



STAGE - ANALYSE DE DONNÉES COMPOSITIONNELLES

Master Informatique - Data Mining

Méthodes et Application à l'Etude du Microbiote de l'air intérieur chez des Patients Asthmatiques

Auteur :
Vishwa ELANKUMARAN

Tutrice Entreprise :
Marta Avalos FERNANDEZ

Tuteur Universitaire :
Julien JACQUES

Collaborateurs :
Laurence DELHAES
Louise-Eva VANDENBORGH

Structure d'accueil : INRIA BORDEAUX SUD-OUEST

2 septembre 2022

Table des matières

Introduction	5
0.1 Présentation	5
0.1.1 Équipe de Recherche	5
0.2 Contexte	6
0.3 Échantillonnage	7
0.4 Objectifs	9
0.4.1 Objectifs Cliniques	9
0.4.2 Objectifs du Stage	9
0.5 Notation	9
1 Données Compositionnelles	11
1.1 Définition	12
1.2 Transformation	15
2 Analyses de Données de Microbiote	17
2.1 Pré-traitement	17
2.2 Analyse de la diversité	17
2.2.1 Diversité α	17
2.2.2 Diversité β	20
2.3 Abondance Différentielle	26
2.4 Réseaux d'inférence	31
3 Application et analyse des données du microbiote chez les personnes asthmatiques	35
3.1 Nettoyage de données	37
3.2 Prévalence	38
3.3 Abondance	39
3.4 Alpha-diversité	40
3.5 Beta-diversité	43
3.6 Abondance différentielle	43
3.7 Réseaux d'inférence	45
Conclusion	46

Annexes	48
3.8 Organigramme	48
3.9 Définitions biologiques	49
3.10 Application à l'étude COBRA-ENV	52

Résumé

Ce stage a pour but de répondre à des questions cliniques à propos du microbiote et mycobiote de l'air intérieur des personnes asthmatiques par le développement d'outils statistiques adaptés. En effet, plusieurs contraintes sont prises en compte : la compositionalité par la définition d'outil arithmétique et géométrie propre à cet espace, la sparsité par la mise en place de modèle avec excès de zéros (ZINB), la grande dimension par des méthodes d'analyses multidimensionnelles telles que la PCoA. C'est ainsi, que nous avons été capable de dire qu'il existe une différence dans le microbiote et mycobiote entre les personnes asthmatiques et les personnes non-asthmatiques. De plus, nous avons décelé chez les asthmatiques que certains facteurs environnementaux tels que l'urbanisation, la saison sont responsables d'un changement dans le microbiote et mycobiote.

Introduction

0.1 Présentation

J'ai réalisé mon stage au centre de recherche de l'INSERM (U1219) de l'Université de Bordeaux dans l'équipe STATISTIQUES POUR LA MÉDECINE TRANSLATIONNELLE (SISTM). Celle-ci est labellisée par l'INSERM et l'INRIA.

0.1.1 Équipe de Recherche

L'équipe de recherche SISTM regroupe des ingénieurs, des biostatisticiens, des doctorants, de praticiens hospitaliers... Ces derniers analysent des données dites omiques (les données génomiques ADN, transcriptomiques ARN) pour répondre à des questions cliniques. En effet, les principaux objectifs de l'équipe sont d'accélérer la mise au point de vaccins en analysant toutes les informations disponibles dans les essais cliniques. Et ce, par le développement de nouvelles approches d'analyse de données et de modélisation pour des données de grande dimension dans des études ayant une taille d'échantillon limitée (petite). Pour répondre à ces questions, cette équipe est structurée en 3 axes correspondant à des plusieurs problématiques (Vous retrouverez en annexe l'organigramme de l'équipe) :

1. **High Dimensional Statistical Learning** : Les recherches biologiques et cliniques ont radicalement changé en raison des progrès technologiques, ce qui a conduit à la possibilité de mesurer beaucoup plus de quantités biologiques qu'auparavant. Cependant, les ensembles de données omiques contiennent plus de paramètres p à estimer que d'observations n . Par conséquent, les méthodes classiques telles que les modèles linéaires sont inefficaces, voire même inapplicables. L'objectif est donc de sélectionner les informations les plus pertinentes en vue d'une meilleure compréhension des données, ainsi qu'une meilleure représentation. Autrement dit, de débloquer l'analyse de données longitudinales de grande dimension en développant des approches statistiques appropriées notamment pour des données à haut débit (microbiome, transcriptome, cytomique ...). Cela passe par l'utilisation de méthodes de clustering afin d'identifier, par exemple les gènes entre les groupes vaccinaux.

2. **Mechanistic Learning** : Cet axe compare et met en œuvre des stratégies de contrôle par des approches appartenant au contrôle statistique et à l'apprentissage par renforcement. De plus, elle étudie la dynamique d'un marqueur à l'aide de modèles basés sur des équations différentielles ordinaires (ODE), permettant de décrire la variabilité qui existe dans les données (modèle d'effets mixtes).
3. **Translational Vaccinology** : L'objectif de cet axe est d'élucider les effets et les mécanismes d'action potentiels des vaccins par des analyses statistiques afin d'accélérer le développements des vaccins.

D'une autre manière, nous pouvons dire que dans l'axe 1, les informations pertinentes sont extraites du Big Data. Celles-ci sont utilisées afin d'estimer les paramètres du modèle mécanistique dans l'axe 2. Ces modèles peuvent ensuite être utilisés pour simuler les stratégies vaccinales optimales à évaluer dans les prochains essais cliniques.

Pour ma part, j'ai réalisé mon travail dans le domaine de l'immunologie au travers de données microbiotes. Dans ce cadre, mon étude s'intéresse aux différentes interactions qu'il peut exister au sein du microbiote présent dans l'air intérieur des patients asthmatiques. Comme il s'agit d'un problème de grande dimension, j'ai rejoint le premier axe de recherche, **High Dimensional Statistical Learning**.

0.2 Contexte

L'asthme est une maladie respiratoire chronique qui se caractérise par une inflammation des voies respiratoires. D'après *OMS*, cette maladie concerne plus de 300 millions de personnes dans le monde. De plus, sa prévalence tend à augmenter dans les pays qui adoptent des modes de vie occidentaux et/ou qui ont une urbanisation croissante. La *Global Initiative for Asthma* [10] définit l'asthme par des symptômes d'essoufflement, d'oppression thoracique, de respiration sifflante. La connaissance des symptômes sont bien connus. En effet, d'après les cliniciens, chez les communautés bactériennes, la constitution du microbiote pulmonaire est déterminée par plusieurs facteurs tels que l'immigration microbienne (due à la dispersion à partir de la muqueuse buccale, aux micro-aspirations et inhalation) ou encore par l'élimination microbienne (toux, système immunitaire). Cela nous amène à penser que le microbiome d'un individu diffère selon plusieurs facteurs. Tandis que chez les communautés fongiques, le mycobiote pulmonaire lui, est présent dans les

voies respiratoires de l'appareil respiratoire. Néanmoins, les causes de cette maladie ne sont pas totalement élucidées mais semblent être liées à l'exosome de l'environnement intérieur par l'inhalation de substances allergènes par exemple (intérieur : acariens ; extérieur : pollens) et d'une prédisposition génétique. En effet, la vaste communauté microbienne présente dans l'environnement intérieur a fait l'objet d'une attention particulière par leur diversité et leur omniprésence. L'étude de cet exposome semble naturel car à chaque inspiration, de milliers de particules sont inhalées. Ainsi, celle-ci représente donc la première porte d'entrée d'un grand nombre de microorganismes. Les contaminations liées à cet environnement sont visibles à l'oeil nu, telles que la présence de moisissures dans certains logements. De plus, au-delà de la présence de microorganismes, l'environnement intérieur peut concerner la présence de pollens, d'acariens ... L'exposition aux microorganismes dans ce environnement peut également provenir de l'urbanisation, du mode de vie, du régime alimentaire ... Et au niveau microbien, cet environnement comprend des bactéries, des champignons et des virus. Ainsi, l'identification du phénotype permet d'anticiper la réponse au traitement.

0.3 Échantillonnage

Une méthode de prélèvement appelé **Electrostatic Dust Collector** ou en français piège à poussières est une méthode de prélèvement par aspiration ou par capteur avec l'utilisation d'une lingette, qui permet d'étudier les communautés de microorganismes présent dans l'air intérieur (appelé aussi microbiote exogène) de patients atteints d'une maladie respiratoire chronique. Cet méthode de prélèvement a l'avantage d'être peu coûteux, donc facilement déployable dans des études contenant un nombre d'individus conséquent. Le microbiote pulmonaire, lui, est étudié à partir d'expectorations des patients. De cette manière, l'étude métagénomique de ces méthodes nous donnera la composition d'un microbiome grâce à la table d'ASV. L'obtention de cette table se fait en deux phases (Figure 2).

1. La métagénomique amplicon ou en anglais metabarcoding repose sur le séquençage d'un gène marqueur qui va servir d'identifiant pour identifier les espèces présentes. De ce fait, ce gène doit être commun à plusieurs espèces tout en présentant des régions suffisamment variables. Dans l'étude des communautés bactériennes, le gène utilisé est celui de l'**ARN 16S** (Figure 1) tandis que dans l'étude des communautés fon-

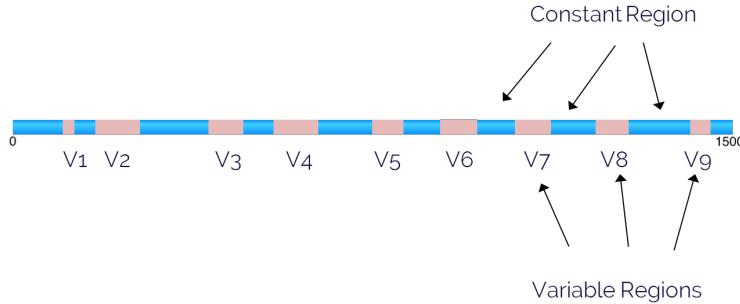


FIGURE 1 – Représentation schématique du gène 16S. Celui-ci est composé d'environ 1500 nucléotides et de neuf régions variables.

giques, le gène utilisé est l'**ITS**. Ces régions variables sont considérées comme des marqueurs reflétant les relations phylogénétiques entre les organismes. [17]

2. Une fois le séquençage terminé, il faut assigner à chaque séquence le nom de la bactérie ou du champignon. C'est ce qu'on appelle l'**assignation taxonomique**). Dans notre cas, un logiciel de séquençage open-source appelé **DADA2** [8] a été utilisé, celle-ci se base sur des modèles d'erreurs. Ainsi, l'objectif de DADA2 est de corriger le bruit introduit dans les séquences. Le principe de résolution est le suivant :
 - *Initialisation* : Les reads appartiennent tous au même groupe. La diversité des reads au sein de ce groupe est ensuite comparée à la diversité attendue uniquement en présence d'erreur de séquençage.
 - *Condition* : Si cette erreur est trop grande, le groupe initial est divisé en sous-groupes plus homogènes.
 - *Boucle* : Le processus est répété dans chacun des sous-groupes jusqu'à obtenir des groupes homogènes.

Cette approche permet d'obtenir des ASV très précis, ne différant que d'un nucléotide. Une fois l'assignation taxonomique réalisée, il suffit de compter le nombre d'espèces présentes dans chaque échantillon et de construire la table des ASV.

0.4 Objectifs

0.4.1 Objectifs Cliniques

Dans cette étude, nous allons nous concentrer sur l'exosome de l'air intérieur afin de considérer l'impact de l'environnement sur la santé, et ainsi mieux comprendre les causes de certaines pathologies. De ce fait, la question principale est de trouver l'origine des microorganismes présents dans l'air intérieur. D'après les cliniciens, les sources de l'exosome microbien de l'air intérieur proviennent de l'intérieur (chauffage, humidité, système de ventilation, murs ...) (Yifan Shan et al. [15]), mais aussi de l'extérieur (localisation géographique ...). Dans cette étude, les cliniciens nous ont proposé des données sur l'environnement intérieur des personnes asthmatiques que nous présenterons plus en détail plus tard dans l'étude. Ces données auront pour but d'expliquer quels sont les microorganismes et les facteurs qui affectent les personnes asthmatiques.

0.4.2 Objectifs du Stage

Afin de répondre aux interrogations des cliniciens, nous allons :

1. Mettre en œuvre des méthodes d'analyses de données multidimensionnelles (telles que les ACP) adaptés aux données microbiotes.
2. Analyser les données COBRA-ENV afin de répondre aux questions posées par les cliniciens, qui ont pour but d'étudier les interactions entre
 - (a) Les composantes bactériennes et fongiques
 - (b) Du microbiote de l'air intérieur chez des patients asthmatiques
 - (c) Du microbiote de l'air intérieur chez des patients asthmatiques avec des mesures étant répétées dans le temps.

De plus, nous développerons un package R, auquel les cliniciens pourront y accéder pour reproduire à la fois les résultats de l'étude COBRA-ENV et de futures analyses. Nous réaliserons et disposerons l'analyse de l'étude COBRA-ENV dans un site web interactif, qui tout au long de notre étude a été notre base de travail.

0.5 Notation

- n le nombre d'individu

- m le nombre d'ASV considéré dans le n -échantillon.
- p le nombre de covariables considérées dans le n -échantillon.
- $Z_{ij} \in \mathbb{N}^+$ l'abondance brute où i représente l'individu et j l'ASV.
- $\tilde{Z}_{ij} \in [0; 1]$ l'abondance relative où i représente l'individu et j l'ASV.
- $Z = (Z_1, \dots, Z_n)^T$ et $\tilde{Z} = (\tilde{Z}_1, \dots, \tilde{Z}_n)^T$ correspondent respectivement aux matrices d'abondance brutes et relatives de taille $n \times m$.
- $y = (y_1, \dots, y_n)^T$ est le vecteur réponse et y_i est la réponse clinique de l'individu i .
- $X = (X_1, \dots, X_n)^T$ représente la matrice de covariables de taille $n \times p$.

1 Données Compositionnelles

La table des ASV est le point de départ des analyses de données métagénomiques. Cette dernière contient le nombre de séquences par ASV et par échantillon. C'est ce qu'on appelle plus brièvement l'abondance. De plus, grâce à cette table, nous pouvons répondre à plusieurs questions cliniques. La Figure 2 montre un exemple de l'obtention d'un tableau d'abondance avec, en ligne, les ASV et en colonne, les individus (un piège à poussière = un individu). Ainsi, chaque cellule du tableau correspond au nombre de séquences pour un ASV et pour un individu. À l'aide d'une base de donnée ARNr (Acide ribonucléique ribosomique) connue, chaque ASV est associé à un rang taxonomique représentant chacun un niveau de classification du monde vivant allant du règne jusqu'à l'espèce (Figure 3a).

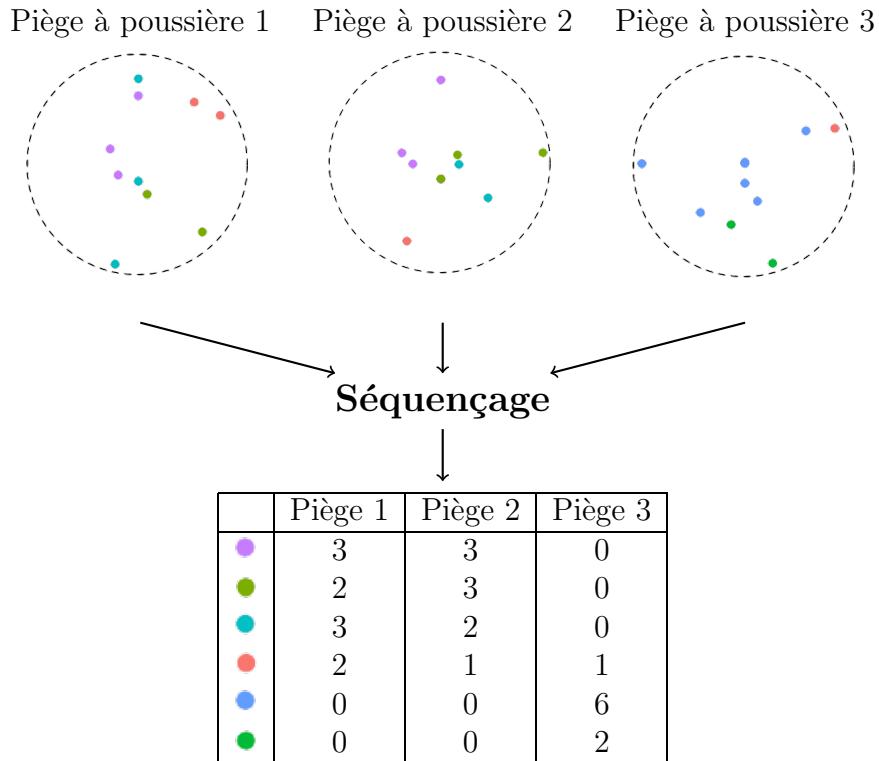
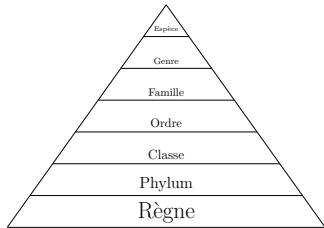


FIGURE 2 – Obtention de la table des comptages des OTUs

L'une des particularités de la table d'abondance est qu'elle contient un

nombre conséquent de zéros. En effet, un ASV qui est présent dans un seul échantillon apparaîtra avec un comptage nul dans les autres échantillons. Ceci est dû au fait que les nombres d'ASV sont dépendants vis-à-vis des individus car la composition microbienne est unique à chaque individu. De plus, deux échantillons peuvent avoir la même profondeur de séquençage et le même vecteur de comptage, mais des biomasses différentes et donc des nombres de séquences bactériennes et fongiques différentes (par exemple, un nombre de lectures de 500 dans un premier échantillon et 1000 dans un autre). Par conséquent, la comparaison directe des comptages entre échantillons est donc impossible. Autrement dit, la comparaison d'abondance brute chez deux individus peut entraîner de fausses conclusions car le nombre total de séquences entre les individus est hétérogène. Le passage aux données compositionnelles est une méthode de résolution possible, et numériquement peu coûteuse. En effet, la normalisation des données permet aux abondances brutes d'être comparables entre échantillon. Néanmoins, les données compositionnelles sont contraignantes et nécessitent donc des méthodes d'analyses adaptées.



(a) Lignée taxonomique

	Kingdom	Phylum	Class	Order	Family	Genus	Species
■	Bacteria	Bacteroidota	Saprospiria	Sapropirales	Sapropiraceae	Aureispira	Aureispira marina
●	Bacteria	Bacteroidota	Sapropiria	Sapropirales	Sapropiraceae	Aureispira	Aureispira maritima
▲	Bacteria	Bacteroidota	Sapropiria	Sapropirales	Sapropiraceae	Rubidimonas	Rubidimonas crustatorum
◆	Bacteria	Bacteroidota	Sapropiria	Sapropirales	Sapropiraceae	Rubidimonas	Rubidimonas crustatorum
◆	Bacteria	Bacteroidota	Sapropiria	Sapropirales	Sapropiraceae	Aureispira	Aureispira marina
◆	Bacteria	Bacteroidota	Sapropiria	Sapropirales	Sapropiraceae	Rubidimonas	Rubidimonas crustatorum

(b) Table taxonomique

FIGURE 3 – Classification des rangs taxonomiques

1.1 Définition

On appelle les données de composition (CoDA), la normalisation des abondances brutes en abondances relatives. La normalisation d'une composition appelée *fermeture* ou (closure) permet de modifier la somme des composantes d'une composition (passer d'une proportion à un pourcentage). Celle-ci engendre un nouvel espace de définition. Cette fermeture est définie pour tout k ,

$$\tilde{Z}_i = \left(\frac{k \cdot Z_{i1}}{\sum_{j=1}^m Z_{ij}}, \dots, \frac{k \cdot Z_{im}}{\sum_{j=1}^m Z_{ij}} \right) \quad i = \{1 \dots, n\} \quad (1)$$

où k représente la somme des composantes.

Définition 1.1 *On appelle simplex, l'espace de probabilité des données compositionnelles (CoDA) telles que pour tout $j \in \{1, \dots, m\}$ on a*

$$S^m = \left\{ \tilde{Z}_i = (\tilde{Z}_{i1}, \dots, \tilde{Z}_{im})^T : \tilde{Z}_{ij} > 0 \middle/ \sum_{j=1}^m \tilde{Z}_{ij} = 1 \right\} \quad (2)$$

Lorsque nous disposons de 3 compositions, on représente souvent ces derniers par un diagramme ternaire dont les sommets représentent les composantes étudiées (3 espèces de bactéries par exemple). Ainsi, la composition globale est définie par le barycentre des trois sommets, pondérées par les poids des composantes (Figure 4).

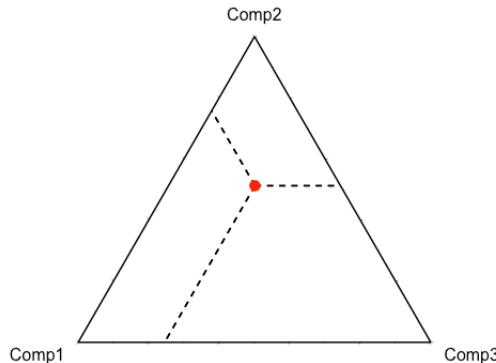


FIGURE 4 – Exemple de diagramme ternaire pour trois compositions

Dans un simplex, nous ne pouvons pas utiliser la géométrie euclidienne pour comparer des compositions. En effet, dans l'Exemple 5 on remarque que la distance euclidienne entre x_1 et x_2 est la même que celle entre x_3 et x_4 alors que la proportion de la première composante a doublé dans le premier cas. Par conséquent, nous avons besoin de définir une nouvelle géométrie dans le simplex. Pawlowsky-Glahn et al. [14] ont défini par leur article LECTURE NOTES ON COMPOSITIONAL DATA ANALYSIS les opérations suivantes :

$$x_1 = (0.1, 0.4, 0.5) \quad x_2 = (0.2, 0.3, 0.5) \quad x_3 = (0.1, 0.6, 0.3) \quad x_4 = (0.2, 0.5, 0.3)$$

EXEMPLE 5 – 4 observations de 3 compositions

Définition 1.2 La perturbation d'une composition $x \in S^m$ par une composition $y \in S^m$,

$$x \oplus y = \left(\frac{x_1 y_1}{\sum_{j=1}^m x_j y_j}, \dots, \frac{x_m y_m}{\sum_{j=1}^m x_j y_j} \right) \quad (3)$$

Définition 1.3 La puissance d'une composition $x \in S^m$ par une constante $\alpha \in \mathbb{R}$,

$$\alpha \odot x = \left(\frac{x_1^\alpha}{\sum_{j=1}^m x_j^\alpha}, \dots, \frac{x_m^\alpha}{\sum_{j=1}^m x_j^\alpha} \right) \quad (4)$$

Définition 1.4 Le produit scalaire de deux compositions $x, y \in S^m$,

$$\langle x, y \rangle = \frac{1}{2m} \sum_{i=1}^m \sum_{j=1}^m \log\left(\frac{x_i}{x_j}\right) \log\left(\frac{y_i}{y_j}\right) \quad (5)$$

Définition 1.5 La distance entre x et y , pour tout $x, y \in S^m$,

$$d(x, y) = \sqrt{\frac{1}{2m} \sum_{i=1}^m \sum_{j=1}^m \left[\log\left(\frac{x_i}{x_j}\right) - \log\left(\frac{y_i}{y_j}\right) \right]^2} \quad (6)$$

Définition 1.6 La norme d'une composition $x \in S^m$,

$$\|x\| = \sqrt{\frac{1}{2m} \sum_{i=1}^m \sum_{j=1}^m \left(\log\left(\frac{x_i}{x_j}\right) \right)^2} \quad (7)$$

A présent, il devient possible de définir des objets géométriques tels que les droites compositionnelles.

De même, les mesures statistiques doivent être redéfinies pour qu'elles puissent prendre en compte sa géométrie.

Définition 1.7 Soit \tilde{Z} un n -échantillon de m -décomposition, on appelle la moyenne

$$\overline{\tilde{Z}}_n = \left(\frac{g_1}{\sum_{j=1}^m g_j}, \dots, \frac{g_m}{\sum_{j=1}^m g_j} \right) \quad (8)$$

avec $g_i = \left(\prod_{i=1}^n \tilde{Z}_{ij} \right)^{1/n}$.

Définition 1.8 La matrice de covariance est définie par

$$Cov = \begin{pmatrix} t_{11} & \dots & t_{1m} \\ \vdots & \ddots & \vdots \\ t_{m1} & \dots & t_{mm} \end{pmatrix} \quad (9)$$

où les $t_{ij} = var\left(\log\left(\frac{\tilde{Z}_i}{\tilde{Z}_j}\right)\right)$.

Ainsi on peut définir la variance totale comme étant,

$$totvar(X) = \frac{1}{2m} \sum_{i=1}^m \sum_{j=1}^m t_{ij} \quad (10)$$

Bien sûr, la corrélation standard ne peut être calculée lorsqu'il s'agit de CoDA car les composantes sont mathématiquement dépendantes. Pour palier à ce problème, on utilise les fonctions de transformations définies par Aitchison [5]. Ces transformations permettent de passer du simplex à un espace euclidien.

1.2 Transformation

1. La transformation ALR (Additive Log Ratio).

Définition 1.9 (*Additive Log Ratio*)

$$alr : S^m \rightarrow \mathbb{R}^{m-1}, alr(\tilde{Z}_i) = \left(\log \frac{\tilde{Z}_{i1}}{\tilde{Z}_{im}}, \dots, \log \frac{\tilde{Z}_{i(m-1)}}{\tilde{Z}_{im}} \right) \quad (11)$$

Une des composantes est choisie comme référence (dénominateur) car la m -ième cordonnée $\log(\tilde{Z}_m)/\tilde{Z}_m$ est nulle. Par conséquent, alr nous renvoie dans un sous-espace \mathbb{R}^{m-1} .

2. La transformation **CLR** (Centered Log Ratio).

Définition 1.10

$$clr : S^m \rightarrow \mathbb{R}^m, \quad clr(\tilde{Z}_i) = \left(\log \frac{\tilde{Z}_{i1}}{g(\tilde{Z}_i)}, \dots, \log \frac{\tilde{Z}_{im}}{g(\tilde{Z}_i)} \right) \quad (12)$$

avec g le centre de décomposition défini $g(\tilde{Z}_i) = \prod_{j=1}^m \tilde{Z}_{ij}^{1/m}$

clr nous renvoie dans un sous-espace \mathbb{R}^d , ce qui le rend plus interprétable. Cependant, elle ne conserve pas les distances tout comme la transformation alr au sens où

$$d(\tilde{Z}_i, \tilde{Z}_{i+1}) \neq ||alr(\tilde{Z}_i) - alr(\tilde{Z}_{i+1})||_2$$

3. La transformation **ILR** (Isometric Log Ratio)

Définition 1.11 Soit (e_1, \dots, e_{m-1}) une base orthogonale de S^m . Une composition $\tilde{Z}_i \in S^m$ peut être exprimée de la façon suivante :

$$\tilde{Z}_i = \bigoplus_{j=1}^{m-1} \tilde{Z}_{ij}^* \odot e_{ij} \quad i = \{1, \dots, n\}$$

avec $\tilde{Z}_i^* = (\tilde{Z}_{i1}^*, \dots, \tilde{Z}_{i(m-1)}^*)$ le vecteur de coordonnées de \tilde{Z}_i . D'où,

$$\tilde{Z}_i^* = ilr(\tilde{Z}_i) = (\langle \tilde{Z}_i, e_{i1} \rangle, \dots, \langle \tilde{Z}_i, e_{i(m-1)} \rangle) \quad (13)$$

Ces transformations impliquent des éléments non nuls (logarithme). Nous avions vu que les données compositionnelles possédaient un grand nombre d'éléments nuls. Une façon de palier à ce problème est d'utiliser une stratégie d'imputation c'est-à-dire que l'on va remplacer les zéros par une petite valeur et modifier les valeurs non nulles afin de respecter les hypothèses des données compositionnelles (somme à 1). Cette méthode appelée **remplacement multiplicatif** est un méthode non-paramétrique proposée par *Martin Fernandez et al.* [13]. Néanmoins, cette méthode peut générer des corrélations artificielles entre les variables.

2 Analyses de Données de Microbiote

L'analyse de données en microbiotes se décompose en plusieurs étapes, correspondant à plusieurs questions.

1. **L'analyse de la diversité** (diversité α et β) : Combien y a-t-il d'espèces différentes dans l'échantillon ? La composition du microbiote est-elle différente selon des groupes d'individus ?
2. **L'analyse différentielle d'abondance** : Quels ASV sont différentiellement abondants entre des groupes d'individus ? Quelles sont les covariables (variables explicatives) significativement associées à la composition du microbiote ? Quelles sont les ASV associés à un statut clinique ?
3. **L'analyse de corrélation** : Existe-t-il des co-occurrences significatives entre espèces ?

2.1 Pré-traitement

Il existe plusieurs manières d'analyser des données microbiotes. En effet, travailler sur notre ensemble de jeu de données est une possibilité, mais pas la plus optimale. Une façon d'analyser ce type de données, qui à la fois nous permet d'éviter des données trop clairsemées, est de supprimer les ASV présents dans moins de 10% des échantillons. De plus, il est plus pertinent d'agréger les ASV à un même niveau taxonomique pour une meilleure interprétation des résultats et pour des raisons cliniques. De cette manière, la dimension du problème et la proportion de zéros sont drastiquement réduits par la réduction du nombre de variables à analyser. L'agrégation permet aussi de rendre l'analyse robuste à l'erreur de séquençage. Cependant, toute perte d'information peut entraîner de fausses conclusions.

2.2 Analyse de la diversité

Évaluer la diversité d'un microbiote est complexe. Nous présentons dans cette partie les méthodes proposées dans la littérature pour explorer la diversité d'un microbiote.

2.2.1 Diversité α

La diversité α permet de mesurer :

- la RICHESSE SPÉCIFIQUE qui, correspond pour un individu i , le nombre d'espèces différentes.
- l'ÉQUITABILITÉ qui elle, représente la régularité de la distribution d'une espèce pour un individu i .

Ainsi, deux échantillons ayant le même nombre d'espèces peuvent avoir une équitabilité différente.

La Figure 2 représente la composition du microbiote de trois individus. Chaque couleur désigne une espèce différente. Le premier et deuxième individu possèdent le même nombre d'espèces différentes mais leur abondance est différente. En effet, le premier individu possède une distribution plus variées d'espèces rouge, turquoise et moins variées d'espèces verte contrairement au deuxième individu. Tandis que le troisième individu possède à la fois un nombre d'espèces différent ainsi qu'une distribution différente.

L'alpha diversité peut être calculé au travers de plusieurs indices :

Proposition 2.1 (Indice de Chao1) *On note S le nombre d'espèces existantes dans un n -échantillon. On estime le nombre d'espèces non observé à partir du nombre d'espèces observé,*

$$\widehat{S}_{Chao1} = s^n + \frac{(n-1)(s_1^n)^2}{2ns_2^n} \quad (14)$$

où s_n^k représente le nombre d'espèces observé k fois dans un échantillon de n individus.

L'indice de Chao1 est un estimateur non-paramétrique qui effectue un comptage afin de savoir le nombre d'espèces présentes dans l'échantillon. Cependant, elle ne tient en compte de l'abondance des espèces.

Proposition 2.2 (Indice de Shannon) *Soit S la richesse spécifique dans un n -échantillon. On appelle l'indice de Shannon, une mesure quantifiant l'hétérogénéité d'un échantillon (entropie).*

$$H = - \sum_{j=1}^S p_j \log(p_j) \quad (15)$$

où p_j représente la probabilité qu'un individu possède l'espèce j , autrement dit la proportion d'une espèce j par rapport au nombre d'individus pour l'espèce j .

L'indice de Shannon, lui prend en compte la richesse (nombre d'espèces) ainsi que l'abondance (relative) des espèces. Ainsi, l'indice H varie en fonction du nombre d'espèce et de l'abondance des différentes espèces. Une grande valeur de H implique une grande diversité. De plus, cela implique la présence d'un grand nombre d'espèces avec des abondances équilibrées. La valeur maximale de l'indice est donnée par $\log(S)$ impliquant que toutes les espèces sont réparties de manière égale. Au contraire, si l'échantillon contient une seule et même espèce alors $H = 0$.

Proposition 2.3 (Indice de Simpson) *Soit S la richesse spécifique dans un n -échantillon. On appelle l'indice de Simpson,*

$$E = \sum_{j=1}^S p_j^2 \quad (16)$$

où p_j est l'abondance relative de l'espèce j .

L'indice de Simpson est une mesure caractérisant la diversité comme étant la probabilité que deux séquences tirées au hasard appartiennent à la même espèce. Elle prend ses valeurs dans l'intervalle $[0, 1]$. Ainsi, lorsque E tend vers 1 la diversité est "minimale" c'est-à-dire s'il existe une seule et unique espèce dans l'échantillon. Et lorsqu'elle vaut $1/S$ la diversité est maximale c'est-à-dire si les différentes espèces ont la même probabilité p_j (équiprobabilité). Ainsi, avec cet indice, plus de poids est appliqué aux espèces abondantes.

Preuve 2.1 *Soit s_n le nombre d'espèces différentes observé dans un n -échantillon. On note n_j le nombre d'individus ayant l'espèce j et p_j la probabilité qu'un individu possède l'espèce j . Par conséquent, la loi de probabilité que l'espèce j soit observée k fois est la loi binomiale. De cette manière, nous pouvons écrire l'espérance du nombre d'espèces observées k fois comme suit :*

$$\mathbb{E}(s_k^n) = \binom{n}{k} \sum_{j=1}^{s^n} p_j^k (1 - p_j)^{n-k}$$

En sachant cela, nous pouvons déduire l'espérance du nombre d'espèces non observé $\mathbb{E}(s_0^n)$.

$$\mathbb{E}(s_0^n) = \sum_{j=1}^{s^n} (1 - p_j)^n$$

Puis $\mathbb{E}(s_1^n)$,

$$\mathbb{E}(s_1^n) = \frac{n!}{(n-1)!} \sum_{j=1}^{s^n} p_j (1-p_j)^{n-1}$$

$$\frac{\mathbb{E}(s_1^n)}{n} = \sum_{j=1}^{s^n} p_j (1-p_j)^{n-1}$$

Et $\mathbb{E}(s_2^n)$,

$$\mathbb{E}(s_2^n) = \frac{n!}{2 \cdot (n-2)!} \sum_{j=1}^{s^n} p_j^2 (1-p_j)^{n-2}$$

$$\frac{2\mathbb{E}(s_2^n)}{n(n-1)} = \sum_{j=1}^{s^n} p_j^2 (1-p_j)^{n-2}$$

On a par l'inégalité de Cauchy-Schwarz,

$$\left(\sum_{j=1}^{s^n} p_j (1-p_j)^{n-1} \right)^2 \leq \left(\sum_{j=1}^{s^n} (1-p_j)^n \right) \left(\sum_{j=1}^{s^n} p_j^2 (1-p_j)^{n-2} \right)$$

$$\frac{\mathbb{E}(s_1^n)^2}{n^2} \leq \mathbb{E}(s_0^n) \frac{2\mathbb{E}(s_2^n)}{n(n-1)}$$

$$\frac{(n-1)\mathbb{E}(s_1^n)^2}{2n\mathbb{E}(s_2^n)} \leq \mathbb{E}(s_0^n)$$

Remarque 2.1 L'indice de Shannon est une moyenne géométrique pondérée des abondances relatives. En effet, on peut réécrire l'indice de Shannon de la manière suivante :

$$H = \log \left(\frac{1}{\prod_{j=1}^S p_j^{p_j}} \right)$$

Ainsi, plus l'abondance est inégale, plus la moyenne pondérée est grande donc en fine, l'indice de Shannon est petite.

2.2.2 Diversité β

La notion de β -diversité introduite par Whittaker [16] permet de comparer la composition des communautés microbiennes des différentes structures entre des groupes d'individus par des calculs de distance. En effet, on considère deux types de distance :

- DISTANCE PHYLOGÉNÉTIQUE : Prise en compte de la co-présente de deux séquences ou individus car les organismes ou individus partagent tous un lien de parenté entre eux. Ainsi, par exemple, deux individus ayant des espèces proches vont avoir une distance inférieure à deux individus ayant des espèces éloignées.

1. Distance de Bray-Curtis :

Proposition 2.4 Soit Z , une matrice d'abondance de taille $n \times m$. Pour $(i, i') \in \{(1, 1), \dots, (n, n)\}$ on a,

$$d_{ii'}^{BC} = 1 - \frac{2 \sum_{j=1}^m \min(Z_{ij}, Z_{i'j})}{\sum_{j=1}^m (Z_{ij} + Z_{i'j})} \quad (17)$$

avec $d_{ii'}$ la distance entre l'individu i et i' .

La distance de Bray-Curtis permet d'évaluer la dissimilarité en termes d'abondances, entre deux individus. Cet indice prend ses valeurs dans l'intervalle $[0, 1]$. En effet, lorsque la valeur de l'indice tend vers 0, cela signifie que les deux individus ont la même composition microbienne c'est-à-dire qu'elles partagent pour chaque espèce plus ou moins le même nombre de séquences, et inversement, lorsque la valeur de l'indice tend vers 1, les deux individus sont dissemblables. De plus, cet indice suppose que les deux individus soient de même taille car si un individu possède plus de taxons qu'un autre individu, alors, logiquement il y a plus d'espèces dans un échantillon que dans un autre.

2. Distance de Jaccard :

Proposition 2.5 Soit Z , une matrice d'abondance de taille $n \times m$. Pour $(i, i') \in \{(1, 1), \dots, (n, n)\}$ on a,

$$d_{ii'}^J = 1 - \frac{2 \sum_{j=1}^m \min(Z_{ij}, Z_{i'j})}{\sum_{j=1}^m \max(Z_{ij}, Z_{i'j})} \quad (18)$$

avec $d_{ii'}$ la distance entre l'individu i et i' .

L'indice de Jaccard permet une comparaison entre deux individus en calculant le rapport entre les espèces communes aux deux individus. De cette façon, cet indice ne se base pas sur l'abondance, mais plutôt sur la présence et l'absence des ASV. Ainsi, le fait

qu'une espèce soit absence ou non contribue à augmenter la dissimilarité. Tout comme l'indice précédent, les valeurs de cet indice varient entre 0 et 1. De ce fait, lorsque la valeur de l'indice vaut 1 alors les deux individus n'ont aucune espèce en commun, et par suite, si celle-ci vaut 0 alors les deux individus ont toutes leurs espèces en commun.

- DISTANCE NON PHYLOGÉNÉTIQUE : Contraire à la distance phylogénétique. De cette manière, deux individus ayant des espèces proches vont avoir la même distance que deux individus ayant des espèces éloignées.

1. Distance UniFrac :

Proposition 2.6 Soit \tilde{Z} , une matrice d'abondance relative de taille $n \times m$. Pour $(i, i') \in \{(1, 1), \dots, (n, n)\}$ on a,

$$d_{ii'}^U = \frac{\sum_{m=1}^M b_m |\tilde{Z}_{im} - \tilde{Z}_{i'm}|}{\sum_{m=1}^M b_m (\tilde{Z}_{im} + \tilde{Z}_{i'm})} \quad (19)$$

avec M le nombre de branches dans l'arbre phylogénétique, b_m la longueur de la branche m .

La distance UniFrac prend en compte des informations sur la parenté des membres, qui elle-même est calculée à l'aide des distances phylogénétiques entre les organismes observés. Cette distance est une version pondérée d'UniFrac, qui tient en compte de l'abondance relative des taxons.

De cette façon, deux individus peuvent avoir une structure phylogénétique différente et ce même, lorsque l'alpha diversité des deux individus est la même. Ces différents indices peuvent être interprétés comme une distance comme vu précédemment et peut aussi être visualiser à l'aide d'une méthode d'ordination.

1. **L'analyse en composantes principales (ACP)** est une technique de réduction de dimension couramment utilisée pour l'analyse multivariée. Cet méthode vise à réduire la dimension d'un ensemble de données afin de faciliter son interprétation et sa visualisation. Celle-ci exploite la structure de dépendance entre les variables pour nous donner les variables qui résument (vis-à-vis de la variance) au mieux les données ou composantes principales. D'un point de vue algébrique, l'ACP peut être considérée comme un problème de factorisation matricielle. A la

différence d'une ACP classique 1, Aitchison a redéfini l'ACP pour les données compositionnelles en centrant et réduisant les données avant de calculer les vecteurs propres de la matrice de covariance.

2. **L'analyse en coordonnées principales** (PCoA) est une généralisation de l'analyse en composantes principales et un cas particulier du positionnement multidimensionnelle (MDS) qui permet de visualiser des dissimilarités d'un jeu de données. En effet, la distance n'est plus euclidienne mais peut être la distance de bray-curtis ou d'autres énoncés plus haut. De même que l'ACP, on obtient des vecteurs de coordonnées principales (ou individus). Ainsi, le but est de projeter les données dans un espace de dimension inférieure en minimisant une fonction de perte (stress). Pour ce faire, cette méthode dépend d'une matrice de dissimilarité permettant de simplifier son interprétation. De cette manière, plus les individus sont proches dans la projection, plus leurs structures phylogénétiques sont similaires. Par conséquent, la PCoA permet de visualiser si les structures phylogénétiques sont différents entre les groupes. Lorsqu'il s'agit de l'échelle multidimensionnelle métrique (mMDS), la fonction de perte à minimiser est une somme résiduelle de carrés tel qui suit :

$$Stress(\tilde{Z}_1, \dots, \tilde{Z}_n) = \sqrt{\sum_{i \neq j=1, \dots, n} (d_{ij} - \|\tilde{Z}_i - \tilde{Z}_j\|)^2} \quad (20)$$

avec d_{ij} , la distance entre l'individu i et j .

Et lorsqu'il s'agit d'une échelle multidimensionnelle non métrique, utilisée pour la prise en compte de l'arbre phylogénétique, la fonction à minimiser est la suivante :

$$\sqrt{\frac{\sum(f(\tilde{Z}) - d)^2}{\sum d^2}} \quad (21)$$

avec d la matrice de dissimilarité et f une transformation monotone de \tilde{Z} .

3. **PLNmodels** est un modèle d'ACP probabiliste, défini par *Julien Chi-quet et al.* [9], basé sur un modèle de l'ACP définie dans un cadre gaussien dans lequel les coefficients sont traités comme des variables aléatoires latentes. Comme il s'agit de variables latentes, les estimations du maximum de vraisemblance peuvent être obtenues grâce à

un algorithme EM. Ce modèle probabiliste permet de combiner la réduction de dimension avec d'autres outils de modélisation tels que la régression. Ainsi, une correction sur les effets des variables peuvent permettre d'éviter la présence de corrélations car les variables observées peuvent être affectées par les variations de ces covariables.

Proposition 2.7 *Soit Y un échantillon de vecteurs d'observation à p dimension. On peut écrire le modèle PLN-PCA dans un cadre hiérarchique qui relie les vecteurs d'observations Y_i à un échantillon de vecteurs de variables latentes W_i à q dimensions pour tout $i \in \{1, \dots, n\}$:*

$$\text{Espace latent : } W_i \text{ i.i.d } W_i \sim \mathcal{N}(0_q, \mathbb{1}_q)$$

$$\text{Espace des paramètres : } Z_i = \mu + B^T W_i$$

$$\text{Espace d'observation : } Y_{ij}|Z_{ij} \sim \mathcal{P}(\exp Z_{ij})$$

Le paramètre μ , lui correspond aux effets fixes, la matrice B de taille $p \times q$ capture la dépendance entre les variables observées et latentes. Z_i est une transformation linéaire de W_i et $Y_i|Z_i$ correspond à du bruit. De plus, les vecteurs latents sont par hypothèse des gaussiennes indépendants avec une variance unitaire garantissant la "non-structure" de cet espace. De plus, la dimension de l'espace latent correspond au nombre d'axes de l'ACP autrement dit elle correspond au rang de la matrice BB^T . Ainsi, le modèle peut être réécrit de la manière suivante :

$$\text{Espace latent : } \mathcal{N}(\mu, \Sigma) \quad \Sigma = BB^T$$

$$\text{Espace d'observation : } Y_{ij}|Z_{ij} \sim \mathcal{P}(\exp Z_{ij})$$

Une fois l'ordination effectuée, l'observation des différences de structure phylogénétique est possible. Néanmoins, des tests statistiques sont à réaliser afin de déterminer s'il existe une différence entre les groupes d'individus.

1. **ANOSIM**, mesure les différences entre des groupes d'individus. Ainsi, si les groupes sont différents alors les échantillons appartenant à chacun de ces groupes devraient avoir des compositions plus semblables que celles provenant de différents groupes. La statistique de test R s'écrit :

$$-1 \leq \frac{r_B - r_w}{\frac{n(n-1)}{4}} \leq 1$$

Matrice de départ	
$X = dist(\tilde{Z})$	
n individus dans \mathbb{R}^m	m taxons dans \mathbb{R}^n
$R = \frac{1}{n} X^T X$	$S = \frac{1}{n} X X^T$
Diagonnalisation	
AXES PRINCIPAUX	COMPOSANTES PRINCIPALES
<i>Valeurs propres</i>	<i>Valeurs propres</i>
Λ_m	Λ_n
<i>Vecteurs propres</i>	<i>Vecteurs propres</i>
U	V
<i>Coordonnées des Individus</i>	<i>Coordonnées des variables</i>
$L = XU$	$C = \frac{1}{\sqrt{n}} X^T V$

TABLE 1 – Etapes d'une analyse d'ordination ACP

avec r_B la moyenne des rangs des dissimilarités d'individus provenant de groupes différents et r_w la moyenne des rangs des dissimilarités d'individus provenant du même groupe. Si la statistique de test R est proche de 1 alors les paires d'individus provenant du même groupe partage les similitudes. Tandis que $R = 0$ implique que les moyennes sont égales. Pour déterminer la p -valeur, on utilise une méthode de permutation.

- PERMANOVA est un test non-paramétrique utilisé pour comparer des groupes et tester l'hypothèse selon laquelle les centroïdes (centre du groupe d'individus) d'un groupe diffère des centroïdes d'autres groupes. La statistique de test s'écrit

$$F = \frac{(SS_T - SS_W)/(G - 1)}{SS_W/(n - G)}$$

avec $SS_T = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{i'} i' = i + 1 n d_{ii'}^2$, et $SS_W = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{i'=i+1}^n d_{ii'}^2 \delta_{ii'}$ où d est la matrice de dissimilarité et $\delta_{ii'}$ prend 1 quand les individus i et i' sont dans le même groupe 0 sinon et G est le nombre de groupes considérés. Une faible valeur de p indique que les groupes ont en moyenne une composition différente.

2.3 Abondance Différentielle

L’analyse de l’abondance différentielle permet d’identifier les différences de composition taxonomique entre les échantillons métagénomiques (ex : groupe témoin vs groupe traitement). Ainsi, l’identification d’un taxon peut conduire à des informations, ainsi qu’une mise en place de traitements car ces taxons peuvent être responsables de changements (néfastes ou non) dans le microbiote. Etant donné la nature des données microbiote (surdispersion, clairsemées ...), il est nécessaire d’utiliser des outils analytiques adaptés pour traiter ces caractéristiques. De plus, le nombre de taxons est généralement important, ce qui nécessite des algorithmes efficaces pour détecter les taxons significatifs. Pour rajouter à cela, il est nécessaire de prendre en compte de la corrélation dans le temps des sujets par la mise en place de modèles longitudinaux. En effet, les études longitudinales contiennent un nombre important de sujets et des mesures temporelles pour chaque sujet. Pour ce faire, divers modèles linéaires généralisés ont été proposés et dans cette étude, nous allons utiliser deux modèles.

1. Modèle mixtes binomial négatif (NBMMs) *Xinyan Zhang et al.*

[18] : Les distributions du type négative binomiale avec ou sans excès de zéros sont très utilisées pour résoudre les problèmes de surdispersion et de sparsité dans les données de comptages. De plus, les modèles mixtes sont des approches standard pour prendre en compte des données longitudinales.

Notons \tilde{Z}_{ijk} le taxon étudié où $i \in \{1, \dots, n\}$ représente les individus, $j \in \{1, \dots, m\}$ les taxons et $k \in \{1, \dots, n_i\}$ correspond au nombre de mesures effectués pour l’individu i (données longitudinales). De plus, nous supposons que les \tilde{Z}_{ijk} suivent une distribution binomiale négative :

$$\tilde{Z}_{ijk} \sim \mathcal{NB}(\tilde{Z}_{ijk} | \mu_{ijk}, \theta)$$

$$\tilde{Z}_{ijk} = \frac{\Gamma(\tilde{Z}_{ijk} + \theta)}{\Gamma(\theta)\tilde{Z}_{ijk}!} \left(\frac{\theta}{\mu_{ijk} + \theta} \right)^{\theta} \left(\frac{\mu_{ijk}}{\mu_{ijk} + \theta} \right)^{\tilde{Z}_{ijk}}$$

où θ est le paramètre de dispersion qui contrôle la surdispersion. De plus, cette distribution peut également être exprimée comme une mélange de lois Gamma-Poisson.

$$\tilde{Z}_{ijk} \sim \mathcal{P}(\tilde{Z}_{ijk} | \mu_{ijk}\varepsilon_{ijk})$$

$\varepsilon_{ijk} \sim Gamma(\theta, \theta)$. Ainsi, nous pouvons déterminer l'espérance et la variance de \tilde{Z}_{ijk} .

$$\text{Espérance : } \mathbb{E}(\tilde{Z}_{ijk}) = \mu_{ijk}$$

$$\text{Variance : } V(\tilde{Z}_{ijk}) = \mu_{ijk} + \frac{\mu_{ijk}^2}{\theta}$$

On remarque que $V(\tilde{Z}_{ijk}) \geq \mathbb{E}(\tilde{Z}_{ijk})$, montrant que le paramètre θ est le paramètre contrôlant la surdispersion. De plus, lorsque $\theta \rightarrow +\infty$, $V(\tilde{Z}_{ijk}) = \mu_{ijk}$, signifiant que notre modèle négatif binomiale converge vers un modèle de poisson (sans surdispersion).

Le paramètre μ_{ijk} , quant à lui, correspond à la moyenne, qui est associé à une fonction de liaison de type logarithmique :

$$\log(\mu_{ijk}) = X_{ijk}\beta + b_{ij} + \log(O_{ijk})$$

où β correspond au vecteur des effets fixes, b est le vecteur des effets aléatoires et le terme $\log(O_{ijk})$ est le paramètre d'offset qui corrige la variation de la séquence totale des ASV. De plus, nous supposons que le vecteur des effets aléatoires suit une loi normale et ainsi évite une quelconque inférence biaisée sur les effets fixes.

$$b_{ij} \sim \mathcal{N}(0, \Psi)$$

avec Ψ la matrice variance covariance.

La résolution de ce modèle se fait grâce à l'algorithme IWLS (Iterative Weighted Least Squares), développé *Xinyan Zhang et al.* [20]. Le principe est d'approximer la vraisemblance du modèle linéaire généralisé par une vraisemblance normale pondérée en mettant à jour les paramètres du modèle normal pondéré. La vraisemblance du modèle négatif binomial peut être approché par la vraisemblance normale pondérée :

$$\mathcal{NB}(\tilde{Z}_{ijk} | \mu_{ijk}, \theta) \approx \mathcal{N}(t_{ik} | \log(\mu_{ijk}), w_{ijk}^{-1})$$

où la pseudo-réponse t_{ik} est :

$$t_{ijk} = \log(\hat{\mu}_{ijk}) - \frac{\partial_{\log(\mu_{ijk})} L(\tilde{Z}_{ijk} | \log(\hat{\mu}_{ijk}), \hat{\theta})}{\partial_{\log(\mu_{ijk})}^2 L(\tilde{Z}_{ijk} | \log(\hat{\mu}_{ijk}), \hat{\theta})}$$

et le pseudo-poids w_{ijk} est :

$$w_{ijk} = -\partial_{\log(\mu_{ijk})}^2 L(\tilde{Z}_{ijk} | \log(\hat{\mu}_{ijk}), \hat{\theta})$$

On peut approximer la pseudo-réponse t_{ijk} par un modèle linéaire mixte avec w_{ijk} comme poids :

$$t_{ijk} \approx \log(\mu_{ijk}) + w_{ijk}^{-1/2} e_{ijk}$$

où $e \sim \mathcal{N}(0, \sigma^2 \mathbb{1})$. De cette manière, les paramètres $(\beta, b, \Psi, \sigma^2)$ sont mis à jour à l'aide de l'algorithme des modèles linéaires mixtes. Nous pouvons résumer cet algorithme sur le Tableau 1

Algorithm 1 Algorithme IWLS

Initialisation :

$\beta \leftarrow \text{value}$

$b \leftarrow \text{value}$

$\theta \leftarrow \text{value}$

while $(\log(\mu_{ijk})^{(j)} - \log(\mu_{ijk})^{(j-1)})^2 \leq \varepsilon$ **do**

En se basant sur les valeurs de $(\beta^{(j-1)}, b^{(j-1)}, \theta^{(j-1)})$

Calculer $t_{ijk}^{(j)}$ et $w_{ijk}^{(j)}$

Mise à jour des paramètres $(\beta^{(j)}, b^{(j)}, \theta^{(j)})$ à l'aide du modèle linéaire mixte

Mise à jour du paramètre θ par l'algorithme de Newton–Raphson

end while

2. Modèle mixte binomial négatif avec excès de zéros (ZINBMMs) :

Ce modèle est très similaire au modèle NBMMs. En effet, nous supposons que les \tilde{Z}_{ijk} suivent une distribution de type binomiale négatif avec excès de zéro (zero-inflated) appelée aussi ZINB. Cette distribution se compose en deux parties, la première partie est un modèle logistique pour prédire les zéros et la deuxième est une distribution négative binomiale pour les dénombrements surdispersés. Ainsi, nous pouvons écrire le modèle ZINB comme suit :

$$\tilde{Z}_{ijk} \sim \begin{cases} 0 & \text{avec probabilité } p_{ijk} \\ \mathcal{NB}(\tilde{Z}_{ijk} | \mu_{ijk}, \theta) & \text{avec probabilité } 1 - p_{ijk} \end{cases}$$

où p_{ijk} est la probabilité que \tilde{Z}_{ijk} soit dans un état zéro, et pour rappel μ_{ijk} et θ sont respectivement les moyennes et le paramètre de dispersion

de la distribution négative binomiale. Nous pouvons réécrire \tilde{Z}_{ijk} pour tout $\xi = (\xi_{ijn_1}, \dots, \xi_{ijn_i})$ et les $j \in \{1, \dots, m\}$:

$$\tilde{Z}_{ijk} = p_{ijk}^{\xi_{ijk}} (1 - p_{ijk})^{1-\xi_{ijk}}$$

avec ξ qui correspond aux différents états (1 pour des excès de zéros et 0 pour une distribution négative binomiale). Nous avons les mêmes hypothèses que le modèle NBMMs. Cependant l'algorithme de résolution du modèle n'est pas totalement le même. En effet, *Xinyan Zhang, Neng-jun Yi* [19] ont proposé et développé l'algorithme EM-IWLS.

La log-vraisemblance de ce modèle s'écrit de la manière suivante :

$$L(\Omega; \tilde{Z}, \xi) = \sum_{i=1}^n \sum_{k=1}^{n_i} \log[p_{ijk}^{\xi_{ijk}} (1 - p_{ijk})^{1-\xi_{ijk}}] + \sum_{i=1}^n \sum_{k=1}^{n_i} [1 - \xi_{ijk}] \log [\mathcal{NB}(\tilde{Z}_{ijk} | \mu_{ijk}, \theta)]$$

où Ω représente tous les paramètres à estimer.

L'étape d'expectation de l'algorithme EM-IWLS remplace les variables indicatrices ξ par leurs valeurs conditionnelles attendues. Tandis que dans l'étape de maximisation, l'algorithme met à jour les paramètres en exécutant une régression logistique. De cette manière, la régression logistique peut être ajustée par l'algorithme IWLS vu précédemment.

3. **Modèle linéaire mixtes (LinDA)** : Ce modèle a été développé par *Huijuan Zhou et al.* [11], pour l'analyse de l'abondance différentielle. Elle se base sur un modèle log-linéaire sur l'abondance :

$$\log(\tilde{Z}_{ij}) = u_i \alpha_j + X_{ij} \beta_j + \varepsilon_{ij}$$

où ε_{ij} est le terme d'erreur. Ce modèle est facilement extensible dans le cadre d'un modèle à effet mixtes. En effet,

$$\log(\tilde{Z}_{ij}) = u_i \alpha_j + X_{ij} \beta_j + b_j + \varepsilon_{ij}$$

où b_j représente l'effet aléatoire. Ce but est de découvrir les taxons qui sont abondants et pour cela, il faut tester pour tout $j \in \{1, \dots, m\}$,

$$H_0 : \alpha_j = 0 \quad vs \quad H_1 : \alpha_j \neq 0$$

Pour s'attaquer au problème de la compositionnalité, une transformation CLR est appliquée. Ainsi, en utilisant cette transformation, le modèle se réécrit de la manière suivante :

$$\log \left(\frac{\tilde{Z}_{ij}}{(\prod_{k=1}^m \tilde{Z}_{im})^{1/m}} \right) = u_i(\alpha_j - \bar{\alpha}) + X_{ij}(\beta - \bar{\beta}) + b_j + \varepsilon_{ij} - \bar{\varepsilon}_i$$

où $\bar{\alpha} = \frac{1}{m} \sum_{j=1}^m \alpha_j$, $\bar{\beta} = \frac{1}{m} \sum_{j=1}^m \beta_j$, $\bar{\varepsilon} = \frac{1}{m} \sum_{j=1}^m \varepsilon_{ij}$ représente les moyennes empiriques.

De plus, ce modèle suppose qu'il existe qu'une petite partie des taxons qui sont différentiels ($\alpha_j = 0$). Sous cette hypothèse, un biais de correction est appliqué, sur $\tilde{\alpha}_j$ qui est un estimateur sans biais de $\alpha_j - \bar{\alpha}$. De cette manière, l'espérance de $\tilde{\alpha}$ doit être proche de $-\bar{\alpha}$. Donc, cela revient à estimer $-\bar{\alpha}$ par $-\tilde{\alpha}$.

$$-\tilde{\alpha} = \sqrt{n}^{-1/2} \operatorname{argmax}_{x \in \mathbb{R}} \frac{1}{mh} \sum_{j=1}^m K\left(\frac{x - \sqrt{n}\tilde{\alpha}_j}{h}\right)$$

où K est une fonction à noyau uniforme avec $\int_{\mathbb{R}} K(y) dy = 1$ et h représente la fenêtre.

Afin de savoir quels taxons sont différentiellement abondants, un test statistique est nécessaire. Une estimation de la variance de $\hat{\alpha}_j$ est

$$V(\hat{\alpha}) = \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^n u_i u_i^T \right)^{-1} \hat{\sigma}_j^2$$

Par suite, nous pouvons déduire la statistique de test grâce à la loi de Student,

$$T_j = \frac{\sqrt{n}\hat{\alpha}_j}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n u_i u_i^T\right)^{-1} \hat{\sigma}_j^2}}$$

Cette statistique de test suit asymptotiquement une loi normale lorsqu'il y a beaucoup d'échantillons.

Suite à ça, une p -valeur peut être définie afin de tester l'hypothèse nulle,

$$p_j = 2F_{n-d-2}(-|T_j|)$$

où F représente la fonction de répartition de la loi de Student à $n-d-2$ degré de liberté et d la dimension de l'espace des covariables.

Algorithm 2 Algorithme LinDA

Initialisation :

Appliquer une transformation CLR aux données de comptages

Etape 1 :

Faire une régression basée sur la transformation effectuée

Calculer $\tilde{\alpha}_j$ et $\hat{\sigma}_j^2$

Etape 2 :

Estimer $\hat{\alpha}_j = \tilde{\alpha}_j - \bar{\alpha}$ (biais corrigé)

Etape 3 :

Calcul des *p*-valeurs

Multiplicité des tests en appliquant la procédure de BH

2.4 Réseaux d’inférence

Le microbiote joue un rôle dans de nombreux processus biologiques, et forme des systèmes complexes. L’analyse des associations entre les ASV est étudiée afin de comprendre les interactions entre les microbiotes. Ce qui permettrait de comprendre les mécanismes qui régissent ces écosystèmes. Ainsi, les approches par réseaux permettent d’identifier des associations statistiques entre les microbiotes, en faisant l’hypothèse que ces associations statistiques reflètent les interactions biologiques. Nous dirons donc par la suite que deux taxons sont en interaction si, et seulement si, une association statistique est observée. Cela implique que si aucune association n’est observée alors ces deux taxons ne sont pas en interaction.

La construction d’un réseau n’est pas simple. En effet, lorsqu’il existe une forte proportion de zéros, les performances des outils de détection d’associations ne sont plus fiables. Les réseaux basés sur les corrélations et les modèles graphiques ne prennent pas toujours en compte la compositionnalité des données. Dans cette étude, nous proposons deux méthodes qui prennent en compte de ces problèmes.

1. Microbial Association Graphical Model Analysis (MAGMA)

est une méthode de détection des interactions entre le microbiote qui prend en compte de la structure des données métagénomiques, impliquant :

- Un excès de zéros

- Une surdispersion des données
- Compositionnelles
- La possibilité d'intégration de covariables.

Arnaud Cougoul et al. [7] a proposé un modèle graphique gaussien combiné avec des distributions marginales GLM (Modèle Linéaire Généralisé). Ce modèle est basé sur l'estimation de données latentes par la médiane des valeurs, permettant de gérer les différents points évoqués précédemment. Comme nous l'avons vu précédemment, les données de comptage ne suivent pas une distribution normale. Pour cela, nous supposons que la distribution de \tilde{Z} peut être transformée à partir d'une variable latente normale multivariée L . D'où,

$$\tilde{Z}_{ij} = F_{ij}^{-1}(\Phi(L_{ij})) \quad (22)$$

où Φ est la fonction de répartition de la loi normale et F_{ij}^{-1} est l'inverse de la fonction de répartition de \tilde{Z}_{ij} . Nous supposons que les abondances sont distribuées selon une loi de ZINB (Zero-Inflated Negative Binomial).

$$F_{ij} \sim \mathcal{ZINB}(\lambda_{ij}, \theta_j, \pi_j) \quad (23)$$

où λ_{ij} est la moyenne de la partie binomiale négative pour l'échantillon i et le taxon j , θ_j est le paramètre de dispersion et π_j est la probabilité des zéros.

$$\log(\lambda_{ij}) = \beta_j + X_i^T \gamma_j + \log(\sigma_i) \quad (24)$$

avec β_j la moyenne des espèces j , γ_j est l'effet de la covariable X sur l'espèce j et σ_i est la profondeur de séquençage de l'échantillon i .

La résolution de ce modèle se fait par le calcul du maximum de vraisemblance. Cependant celle-ci doit être approximée. Nous transformons les données de comptage en utilisant le point médian de la distribution L des valeurs accessibles,

$$\tilde{L}_{ij} = \Phi^{-1} \left(\frac{\widehat{F}_{ij}(Z_{ij} - 1) + \widehat{F}_{ij}(Z_{ij})}{2} \right) \quad (25)$$

Pour estimer σ_i , on utilise cette fois-ci, la moyenne géométrique (GMPR). Il s'agit de la différence moyenne entre l'abondance d'une ASV trouvée dans l'échantillon i et de l'abondance dans les autres échantillons.

$$\hat{\sigma}_i = \left(\prod_{i'=1}^n r_{ii'} \right)^{1/n} \quad (26)$$

avec $r_{ii'} = mediane_j(\frac{\tilde{Z}_{ij}}{\tilde{Z}_{i'j}})$, la médiane des rapports des comptages non nuls. Par exemple, si $r_{ii'} = 2$, les taxons de l'échantillon i auront en moyenne 2 fois plus de séquences que ceux de l'échantillon i' . Enfin, nous pouvons déduire le réseau d'association à partir des données transformées \tilde{L} de la variable observée \tilde{Z} . Nous utilisons l'inférence graphique de lasso (GLasso) pour estimer une matrice de précision clairsemée Ω . L'estimation de cette matrice implique de maximiser le maximum de vraisemblance pénalisé.

$$L_{pen}(\tilde{L}, \Omega) = \log |\Omega| - Tr(S\Omega) - \rho ||\Omega||_1 \quad (27)$$

où S est la covariance de \tilde{L} , ρ le paramètre de pénalité. Comme précédemment, afin de sélectionner le modèle optimal, on utilise un critère (BIC, StARS...)

Algorithm 3 Algorithme MAGMA

Etape 1 :

Ajuster les abondances des ASV aux distributions ZINB en utilisant les équations (22) et (23).

Etape 2 :

Approximation des données latentes \tilde{Z} (Equation 25).

Etape 3 :

Estimer une matrice de précision (Equation 27).

Etape 4 :

Sélectionner la pénalité selon un critère (BIC, StARS...).

2. **PLNnetwork** [12] : Le modèle de réseau Poisson Log-Normal (PLN) pour les données multivariées peut être considéré comme un modèle PLN [4] avec une contrainte sur les coefficients de Ω . Nous supposons que la matrice de précision Ω est clairsemée et que le modèle de PLN-network est :

$$\begin{aligned} \text{Couche latente : } L_i &\sim \mathcal{N}(\mu_i, \Omega^{-1}) & \|\Omega\|_1 < c \\ \text{Espace d'observation : } \tilde{Z}_{ij} | L_{ij} &\sim \mathcal{P}(\exp[o_{ij} + X_i \beta_j + L_{ij}]) \end{aligned}$$

où μ correspond aux principaux effets, la l_1 -pénalité sur Ω vise à identifier les interactions directes entre les taxons.

Le but est donc d'estimer les paramètres $\theta = (\beta, \Omega^{-1})$. L'évaluation de la probabilité logarithmique $p_\theta(\tilde{Z}) = \log \int p_\theta(\tilde{Z}, L) dL$ est insoluble. Pour contourner ce problème, nous avons recours à une approximation variationnelle, qui consiste à trouver une approximation de la distribution conditionnelle $p_\theta(L_i | \tilde{Z}_i)$. Cette approche s'appuie sur une mesure de divergence entre la vraie distribution conditionnelle et la distribution approchée, choisie dans une classe Q de distributions simples, ici l'ensemble des distributions gaussiennes. Chaque distribution conditionnelle est approximée par une distribution gaussienne multivariées q_i . Le choix de la divergence de Kullback-Leibler pour mesurer la qualité de l'approximation conduit à l'algorithme EM "variationnel" (VEM), qui vise à maximiser la limite inférieure de la log-vraisemblance des données observées. Ainsi, la fonction à optimiser est la suivante :

$$\mathcal{J}(\tilde{Z}; \mu, S, \theta) - \lambda |\Omega|_{1,0} \quad (28)$$

où $|\Omega|_{1,0}$ est la somme des valeurs absolues des termes non diagonals de Ω . Le paramètre λ contrôle le nombre d'arêtes dans le réseau (un plus grand λ donne moins d'arêtes). μ et S représente respectivement l'approximation de la moyenne μ_i et l'approximation de la matrice de covariance diagonale par l'utilisation de l'inférence variationnelle EM. Puis comme avec le modèle précédent (MAGMA), nous nous appuyons sur l'approche de stabilité de la sélection de la régularisation (StARS) pour sélectionner une valeur optimale de λ . Bien évidemment, StARS n'est pas la seule méthode de sélection de modèle, nous pouvons extraire celui ci en utilisant par exemple le BIC.

3 Application et analyse des données du microbiote chez les personnes asthmatiques

Les données de la **Cohorte of Bronchial Obstruction and Asthma on the Environment (COBRA-ENV)** sont des données de microbiotes de patients asthmatiques. Elle caractérise les bactéries et les champignons. Ce jeu de données concerne l'échantillonnage de l'environnement intérieur de 43 patients asthmatiques et de 20 témoins. L'obtention de ces données ont été obtenues grâce à des pièges à poussières, placés au domicile des différents acteurs de l'étude sur la même période. Ces pièges à poussières ont été prélevés sur deux périodes distinctes, en hiver (du 21 décembre 2017 au 1er mars 2018) et au printemps 2019 (du 15 mars 2019 au 24 mai 2019). Cependant, le prélèvement des données des patients témoins ont été seulement réalisés durant le deuxième prélèvement (période du printemps 2019). Ainsi, l'analyse de ces échantillons au travers d'une étude métagénomique (séquençage ASV) a permis l'obtention d'une table d'abondance (Table 2). De plus, ce jeu de données inclut des variables informatives et cliniques (Table 3) qui correspondent :

1. *Information sur l'individu :*

- **Samples** : Nom de code pour identifier les patients en fonction de la saison → P_{xx} = Patient, T_{xx} = Témoin, P = Printemps, H = Hiver
- **Individus** : Numéro d'identification des individus
- **Type** : Distingue les patients des témoins
- **Etude** : Dans quelle étude provient l'individu COBRA1 = Hiver et COBRA2 = Printemps
- **Saison** : Printemps ou Hiver
- **Type2** : Les différents types d'asthme (léger, modéré, sévère)
- **ddn** : La date de naissance de l'individu *i*
- **sexé** : Le sexe de l'individu *i*
- **Poids** : Le poids de l'individu au temps *i*
- **Taille** : La taille de l'individu *i*

2. *Spatiale :*

- **Commune** : La commune à laquelle un individu *i* habite

- **Code_postaux** : Le code postal de la commune auquel l'individu *i* habite

3. Mesures cliniques :

- **FeNO** : La fraction exhalée du NO (FeNO) est une mesure permettant de déterminer l'inflammation pulmonaire d'un individu *i*. Elle est exprimée par parts per billion (ppb)
- **group_FeNO** : Seuil de référence chez un adulte. Un niveau de FeNO inférieur à 25 ppb est considéré comme normal, supérieur à 25 ppb considéré comme intermédiaire et élevé au-delà de 50 ppb
- **VEMS%** : Le VEMS ou volume expiratoire maximal par seconde correspond au volume d'air expiré pendant la première seconde d'une expiration. Ici, elle est exprimée en pourcentage
- **group_VEMS** : Valeur de référence pour VEMS%
- **CVF%** : La capacité vitale forcée (CVF) est le volume de gaz exhalé au cours d'une expiration effectuée aussi fort, rapidement et complètement que possible en partant d'une inspiration complète. De même, elle est exprimée en pourcentage.
- **Tiff** : Le coefficient de Tiffeneau permet d'évaluer le degré d'obstruction bronchique. Elle est obtenu par le rapport VEMS et CVF.
- **group_Tiff** : Valeur de référence. Le coefficient est considéré comme anormal lorsqu'il est inférieur à 70%.
- **eosino** : C'est un type de globule blanc qui joue un rôle dans la réponse de l'organisme (défense contre certains parasites ou même inflammation). Elle est exprimée en cells/mcL.
- **group_eosino** : Valeur de référence pour eosino.
- **DEMM%** : Volume expiré entre 25-75% de la capacité vitale. Appelé aussi DEM 25-75.

Dans cette étude, nous chercherons à expliquer quels sont les microorganismes et les facteurs qui affectent les personnes asthmatiques. Pour cela, nous nous baserons sur les méthodes évoquées tout au long de ce rapport. Avant d'effectuer une analyse sur ces données, nous allons procéder à un nettoyage/préprocessing.

P5451.P	P6.P	P7.P	P98.P	P109.P	P109.H	
0	0	0	0	0	0	...
0	0	0	0	0	0	...
0	0	0	0	0	0	...
0	0	0	13	0	0	...
0	0	0	0	0	0	...
0	0	0	0	0	0	...

TABLE 2 – Table d’abondance

Samples	Individus	Type	Etude	Saison	...
P2.P	P2	Patient	COBRA2	Printemps	...
P3.P	P3	Patient	COBRA2	Printemps	...
P5.P	P5	Patient	COBRA2	Printemps	...
P6.P	P6	Patient	COBRA2	Printemps	...
P7.P	P7	Patient	COBRA2	Printemps	...
P98.P	P98	Patient	COBRA2	Printemps	...

TABLE 3 – Données cliniques

3.1 Nettoyage de données

En général, lorsqu’on travaille avec les données microbiotes, nous comparons les résultats selon la même famille taxonomique afin d’effectuer une comparaison juste. Chez les bactéries, nous nous placerons sur la taxonomie des genres et pour les champignons, nous nous placerons sur la taxonomie des espèces. Ce choix n’est pas anodin. En effet, nous arrivons à identifier plus de différences entre les espèces de champignons, tandis que chez les bactéries, cette différence est très faible, les espèces de bactéries ont tendance à se ressembler. En effet, la portion du gène amplifié pour discriminer les communautés bactériennes (région V3-V4 de l’ADNr 16S) n’est pas suffisamment résolutif car elle ne contient pas suffisamment de diversité nucléique pour obtenir une information fiable à l’espèce tandis que pour les champignons, la région inter-génique (ITS2) de l’ADNr 18S permet une résolution à l’espèce. Ainsi, le nombre de taxon donc variable va considérablement réduire, passant de 21855 taxons à 637 taxons chez les bactéries.

Afin d’éviter des redondances, dans cette partie, nous parlerons uniquement des bactéries et non des fongiques. De plus, nous n’allons pas aborder

et explorer toutes les variables. Cependant, vous retrouverez en annexe ainsi que sur le site web interactif tout les résultats y compris ceux des espèces fongiques ainsi que les autres analyses effectuées.

Afin de compléter l'analyse, nous avons créer des variables supplémentaire.

1. *Information Patient* :

- **Age** : L'âge de l'individu déterminé par la date de naissance.
- **AgeGroup** : Catégorisation de l'âge en 3 groupes (moins de 50ans, entre 50 et 70ans et plus de 70ans).
- **Active** : Activité professionnelle d'un individu basé sur si la personne est actif professionnellement ou pas (déterminé par l'âge).
- **IMC** : Classification du poids de l'individu i à partir de l'indice de masse corporel
- **State** : Changement de type d'asthme selon la saison (ex : "High" en hiver → "Low" en printemps)

2. *Spatiale* :

- **CoastalWetlands** : L'individu i habite près de la mer, dans un rayon de 10km.
- **ForestationRate** : La surface boisé de la commune de l'individu i définie Institut National de l'Information Géographique et Forestière [2].
- **Zone** : Classification en zone urbanisé des communes définie par l'INSEE [3].
- **Humidité** : Classification en zone humides des communes. Un milieu humide peut être ou avoir été en eau, inondé ou gorgé d'eau de façon permanente ou temporaire. L'eau peut y être stagnante ou courante, douce, salée ou saumâtre. Ces données proviennent des Eaux de France [1].

3.2 Prévalence

Dans un premier temps, nous nous sommes intéressés à la prévalence car nous remarquons dans la table d'abondance 2, un grand nombre de zéros. La prévalence est une mesure qui décrit la présence d'un taxon dans un échantillon. Autrement dit, c'est la fréquence à laquelle un taxon apparaît chez un individu.

	Prévalence
Physcomitrella	1
Propionibacterium	1
Kocuria	1
Skermanella	1
Roseomonas	1
Methylobacterium	1

TABLE 4 – Prévalence de chaque genre taxonomique par ordre décroissant au printemps

Au printemps, *Roseomonas* apparaît chez tous les individus (Table 4).

La prévalence nous permet aussi de choisir des axes d'études, microbiote majoritaire et microbiote minoritaire. Car en effet, la prévalence diminue rapidement. Ainsi, beaucoup d'individus partagent des taxons communs et inversement, peu d'individus partagent le même taxon (Figure 6).

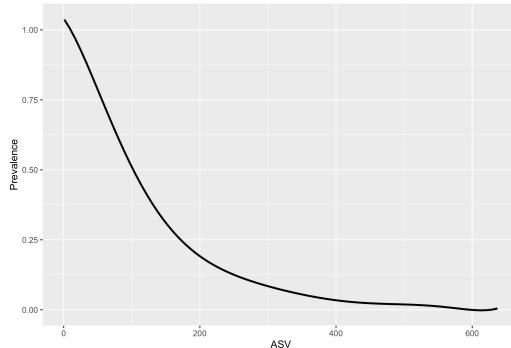


FIGURE 6 – Evolution de la prévalence par ordre décroissant des genres bactériens au printemps

On pourrait se demander si la prévalence a un lien avec l'abondance ? Est-ce que les plus abondants sont les plus prévalents ?

3.3 Abondance

Nous pouvons représenter l'abondance sous plusieurs formes. Nous pouvons à l'aide d'un diagramme en barre visualiser l'abondance. Plus particulièrement des taxons les plus abondants. On retrouve parmi les taxons

les plus abondant Roseomonas, qui est un taxon apparaissant chez tout les individus. Parmi eux, *Paracoccus* est le taxon le plus abondant de notre échantillon (Figure 7).

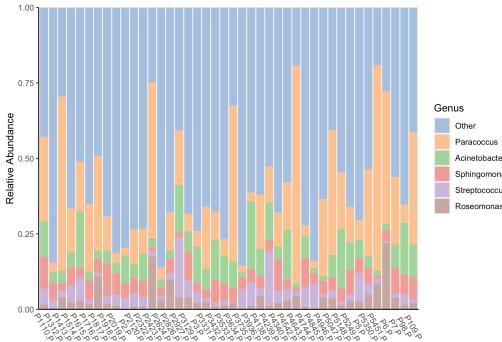


FIGURE 7 – Les 5 genres bactériens les plus abondants au printemps

Afin de savoir s'il existe une différence de diversité et composition, nous devons utiliser des outils statistiques.

3.4 Alpha-diversité

Nous calculons l'alpha diversité en prenant plusieurs indices de diversité, qui seront *Chao1*, *Shannon* et *Simpson*. Bien sûr, une représentation serait plus agréable que de simple nombres. C'est pourquoi nous représenterons l'alpha diversité sous la forme d'une boîte à moustache. Dans un premier, nous pouvons regarder s'il existe une différence de diversité entre un patient et un témoin. Afin de déterminer s'il existe une différence entre les deux groupes, un test non-paramétrique de Wilcoxon est appliqué. Ce test suppose sous l'hypothèse nulle H_0 que les distributions des deux groupes sont similaires. Nous remarquons qu'il existe une différence de diversité au risque de 5% entre les personnes asthmatiques et non-asthmatiques (Figure 8). Ainsi, les personnes non-asthmatiques possèdent moins de genres bactériens. De plus, ces taxons sont répartis de manière inégales. Tandis que les personnes ayant de l'asthme ont un nombre d'espèces est plus conséquent, et leur abondance est plus équilibrée (ne présente que peu d'inégalité).

Si nous voulons savoir s'il existe une différence entre les types d'asthmes selon la saison à l'aide d'un test, nous devons mettre en place un modèle. Un simple modèle linéaire mixte peut suffire.

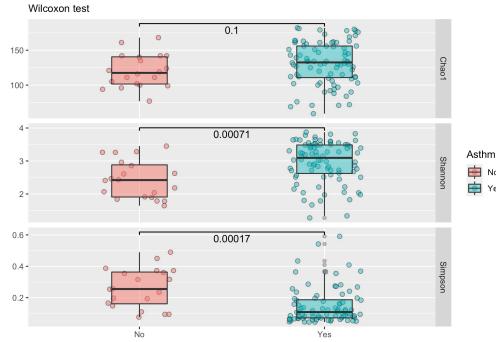


FIGURE 8 – L’alpha-diversité des genres bactériens entre les personnes asthmatiques et les personnes non-asthmatiques

	Estimate	Std. Error	df	t value	$Pr(> t)$
(Intercept)	14997.450	1636.455	82.12713	9.164594	0.0000000
Type2High	5112.713	2178.037	87.92130	2.347395	0.0211507
Type2Low	4809.402	2146.306	87.09531	2.240781	0.0275840
SeasonWinter	-3747.153	1145.682	51.09006	-3.270675	0.0019251

TABLE 5 – Effets fixes du modèle linéaire mixte (Chao1)

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{Type2}_{ij} + \beta_2 \cdot \text{Season}_j + b_i + \varepsilon_i$$

où les Y_{ij} représente l’indice étudié (Chao1 par exemple), b est le vecteur des effets aléatoires, σ est l’erreur de mesure.

Ainsi, on remarque que la saison et le type d’asthme ont un effet sur la diversité. En effet, les personnes asthmatiques ont un nombre d’espèces plus élevé que les personnes non-asthmatiques. De plus, cette augmentation du nombre d’espèce se voit aussi selon le type d’asthme. En effet, plus elle est sévère, plus le nombre d’espèces augmente (Table 5). Ces résultats coïncident avec la boîte à moustache observée dans la Figure 10.

S’intéresser à l’urbanisation est intéressant. En effet, ces dernières années, le rythme d’urbanisation n’a fait qu’augmenter. Aujourd’hui, plus de la moitié de la population mondiale vit dans un environnement urbain. Cependant, nous remarquons une différence de diversité au risque de 5% entre les personnes asthmatiques habitant dans une zone urbaine et les personnes habitant dans une zone rurale. Lorsqu’on habite dans une zone rurale, les

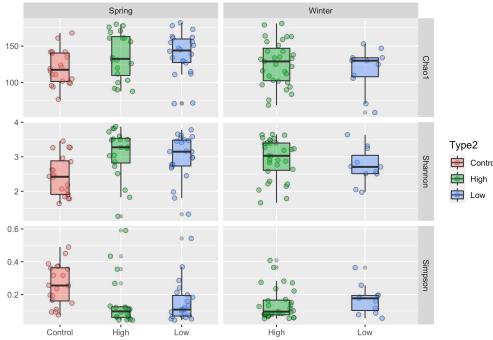


FIGURE 9 – L’alpha-diversité des genres bactériens entre les asthmatiques ayant ou non un asthme sévère

personnes asthmatiques possèdent plus d’espèces avec des abondances plus équilibrées contrairement aux personnes asthmatiques habitant dans les zones urbaines. Ainsi, dans un milieu urbain, les patients asthmatiques passent plus de temps dans un environnement clos ou intérieur, ce qui implique donc que ces patients sont exposées à une plus faible diversité microbienne, impactant par la même occasion notre système immunitaire. De plus, ces résultats confirment ceux de *Sharma et al.* [6].

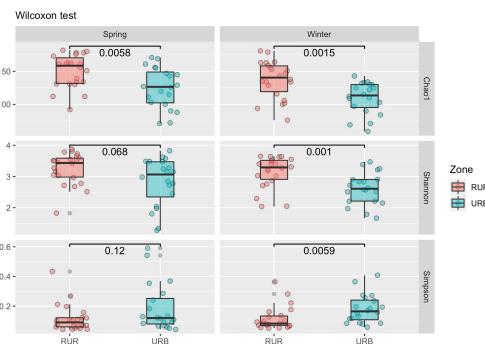


FIGURE 10 – L’alpha-diversité des genres bactériens entre les asthmatiques habitant dans une zone urbaine ou rurale

Est ce que cette différence de diversité implique une différence de composition microbienne ?

3.5 Beta-diversité

Dans cette partie, nous utiliserons une PCoA avec la distance de bray-curtis. De plus, afin de différencier de manière significative, nous effectuerons un test de PERMANOVA.

Nous remarquons qu'il existe au risque de 5%, une différence de composition microbienne entre les personnes asthmatiques et les personnes non-asthmatiques (Table 6). Cependant, les deux groupes ne sont pas répartis de manière homogène, mais la Figure 11 montre deux groupes, l'un inclus dans l'autre. Tandis que, la saisonnalité entraîne un changement de composition microbien. En effet, on identifie deux clusters distincts (Figure 12).

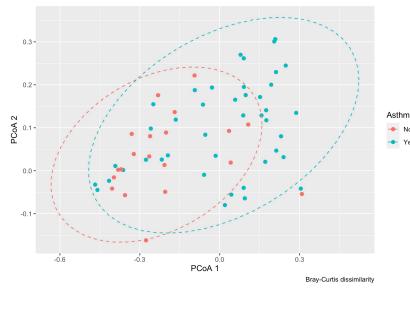


FIGURE 11 – Graphe des individus

	No-Yes
betadisper	0.005
PERMANOVA	0.002

TABLE 6 – p -valeur

Beta-diversité des genres bactériens entre les personnes asthmatiques et non asthmatiques

3.6 Abondance différentielle

Dans cette partie, nous utiliserons un modèle linéaire mixte (LinDA). Nous représenterons seulement les taxons qui sont significatifs au risque de 5% (en prenant en compte de la multiplicité des tests) à l'aide d'un diagramme. De plus, nous représentons la proportion de "différence" en utilisant une mesure appelée log2FoldChange. C'est une mesure très utilisée dans ce contexte car elle décrit le changement de quantité en terme d'abondance entre les deux groupes comparés. En outre, il s'agit d'un rapport de deux valeurs.

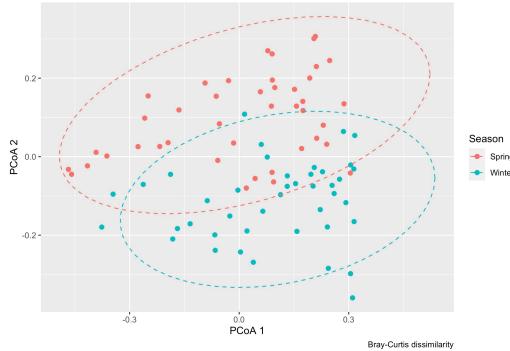


FIGURE 12 – Beta-diversité des genres bactériens en hiver et printemps

Malheureusement ou non, nous n'injecterons pas tous les taxons au modèle, mais seulement un sous-échantillon de celui-ci. En effet, selon les cliniciens, il est préférable de s'intéresser au microbiote majoritaire, c'est-à-dire aux taxons qui ont une abondance relative supérieure à 10% ainsi qu'une prévalence à plus de 20%. De cet fait, nous remarquons que *Finegoldia* est environ 8 (2^3) fois plus abondant chez les personnes asthmatiques en comparaison avec les personnes non-asthmatiques. De même, pour le genre bactérien *Bronchothrix* qui est environ 5.6 fois plus abondant chez les personnes asthmatiques. Quant au genre *Paracoccus*, lui, est environ 3 ($2^{-1.5}$) fois moins abondant chez les personnes asthmatiques (Figure 13)

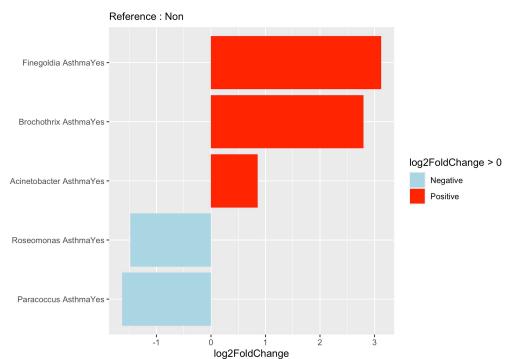


FIGURE 13 – Abondance différentielle des genres bactériens entre les personnes asthmatiques et non-asthmatiques au printemps

3.7 Réseaux d'inférence

Afin d'étudier les différentes associations qu'il peut y exister entre les personnes asthmatiques et non asthmatiques, nous allons créer notre réseau PLN. Ici, nous avons seulement expliqué l'abondance sans prendre en compte des covariables afin de pouvoir distinguer les différentes associations entre les deux groupes (asthmatiques et témoins). Nous représenterons tout naturellement notre modèle sous la forme d'un réseau où les arêtes représenteront les différentes associations et leurs épaisseurs, la force de ces associations (poids). De plus, comme précédemment, nous étudierons le microbiote majoritaire.

Au printemps, chez les genres bactériens, nous remarquons chez les personnes asthmatiques que *Brevundimonas* et *Finegoldia* sont négativement corrélés, nous disant que la présence de ces deux taxons entraîne une baisse de l'abondance de ces derniers (moins il y a de *Finegoldia* moins il y a de *Brevundimonas*). De plus, *Corynebacterium* est positivement associé à *Ezakiella*, ce qui veut dire que si l'abondance de l'un augmente l'abondance de l'autre augmente de même. Tandis que chez les personnes non asthmatiques, le nombre d'associations est moins élevé, et on remarque que *Finegoldia* est cette fois positivement corrélé non pas avec *Brevundimonas* mais avec *Clostridiisalibacter* (Figure 14).

Ainsi, les genres bactériens *Finegoldia* et *Ezakiella* ont un rôle différent dans chacun des deux groupes.

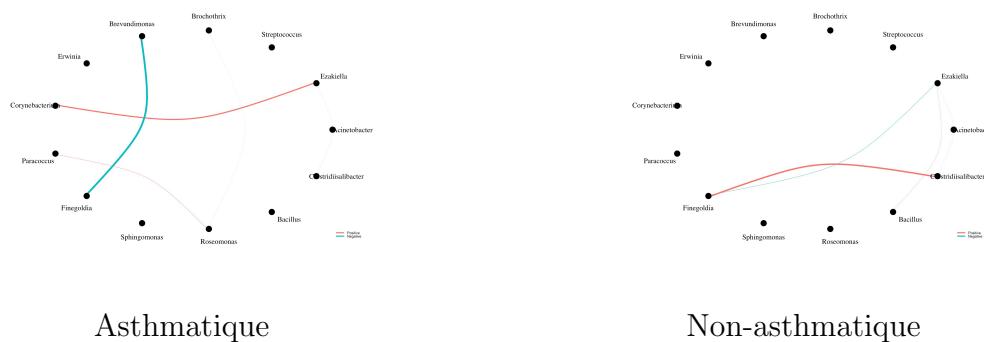


FIGURE 14 – Association des genres bactériens majoritaires entre les personnes asthmatiques et non-asthmatiques au printemps

Conclusion

L'étude des données de microbiotes est aujourd'hui devenu un enjeu majeur à la compréhension du fonctionnement et l'impact des microorganismes sur le corps humain et sur son environnement. Hélas, ces données ont pour la particularité d'être sparses (grand nombre de zéro), surdispersées (valeurs d'abondance large) et compositionnelles (abondance relative). De plus, ces données sont en grande dimension, c'est-à-dire qu'elles contiennent plus de variables que d'échantillons, rendant l'analyse de celle-ci, une quête ardue. C'est pourquoi, la mise en place d'outils statistiques adaptés est nécessaire. Ces outils permettent d'analyser à la fois le microbiote d'un individu dans sa globalité ou de façon précise, à un rang taxonomique, voire même, au niveau d'un taxon. En effet, les outils d'analyses statistiques exploratoires (analyse de diversité) permettent d'avoir un aperçu globale d'un échantillon. Les analyses en statistiques avancées font intervenir des méthodes complexes permettant de comprendre les interactions ou encore les différences pouvant exister au sein des différents microorganismes grâce à des réseaux d'inférence, des méthodes de classifications ou encore des méthodes d'analyse de corrélation. Tous ces outils m'ont permis, dans le cadre de mon stage, à répondre aux différentes interrogations des cliniciens. Il est clair que le microbiote de l'air intérieur a un impact sur les personnes asthmatiques. La temporalité plus particulièrement la saison ainsi que le fait d'habiter dans une zone urbaine ou rurale sont les facteurs les plus importants d'après notre étude, qui modifient à la fois la diversité et la composition du microbiote d'un individu. Nous identifions *Finegoldia*, *Brevundimonas*, *Clotridiisalibacter*, *Ezakiella* comme étant des genres bactériens, ayant un impact chez les personnes asthmatiques. Tandis que pour les espèces fongiques, nous avons trouvé peu de facteurs qui influent sur les personnes asthmatiques. On retrouve également la saisonnalité qui est un facteur qui influent le mycobiote des personnes asthmatiques. Il est normal que l'on retrouve que de facteurs influant les personnes asthmatiques. En effet, les champignons ont la particularité de résister à des conditions environnementales extrêmes et leurs spores se déplacent dans l'air à des distances importantes. Ainsi la localisation détermine la biomasse de l'air intérieur, qui évoluera en fonction des saisons. Nous pouvons expliquer l'impact de la saison par le type de chauffage utilisé par une diminution de la diversité fongique et une augmentation de l'espèce *Epicoccum nigrum*. Malgré ça, nous pouvons émettre certaines limites à cette étude. L'environne-

ment, notamment la localisation semblerait jouer un rôle dans le microbiote de l'air intérieur. Or, les données issues de COBRA-ENV proviennent uniquement de la région Nouvelle-Aquitaine. De plus, l'assignation taxonomique n'est pas assez précise, car peu d'affectations d'espèces ont été réalisées à cause de la couverture du gène marqueur. Une suite à cette étude serait d'approfondir les résultats avec d'autres tests statistiques (ANOSIM, MiRKAT, ...), une utilisation d'autres métriques pour la PCoA (indice de Jaccard). Ainsi, l'apport de nouvelles données permettrait d'apporter plus d'informations, et d'approfondir l'étude.

Annexes

3.8 Organigramme

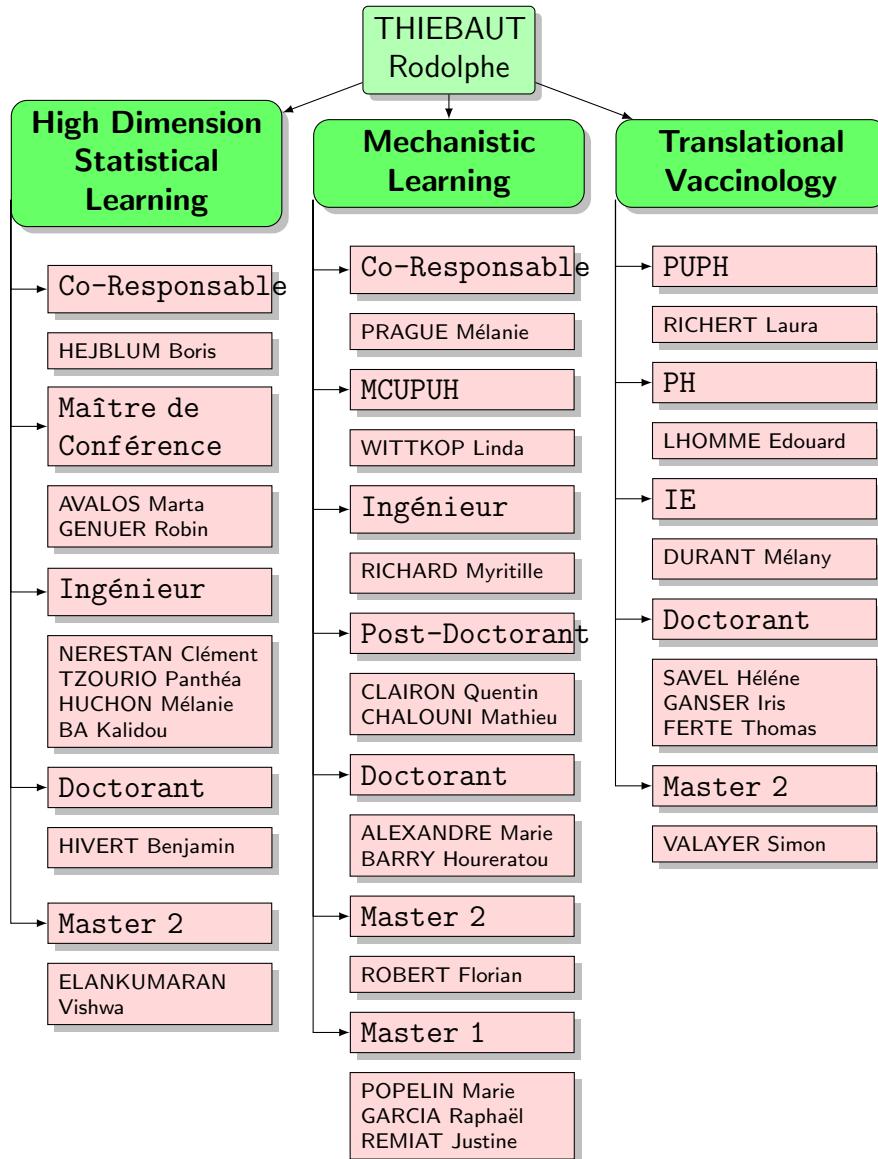


FIGURE 15 – Organigramme de l'entreprise

3.9 Définitions biologiques

Bactérie

Une bactérie est un micro-organisme formé d'une seule cellule, sans noyau.

Cohorte

Désigne un groupe de personnes.

Dybiose

La dysbiose se traduit par un déséquilibre. Il peut s'agir d'une diminution du nombre de bonnes bactéries ou d'une hausse des mauvaises bactéries.

Eubiose

A l'opposé de dybiose, elle caractérise un équilibre du microbiote.

Exposome

L'exposome est un terme qui définit l'ensemble des expositions environnementales rencontrées par un individu au cours de sa vie.

Gène

Un gène est une séquence de nucléotides dont l'expression affecte les caractères d'un organisme.

Génome

C'est l'ensemble des gènes d'un individu.

Métagénomique

C'est la méthode qui vise à étudier le microbiome. C'est une technique de séquençage et d'analyse de l'ADN contenu dans un milieu. Elle séquence les génomes de plusieurs individus d'espèces différentes dans un milieu donné. Une analyse nous donnera la composition d'un microbiome c'est-à-dire quelles espèces sont présentes, leurs abondances et leurs diversités.

Microbiome

On parle de microbiome lorsqu'on fait référence aux génomes (données génétique) du microbiote.

Microbiote

Le mot microbiote désigne l'ensemble des micro-organismes (bactéries, virus, parasites champignons non pathogènes) qui vivent dans un environnement spécifique. Dans la littérature scientifique, on dit qu'un microbiote est une bactérie et un mycobiote un champignon.

Phénotype

Ensemble des caractères observables d'un organisme.

Pipeline

Enchaînement d'étapes informatique qui a pour but de donner sens à un fichier de séquençage.

Prévalence

Mesure dénombrant le nombre de cas de maladies à un instant t , permettant de connaître l'état de santé d'une population.

Probiotique

Micro-organisme vivant qui, ingéré en quantité suffisante, a un effet bénéfique sur la santé.

Richesse Microbienne

C'est la diversité des groupes présents dans le microbiote. Plus un microbiote est riche, meilleure est sa santé.

Séquençage

Consiste à déterminer l'ordre d'enchaînement des nucléotides, aussi appelé séquence, pour un fragment d'ADN donné. Appelé aussi des reads.

Sputum

Sputum ou expectoration en français est un mélange de salive et de mucus toussé par les voies respiratoires, généralement à la suite d'une infection ou d'une autre maladie.

Structure phylogénétique

On parle de structure phylogénétique lorsqu'un microbe a une histoire évolutive et des informations sur ses liens de parenté sont disponibles.

Taxonomie

La taxonomie désigne la classification des bactéries selon le phylum, la classe, l'ordre, la famille, le genre, l'espèce et la souche.

Unité Taxonomique Opérationnelle (OTU)

Une OTU est un regroupement d'individus d'une même espèce (organismes avec des caractéristiques proches) dont les séquences d'ARNr 16S (pour les microbiotes) et les séquences d'ARNr ITS (pour les mycobiotes) présentent une similitude de plus de 97%. Les OTUs peuvent être utilisés pour classifier les espèces taxonomiques bien qu'elles ne représentent pas les espèces bactériennes. La table des OTUs est un tableau à double entrées contenant le nombre de séquences par OTU et par échantillon. On parle d'abondance.

Voies Respiratoires

- Les voies supérieures (extra-thoracique) sont constituées des fosses nasales, de la bouche, du pharynx et du larynx. Elles ont le rôle de conduction en permettant l'humidification, la filtration, le réglage de la température de l'air inspiré et le transport de l'oxygène vers les poumons.
- Les voies inférieures (intra-thoracique) sont une zone de conduction composée de la trachée, des bronches et bronchioles, associée à une zone d'échanges gazeux, des poumons aux alvéoles pulmonaires.

3.10 Application à l'étude COBRA-ENV

Il n'y a pas de différence de diversité chez les espèces fongiques entre les personnes asthmatiques et non-asthmatiques (Figure 16)

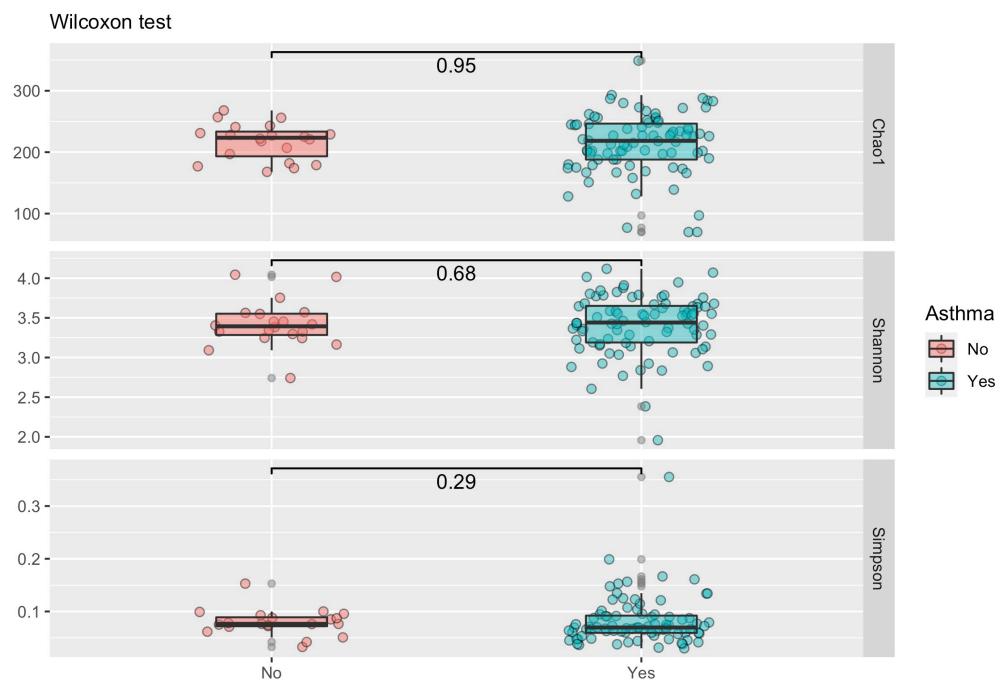


FIGURE 16 – Alpha diversité des espèces fongiques entre les personnes asthmatiques et non-asthmatiques

Les variables cliniques, représentant des mesures cliniques n'ont pas d'impact sur la diversité (Figure 17)

Nous pouvons dire que chez les espèces fongiques qui partagent le mycobiote d'un individu ne diffère pas lorsqu'on est asthmatique ou non (Figure 18).

Il est clair qu'il y a une différence de composition microbienne et mycobienne entre les deux saisons, hiver et printemps (Figure 19).

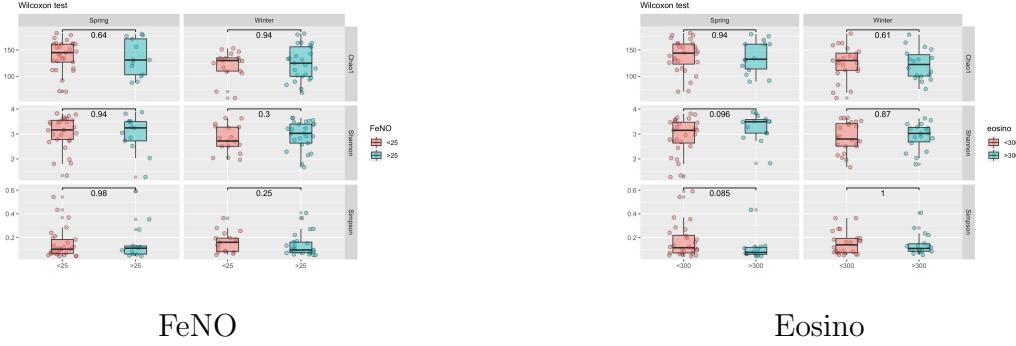


FIGURE 17 – Alpha diversité des genres bactériens chez les personnes asthmatiques

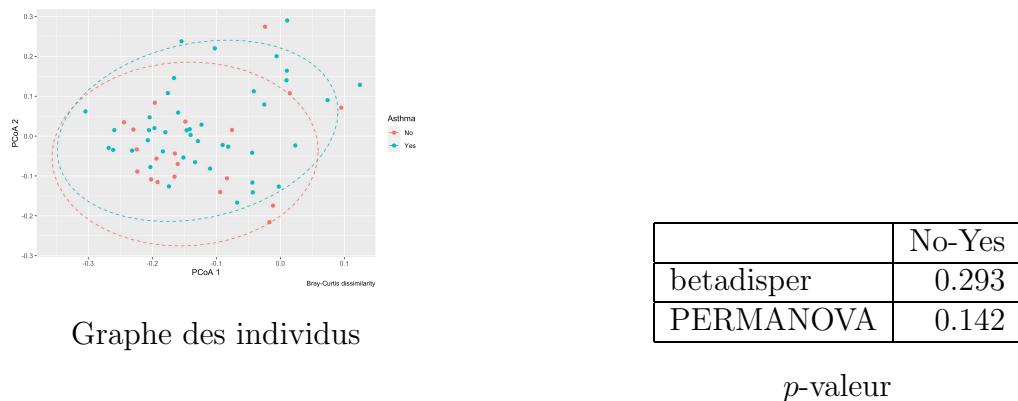
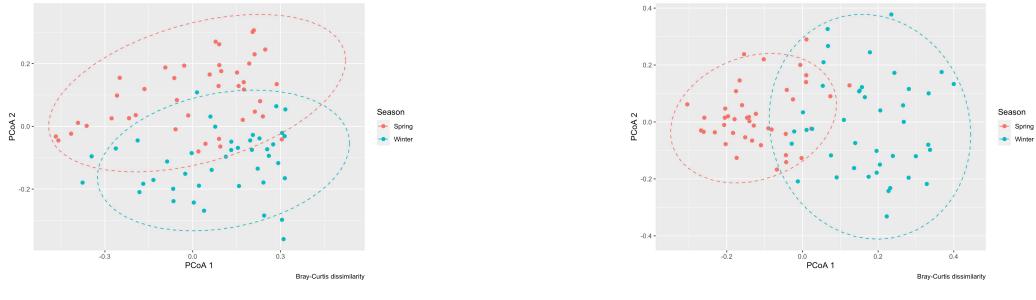


FIGURE 18 – Beta diversité des espèces fongiques entre les personnes asthmatiques et les personnes non asthmatiques



Genres bactériens

Espèces fongiques

FIGURE 19 – Beta diversité entre le printemps et l'hiver



Genres bactériens

Espèces fongiques

FIGURE 20 – Taxon abondant différemment entre le printemps et l'hiver



FIGURE 21 – Réseaux d’association entre les personnes habitant dans des zones

Références

- [1] Eaux de france. <http://www.zones-humides.org/les-prelocalisations-et-inventaires-de-milieux-humides>.
- [2] Grille communale de densité. <https://inventaire-forestier.ign.fr/spip.php?article646>.
- [3] Institut national de l'information géographique et forestière. <https://www.observatoire-des-territoires.gouv.fr/grille-communale-de-densite>.
- [4] J. Aitchison and C. H. Ho. The multivariate poisson-log normal distribution. 1989.
- [5] John Aitchison. *The Statistical Analysis of Compositional Data*. 1986.
- [6] Jack A Gilbert Anukriti Sharma. Microbial exposure and human health. 2018.
- [7] Xavier Bailly Arnaud Cougoul and Ernst C. Wit. Magma : inference of sparse microbial association networks. 2019.
- [8] Benjamin J Callahan and al. *DADA2 : High Resolution Sample Inference from Illumina Amplicon Data*. 2016.
- [9] Mahendra Mariadassou Chiquet, Julien and Stéphane Robin. Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 2018.
- [10] Global Initiative for Asthma. *Traitemen et Prévention de l'Asthme pour les adultes et les enfants de 5ans et plus*. 2019.
- [11] Jun Chen Xianyang Zhang Huijuan Zhou, Kejun He. Linear models for differential abundance analysis of microbiome compositional data. *Genome Biology*, 2022.
- [12] Stéphane Robin Julien Chiquet, Mahendra Mariadassou. Variational inference of sparse network from count data. 2018.
- [13] Pawlowsky-Glahn Martin-Fernandez, Barcelo-Vidal. *Dealing With Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation*. 2003.
- [14] Vera Pawlowsky-Glahn, Juan Jose Egozcue, and Raimon Tolosana-Delgado. *Lecture Notes on Compositional Data Analysis*. 01 2007.

- [15] Yifan Shan, Weidong Wu, Wei Fan, Tari Haahtela, and Guicheng Zhang. House dust microbiome and human health risks. *International Microbiology*, 2019.
- [16] Robert Harding Whittaker. Vegetation of the siskiyou mountains, oregon and california. 1960.
- [17] C R Woese, O Kandler, and M L Wheelis. Towards a natural system of organisms : proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 1990.
- [18] Nengjun Yi Xinyan Zhang. Nbzimm : negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC Bioinformatics*.
- [19] Nengjun Yi Xinyan Zhang. Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics*, 2020.
- [20] Zaixiang Tang Lei Zhang Xiangqin Cui Andrew K. Benson Nengjun Yi Xinyan Zhang, Himel Mallick. Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, 2017.