCoA represents 5–10% of the explained variance and systematically separates the taxonomy from the correlation tree. Moreover, the taxonomy is neither in the bootstrap confidence region of the correlation tree, nor closer to it than a randomized tree.

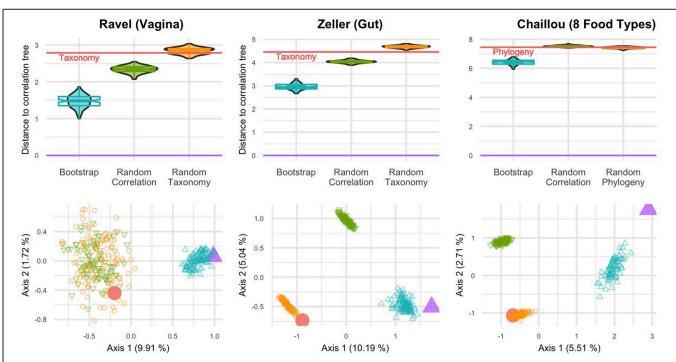
The only exception is the Chlamydia dataset (Caporaso et al., 2011), where the phylogeny is within the confidence region of the correlation (Figure S1). Note however that this dataset is very small (26 samples) and has many taxa with low abundances, resulting in an extremely large confidence region for the correlation tree. It is also the only one that covers environments ranging from stool to soil and freshwater and thus, for which ecological niche and taxonomy may overlap (Philippot et al., 2010).

In light of these results, we find that the phylogeny is different from the correlation tree, especially when focusing on a single biome. In other words, taxa with similar abundance profiles are not clustered in the phylogeny and the phylogeny may therefore not be a good proxy to find groups of differentially abundant taxa.

Similar results are observed when using RF distance instead of BHV distance (Figure S2).

## 3. RESULTS AND DISCUSSION

We first examine the relation between the correlation tree and the phylogeny (or taxonomy) using the Ravel (vaginal microbiome), Zeller (gut microbiome) and Chaillou (food microbiome) datasets. As they contain a high number of samples, they are the best suited for bootstrapping correlation trees. Since phylogeny and correlation-based tree have very different topologies, we perform two simulations studies to compare a hierarchical procedure (z-score smoothing) based on (i) the phylogeny or (ii) the correlation-based tree to (iii) a standard non-hierarchical



**FIGURE 3 |** BHV distances between various trees for three datasets: Ravel (left), Zeller (center) and Chaillou (right). **(Top)** Violinplots and notched boxplots of distances to the correlation tree. The distance between taxonomy (or phylogeny) and correlation is indicated by the red line. **(Bottom)** PCoA projection of all distances on the principal plane. The correlation tree is in purple (Δ), taxonomy (or phylogeny) in red (○), bootstrapped trees in blue, random correlation trees and random taxonomies (or phylogenies) in green and orange respectively.

is estimated from the data and thus sufficient data should be available to build a robust correlation tree. This may be a problem in the microbiome field where the number of samples is usually smaller (sometimes much smaller) than the number of taxa. This is also problematic for rare taxa, where shared zeroes may distort the correlation. The problem is usually alleviated by filtering out taxa with low abundance and/or prevalence. However, such filters disproportionately affect rare taxa and lead to a severe underestimation of the ecological role played by rare taxa (see Jousset et al., 2017 for a review).

Second, since the same data are used to build the correlation tree and to test differential abundance, some care should be taken not to overfit the data. For example, permutation-based tests are valid because the group labels are not used during the tree construction and are thus independent of the hierarchical structure (Goeman and Finos, 2012) but other tests should be used with caution.

### 3.3. Simulation Study

#### 3.3.1. Non-parametric Simulations

Note first that z-smoothing numerically failed and did not produce any results for 4% of the simulations (ranging from 2% for the randomized correlation trees to 8% for the correlation trees). Second, the hyperparameters  $k$  and  $\rho$  controlling the level of smoothing are often very far from 1 (below and above, respectively) resulting in little to no smoothing. Figure 4 shows the impact of smoothing on z-scores: in more than half of the

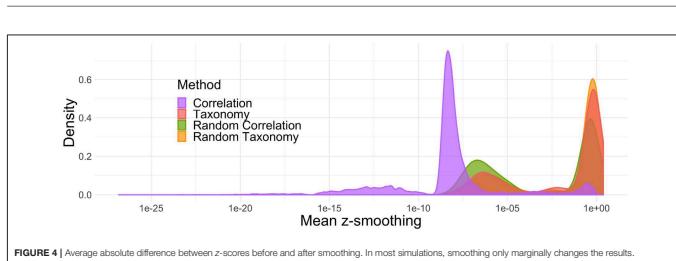
simulations, the z-scores were shifted by less than  $10^{-2}$  units in either direction. Among the different topologies tested, the phenomenon was the strongest for the correlation trees: the z-scores were shifted by more than  $10^{-2}$  units in less than 5% of the simulations.

Concerning FDR control, the standard BH procedure was the only one that achieved a nominal FDR rate below 5% across different fold changes and proportions of null hypothesis (Figure 5, bottom row). All other procedures exceeded the target rate, reaching nominal rates of up to 7%, when the number of null hypothesis grew beyond 90%.

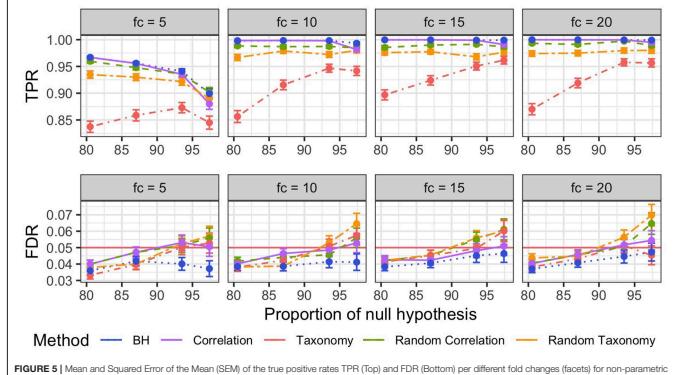
BH was similarly the most powerful method across all fold changes and proportions of null hypothesis (Figure 5, top row), with correlation tree and randomized correlation trees coming close second and third. BH, correlation tree and randomized trees outperformed the taxonomy in all settings, resulting in TPR increase of up to 0.15.

The comparatively bad result of the taxonomy is also expected from our simulation settings as the taxonomy is independent from simulated differential abundance. Forcing the discoveries to be close in the tree therefore introduces a systematic bias and results in a loss of power, especially for differential taxa that are isolated, and an increase in false discoveries, especially for non-differential taxa that are close to differential ones.

The better results of *a priori* uninformative random trees compared to the taxonomy were however more surprising,



**FIGURE 4 |** Average absolute difference between z-scores before and after smoothing. In most simulations, smoothing only marginally changes the results.



**FIGURE 5 |** Mean and Squared Error of the Mean (SEM) of the true positive rates TPR (Top) and FDR (Bottom) per different fold changes (facets) for non-parametric simulations. The different FDR control procedures are color-coded. Mean and SEM are computed over 600 replicates.

especially in light of the similar levels of smoothing for all those trees. It turned out that the random trees were, on average, closer to the correct correlation structure of differential taxa than the taxonomy and therefore had a lesser negative impact on the detection power.

It is clear from these results that using a tree reflecting the abundance data true structure, such as the correlation tree, does not increase the number of discoveries but does not degrade the performance of the method either. In contrast, using a wrong structure degrades the detection power from only slightly at best (for random trees) to quite a lot (taxonomy).

### 3.3.2. Parametric Simulations

Parametric simulations showed exactly the same patterns as non-parametric ones. Z-scores smoothing was limited in most replicates and almost always null when using the correlation tree (Figure S3). BH was the only procedure with a nominal FDR below the target rate of 5% in all settings and all trees led to nominal above the threshold when the proportion of differential taxa was low (Figure S4, bottom row). Finally, BH had the highest TPR among all methods (Figure S4, top row).

The results differed from the non-parametric ones in one important aspect: all methods had low TPR, below 0.15, whereas they achieve TPR higher than 0.85 in the non-parametric setting.

This difference is mainly due to the parametric simulation scheme, reused from Xiao et al. (2017): differential taxa are not pre-filtered based on their prevalence and can thus have a very high proportion of zeros in the worst case. Multiplication by a fold-change, no matter how high, leaves those zeros and their corresponding ranks unchanged. This in turn strongly degrades the ability of the rank-based Wilcoxon test, to find differences between groups among those taxa.

### 3.4. Analysis of Real Datasets

#### 3.4.1. Reanalysis of Chlamydia Dataset

The Chlamydia dataset consists of 26 samples distributed over 9 very different environments (feces, freshwater, human skin, sea, ...). Differential abundance of the OTUs across the environment was tested using the same parameters as in the original article (hFDR on the phylogeny,  $\alpha = 0.1$ ). The test identified 8 differential OTUs with a global *a posteriori* FDR of  $\alpha' = 0.32$ . Substituting the correlation tree to the phylogeny in this analysis led to the detection of 3 additional OTUs, at a comparable global FDR of  $\alpha' = 0.324$ .

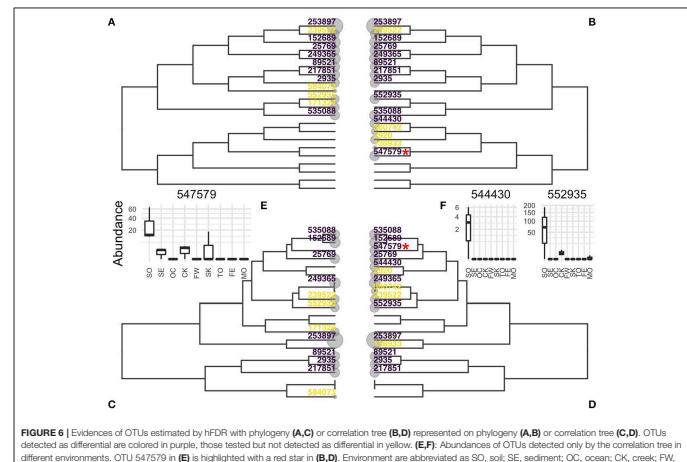
Abrundance boxplots of these three additional OTUs (Figures 6E,F, insets) show that these OTUs are much more abundant in soil samples and almost specific to that environment, validating their differentially abundant status. In that example,

the correlation tree reflected the structure of the data better than the phylogeny and increases the power at no cost to the nominal FDR.

Figure 6 shows the location of evidences ( $e = -\log_{10}(p)$ ) and differential OTUs on both the phylogeny and correlation trees. OTU 547579, highlighted with a red star, is one of the three additional OTUs. It was not tested with the phylogeny because it is the only differential taxa in its clade (Figure 6B) and its top-most ancestor was not rejected. In contrast, it belongs in the correlation tree to a group of soil-specific taxa and the hierarchical procedures sequentially rejected all its ancestors so that it was also tested and rejected.

With this top-down approach, the correlation tree is a better candidate hierarchy than the phylogeny. Indeed, the signals of differential OTUs can be averaged out with noise and/or conflicting signal in the phylogeny, they are pooled together in the correlation tree. This makes it easier to reject high level internal nodes and descend the tree toward differential OTUs.

It should be noted however that the *a posteriori* global FDR is quite high at 0.324. Using the standard BH with a FDR of 0.324 results in 4 new discoveries, for a total of 15. hFDR, with either the correlation or the phylogeny, does not outperform the classical BH procedure. This discrepancy might be explained by the global FDR computation used in hFDR which controls the



**FIGURE 6 |** Evidences of OTUs estimated by hFDR with phylogeny (A,C) or correlation tree (B,D) represented on phylogeny (A,B) or correlation tree (C,D). OTUs detected as differential are colored in purple, those tested but not detected as differential in yellow. (E,F) Abundances of OTUs detected only by the correlation tree in different environments. OTU 547579 in (E) is highlighted with a red star in (B,D). Environment are abbreviated as SC, soil; SE, sediment; OC, ocean; CK, creek; FW, freshwater; SK, skin; TO, tongue; FE, feces; MO, mock.

FDR in the worst case scenario. The actual global FDR could be much lower than this pessimistic bound.

#### 3.4.2. Analysis of Chaillou Dataset

The Chaillou dataset consists of 64 samples uniformly distributed across 8 food types (ground veal, ground beef, poultry sausages, sliced bacon, shrimps, cod fillet, salmon fillet, smoked salmon). Differential abundances of OTUs from the Bacteroidetes phylum (97 OTUs) across food types was tested with hFDR procedure ( $\alpha = 0.01$ , both phylogeny and correlation tree). The test had a global *a posteriori* FDR of 0.04 for both the phylogeny and the correlation tree and detected 28 differential OTUs with the phylogeny and 34 with the correlation tree. Similarly, with a 0.04 FDR level, vanilla BH leads to 55 discoveries.

Unlike the Chlamydia dataset, only 22 OTUs were detected by both methods. Careful examinations of those 22 show that each of them (i) is missing, or below the detection level, in at least one of the 8 food type of the studies whereas (ii) has high prevalence ( $>0.75\%$ ) and abundance in at least one other food type. We can thus classify those 22 as true positives rather than the false positives.

The abundance profiles of the 18 OTUs found only by the correlation tree (hereafter cor-OTUs) or the phylogeny (phy-OTUs) (Figure S5) show marked differences across the 8 food types, validating their differential status. As was the case in the Chlamydia dataset, cor-OTUs are often isolated in the phylogeny (Figure S6) and not even tested during the hierarchical procedure as they are averaged with low-signal taxa.

In contrast, phy-OTUs are often close to detected taxa in the correlation-tree but not detected because of the *F*-test implemented in *StructSSI*. For example, the three phy-OTUs 0656, 1495, and 0241 belong to a cluster of five shrimp-specific OTUs but the two others (0516 and 0519) have some outlier counts and comparatively higher counts than the three phy-OTUs (Figure S7, right). Aggregation at internal nodes leads to high variance which decreases the significance of the *F*-test: *p*-values at the internal nodes do not pass the threshold and the leaves are not tested. Replacing the *F*-test with the Kruskal-Wallis test,

which is more robust to outliers, led to the detection of all OTUs (Figure S7, left).

#### 3.4.3. Analysis of Genera in Zeller Dataset

The Zeller dataset consists of gut microbiomes from 199 subjects that are healthy ( $n = 66$ ), suffer from adenomas ( $n = 42$ ) or from colorectal cancer ( $n = 91$ ). Differential abundances of genera across medical conditions was tested with *s*-score smoothing using several tree (no tree or standard BH, taxonomy, correlation tree, randomized correlation tree and randomized taxonomy) and several FDR threshold levels.

Figure 7 (left panel) shows the number of genera detected by each tree at each threshold. While the correlation tree detects the most taxa and BH the least at almost all threshold values, the differences between all trees are very small (one or two taxa only). In particular, at  $\alpha = 0.05$ , all methods detected either 14 or 16 genera.

In this example, the algorithm estimated  $\rho > 40$  for the random trees and  $k < 10^{-7}$  for the correlation tree, effectively resulting in no smoothing of the *s*-scores. The corresponding values are  $\rho = 0.26$  and  $k = 0.37$  for the taxonomy. The *s*-scores were thus smoothed to a higher extent but this had almost no impact on the number of detected genera.

#### 3.4.4. Analysis of MSPs in Zeller Dataset

Repeating the same analysis at the MSP, rather than genus, level gave similar results. Among the 878 MSP and using  $\alpha = 0.05$ , 234 were detected without correction, 90 with the correlation tree, 85 with standard BH and 77 with a random tree. Neither the taxonomy nor the phylogeny were available for the MSP and they were therefore not compared to the other methods.

In that example  $k = 1.3 \times 10^{-7}$  and the tree has almost no impact on the *s*-scores and the corrected *p*-values (Figure S8, bottom row). The 5 additional taxa detected with the correlation tree are indeed not clustered with other detect taxa and have BH-corrected *p*-values between 0.0505 and 0.0540 (Figure S8, left row). The main differences between the two procedures does not lie in the use of a hierarchical structure rather than the way corrected *p*-values are computed: using permutations for

the correlation and analytic formula for BH. It coincides with previous findings that permutation-based FDR control improves detection of differentially abundant taxa (Jiang et al., 2017).

## 4. CONCLUSION AND PERSPECTIVES

In this work, we investigated the relevance of incorporating a *priori* information in the form of a phylogenetic tree in microbiome differential abundance studies. Doing so was reported to increase the detection rate in recent work (Sankaran and Holmes, 2014; Xiao et al., 2017).

The rationale rests upon the assumption that evolutionary similarity reflects phenotypic similarity. Taxa from the same clade should therefore be more likely to be simultaneously associated to a given outcome than distantly related taxa. Although this assumption sounds natural and supported by evidence for high level taxa such as phylum (Philippot et al., 2010), there are also many arguments against it for low level taxa such as species and strains. Previous work (Harris et al., 2015) even showed some degree of equivalence between species in the gut, i.e., species within the same ecological guild could replace each other during the assembly process.

We considered here whether the phylogeny and taxonomy were good *a priori* trees to capture the structure of the abundance data, as captured by the correlation tree. In all the environments we studied, we found that the taxonomy and/or the phylogeny were significantly different from the correlation tree. Taxa with very similar abundance profiles could be widely spread in the phylogeny and vice-versa. The phylogeny was on average no closer to the correlation tree than a random tree, and thus not a good proxy of the abundance data structure.

We further studied the impact of tree misspecification on two recently published tree-based testing procedures, *s*-score smoothing (Xiao et al., 2017) and hFDR top-down rejection (Yekutieli, 2008).

Concerning *s*-score smoothing, we showed on synthetic data that substituting the correlation tree to the phylogeny increased the detection rate. Quite surprisingly, replacing the phylogeny with a random tree also increased the detection rate (Figure 5), questioning the use of the phylogeny in the first place. The results were even more disappointing on real datasets where all trees led to similar detection rates and none of them significantly outperformed standard BH (Figure 7). In the Zeller MSP dataset, the differences between procedures were limited (Figure S7) and stemmed mostly from the way *p*-values were computed: i.e., using permutations for *s*-score smoothing and closed formula for BH. Overall, using phylogenetic information to smooth *s*-scores degrades the detection rate (at worst) or leaves it unchanged (at best).

Top-down rejection (hFDR) gave more interesting results. Replacing the phylogeny or taxonomy with the correlation tree increased the detection rate, while preserving the global *a posteriori* FDR. In general, taxa detected with the correlation

tree but not with the phylogeny belonged to clades of mostly non-differential taxa in the phylogeny (Figure 6). Their signal was thus averaged with noise and they discarded early-on in the hierarchical procedure. In contrast, they were salvaged on the correlation tree as they belonged clades of taxa with similar abundance profiles. Unfortunately, hFDR suffers from two limitations. First, it has a lower detection rate than standard BH at the same global FDR level. This is likely a side effect of the definition of the global FDR in hFDR, i.e., FDR in the absolute worst case scenario. Second, the current implementation of hFDR in *StructSSI* is limited to *F*-test, which are ill-suited to highly non-gaussian microbiome data.

Our findings are puzzling as the use of prior information should intuitively increase the statistical power and certainly not degrade it. In our opinion, three elements limits the hierarchical methods. First, the lack of flexibility: the limitation of hFDR to *F*-test is a problem which can be alleviated by substituting it with more powerful tests (generalized linear model, omnibus tests, etc.). Second, the inadequacy of the phylogeny as a hierarchical prior. While informative priors can certainly lead to increased statistical power, priors that impose a non-informative structure, or worse a structure that conflicts with the genuine data structure, can hamper the testing procedure by increasing significance when it should decrease it and vice-versa. Replacing the phylogeny with the correlation tree mitigates this effect but only insofar as the correlation is not too noisy. Finally, the good theoretical properties of hFDR were proved under the assumption of independence between a *p*-value and its ancestor. It's unlikely to be the case in practice. Yekutieli (2008) reported that dependence within the families of tested hypotheses and across the tree seemed to result in higher FDR values than under independence ( $p < 0.14$ ). Hierarchical procedures are seducing in theory but hard to implement in practice.

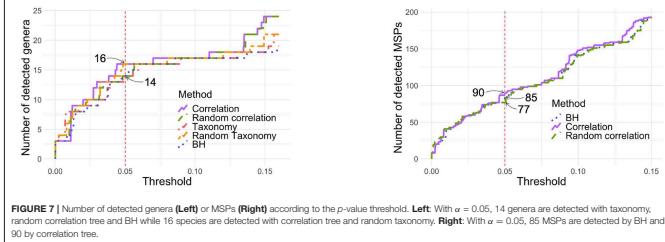
Our conclusions are two-fold. First, the phylogeny does not capture the structure of the abundance data and should be replaced by a better hierarchical structure such as the correlation tree. Second, hierarchical methods in their current state do a poor job of leveraging the hierarchical information to increase the detection rates. Until better hierarchical methods are available (e.g., hFDR with support for more complex tests), we recommend sticking to the time-tested BH procedure for differential abundance analysis.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: [https://github.com/abichat/correlationtree\\_analysis](https://github.com/abichat/correlationtree_analysis).

## AUTHOR CONTRIBUTIONS

MM, CA, and JP designed and directed the study. AB, MM, CA, and JP wrote the manuscript. AB created the synthetic datasets. AB performed all the analyses with substantial input from MM, CA and JP.



**FIGURE 7 |** Number of detected genera (Left) or MSPs (Right) according to the *p*-value threshold. **Left:** With  $\alpha = 0.05$ , 16 genera are detected with taxonomy, random correlation tree and BH while 14 genera are detected with correlation tree and random taxonomy. **Right:** With  $\alpha = 0.05$ , 85 MSPs are detected by BH and 90 by correlation tree.

CA and JP. All authors discussed the results and commented on the manuscript.

## FUNDING

This work was funded by Enterome and the ANRT (Association Nationale de la Recherche et de la Technologie) via the grant CIFRE 2017/0518. The funder (Enterome) was not involved in the funding acquisition.

## REFERENCES

- Bartoli, C., Frachon, L., Barret, M., Rigal, M., Huard-Chauveau, C., Mayonnaud, B., et al. (2018). In situ relationships between microbiota and potential pathobiont in *Arabidopsis thaliana*. *ISME J.* 12, 2024–2038. doi: 10.1038/s41396-018-0152-7
- Behrouzi, A., Nafari, A. H., and Sadat, S. D. (2019). The significance of microbiome in personalized medicine. *Clin. Transl. Med.* 8:16. doi: 10.3890/ctm.801019-019-0232
- Bernardo, L., Morcillo, C., Carletti, P., Ghizzoni, R., Badeck, F., Ruiz, F., et al. (2017). Proteomic insight into the mitigation of wheat root drought stress by arbuscular mycorrhizae. *J. Proteomics* 169, 21–32. doi: 10.1016/j.jprot.2017.03.024
- Billerla, L., Holmes, F., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* 27, 733–767. doi: 10.1006/aama.2001.0579
- Britto, I., Yilmaz, S., Huang, K., Xu, L., Jupiter, S., Dade, A., Jenkins, A. P., et al. (2016). Mobile genes in the human microbiome are structured from global to individual. *Nature* 538:436–439. doi: 10.1038/nature19927
- Buddehl, B. (2014). *Blower: A Fast, Accurate, Splice-Aware Aligner*. Technical report, Lawrence Berkeley National Lab, Berkeley CA.
- Callahan, B. J., et al. (2010). High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 10:852–853. doi: 10.1038/nmeth.1496
- Caporaso, J. G., Kuczynski, J., Stombaum, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7:335. doi: 10.1038/nmeth.1458
- Carriero, J., Lai, C., Wang, C., Wu, J., Wang, X., Li, J., and Holmes, S. (2013). Nonparametric Statistical Methods. New York: Springer.
- Carriero, J., Lai, C., Wang, C., Wu, J., Wang, X., Li, J., and Holmes, S. (2014). Nonparametric Methods for the Analysis of Next-Generation Sequencing Data. *PLoS Comput. Biol.* 10:e1003807. doi: 10.1371/journal.pcbi.1003807
- Carroll, R. J., Walsum, R. L., Muller, S., and Garcia, T. P. (2014). Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics* 30, 831–837. doi: 10.1093/bioinformatics/bts070
- Challou, S., Chauvel-Talon, A., Caekebeke, H., Cardinal, M., Christiansen, S., Denis, C., et al. (2015). Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. *ISME J.* 9:341. doi: 10.1038/ismej.2014.202
- Chen, L., Ren, J., Zhang, L., Huang, S., Wang, X., Chen, J. (2018). Comp: a robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6:e4600. doi: 10.7717/peerj.4600
- Dilles, M.-A., Rao, A., Aubert, J., Hennequel-Antier, C., Jeannougn, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* 14, 671–683. doi: 10.1093/bib/bbs046
- Eren, A. M., Morton, H. G., Lessard, J., Reveillard, J., Vimeis, J. H., and Sogin, M. L. (2016). Microbial community decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 9, 968–979. doi: 10.1038/ismej.2014.195
- Escudie, F., Auer, A., Bernard, M., Mariadassou, M., Caupil, L., Vital, K., et al. (2017). FROGS: find, rapidly, OTUs with galaxy solution. *Bioinformatics* 34, 1287–1294. doi: 10.1093/bioinformatics/btx291
- Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791. doi: 10.1111/j.1558-5646.1985.tb0420x
- Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., et al. (2009). The NCBI biosystems database. *Nucleic Acids Res.* 38(Suppl. 1):D492–D496. doi: 10.1093/nar/gkp858
- Goeeman, J. J., and Firth, D. (2012). The inheritance problem: multiple tests of association. *Statist. Appl. Genet. Mol. Biol.* 11, 1–18. doi: 10.15373/g3.53.3-4.325
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53, 325–338.
- Harris, K., Parsons, T., Jia, U., Zalhi, L., Lahti, L., Holmes, I., and Quince, C. (2015). Linking statistical and ecological theory: Bhullar's unified neutral theory of biodiversity as a hierarchical Dirichlet process. *Proc. IEEE* 105, 516–529.
- Holland, M., and Wolfson, D. A. (1973). *Nonparametric Statistical Methods*. New York: Wiley.
- Jain, L., Amit, A., Jordan, T. J., Heller, K., Aras-Kastiroglu, E., and Knight, R. (2017). Discrete false-discovery rate improves identification of differentially abundant microbes. *msystems* 2:e00292-17. doi: 10.1128/msystems.00292-17
- Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). treospace: statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Resour.* 17, 1385–1392. doi: 10.1111/1755-0994.02676
- Jousser, A., Biemond, C., Chatzitassis, A., Gallien, L., Gobet, A., Kurm, V., et al. (2017). Where less may be more: how the rare biogeophiles pell ecosystems strings. *ISME J.* 11, 861–862. doi: 10.1038/ismej.2016.174
- Kazan, C., and Mariadassou, M. (2016). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. USA* 103:8089–8094. doi: 10.1073/pnas.1601266113
- Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 32, 834–841. doi: 10.1038/nbt.2942
- Lynch, S. V., and Pedersen, O. (2016). The human intestinal microbiome in health and disease. *N. Engl. J. Med.* 375, 2369–2379. doi: 10.1056/NEJMoa1600266
- Mahé, F., Rogues, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 3:e1420. doi: 10.7717/peerj.1420
- Martinez, E., Lai, C., Lennon, J. T., and Martiny, A. C. (2015). Microbiomes in light of traits: a phylogenetic perspective. *Science* 350:eaac9323. doi: 10.1126/science.aac9323
- Matsen, F. A. IV, and Evans, S. N. (2013). Edge principal components and squash clustering: Using the special structure of phylogenetic placement data for sample comparison. *PLoS ONE* 8:e65659. doi: 10.1371/journal.pone.0065659
- Mendes, J., Kruijt, M., de Brujin, I., Delkens, E., van der Voort, M., Schneider, J. H. M., et al. (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* 332, 1097–1100. doi: 10.1126/science.1203980
- Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Dewey, C. M., Ward, D. V., et al. (2012). Microbial community decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 6, 1837–1847. doi: 10.1038/ismej.2012.13-9-679
- Nielsen, H. B., Almeida, M., Junker, A. S., Rasmussen, S., Li, J., Sunagawa, S., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* 32, 822–828. doi: 10.1038/nbt.2939

Opstelten, J. L., Plassais, J., van Mil, S. W., Achouri, E., Pichaud, M., Siersma, P. D., et al. (2016). Gut microbial diversity is reduced in smokers with Crohn's disease. *Inflamm. Bowel Dis.* 32, 2670–2677. doi: 10.1002/ibd.23069

Pasilli, E., Schäffer, L., Manghi, P., Rezen, A., Krishnamoorthy, V., Truong, D. T., et al. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* 14, 1023–1024. doi: 10.1038/nmeth.4468

Philippot, L., Andersson, S. G., Battin, T. J., Prosser, J. I., Schmid, J. P., Whitman, W. B., et al. (2010). The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.* 8, 85–92. doi: 10.1038/nrmicro2367

Plaza-Oriate, F., Le Chatelier, E., Almeda, M., Cervino, A. C., Lautenbacher, F., Magoulès, F., et al. (2018). MSPMiner: abundance-based reconstruction of microbial pan-genomes from shotgun metagenomic data. *bioRxiv*, 35, 1544–1554. doi: 10.1101/154409/biorxiv.353432

Priest, M. N., Dahal, P. S., and Arkin, A. P. (2010). FastTree 2—approximately maximum likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi: 10.1038/nature13560

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koeg, S. S., McCallie, S. E., et al. (2011). Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1):E4680–E4687. doi: 10.1073/pnas.1011073108

Robinson, D. F., and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147. doi: 10.1016/0025-5564(81)90043-2

Routy, B., Le Chatelier, E., Derosa, L., Duong, C. P., Alou, M. T., Daillère, R., et al. (2018). Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* 359, 91–97. doi: 10.1126/science.aan3706

Sankaran, K., and Holmes, S. (2014). structSSI: simultaneous and selective inference for grouped or hierarchically structured data. *J. Stat. Softw.* 59:1. doi: 10.18637/jss.v059.i01

Sonesen, C., and Mariadassou, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics* 14:91. doi: 10.1186/1471-2105-14-91

Trivedi, P., Schenk, P. M., Wallenstein, M. D., and Singh, B. K. (2017). Tiny microbes, big yields: enhancing food crop production with biological solutions. *Microb. Biotechnol.* 10, 999–1003. doi: 10.1111/1751-7915.12804

Washburne, A. D., Silverman, J. D., Leff, J. W., Bennett, D. J., Darcy, J. L., Mukherjee, S., et al. (2017). Phylogenetic factorization of compositional data reveals mean-level associations in microbiome datasets. *PeerJ* 5:e2969. doi: 10.7717/peerj.2969

Wilkensbach, J., Cai, Huang, W., and Gilligan, K. A. (2017). Visualizing phylogenetic tree landscapes. *BMC Bioinformatics* 18:85. doi: 10.1186/1471-2105-18-85

Wu, G. D., Chen, J., Hoffmann, C., Bittiger, K., Chen, Y.-T., Kelbaugh, S. A., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334, 105–108. doi: 10.1126/science.1203844

Xiao, J., Cao, H., and Chen, J. (2018). False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics* 35, 2872–2881. doi: 10.1093/bioinformatics/btw311

Xiao, J., Lin, P., Johnson, S., Zhang, X., and Chen, J. C. (2018). Predictive modeling of microbiome data using a phylogeny-regulated generalized linear mixed model. *Front. Microbiol.* 9:1391. doi: 10.3389/fmicb.2018.01391

Xiao, L., Estelle, J., Kucherlik, P., Ramay-Caldas, Y., Xia, Z., Feng, Q., et al. (2016). A reference gene catalogue of the pig gut microbiome. *Nat. Microbiol.* 1:1616. doi: 10.1038/nmicr.2016.161

Yekutieli, D. (2008). Hierarchical false discovery rate-controlling methodology. *J. Am. Stat. Assoc.* 103, 309–316. doi: 10.1198/01621450700000001373

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kulima, J. R., Costa, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:676. doi: 10.1525/msb.0145645

**Conflict of Interest:** AB and JP were employed by Enterome.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Bichat, Plassais, Ambroise and Mariadassou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

## Supplementary Material

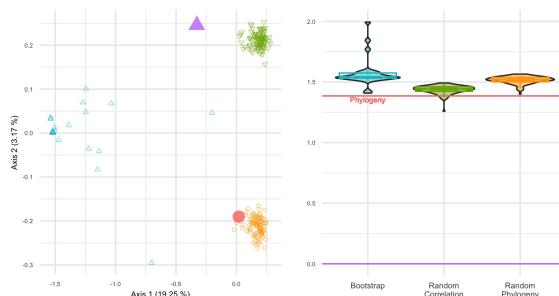
### 1 REPRODUCIBILITY

A R-package was made to provide helpful functions to support the analysis performed in this work. It is available on the GitHub repository: <https://github.com/abichat/correlationtree>.

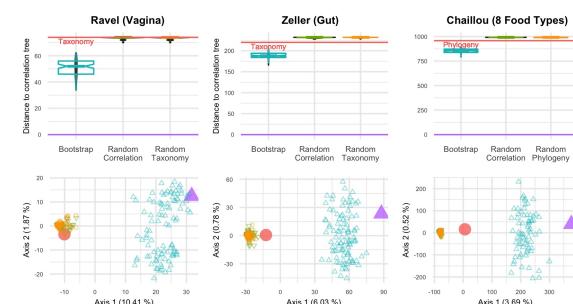
All codes for analysis and figures are also available on GitHub: [https://github.com/abichat/correlationtree\\_analysis](https://github.com/abichat/correlationtree_analysis). Some results or figures might be slightly different due to a different random number seed choice or a different number of simulations.

### 2 SUPPLEMENTARY TABLES AND FIGURES

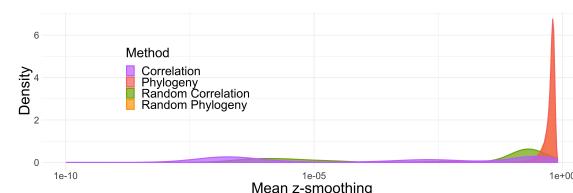
#### 2.1 Figures



**Figure S1.** BHV distances on forest of trees generated on the Chlamydiae dataset. Left: PCoA of pairwise distances. Right: distance to the correlation tree. Unlike other datasets studied in the analysis, bootstrap replicates of the correlation tree are quite far from the original one, and no closer than the phylogeny.



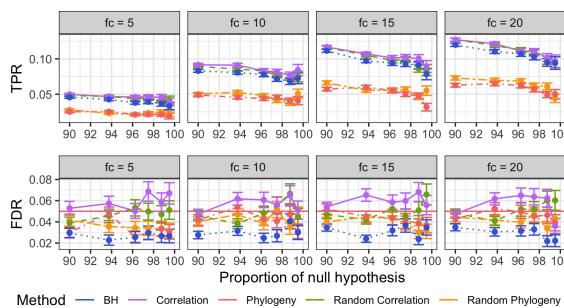
**Figure S2.** RF distances between the correlation tree and various other trees for three datasets: Ravel (left), Zeller (center) and Chaillou (right). Top row: violinplots and notched boxplots of distances to the correlation tree. The distance between taxonomy (or phylogeny, depending on the dataset) and correlation corresponds to the red line. Bottom row: PCoA projection of all distances on the principal plane. The correlation tree is in purple ( $\triangle$ ), taxonomy (or phylogeny) in red ( $\circ$ ), bootstrapped trees in blue, random correlation trees and random taxonomies (or phylogenies) in green and orange respectively. The first axis of the PCoA always separates the taxonomy / phylogeny from the correlation tree.



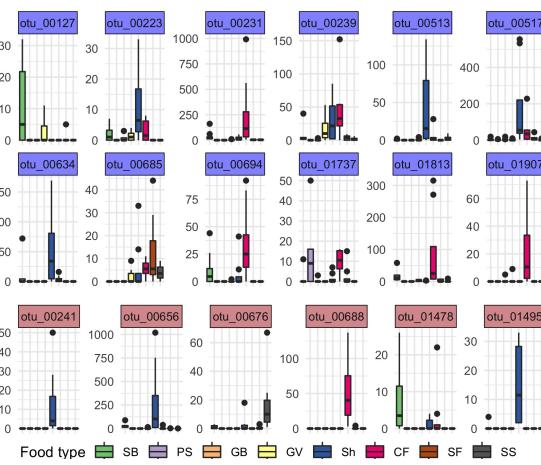
**Figure S3.** Average absolute difference between  $z$ -scores before and after smoothing for parametric simulations. Phylogeny and random phylogenies densities perfectly overlap. In most simulations, the  $z$ -scores are barely changed by the smoothing procedures.

1

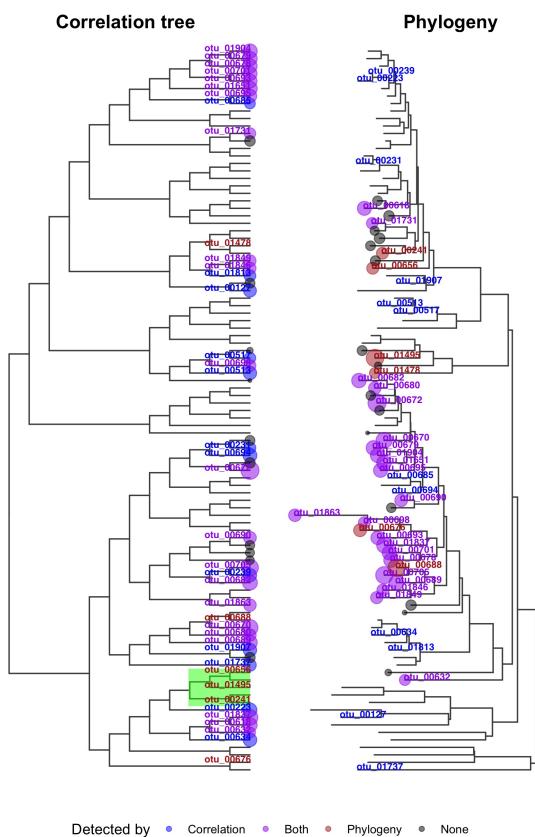
2



**Figure S4.** Mean and Squared Error of the Mean (SEM) of the true positive rates (TPR, top) and FDR (bottom) per different fold changes (facets) for parametric simulations. The different FDR control procedures are color-coded. Mean and SEM are computed over 600 replicates. BH and the correlation tree always outperform the phylogeny but BH is the only one to achieve a nominal FDR below 0.05.



**Figure S5.** Abundances for OTUs detected only by the correlation tree (topmost rows, blue strip background) or by phylogeny (bottom row, red strip background) in the Chaillou dataset. Food types are abbreviated as SB: sliced bacon, PS: poultry sausage, GB: ground beef, GV: ground veal, Sh: shrimp, CF: cod fillet, SF: salmon fillet, SS: smoked salmon. All 18 OTUs have very different abundances across food types.

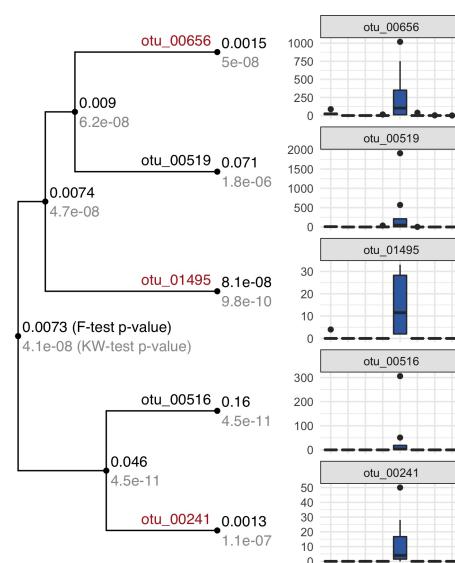


Detected by • Correlation • Both • Phylogeny • None

**Figure S6.** Position of the detected OTUs on both the correlation tree (left) and the phylogeny (right). Point sizes are proportional to evidences ( $-\log_{10}(p\text{-values})$ ) and are absent for OTUs that were not tested during by hFDR. OTUs detected only on one tree using hFDR are clustered in trees and usually quite dispersed in the other one. The green rectangle corresponds to a clade with many taxa detected by the phylogeny but not the correlation tree.

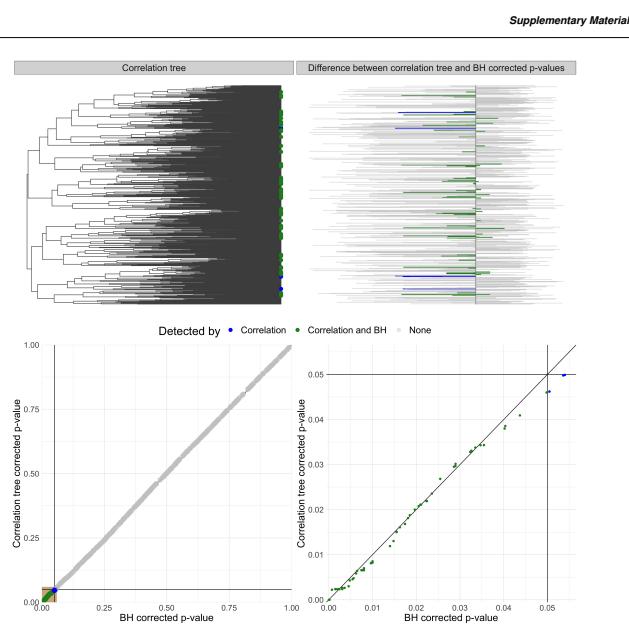
Frontiers

5



**Figure S7.** Focus on the green clade highlighted in Sup. Fig. S6. Left: Tree topology and  $p$ -values from a Fisher test (top) and a Kruskal-Wallis test (bottom) at each tip and each internal node of the correlation tree. Abundances of internal nodes were computed by summing the abundances of all leaves in their subtree. Right: Abundance profile of each OTU across food types. Food types are abbreviated as SB: sliced bacon, PS: poultry sausage, GB: ground beef, GV: ground veal, Sh: shrimp, CF: cod fillet, SF: salmon fillet, SS: smoked salmon.  $F$ -test are not significant due to (i) outlier counts in many abundance profiles and (ii) the obvious differences in variance across food types.

6



**Figure S8.** MSP from Zeller detected only by the correlation tree (blue) or also by the BH correction (green). Top left: Detected OTUs and their location on the tree. Top right: difference between  $p$ -values corrected by the correlation tree and by BH. Bottom left:  $p$ -values adjusted with the correlation tree against those adjusted with the BH correction. Bottom right: zoom on significant  $p$ -values (tan box).  $z$ -scores are virtually uncorrected ( $k = 1.3 \times 10^{-7}$ ) leading to almost identical  $p$ -values. The difference between the two sets of  $p$ -values lies in the use of different correction methods: permutation-based for the correlation tree, formula-based for BH.

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

Cet article devrait être soumis dans *Statistics and Computing* et est disponible sur *arXiv*.

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

Antoine Bichat<sup>1,2</sup>,

Christophe Ambroise<sup>1</sup> and Mahendra Mariadassou<sup>3,\*</sup>

<sup>1</sup>LaMME, Université d’Évry val d’Essonne, 91000 Évry, France

<sup>2</sup>Enterome, 94-96 Avenue Ledru Rollin, 75011 Paris, France

<sup>3</sup>MaIAGE, INRAE, Université Paris-Saclay, 78350, Jouy-en-Josas, France

## Abstract

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

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

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

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

## 1 Introduction

In many fields, statistical testing is classically used as an exploratory tool to look for the association between a variable of interest and many possible explanatory variables. For example, in transcriptomics, the link between a phenotype and the expression of tens of thousands of genes is tested (McLachlan et al., 2005), in Genome Wide Association Studies (GWAS) the association between millions of markers and a phenotype is tested (Bush and Moore, 2012), in functional Magnetic Resonance Imaging (fMRI), the goal is to identify voxels that are significantly activated in two different conditions (Cremers et al., 2017).

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

False Discovery Rate (FDR) (Benjamini and Hochberg, 1995), defined as the expected proportion of false positives among detections. Most of the correction procedures for controlling FWER or FDR rely on independence, or some form of weak dependence, among the hypothesis, which is rarely observed in practice. Multiple testing under dependence is a difficult problem occurring in many fields. In transcriptomics, differential analysis has to deal with gene expressions that are often highly correlated. When performing GWAS, the linkage desequilibrium imposes a strong spatial dependence between markers, and in Functional Magnetic Resonance Imaging (fMRI), two spatially close voxels have often comparable activation.

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

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

We propose to model the hierarchical structure of the multiple tests through an Ornstein-Uhlenbeck process on a tree. The process correlation structure is used for smoothing the p-values, after conversion to z-scores, similarly to the algorithm proposed in Xiao et al. (2017) but with an explicit underlying model. We resort to an  $\ell_1$  penalized estimation of the mean of the Ornstein-Uhlenbeck process, followed by a debiasing procedure (Javanmard and Montanari, 2013, 2014; Zhang and Zhang, 2014) for p-value computation. Eventually, we use a tuning proposed by Javanmard et al. (2019) to control the FDR.

Model selection is achieved via a Bayesian Information Criterion (BIC). We provide some background on hierarchical procedures in Section 2. We introduce the model and statistical procedure in Section 3 and detail the computational steps in Section 4. The performances of the algorithm are assessed via simulations in Section 5. The use of the proposed model is illustrated in Section 6, where we demonstrate its ability to discover novel associations in a metagenomic dataset.

## 2 Background

Hypothesis testing has become a standard in scientific literature to accept or reject a hypothesis under uncertainty. This type of procedure aims to make a decision by controlling the risk of error. Classically two types of errors are distinguished: type I errors (also called false discoveries or False Positive, FP), which wrongly reject the null hypothesis and the types II errors (False Negative,

the nominal FDR much smaller than the target, resulting in many FN. Permutation procedures are an appealing alternative that can automatically adapt to the dependence structure of the p-values (Tusher et al., 2001) but may fail when confronted to unbalanced design or correlated data. Knowledge of the correlation structure can be leveraged to increase the power while still controlling the FDR below a given target. Several approaches have been developed along those lines when the tests are organized along a hierarchical structure, typically encoded in a tree.

### 2.3 Hierarchical testing

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

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

$$\begin{aligned} z \mid \mu &\sim \mathcal{N}_m(\mu, \sigma^2 \mathbf{I}_m) \\ \mu &\sim \mathcal{N}_m(\gamma \mathbf{1}_m, \tau^2 \mathbf{C}_p) \end{aligned}$$

where  $\mu$  captures the effect size of each taxon and  $z$  is a noisy observation of  $\mu$ . The maximum a posteriori estimator  $\mu^*$  of  $\mu$  is given by

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

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

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

FN) which wrongly accept the null hypothesis.

In many fields (genomics, fMRI, ...), the number of hypotheses to be tested is very large (Goeman and Solari, 2014; Nichols, 2012), both in absolute terms and compared to the number of samples. Applying the control strategy of a single test when many hypotheses are examined simultaneously is often not desirable, as it leads to many errors. To limit and control the number of errors in the context of multiple hypothesis problem, two main strategies are used, separately or in combination. The first and by far most common strategy is to control, or at least quantify, the number of errors made on many tests carried out simultaneously. Two main quantification criteria are often considered, the FDR and the FWER. The second strategy is to reduce the number of tests by aggregating certain hypotheses. Aggregation strategies vary and can be based on a prior knowledge (*e.g.* metabolic pathways, functional modules of genes) or on clustering algorithms.

### 2.1 Examples of multiple testing strategies

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

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

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

### 2.2 Independence assumption

The assumption of independence of tests is convenient as it provides for both exact analyses and simple error bounds for classical procedures (Benjamini and Hochberg, 1995, *e.g.*). It is however unrealistic in practice. In many fields, including all the previous examples, measurements typically exhibit strong correlations. Some correction procedure, like the one proposed by Benjamini and Yekutieli (2001), make few assumptions while guaranteeing control of the FDR. Those general guarantees come with a high cost in terms of statistical power:

node  $i$

$$U_i(t) = \left| \frac{\sum_{k \in B(i)} s_k \mathbb{I}_{\{p_k \leq t\}}}{\#B(i)} \right|$$

where  $B(i)$  is the set of descendants of node  $i$ ,  $p_k$  and  $s_k \in \{-1, -1\}$  are the p-value of the node  $k$  and the sign of the associated effect, and  $t$  is a tuning parameter. A node  $i$  will be considered as candidate if  $U_i(t) \approx 1$  and  $p_i < \alpha$ . This ensure that all descendants are (i) significant at level  $t$  with (ii) effects of coherent sign. At the end, multiplicity correction is only done on nodes (including leaves) that do not descend from another candidate.

## 3 Models and algorithms

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

### 3.1 Ornstein-Uhlenbeck process on a tree

An Ornstein-Uhlenbeck (OU) process  $(W_t)$  with optimal value  $\beta_{\text{ou}}$ , selection strength  $\alpha_{\text{ou}}$  and drift parameter  $\sigma_{\text{ou}}$  is a Gaussian process that satisfies the stochastic differential equation:

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

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

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

common ancestor of  $i$  and  $j$  and note  $d_{ij} = t_i - t_j - 2t_{ij}$  the distance in the tree between nodes  $i$  and  $j$ . The distribution of the trait at node  $i$  is given by:

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

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

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

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

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

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

### 3.2 Procedure

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

6

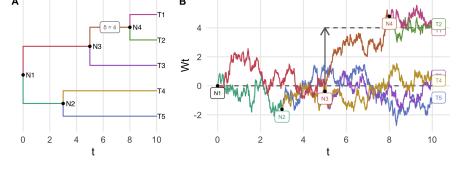


Figure 1: (A) Phylogenetic tree with 5 tips and 4 internal nodes (root  $N_1$  included). A shift occurs on the branch leading to  $N_4$ . (B) Ornstein-Uhlenbeck process with shifts on the tree defined in the left panel. At each node, the process spawns two independent process with the same initial value. The shifts on the optimal value on the branch leading to  $N_4$  results in a different mean value for  $N_4$  and all its offsprings ( $T_1$  and  $T_2$ ).

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

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

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

(A1) Under  $H_1$ ,  $\mathfrak{z}_i \sim \mathcal{N}(\mu_i, 1)$  where  $\mu_i \leq 0$ ;

(A2)  $\mathfrak{z}$  arises from a shifted Ornstein-Uhlenbeck process on a  $\mathcal{T}$  with parameters  $\alpha_{on}$ ,  $\Delta_{on}$  and  $\Delta$ .

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

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

where  $\Sigma$  is fully specified by the parameters  $\sigma_{on}$  and  $\alpha_{on}$ . Note that the diagonal coefficients of  $\Sigma$  are all equal to  $\sigma_{on}^2 / 2\alpha_{on} (1 - 2e^{-2\alpha_{on}h})$ . As they correspond to marginal variances, this forces the equality  $\sigma_{on}^2 = (1 - 2e^{-2\alpha_{on}h}) / 2\alpha_{on}$ , so that  $\Sigma$  depends only on  $\alpha_{on}$ , *i.e.*  $\Sigma = \Sigma(\alpha_{on})$ . Finally, the decomposition  $\mu = T\Delta$ , where  $T$  acts as a phylogenetic design matrix, ensures that alternative hypotheses are likely to form clades, *i.e.* groups of tips obtained by cutting a single branch in the tree.

7

$$T = \begin{pmatrix} N_0 & N_1 & N_2 & N_3 & T_1 & T_2 & T_3 & T_4 & T_5 \\ T_1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ T_2 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ T_3 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ T_4 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ T_5 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \Delta = \begin{pmatrix} b_{N_1} & 0 \\ b_{N_2} & 0 \\ b_{N_3} & 0 \\ b_{N_4} & \delta \\ b_{T_1} & 0 \\ b_{T_2} & 0 \\ b_{T_3} & 0 \\ b_{T_4} & 0 \\ b_{T_5} & 0 \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_{T_1} & \delta \Delta_{N_4} \\ \mu_{T_2} & \delta \Delta_{N_4} \\ \mu_{T_3} & 0 \\ \mu_{T_4} & 0 \\ \mu_{T_5} & 0 \end{pmatrix}$$

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

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

#### 3.2.1 Parameter Estimation

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

$$\hat{\Delta} = \underset{\Delta \in \mathbb{R}^n \text{ s.t. } T\Delta \in \mathbb{R}^m}{\operatorname{argmin}} \frac{1}{2} \|y - X\Delta\|_2^2 + \lambda \|\Delta\|_1. \quad (4)$$

Intuitively, the decomposition together with the  $\ell_1$  penalty works as a nested group lasso penalty for the components of  $\mu$ , where the groups correspond to clades of  $\mathcal{T}$ , while the constraint  $T\Delta \in \mathbb{R}^m$  forces components of  $\mu$  to be non positive. For compactity, we note the feasible set  $\mathcal{D} = \{\Delta \in \mathbb{R}^n \text{ s.t. } T\Delta \in \mathbb{R}^m\}$ . Finally, we use the Cholesky decomposition  $\Sigma^{-1} = R^T R$  to simplify the problem into the very well studied optimisation problem:

$$\hat{\Delta} = \underset{\Delta \in \mathcal{D}}{\operatorname{argmin}} \frac{1}{2} \|y - X\Delta\|_2^2 + \lambda \|\Delta\|_1 \quad (5)$$

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

**Estimation of  $\hat{\Sigma}$  and tuning of  $\lambda$ .** Remember first that  $\Sigma$  is completely determined by  $\alpha_{on}$  because of the link between  $\alpha_{on}$  and  $\sigma_{on}^2$ . There are no closed-form expression for the maximum likelihood estimator of  $\alpha_{on}$ . We therefore

resort to numerical optimisation. To tune the parameter  $\lambda$ , we test several values to estimate models with different sparsity levels and select the best one using a BIC criterior.

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

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

#### 3.2.2 Confidence intervals

Lasso procedures are known to produce biased estimators and do not return confidence intervals for the point estimate  $\hat{\mu}$ . Instead of simply returning all negative components of  $\hat{\mu} = T\hat{\Delta}$ , we first unbias the estimates and construct confidence intervals for the components of  $\Delta$ , and in turn of  $\hat{\mu}$ , using the debiasing procedure of Javanmard and Montanari (2013, 2014); Zhang and Zhang (2014).

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

$$(\hat{\Delta}^{(\text{init})}, \hat{\sigma}) = \underset{\Delta \in \mathcal{D}, \sigma > 0}{\operatorname{argmin}} \frac{\|y - X\hat{\Delta}\|_2^2}{2\sigma m} + \frac{\sigma}{2} + \lambda \|\Delta\|_1 \quad (7)$$

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

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

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

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

but the decorrelation matrix  $S$  is computed differently: by inverting  $\hat{\Sigma}$  in a colwise fashion. Column  $s_j$  is solution of the optimization problem:

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

8

9

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

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

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

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

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

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

$$p_i^{\text{ss}} = \Phi \left( \frac{t_i^T \hat{\Delta}}{(t_i^T V t_i)^{1/2}} \right). \quad (11)$$

### 3.2.3 FDR control

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

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

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

$$q_i^{\text{ss}} := \frac{p_i^{\text{ss}} \alpha}{\Phi(-t^*)} \leq \alpha. \quad (12)$$

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

10

### 3.2.4 Algorithm

The algorithm 1 summarise our procedure.

#### Algorithm 1 Zazou procedure

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

## 4 Sign-constrained lasso

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

$$\hat{\Delta} = \underset{\Delta \in \mathbb{R}^m}{\operatorname{argmin}} \frac{1}{2} \|y - X\Delta\|_2^2 + \lambda \|\Delta\|_1$$

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

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

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

Let  $I_+ = \{i : v_i > 0\}$  and  $I_- = \{i : v_i < 0\}$  and note  $\theta_{\max} = \min_{I_+} \{-u_i/v_i\}$  and  $\theta_{\min} = \max_{I_-} \{-u_i/v_i\}$  with the usual conventions that  $\max(\emptyset) = -\infty$  and

11

4. apply a fold-change to the observed abundance of differentially abundant taxa in group B.

This non-parametric simulation scheme was previously used in Bichat et al. (2020). We considered two variants for step 3, respectively called *positive* and *negative*. In the negative variant, differentially abundant taxa were selected randomly across the tree, so the phylogeny is not informative. In the positive variant, taxa are instead selected in a phylogenetically consistent manner. Formally, the phylogeny was first used to compute the cophenetic (Sneath et al., 1973) distance matrix between taxa. A partioning around medoids algorithm was then used to create cluster of related species. One or more clusters were then picked at random and all species in those clusters were selected as differentially abundant.

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

**Positive simulations.** The results of positive simulations (*i.e.* where the phylogeny is informative) are shown in Figure 3. All correction methods have control the FDR at the target rate or below when the fold change is larger than 5. For smaller fold changes, both SS and CI variations of zazou exhibit nominal FDR slightly above the target level (up to 9% in the worst case). In all settings, BY had the lowest TPR, whereas TF was comparable to vanilla BH, in line with results of Bichat et al. (2020). Finally, zazou (both SS and CI variations) had the best overall TPR, with largest gains observed in the lowest fold-change setting.

The higher than intended FDR of zazou methods suggests that the problem of finding an adequate threshold for  $p_i^{\text{ss}}$  not completely solved by Javanmard et al. (2019) procedure. To assess the performance of zazou in a threshold-independent manner, we also compared the AUC of all procedures. Fig. shows that zazou (both variants) have higher AUC than all other methods. As reported previously, TF and BH have are at the same level and BY has the lowest ROC curve. Focus on the beginning of left hand side of the curve shows that zazou is more efficient starting from the first discoveries.

**Negative simulations.** The negative simulations are designed to assess the robustness of our algorithm with respect to uninformative phylogenies, or equivalently misspecified hierarchies. Fig. 5 shows that, as expected, standard BH outperforms competing methods (in terms of AUC) when the tree is misspecified. Forcing an inadequate tree structure results in AUC losses ranging from 15 to 20 points compared to no structure. The puzzling lack of AUC loss for the TF procedure is explained by an implementation trick: TF always performs BH

min( $\emptyset$ ) =  $+\infty$ . Problem (13) is feasible if and only if (i)  $\theta_{\min} \leq \theta_{\max}$  and (ii) for all  $i$ ,  $v_i = 0 \Rightarrow u_i \leq 0$ , in which case the feasible region is  $[\theta_{\min}, \theta_{\max}]$ . Computing the subgradient  $\partial h(\theta)$  of  $h$  and looking for values  $\theta$  such that  $0 \in \partial h(\theta)$  leads to the usual shrinked estimates:

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

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

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

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

## 5 Synthetic Data

### 5.1 Metagenomics

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

Most DAT procedures proceed with univariate tests (one test per species) followed by a correction procedure. In the synthetic datasets, we consider discrete covariates only. Since our goal is to compare correction procedures, we always use Wilcoxon or Kruskall-Wallis tests for the first step.

### 5.2 Simulations

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

1. start with an homogeneous dataset,
2. assign each sample to group A or B at random
3. select differentially abundant taxa in a phylogenetically consistent manner (differentially abundant taxa)

12

13

## Packages R

### yatah

Lorsque l'on travaille avec les données taxonomiques, il est courant d'avoir des lignes taxonomiques écrites sous la forme `k_Bacterial|p_Firmicutes|c_Bacilli` pour décrire les différents clades auxquels appartient un taxon donné. `{yatah}` fournit une série de fonctions basées sur des expressions régulières pour manipuler de telles lignées : filtrer les lignées appartenant à un clade spécifique, ne conserver que le dernier rang ou encore créer la table et l'arbre taxonomiques associés. Au moment de la rédaction de ce manuscrit, la version 0.1.0 est disponible sur le CRAN.

### evabic

`{evabic}`, pour *EVAluation of BIrary Classifiers*, permet de calculer facilement des métriques associées à des procédures de classification binaires, typiquement rejet ou non de l'hypothèse nulle. Il a été conçu pour bien s'intégrer dans une logique de code orienté `{dplyr}` et `{tidyverse}` (Wickham et al., 2019). Au moment de la rédaction de ce manuscrit, la version 0.0.3 est disponible sur le CRAN.

### correlationtree

La création de l'arbre des corrélations proposée dans la section 3.3.3 a été implémentée dans le package `{correlationtree}`. Au moment de la rédaction de ce manuscrit, la version 0.0.3 est disponible sur GitHub.

### zazou

L'algorithme *zazou*, décrite dans la section 4.2, est implémenté dans le package éponyme. Il bénéficie du package `{ggtree}` (Yu, Smith, Zhu, Guan, & Lam, 2017) pour mettre en regard l'arbre et différentes variables associées aux feuilles. Un exemple d'utilisation est donnée dans l'annexe C.



# Annexe C

## Vignette de zazou

Nous commençons par charger les *packages* qui seront utilisés dans cette vignette.

```
library(zazou)
library(evabic)
library(dplyr)
library(ggplot2)
set.seed(20201209)
```

On récupère ensuite le jeu de données de Wu et al. (2011), où l'on ne conserve que les individus avec une faible consommation d'alcool. On modifie les longueurs de branches de la phylogénie pour que celle-ci soit ultramétrique.

```
data("alcohol")
abund <- alcohol$X[, alcohol$Y == "Low"]
tree <- force_ultrametric(alcohol$tree)
```

On filtre ensuite les OTUs qui sont présentes dans au moins 25 échantillons parmi les 49.

```
abund <- abund[rowSums(abund > 0) > 25, ]
tree <- ape::drop.tip(tree, setdiff(tree$tip.label, rownames(abund)))
```

Il en reste alors 91.

Nous assignons ensuite chaque échantillon aléatoirement à un groupe *A* ou *B*.

```
groups <- sample(c("A", "B"), size = ncol(abund), replace = TRUE)
pvalues_original <- test_wilcoxon(abund, groups)$p.value
zscores_original <- p2z(pvalues_original)
sum(pvalues_original < 0.05)
```

```
[1] 7
```

Certaines OTUs sont déjà considérées comme différentiellement abondantes.

L'ensemble des OTUs est partitionné en 20 groupes cohérents avec la phylogénie, comme expliqué dans la section 4.3.1. Puis 5 de ces groupes et les OTUs qu'ils contiennent sont tirés aléatoirement pour être différentiellement abondants.

```
clustering <- create_clusters(tree, N_clusters = 20,
                               method = "paraphyletic")
clusters_da <- sample(20, 5)
clusters_da
```

```
[1] 5 4 6 1 17
```

```
otus_da <- names(clustering[which(clustering %in% clusters_da)])
1 - length(otus_da) / nrow(abund)
```

```
[1] 0.6483516
```

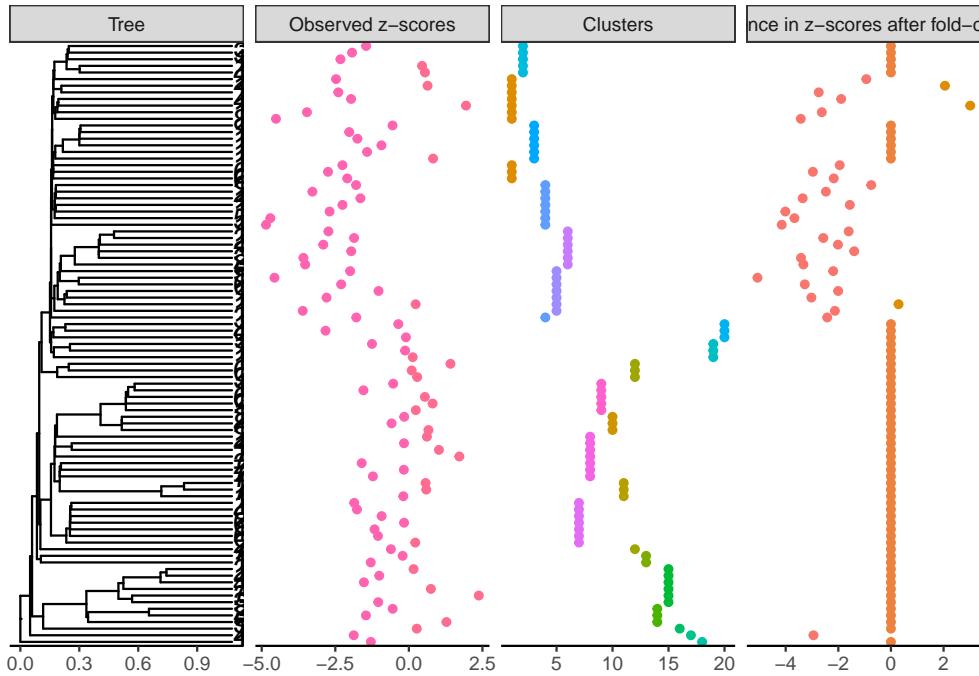
On applique ensuite un *fold-change* de 10 aux OTUs différentiellement abondants dans le groupe *B*.

```
abund[otus_da, groups == "B"] <- 10 * abund[otus_da, groups == "B"]
```

On effectue un test de Wilcoxon sur ce nouveau jeu de données.

```
pvalues <- test_wilcoxon(abund, groups)$p.value
zscores <- p2z(pvalues)
plot_shifts(tree, NA, obs_scores = zscores,
            sup_scores = list(
              list(scores = clustering,
                   title = "Clusters",
                   color = as.character(clustering)),
              list(scores = zscores - zscores_original,
                   title = "Difference in z-scores after fold-change",
                   color = as.character(sign(zscores - zscores_original)))
            ))
```

Warning: Removed 181 rows containing missing values (geom\_label).



Nous donnons ensuite nos  $z$ -scores à la fonction `estimate_shifts` pour déterminer la position idéale des sauts avec une régression lasso. La grille des  $\alpha_{\text{ou}}$  est également spécifiée mais celles des  $\lambda$  est déterminée automatiquement.

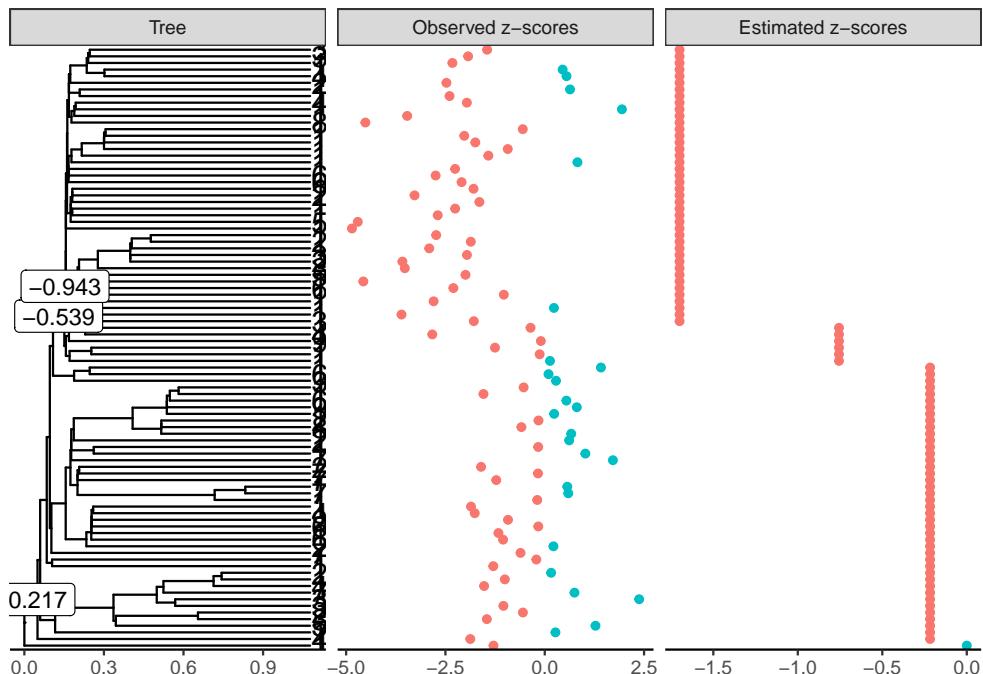
```
estimation_lasso <- estimate_shifts(zscores = zscores, tree = tree,
                                         alphaOU = c(0.2, 0.5, 1, 2, 5),
                                         method = "lasso")
estimation_lasso
```

```
Tree is binary with 91 leafs and 180 branches
Covariance matrix has been estimated from an OU with alpha = 5 and sigma = 3.162
---
Method: lasso with model selection
Regularization parameter: lambda = 14.068
Objective value: 100.466
BIC: 333.88
pBIC: 372.38
---
Estimated shifts: 0 -0.217 0 0 0 -0.539 -0.943 0 0 ...
3 shifts have been identified (ie 98.3 % of sparsity)
A parsimonious solution would involve 3 shifts
---
```

```
Observed z-scores: -2.474 -1.929 -1.451 -2.324 0.549 0.45 -2.397 0.636 -3.463 1.9
Estimated z-scores: -1.7 -1.7 -1.7 -1.7 -1.7 -1.7 -1.7 -1.7 -1.7 -1.7 -1.7
90 z-scores have been shifted (ie 1.1 % of sparsity)
```

```
plot(estimation_lasso)
```

Warning: Removed 178 rows containing missing values (geom\_label).



Trois sauts sur l'arbre ont été détectées, sur des branches ayant beaucoup de descendants.

Si nous voulons avoir des  $p$ -valeurs et des intervalles de confiance, il faut utiliser une régression *scaled lasso* avant d'appliquer la fonction `estimate_confint`.

```
estimation_scaledlasso <-
  estimate_shifts(zscores = zscores,
                  tree = tree, alphaOU = c(0.2, 0.5, 1, 2, 5),
                  method = "scaled lasso")
estimation_scoresystem <-
  estimate_confint(estimation_scaledlasso,
                  method = "score system")
```

`t_star` is not feasible, falling back to default value.

```
estimation_scoresystem
```

Tree is binary with 91 leafs and 180 branches

Method: score system

Confidence threshold: 0.05

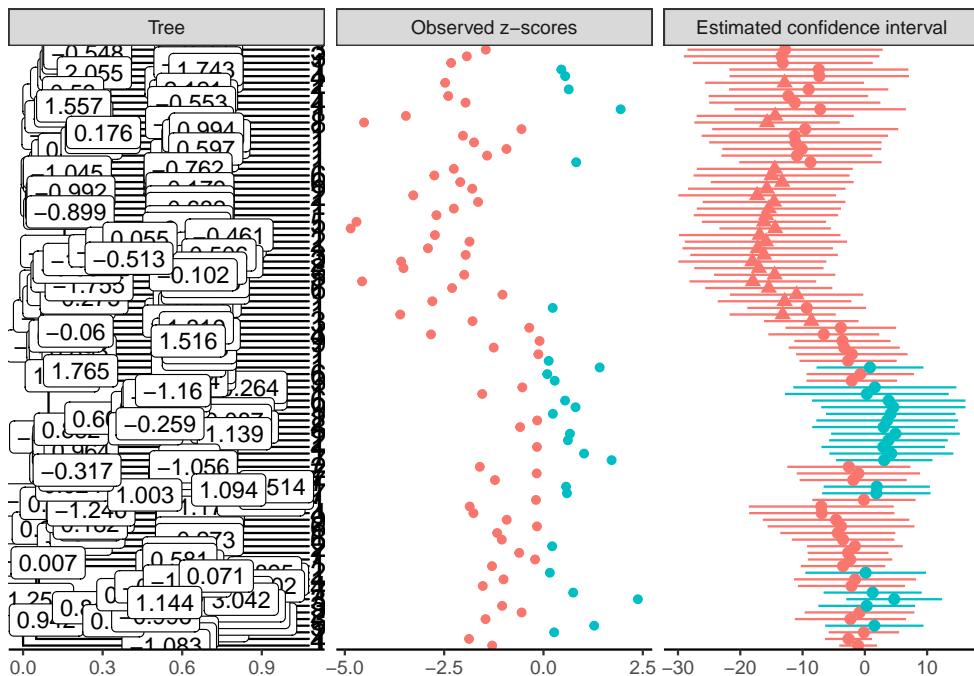
---

z-scores:

	leaf	estimate	lower	upper	pvalue	qvalue
1	283	-12.887241	-25.644441	-0.1300754	0.02385467	0.8941353
2	3494	-13.363635	-29.06041	2.3331434	0.04759456	1.0000000
3	3470	-12.803510	-28.47336	2.8663408	0.05463910	1.0000000
4	1661	-13.171378	-27.61457	1.2718194	0.03693836	1.0000000
5	4206	-7.310320	-21.70978	7.0891438	0.15985974	1.0000000
6	10038	-7.386608	-21.75859	6.9853783	0.15688556	1.0000000
		...				

```
plot(estimation_scoresystem)
```

Warning: Removed 1 rows containing missing values (geom\_label).

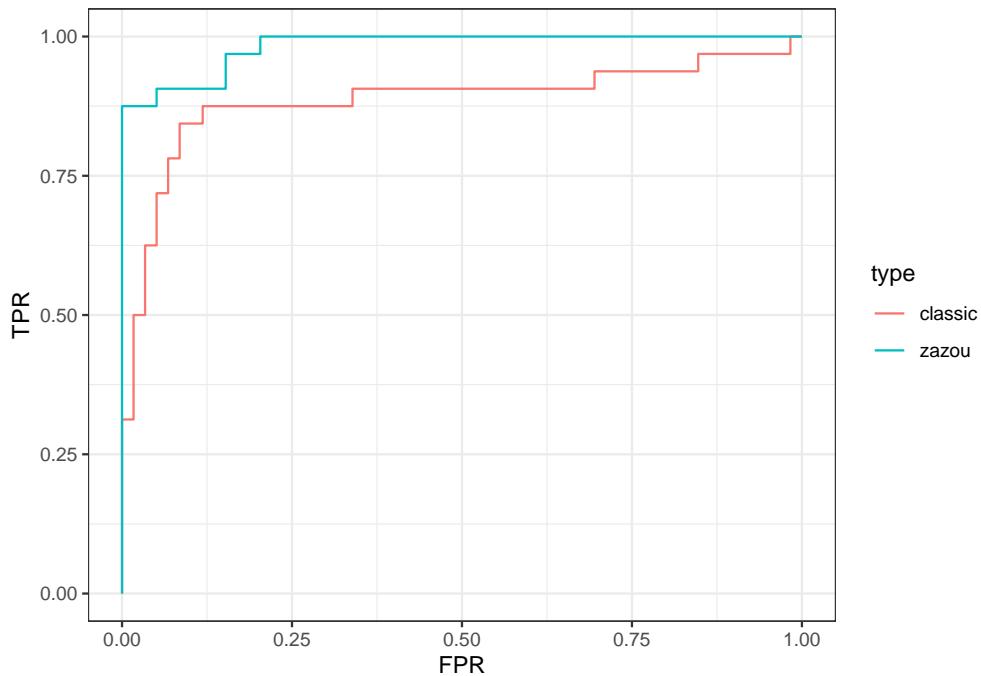


```
pvalues_smoothed <- pull_pvalues(estimation_scoresystem)
```

Après le débiaisage, les sauts dans les branches ne sont plus parcimonieux.

Nous allons regarder les performances de notre prédicteurs via une courbe ROC.

```
df_measures_zazou <-
  ebc_tidy_by_threshold(detection_values = pvalues_smoothed,
                        true = otus_da, all = rownames(abund))
df_measures_classic <-
  ebc_tidy_by_threshold(detection_values = pvalues,
                        true = otus_da, all = rownames(abund))
bind_rows(mutate(df_measures_zazou, type = "zazou"),
          mutate(df_measures_classic, type = "classic")) %>%
  ggplot() +
  aes(x = FPR, y = TPR, color = type) +
  geom_line() +
  theme_bw()
```



La courbe ROC de *zazou* est au dessus de celle de la méthode standard. Et ceci se traduit dans en terme d'AUC :

```
ebc_AUC_from_measures(df_measures_zazou)
```

```
[1] 0.9825212
```

```
ebc_AUC_from_measures(df_measures_classic)
```

```
[1] 0.8850636
```

Version de {zazou} utilisée dans cette analyse :

```
packageVersion("zazou")
```

```
[1] '0.0.1'
```



# Références

- Abrahamsson, T. R., Jakobsson, H. E., Andersson, A. F., Björkstén, B., Engstrand, L., & Jenmalm, M. C. (2012). Low diversity of the gut microbiota in infants with atopic eczema. *Journal of Allergy and Clinical Immunology*, 129(2), 434–440.
- Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society : Series B (Methodological)*, 44(2), 139–160.
- Aitchison, J. (1986). The statistical analysis of compositional data. Monographs on statistics and applied probability (reprinted in 2003).
- Albarède, F. (1996). *Introduction to geochemical modeling*. Cambridge University Press.
- Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., ... others. (2020). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, 1–10.
- Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, 1–1.
- Araya, J. P., González, M., Cardinale, M., Schnell, S., & Stoll, A. (2020). Microbiome dynamics associated with the atacama flowering desert. *Frontiers in Microbiology*, 10, 3160.
- Aronson, H. S., Zellmer, A. J., & Goffredi, S. K. (2017). The specific and exclusive microbiome of the deep-sea bone-eating snail, rubyspira osteovora. *FEMS Microbiology Ecology*, 93(3), fiw250.
- Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., ... others. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346), 174–180.
- Bastide, P., Ané, C., Robin, S., & Mariadassou, M. (2018). Inference of adaptive shifts for multivariate correlated traits. *Systematic Biology*, 67(4), 662–680.

- Bastide, P., Mariadassou, M., & Robin, S. (2017). Detection of adaptive shifts on phylogenies by using shifted stochastic processes on a tree. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 79(4), 1067–1093.
- Baudry, J.-P., Maugis, C., & Michel, B. (2012). Slope heuristics : Overview and implementation. *Statistics and Computing*, 22(2), 455–470.
- Bedarf, J. R., Hildebrand, F., Coelho, L. P., Sunagawa, S., Bahram, M., Goeser, F., ... Wüllner, U. (2017). Functional implications of microbial and viral gut metagenome changes in early stage l-dopa-naïve parkinson's disease patients. *Genome Medicine*, 9(1), 1–13.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society : Series B (Methodological)*, 57(1), 289–300.
- Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 1165–1188.
- Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., & Lemaitre, C. (2016). Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*, 2, e94.
- Bichat, A., Plassais, J., Ambroise, C., & Mariadassou, M. (2020). Incorporating phylogenetic information in microbiome differential abundance studies has no effect on detection power and fdr control. *Frontiers in Microbiology*, 11, 649. <http://doi.org/10.3389/fmicb.2020.00649>
- Billera, L. J., Holmes, S. P., & Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4), 733–767.
- Blander, J. M., Longman, R. S., Iliev, I. D., Sonnenberg, G. F., & Artis, D. (2017). Regulation of inflammation by microbiota interactions with the host. *Nature Immunology*, 18(8), 851–860.
- Bokulich, N. A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., ... others. (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Science Translational Medicine*, 8(343), 343ra82–343ra82.
- Brady, A., & Salzberg, S. L. (2009). Phymm and phymmbl : Metagenomic phylogenetic classification with interpolated markov models. *Nature Methods*, 6(9), 673–676.
- Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., ... others. (2016). Mobile genes in the human microbiome are structured from global to

- individual scales. *Nature*, 535(7612), 435–439.
- Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11(12), 2639–2643.
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2 : High-resolution sample inference from illumina amplicon data. *Nature Methods*, 13(7), 581.
- Canani, R. B., Di Costanzo, M., Leone, L., Pedata, M., Meli, R., & Calignano, A. (2011). Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. *World Journal of Gastroenterology : WJG*, 17(12), 1519.
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... others. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5), 335.
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., ... Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, 108(Supplement 1), 4516–4522.
- Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics*, 19(3 Pt 1), 233.
- Cekanaviciute, E., Yoo, B. B., Runia, T. F., Debelius, J. W., Singh, S., Nelson, C. A., ... others. (2017). Gut bacteria from multiple sclerosis patients modulate human t cells and exacerbate symptoms in mouse models. *Proceedings of the National Academy of Sciences*, 114(40), 10713–10718.
- Chaillou, S., Chaulot-Talmon, A., Caekebeke, H., Cardinal, M., Christieans, S., Denis, C., ... others. (2015). Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. *The ISME Journal*, 9(5), 1105–1118.
- Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., & Ballman, K. (2018). An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics*, 34(4), 643–651.
- Chene, L., Sader, C. D., Magalhaes, J., Strozzi, F., Tibaldi, L., Mendez, C., ... Bonny, C. (2019). Microbiome derived peptides stimulate strong immune response against tumor associated antigens and trigger in vivo tumor regression after vaccination. AACR.

- Chong, P. P., Chin, V. K., Looi, C. Y., Wong, W. F., Madhavan, P., & Yong, V. C. (2019). The microbiome and irritable bowel syndrome—a review on the pathophysiology, current research and future therapy. *Frontiers in Microbiology*, 10, 1136.
- Coelho, L. P., Alves, R., Monteiro, P., Huerta-Cepas, J., Freitas, A. T., & Bork, P. (2019). NG-meta-profiler : Fast processing of metagenomes using ngless, a domain-specific language. *Microbiome*, 7(1), 84.
- Cuthbertson, L., Rogers, G. B., Walker, A. W., Oliver, A., Hafiz, T., Hoffman, L. R., ... Van Der Gast, C. J. (2014). Time between collection and storage significantly influences bacterial sequence composition in sputum samples from cystic fibrosis respiratory infections. *Journal of Clinical Microbiology*, 52(8), 3011–3016.
- David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., ... others. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484), 559–563.
- Deorowicz, S., Kokot, M., Grabowski, S., & Debdaj-Grabysz, A. (2015). KMC 2 : Fast and resource-frugal k-mer counting. *Bioinformatics*, 31(10), 1569–1576.
- DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with arb. *Applied and Environmental Microbiology*, 72(7), 5069–5072.
- Ding, X., Zhang, F., Li, Q., Ting, Z., Cui, B., & Li, P. (2019). Selective microbiota transplantation is effective for controlling tourette's syndrome. *Gastroenterology*, 156(6), S–456.
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19), 2460–2461.
- Edgar, R. C. (2016). UCHIME2 : Improved chimera prediction for amplicon sequencing. *BioRxiv*, 074252.
- Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), 279–300.
- Ekekezie, C., Perler, B. K., Wexler, A., Duff, C., Lillis, C. J., & Kelly, C. R. (2020). Understanding the scope of do-it-yourself fecal microbiota transplant. *American Journal of Gastroenterology*, 115(4), 603–607.

- Eloe-Fadrosh, E. A., McArthur, M. A., Seekatz, A. M., Drabek, E. F., Rasko, D. A., Sztein, M. B., & Fraser, C. M. (2013). Impact of oral typhoid vaccination on the human gut microbiota and correlations with s. Typhi-specific immunological responses. *PloS One*, 8(4), e62026.
- Eren, A. M., Borisy, G. G., Huse, S. M., & Welch, J. L. M. (2014). Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Sciences*, 111(28), E2875–E2884.
- Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., & Sogin, M. L. (2013). Oligotyping : Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, 4(12), 1111–1119.
- Fan, D., Coughlin, L. A., Neubauer, M. M., Kim, J., Kim, M. S., Zhan, X., ... Koh, A. Y. (2015). Activation of hif-1 $\alpha$  and ll-37 by commensal bacteria inhibits candida albicans colonization. *Nature Medicine*, 21(7), 808.
- Felsenstein, J. (1985). Confidence limits on phylogenies : An approach using the bootstrap. *Evolution*, 39(4), 783–791.
- Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets : Characterizing rna-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2(1), 15.
- Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P., & Forano, E. (2012). Microbial degradation of complex carbohydrates in the gut. *Gut Microbes*, 3(4), 289–306.
- Foster, J. A., & Neufeld, K.-A. M. (2013). Gut–brain axis : How the microbiome influences anxiety and depression. *Trends in Neurosciences*, 36(5), 305–312.
- Freckleton, R. P., Harvey, P. H., & Pagel, M. (2003). Bergmann's rule and body size in mammals. *The American Naturalist*, 161(5), 821–825.
- Fu, W. J. (1998). Penalized regressions : The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3), 397–416.
- Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., ... Bryant, S. H. (2010). The ncbi biosystems database. *Nucleic Acids Research*, 38(suppl\_1), D492–D496.
- Gibson, G. R., Hutkins, R., Sanders, M. E., Prescott, S. L., Reimer, R. A., Salminen, S. J., ... others. (2017). Expert consensus document : The international scientific association for probiotics and prebiotics (isapp) consensus statement

- on the definition and scope of prebiotics. *Nature Reviews Gastroenterology & Hepatology*, 14(8), 491.
- Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egoscue, J. J. (2017). Microbiome datasets are compositional : And this is not optional. *Frontiers in Microbiology*, 8, 2224.
- Gloor, G. B., & Reid, G. (2016). Compositional analysis : A valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, 62(8), 692–703.
- Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., & Egoscue, J. J. (2016). It's all relative : Analyzing microbiome data as compositions. *Annals of Epidemiology*, 26(5), 322–329.
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4), 325–338.
- Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M., & Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to japanese gut microbiota. *Nature*, 464(7290), 908–912.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.
- Holmes, I., Harris, K., & Quince, C. (2012). Dirichlet multinomial mixtures : Generative models for microbial metagenomics. *PloS One*, 7(2), e30126.
- Huang, R., Soneson, C., Germain, P.-L., Schmidt, T., Mering, C. von, & Robinson, M. (2020). TreeclimbR pinpoints the data-dependent resolution of hierarchical hypotheses. <http://doi.org/10.1101/2020.06.08.140608>
- Jaglin, M., Rhimi, M., Philippe, C., Pons, N., Bruneau, A., Goustad, B., ... Rabot, S. (2018). Indole, a signaling molecule produced by the gut microbiota, negatively impacts emotional behaviors in rats. *Frontiers in Neuroscience*, 12, 216.
- Javanmard, A., Javadi, H., & others. (2019). False discovery rate control via de-biased lasso. *Electronic Journal of Statistics*, 13(1), 1212–1253.
- Javanmard, A., & Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional statistical models. In *Advances in neural information processing systems* (pp. 1187–1195).
- Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Re-*

- search, 15(1), 2869–2909.
- Jiang, L., Amir, A., Morton, J. T., Heller, R., Arias-Castro, E., & Knight, R. (2017). Discrete false-discovery rate improves identification of differentially abundant microbes. *MSystems*, 2(6).
- Jombart, T., Kendall, M., Almagro-Garcia, J., & Colijn, C. (2017). Treospace : Statistical exploration of landscapes of phylogenetic trees. *Molecular Ecology Resources*, 17(6), 1385–1392.
- Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., ... others. (2017). Where less may be more : How the rare biosphere pulls ecosystems strings. *The ISME Journal*, 11(4), 853–862.
- Kates, A. E., Jarrett, O., Skarlupka, J. H., Sethi, A., Duster, M., Watson, L., ... Safdar, N. (2020). Household pet ownership and the microbial diversity of the human gut microbiota. *Frontiers in Cellular and Infection Microbiology*, 10, 73.
- Kelly, T. N., Bazzano, L. A., Ajami, N. J., He, H., Zhao, J., Petrosino, J. F., ... He, J. (2016). Gut microbiome associates with lifetime cardiovascular disease risk profile among bogalusa heart study participants. *Circulation Research*, 119(8), 956–964.
- Kembel, S. W., Wu, M., Eisen, J. A., & Green, J. L. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol*, 8(10), e1002743.
- Khabbazian, M., Kriebel, R., Rohe, K., & Ané, C. (2016). Fast and accurate detection of evolutionary shifts in ornstein–uhlenbeck models. *Methods in Ecology and Evolution*, 7(7), 811–824.
- Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge : Rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12), 1721–1729.
- Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621.
- Kultima, J. R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D. R., ... others. (2012). MOCAT : A metagenomics assembly and gene prediction toolkit. *PloS One*, 7(10), e47656.
- Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124(4), 837–848.
- Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., & Pop, M. (2011). Accurate and

- fast estimation of taxonomic profiles from metagenomic shotgun sequences. *Genome Biology*, 12(1), 1–27.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12), 550.
- Mai, V., Young, C. M., Ukhanova, M., Wang, X., Sun, Y., Casella, G., ... others. (2011). Fecal microbiota in premature infants prior to necrotizing enterocolitis. *PloS One*, 6(6), e20647.
- Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker Jr, C. T., Saxman, P. R., Stredwick, J. M., ... others. (2000). The rdp (ribosomal database project) continues. *Nucleic Acids Research*, 28(1), 173–174.
- Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., & Woese, C. R. (1997). The rdp (ribosomal database project). *Nucleic Acids Research*, 25(1), 109–110.
- Maillet, N., Collet, G., Vannier, T., Lavenier, D., & Peterlongo, P. (2014). COM-MET : Comparing and combining multiple metagenomic datasets. In *2014 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 94–98). IEEE.
- Maillet, N., Lemaitre, C., Chikhi, R., Lavenier, D., & Peterlongo, P. (2012). Compareads : Comparing huge metagenomic experiments. In *BMC bioinformatics* (Vol. 13, p. S10). Springer.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60.
- Mathieu-Daudé, F., Welsh, J., Vogt, T., & McClelland, M. (1996). DNA rehybridization during pcr : The ‘c o t effect’and its consequences. *Nucleic Acids Research*, 24(11), 2080–2086.
- McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., ... others. (2018). American gut : An open platform for citizen science microbiome research. *Msystems*, 3(3), e00031–18.
- McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.
- McMurdie, P. J., & Holmes, S. (2014). Waste not, want not : Why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4), e1003531.
- Meyerhans, A., Vartanian, J.-P., & Wain-Hobson, S. (1990). DNA recombination during pcr. *Nucleic Acids Research*, 18(7), 1687–1691.

- Modolo, L., & Lerat, E. (2015). UrQt : An efficient software for the unsupervised quality trimming of ngs data. *BMC Bioinformatics*, 16(1), 137.
- Mohty, M., Malard, F., Vekhoff, A., Lapusan, S., Isnard, F., d'Incan, E., ... others. (2018). The odyssee study : Prevention of dysbiosis complications with autologous fecal microbiota transfer (fmt) in acute myeloid leukemia (aml) patients undergoing intensive treatment : Results of a prospective multicenter trial. In *60th annual meeting of the american-society-of-hematology (ash)* (Vol. 132, p. 4). AMER SOC HEMATOLOGY.
- Morgan, X. C., & Huttenhower, C. (2012). Human microbiome analysis. *PLoS Comput Biol*, 8(12), e1002808.
- Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., ... others. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9), R79.
- O'Toole, P. W., & Claesson, M. J. (2010). Gut microbiota : Changes throughout the lifespan from infancy to elderly. *International Dairy Journal*, 20(4), 281–291.
- Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK : Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(1), 236.
- Owen, M., & Provan, J. S. (2010). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1), 2–13.
- Palleja, A., Mikkelsen, K. H., Forslund, S. K., Kashani, A., Allin, K. H., Nielsen, T., ... others. (2018). Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nature Microbiology*, 3(11), 1255–1265.
- Park, M. Y., Hastie, T., & Tibshirani, R. (2007). Averaged gene expressions for regression. *Biostatistics*, 8(2), 212–227.
- Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., ... others. (2017). Accessible, curated metagenomic data through experimenthub. *Nature Methods*, 14(11), 1023.
- Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12), 1200–1202.
- Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana Delgado, R. (2007). Lecture notes on compositional data analysis.

- Philippot, L., Andersson, S. G., Battin, T. J., Prosser, J. I., Schimel, J. P., Whitman, W. B., & Hallin, S. (2010). The ecological coherence of high bacterial taxonomic ranks. *Nature Reviews Microbiology*, 8(7), 523–529.
- Pinto-Sanchez, M. I., Hall, G. B., Ghajar, K., Nardelli, A., Bolino, C., Lau, J. T., ... others. (2017). Probiotic bifidobacterium longum ncc3001 reduces depression scores and alters brain activity : A pilot study in patients with irritable bowel syndrome. *Gastroenterology*, 153(2), 448–459.
- Pistollato, F., Sumalla Cano, S., Elio, I., Masias Vergara, M., Giampieri, F., & Battino, M. (2016). Role of gut microbiota and nutrients in amyloid formation and pathogenesis of alzheimer disease. *Nutrition Reviews*, 74(10), 624–634.
- Plaza Oñate, F., Le Chatelier, E., Almeida, M., Cervino, A. C., Gauthier, F., Magoulès, F., ... Wren, J. (2018). MSPminer : Abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics*.
- Polyak, B. T. (1987). Introduction to optimization. Optimization software. Inc., Publications Division, New York, 1.
- Pons, N., Batto, J.-M., Kennedy, S., Almeida, M., Boumezbeur, F., Moumen, B., & others. (2010). METEOR, a platform for quantitative metagenomic profiling of complex ecosystems. *Journées Ouvertes En Biologie, Informatique et Mathématiques* <Http://Www.Jobim2010.Fr/Sites/Default/Files/Presentations/27Pons.Pdf>.
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS One*, 5(3), e9490.
- Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., ... others. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418), 55–60.
- Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., ... others. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513(7516), 59–64.
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O. (2012). The silva ribosomal rna gene database project : Improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1), D590–D596.
- Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9), 833–844.

- Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., ... others. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(Supplement 1), 4680–4687.
- R Core Team. (2020). *R : A language and environment for statistical computing*. Vienna, Austria : R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Reardon, S. (2018). Faecal transplants could help preserve vulnerable species. *Nature*, 558(7709), 173–175.
- Regier, Y., Komma, K., Weigel, M., Kraiczy, P., Laisi, A., Pulliainen, A. T., ... Kempf, V. A. (2019). Combination of microbiome analysis and serodiagnostics to assess the risk of pathogen transmission by ticks to humans and animals in central germany. *Parasites & Vectors*, 12(1), 11.
- Reynolds, A. P., Richards, G., Iglesia, B. de la, & Rayward-Smith, V. J. (2006). Clustering rules : A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4), 475–504.
- Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2), 131–147.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). EdgeR : A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140.
- Robinson, M. D., & Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21), 2881–2887.
- Rowland, I., Gibson, G., Heinken, A., Scott, K., Swann, J., Thiele, I., & Tuohy, K. (2018). Gut microbiota functions : Metabolism of nutrients and other food components. *European Journal of Nutrition*, 57(1), 1–24.
- Sakwinska, O., Berger, B., Zolezzi, I. S., & Holbrook, J. (2017). Prebiotics for reducing the risk of obesity later in life. WO2016026684A1.
- Sankaran, K., & Holmes, S. (2014). StructSSI : Simultaneous and selective inference for grouped or hierarchically structured data. *Journal of Statistical Software*, 59(13), 1.
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... others. (2009). Introducing mothur : Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23), 7537–7541.

- Schretter, C. E., Vielmetter, J., Bartos, I., Marka, Z., Marka, S., Argade, S., & Mazmanian, S. K. (2018). A gut microbial factor modulates locomotor behaviour in drosophila. *Nature*, 563(7731), 402–406.
- Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8), 811–814.
- Sergeant, M. J., Constantinidou, C., Cogan, T., Penn, C. W., & Pallen, M. J. (2012). High-throughput sequencing of 16S rRNA gene amplicons : Effects of extraction procedure, primer length and annealing temperature. *PloS One*, 7(5), e38094.
- Sharon, G., Segal, D., Ringo, J. M., Hefetz, A., Zilber-Rosenberg, I., & Rosenberg, E. (2010). Commensal bacteria play a role in mating preference of drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 107(46), 20051–20056.
- Silverman, J. D., Washburne, A. D., Mukherjee, S., & David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *Elife*, 6, e21887.
- Sneath, P. H., Sokal, R. R., & others. (1973). *Numerical taxonomy. The principles and practice of numerical classification*.
- Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendograms by objective methods. *Taxon*, 11(2), 33–40.
- Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R., & Schmidt, T. M. (2015). Rrn db : Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research*, 43(D1), D593–D598.
- Stokholm, J., Blaser, M. J., Thorsen, J., Rasmussen, M. A., Waage, J., Vinding, R. K., . . . others. (2018). Maturation of the gut microbiome and risk of asthma in childhood. *Nature Communications*, 9(1), 1–10.
- Sun, T., & Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4), 879–898.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*, 58(1), 267–288.
- Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., . . . Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic

- profiling. *Nature Methods*, 12(10), 902–903.
- Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., ... others. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228), 480–484.
- Valdez, Y., Brown, E. M., & Finlay, B. B. (2014). Influence of the microbiota on vaccine effectiveness. *Trends in Immunology*, 35(11), 526–537.
- Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., ... others. (2017). Quantitative microbiome profiling links gut community variation to microbial load. *Nature*, 551(7681), 507–511.
- Van Nood, E., Vrieze, A., Nieuwdorp, M., Fuentes, S., Zoetendal, E. G., Vos, W. M. de, ... others. (2013). Duodenal infusion of donor feces for recurrent clostridium difficile. *New England Journal of Medicine*, 368(5), 407–415.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <http://doi.org/10.21105/joss.01686>
- Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics* (pp. 196–202). Springer.
- Wilgenbusch, J. C., Huang, W., & Gallivan, K. A. (2017). Visualizing phylogenetic tree landscapes. *BMC Bioinformatics*, 18(1), 85.
- Wood, D. E., & Salzberg, S. L. (2014). Kraken : Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3), 1–12.
- Wright, E. S., Yilmaz, L. S., & Noguera, D. R. (2012). DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Applied and Environmental Microbiology*, 78(3), 717–725.
- Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., ... others. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052), 105–108.
- Xia, Y., Sun, J., & Chen, D.-G. (2018). *Statistical analysis of microbiome data with r*. Springer.
- Xiao, J., Cao, H., & Chen, J. (2017). False discovery rate control incorporating phylogenetic tree increases detection power in microbiome-wide multiple testing. *Bioinformatics*, 33(18), 2873–2881.
- Xinyan, Z., Himel, M., & Nengjun, Y. (2016). Zero-inflated negative binomial

- regression for differential abundance testing in microbiome studies. *Journal of Bioinformatics and Genomics*, (2), 1–1.
- Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., ... others. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402), 222–227.
- Yekutieli, D. (2008). Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association*, 103(481), 309–316.
- Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1), 28–36.
- Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., ... others. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11), 766.
- Zhang, C.-H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 76(1), 217–242.
- Zhang, F., Luo, W., Shi, Y., Fan, Z., & Ji, G. (2012). Should we standardize the 1,700-year-old fecal microbiota transplantation ? *American Journal of Gastroenterology*, 107(11), 1755.
- Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A. K., & Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 18(1), 4.
- Zheng, P., Zeng, B., Liu, M., Chen, J., Pan, J., Han, Y., ... others. (2019). The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-gaba cycle and schizophrenia-relevant behaviors in mice. *Science Advances*, 5(2), eaau8317.
- Zhernakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., ... others. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285), 565–569.



**Titre :** Prise en compte de l'organisation hiérarchique des espèces pour la découverte de signatures métagénomiques multi-échelles

**Mots clés :** Statistique – Apprentissage – Métagénomique – Arbre phylogénétique – Tests multiples – Processus stochastiques

**Résumé :** Cette thèse porte sur l'inclusion d'informations hiérarchiques dans des procédures de détection d'abondance différentielle sur des données métagénomiques. Les différents taxons qui composent le microbiote sont généralement accompagnés d'un arbre, comme la taxonomie ou la phylogénie, qui traduit une proximité biologique entre eux. Il est alors naturel de vouloir tirer parti de cette information hiérarchique afin d'augmenter la puissance des tests de détection de taxons différemment abondants. Dans un premier temps, nous nous

sommes intéressés aux performances des procédures hiérarchiques existantes et à l'impact du choix de l'arbre sur celles-ci. Dans un second temps, nous avons développé notre propre méthode hiérarchique de détection d'abondance différentielle. Celle-ci modélise les  $z$ -scores associés à chaque taxon comme la réalisation d'un processus d'Ornstein-Uhlenbeck sur arbre avec sauts dans la valeur optimale du processus puis effectue une régression de type lasso pour déterminer les positions et intensités optimales des sauts.

**Title:** Discovering multi-scale metagenomic signatures through hierarchical organization of species

**Keywords:** Statistics – Machine learning – Metagenomics – Phylogenetic tree – Multiple testing – Stochastic processes

**Abstract:** This thesis deals with the use of hierarchical information in differential abundance analyses in metagenomics. Taxa that make up the microbiome are usually associated with a tree, like the taxonomy or the phylogeny, that reflects a biological link between them. It is therefore natural to exploit this hierarchical information to increase the statistical power of differential abundance techniques. We first investigated the efficiency of existing hierarchi-

cal differential abundance detection procedures and the impact of tree choice on those. We then developed our own hierarchical differentially abundance detection procedure. It models the taxa associated  $z$ -scores as realization of an Ornstein-Uhlenbeck process on a tree with shifts on its optimal value then a lasso-like regression is used to identify optimal positions and intensities of the shifts.