République Tunisienne
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de **C**arthage

Ecole **S**upérieure de la **S**tatistique et de l'**A**nalyse de l'**I**nformation

Rapport de Projet de Fin d'Etudes présenté pour l'obtention du

Diplôme National d'Ingénieur en Statistique et Analyse de l'Information

Sana ZAGHOUANI

Fouille de données pour l'exploration du lien entre microbiote intestinal et santé

Soutenu le 26 octobre 2018 devant le Jury composé de:

- Mme Aïcha El-Golli Jabbes, Maître assistante, ESSAI (Président)
- Mme Héla Ouaili-Mallek, Assistante, ESSAI (Rapporteur)
- Mme Magali Berland, Ingénieure de recherche, INRA (Encadrant entreprise)
- M. Dhafer Malouche, Maître de Conférences, ESSAI (Encadrant universitaire)

Stage de Fin d'Etudes effectué à

Année universitaire 2017/2018

Avec tout respect et amour je dédie ce modeste travail,

À mes parents.

Remerciements

Peu de mots peuvent traduire ma profonde gratitude envers les personnes qui m'ont aidé à la réalisation de ce travail.

Je tiens à remercier mes encadrants à l'INRA de Jouy-en-Josas, Mme Magali Berland et M. Mahendra Mariadassou pour leur patience, pour le temps qu'ils m'ont consacré tout au long de cette période, de m'avoir dirigé et guidé durant la réalisation de ce travail, pour leur assistance et leurs précieuses recommandations.

Je témoigne ma gratitude à tous les membres de l'équipe « InfoBioStat » dans laquelle j'ai travaillé, pour leur extrême sympathie, les moments partagés ensemble et pour l'atmosphère conviviale et très professionnelle qui règne dans leurs bureaux.

Je voudrais également remercier M. Dhafer Malouche, pour la qualité et la pertinence de ses conseils.

Enfin, mes remerciements sont adressés à tous les professeurs et tout le personnel contribuant à offrir, au sein de l'École Supérieure de la Statistique et de l'Analyse de l'Information (ESSAI), cette qualité de formation.

Résumé

Dès la naissance, nos intestins sont colonisés par des milliards de milliards de micro-organismes : bactéries, archées, virus, champignons. Cet écosystème microbien, aussi appelé microbiote intestinal, interagit avec nos cellules et notre santé tout au long de notre vie. Son exploration offre de nouvelles pistes dans la compréhension de certaines maladies et pour de futurs traitements. Il est difficile de le faire avec les méthodes biomoléculaires classiques, c'est pourquoi différentes variantes de métagénomique se sont fortement développées ces dernières années. Parmi elles, la métagénomique quantitative et fonctionnelle qui consiste à quantifier l'ensemble des gènes et leurs fonctions provenant de tous les micro-organismes détectés au sein d'échantillons d'individus qui, dans un contexte clinique, sont sains ou malades. Cependant, ces données massives nécessitent un traitement statistique afin d'extraire les associations significatives entre les données de gènes et les données cliniques. Le travail de recherche présenté dans ce manuscrit, consiste à réaliser des analyses multivariées sur ces données métagénomiques de comptages, qui présentent certaines particularités dont une forte inflation de zéro. Plusieurs méthodes de réductions de dimension non supervisées (ordination, Probabilistic Poisson PCA), et supervisées (LDA Poisson log normal) sont alors utilisées. Ces méthodes ont permis d'identifier et de visualiser des structures de microbiote intestinal non liées au statut « sain /malade » mais plutôt à la richesse en espèces bactériennes dans les échantillons.

Mots clés : métagénomique, microbiote intestinal humain, données de comptages, analyse multivariée non supervisée, ordination, Probabilistic Poisson PCA

Table des matières

Liste des tableaux

Liste des figures

Glossaire

- **ADN :** macromolécule biologique présente dans toutes les cellules. L'ADN est le support stable et transmissible de l'information génétique ; il définit les fonctions biologiques d'un organisme. Les molécules d'ADN sont composées de quatre nucléotides possibles : adénine (A), cytosine (C), guanine (G) ou thymine (T).

- **Gène :** séquence de la molécule d'ADN qui code pour une molécule qui a une fonction. Un gène est caractérisé à la fois par sa position et par l'ordre de ses nucléotides.

- **Génome :** l'ensemble des gènes d'un organisme.

- **Séquençage :** le séquençage de l'ADN consiste à déterminer l'ordre d'enchaînement des nucléotides pour un fragment d'ADN donné**.** Le séquenceur produit un texte numérique composé d'une succession de lettres dans l'alphabet {A, C, G, T}.

- **Lecture ou « read » : le** texte numérique produit par le séquenceur consiste en des millions de courtes séquences d'ADN. Chacune de ces séquences est appelée « lecture » issue du terme anglais « read ».

- **Alignement (Mapping) :** le mapping consiste à aligner chaque lecture issue du séquençage sur une base de données de gènes de référence. Il peut s'agir d'un l'alignement complet (global) de la lecture, ou bien d'un alignement local. Un alignement d'une qualité suffisante permet d'attester de la présence d'un gène donné.

- **Profondeur de séquençage :** le nombre de lectures généré lors d'un séquençage. Il peut différer entre les échantillons pour des raisons techniques liées au séquenceur.

- **Richesse :** l'indice de biodiversité le plus courant, et le plus utilisé en biologie, il est égal au nombre d'espèces différentes présentes dans un échantillon.

- **Espèce :** le plus bas niveau d'une taxonomie (classification hiérarchique) qui contient plusieurs niveaux : Pour l'homme, ces niveaux sont, par ordre croissant, l'espèce (*Homo sapiens*), le genre (Homo), la famille (Hominidés), l'ordre (Primates), la classe (Mammifères), l'embranchement (Chordés), et le règne (Animaux).

Introduction

Ce document constitue le rapport de mon stage de fin d'études au sein de l'équipe de recherche et développement InfoBioStat de l'unité MetaGenoPolis de l'INRA de Jouy-en-Josas. Ce projet marque la fin de ma formation d'ingénieur à l'École supérieure de la Statistique et de l'Analyse de l'Information et représente l'occasion de mettre en pratique toutes les connaissances acquises durant ces dernières années. Il s'agit d'une activité de recherche dans le domaine de la biostatistique sur des problématiques de data mining sur des données de comptages métagénomiques en grande dimension et dont l'objectif est d'explorer le lien entre le microbiote intestinal et la santé humaine.

La santé humaine est fortement liée au microbiote intestinal humain. Il est difficile de l'étudier avec les méthodes classiques, c'est pourquoi la métagénomique s'est fortement développée ces dernières années. Actuellement, beaucoup de données de comptages sont générées par la métagénomique. En effet, l'unité MetaGenoPolis de l'INRA, Jouy-en-Josas manipule des matrices d'abondances de plusieurs millions de gènes ou plusieurs milliers d'espèces bactériennes et leurs fonctions pour un nombre d'individus donné. L'un des challenges pour l'analyse de ces données est d'être capable de les normaliser et de les explorer, afin de trouver les structures cachées du microbiote intestinal chez des individus sains et/ou malades de différentes cohortes.

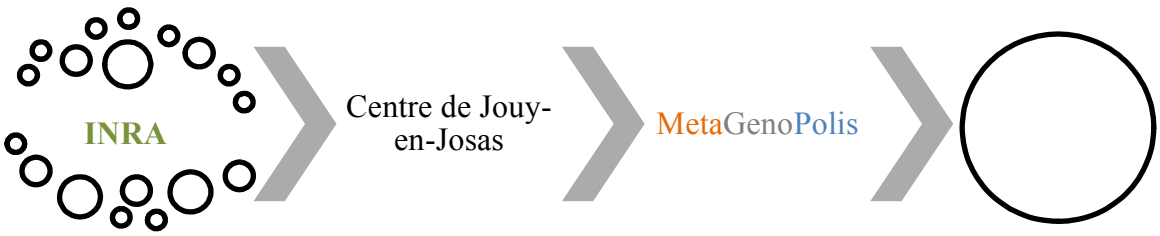
Dans ce projet, nous nous intéresserons particulièrement à l'exploration de grands volumes de données de comptages métagénomiques sparses provenant d'échantillons du microbiote intestinal humain en appliquant des méthodes d'apprentissage non supervisées afin d'identifier les similarités et les différences au sein de ces données. C'est un problème de réductions de dimension et de recherche de structure discrète. Pour cela, deux bases de données de comptages produites sur des individus sains et atteints d'une cirrhose du foie ont été utilisées, notamment des comptages d'espèces bactériennes pour la première et des comptages des fonctions des gènes pour la deuxième.

Chapitre I
Cadre général du projet

1 Présentation générale du projet

Ce chapitre a pour objectif de situer le projet dans son cadre général et d'exposer le contexte et les objectifs à atteindre durant ce stage de projet de fin d'études. Nous allons débuter par une présentation de l'organisme d'accueil avec un bref aperçu de son histoire pour arriver à ses multiples évolutions et succès. Par la suite, nous allons faire une description du projet ainsi que la méthodologie et le formalisme adopté.

1.1 Présentation de l'organisme d'accueil



1.1.1 L'INRA

Le premier institut de recherche agronomique en Europe avec 8 165 chercheurs, ingénieurs et techniciens permanents, au deuxième rang mondial pour ses publications en sciences agronomiques, l'INRA contribue à la production de connaissances et à l'innovation dans l'alimentation, l'agriculture et l'environnement.

L'Institut déploie sa stratégie de recherche en mobilisant ses 13 départements scientifiques et en s'appuyant sur un réseau unique en Europe, fort de plus de 200 unités de recherche et de 45 unités expérimentales implantées dans 17 centres en région.

L’ambition est, dans une perspective mondiale, de contribuer à assurer une alimentation saine et de qualité, une agriculture compétitive et durable ainsi qu’un environnement préservé et valorisé.

Le centre de Jouy-en-Josas :

Le centre de Jouy-en-Josas fait partie d’un des 17 sites INRA en France. Il a pour mission de produire des connaissances scientifiques et de contribuer à l’innovation. En accueillant plus de 150 doctorants, il participe également activement à la formation des jeunes par et pour la recherche. Les recherches sur la biologie animale, la microbiologie et les sciences de l’aliment constituent le cœur de son activité ; elles mobilisent mathématiciens, statisticiens et informaticiens pour appréhender la complexité des mécanismes biologiques étudiés et prédire le fonctionnement de systèmes vivants à différentes échelles et pour des finalités variées.

Les équipes du centre se mobilisent pour allier connaissances fondamentales et appliquées, explorant la diversité et la variabilité, de la molécule à l’individu dans son environnement.

Figure 1-1-Chiffres clés du centre INRA de Jouy-en-Josas

1.1.2 L’unité MetaGenoPolis (MGP)

MetaGenoPolis « MGP » est un projet de démonstrateur préindustriel financé par le programme français des Investissements d'Avenir dont l'objectif est de démontrer l'impact du microbiote intestinal humain sur la santé et la maladie en ayant recours à la métagénomique quantitative et fonctionnelle.

La caractérisation du microbiote intestinal humain permet d'étudier ses variations selon le type de population, le génotype, les maladies, l'âge, les habitudes alimentaires, la prise de médicament et l'environnement. Cette approche ouvre la perspective de le moduler afin d'améliorer la santé et le bien-être de chaque individu. MGP a regroupé différents domaines d'expertise et savoir-faire en 4 plateformes scientifiques et une plateforme d'éthique :

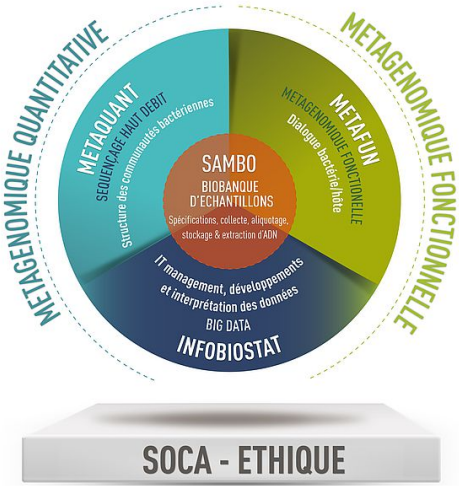


Figure 1-2- Les plateformes de MetaGenoPolis

1.1.3 La plateforme InfoBioStat (IBS)

C'est au sein de la plateforme InfoBioStat que j'ai effectué mon stage. Cette plateforme occupe une place centrale dans les activités de R&D de l'unité MGP dans le domaine de la métagénomique quantitative (analyse quantitative des gènes et espèces présents dans la flore bactérienne intestinale).

En collaboration avec les autres plateformes de l'unité – en particulier Sambo et MetaQuant – InfoBioStat développe une forte expertise dans l'analyse du contenu génétique du microbiote humain, dans la santé et la maladie, et la modélisation de sa diversité et sa dynamique pour :

- Stratifier les individus selon la composition de leur microbiote.
- Identifier des biomarqueurs spécifiques d'un état de santé donné.
- Modéliser in silico des tests de diagnostic et de pronostique.

Des outils et technologies informatiques de pointe sont nécessaires pour traiter et analyser les très larges volumes de données générées par les approches de métagénomique. Dans ce contexte, InfoBioStat met en œuvre une infrastructure optimisée pour :

- Le stockage, la sécurisation et l'accès aux larges volumes de données (Big Data) ;
- Le calcul informatique haute performance (HPC) ;
- La gestion de bases de données (relationnelles et/ou analytiques en temps réel) ;
- Le développement de logiciels dans un contexte d'industrialisation des procédés ;
- Le développement pour les analystes de la plateforme d'outils adaptés au Big Data utilisant les architectures GPU/HPC/Xeon Phi.

1.2 Présentation de la mission :

1.2.1 Problématique

La santé humaine est fortement liée au microbiote intestinal humain. L'exploration de ce dernier par la métagénomique permet d'optimiser la santé et le bien-être de chacun. La métagénomique au sein de MetaGenoPolis repose sur l'obtention et la manipulation de données de comptages à grande dimension. En effet, les statisticiens ont un rôle important à jouer dans l'exploration et l'analyse de ces données de comptages, bruitées, à inflation de zéros et recueillies à différentes échelles. Dans notre étude, le rôle du statisticien consiste principalement à proposer ou à trouver les méthodes multivariées non supervisées les plus adaptées afin d'extraire la structure existante entre les communautés microbiennes (échantillons), i.e., la façon dont les données sont organisées. On y retrouve l'analyse en composantes principales (ACP), et le positionnement multidimensionnel (MDS). Cependant, la situation idéale pour appliquer ces deux méthodes d'analyses multivariées, est une distribution multivariée gaussienne, ce qui n'est pas le cas avec nos matrices de comptages métagénomiques creuses. Pour tenter de pallier à ce problème, une nouvelle approche mathématique « l'ACP probabiliste de poisson log-normal », a été développée au sein de l'équipe MaIAGE de l'INRA de Jouy-en-Josas. Cependant, cette méthode n'a jamais été testée dans un contexte biomédical.

1.2.2 Objectifs du stage

L'objectif de ce stage est d'appliquer et de comparer plusieurs méthodes d'analyse multivariée non supervisées (d'ordination et probabilistes) sur des données de comptages d'espèces bactériennes et de modules fonctionnels chez des individus en bonne santé et des individus malades. Ces données ont la particularité d'être non gaussiennes, en grande dimension et sparses (plus de 80 % de zéros). Le but sera d'évaluer leur pertinence dans l'exploration de structures d'intérêt du microbiote intestinal humain.

Cette étude comporte plusieurs étapes :

- Comparaison entre les résultats de différentes mesures d'ordination, c'est-à-dire de réduction de dimension, appliquées à plusieurs indices de dissimilarité : Bray-Curtis, Jaccard, …

- Prise en main de l'ACP probabiliste basée sur un modèle poisson log-normal, avec l'intégration des covariables afin de tester leurs influences sur la structure des données.

- Interprétation des structures trouvées dans les jeux de données de comptages d'espèces bactériennes et des modules fonctionnels.

- Comparaison de la cohérence des signaux entre les différents jeux de données.

1.3 Organisation du rapport

Ce rapport est organisé en 4 chapitres :

- **Chapitre 2** : le deuxième chapitre est consacré à l'introduction du contexte biologique, la présentation de la méthodologie utilisée pour le traitement des données, et l'explication des principes généraux de fouille de données utilisés tout au long de notre travail. Le lecteur pourra ainsi se référer à ce chapitre dès lors qu'il en ressentira le besoin.

- **Chapitre 3** : ce chapitre est dédié aux différentes expérimentations effectuées pour l'exploration des données de comptages de deux cohortes humaines afin d'évaluer plusieurs méthodes de data mining dans un contexte métagénomique et de comparer leurs pertinences. A savoir les méthodes d'ordinations en utilisant différents indices de dissimilarités (Bray-Curtis, Jaccard…) et une méthode probabiliste (ACP probabiliste de poisson).

- **Chapitre 4 :** le dernier chapitre est consacré à la Data-Visualisation : Développement d'une application Rshiny pour diffuser et rendre disponible ces méthodes pour

l'équipe InfoBioStat. Ces visualisations ont l'avantage de rendre les résultats plus lisibles et de permettre d'étudier les données en fonctions de plusieurs covariables (les paramètres cliniques).

Chapitre II

Approche théorique et méthodes

2 Approche théorique et méthodes

2.1 Contexte biologique

2.1.1 Le microbiote humain

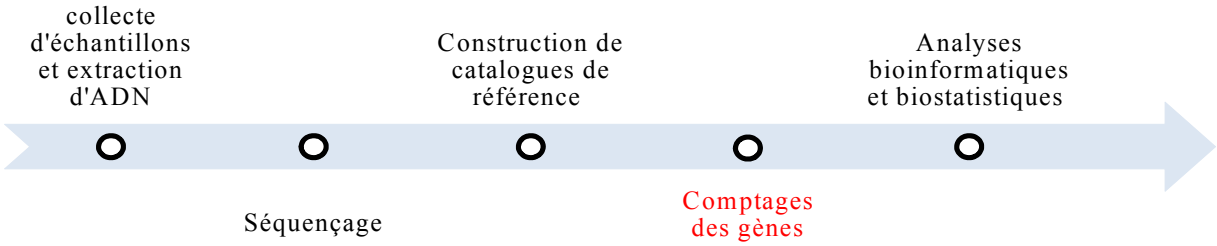
Dès la naissance, notre organisme est colonisé par des milliards de milliards de micro-organismes : bactéries, archées, virus, champignons. Ces différentes espèces s'installent sur la peau, dans la bouche, les intestins, les poumons… Tout au long de notre vie, ces colonies interagissent avec nos cellules. Pour les définir, les biologistes parlent de « microbiote ».

Le microbiote intestinal est le plus important d'entre eux. Son exploration offre de nouvelles pistes dans la compréhension de certaines maladies et pour de futurs traitements. Notre microbiote intestinal n'abrite pas moins de $\mathbf{10^{12}}$ **à** $\mathbf{10^{14}}$ micro-organismes, soit un potentiel fonctionnel bien plus important que celui des cellules qui constituent notre corps.

2.1.2 La métagénomique

La métagénomique est l'étude du contenu génétique des micro-organismes présents dans un échantillon issu d'un environnement complexe. La métagénomique permet donc l'analyse du microbiote intestinal. Alors que la génomique consiste à étudier un unique génome, la métagénomique étudie les génomes de plusieurs espèces différentes présentes dans un écosystème.

2.1.3 Étapes nécessaires à l'obtention de données métagénomiques :



L'étude du microbiote intestinal tel qu'il est effectué à MGP commence par la réception d'**échantillons de selles** issus de cohortes ou d'études cliniques. L'ADN contenu dans un échantillon est extrait et préparé pour **le séquençage**. Plusieurs millions de courtes séquences, ou « reads », sont obtenues et alignées, ou « mappées », sur un catalogue de référence composé de ***m* gènes** bactériens présents dans l'intestin humain. Le dernier catalogue de référence compte m = 10.4 millions de gènes. En comptant le nombre de reads alignés sur chaque gène, un vecteur d'abondance de **longueur *m*** est obtenu pour chaque individu. Des analyses biostatistiques se font par la suite, afin d'explorer ces données et d'identifier les espèces bactériennes associées à un phénotype clinique (par exemple : individu sain ou malade). La figure 2-1 (d'après Ehrlich SD.C R Biol. [2016]) et la figure 2-2 décrivent en détail ce processus de production des données de comptages des gènes.

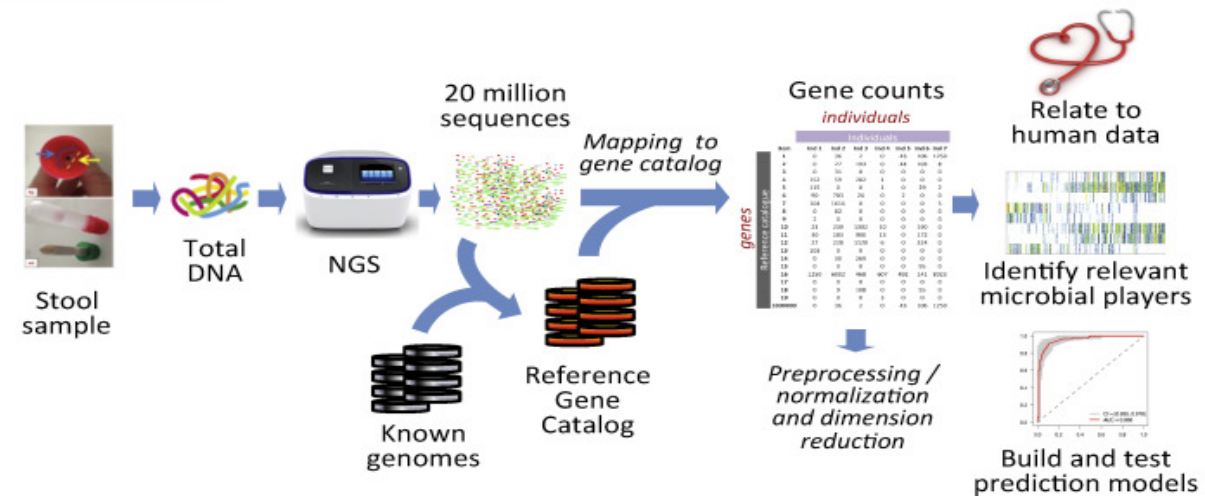


Figure 2-3- Une expérience de séquençage d'ADN au sein de MGP

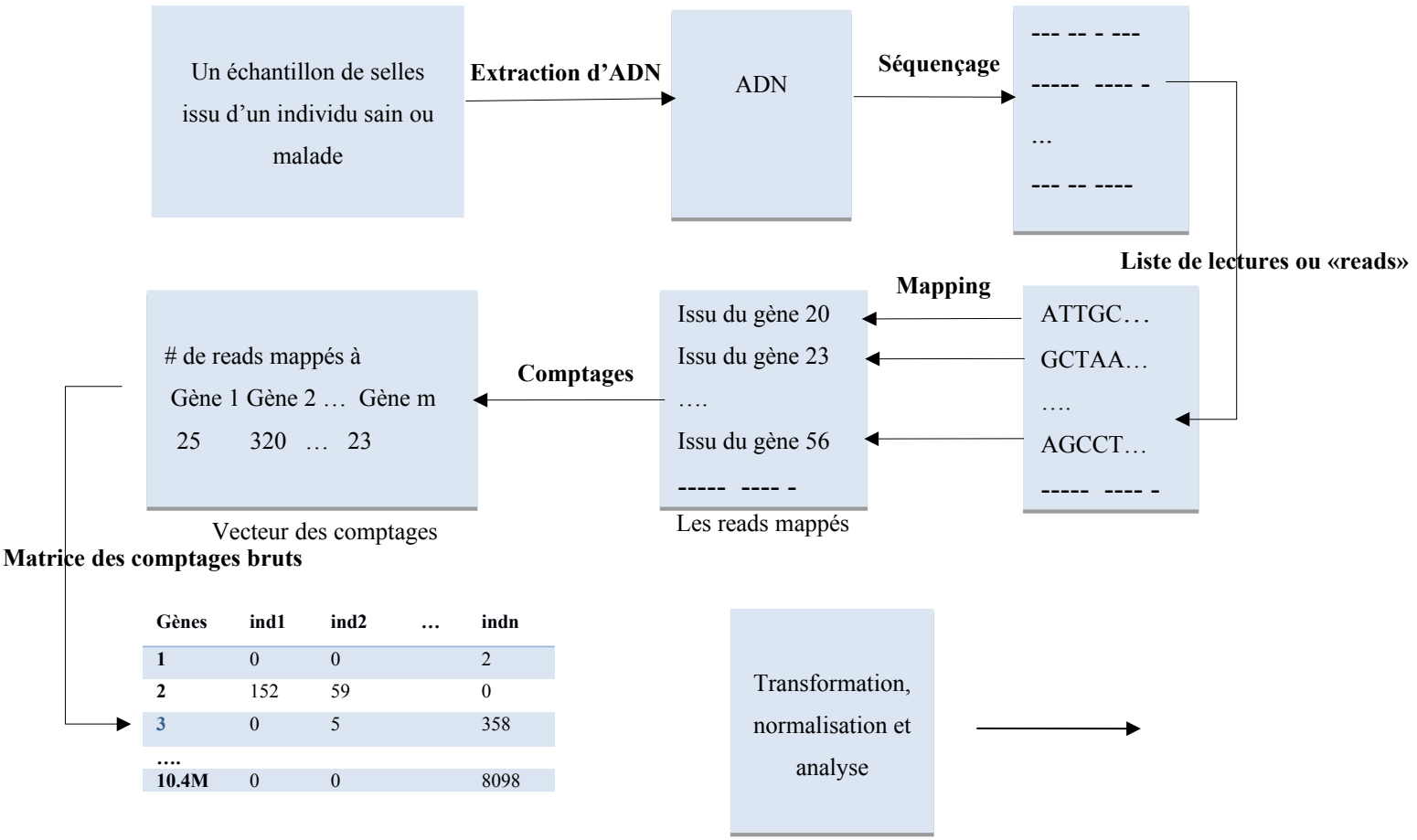


Figure 2-4- Processus de création de la matrice de comptages des gènes

2.1.4 Caractéristiques de la matrice de comptages bruts des gènes :

- Creuse : contient beaucoup de comptages nuls (sur les jeux de données utilisés durant mon stage la proportion de zéros dans la table de comptages des MGS variait de 70% à 90%).
- De grande dimension : plusieurs millions de variables (comptage des gènes) pour seulement quelques centaines d'observations (individus). Exemple :10.4M de gènes pour 237 patients.
- Dépendance entre les variables, liée aux interactions écologiques entre espèces (par exemple, le mutualisme entre deux espèces induit une corrélation positive entre leurs

abondances) ou à l'appartenance à une entité biologique commune (deux gènes issus d'un même génome ont des abondances similaires).

- Grande dispersion des comptages.

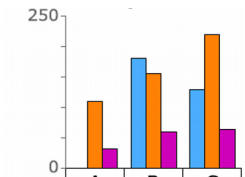
2.1.5 Prétraitement de données de comptages des gènes :

Raréfaction (downsizing) :

La table de comptages des gènes contient le nombre de lectures attribuées à chaque gène pour chaque échantillon. Le nombre total de lectures issues du séquençage peut varier d'un échantillon à l'autre (notamment à cause de la variabilité technique des séquenceurs). Comparer directement deux échantillons avec un nombre total de lectures différentes risque de révéler des différences de proportions des gènes qui sont dues à cette variation de profondeur de séquençage, et non à une différence biologique réelle. Par exemple dans la Figure2-3-a le gène2 semble moins abondant dans l'échantillon A que dans l'échantillon C. Cette différence est uniquement due au fait que l'échantillon A a un nombre total de lectures bien plus faible (8000) que l'échantillon C (20000). Raréfier les comptages entre échantillons est une façon de les rendre comparables.

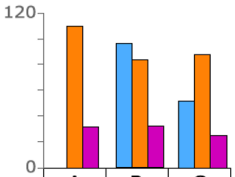
Ce type de prétraitement est utilisé au sein de MGP : la raréfaction, ou downsizing en anglais, consiste à sous-échantillonner les comptages pour que tous les échantillons aient le même comptage total (Figure 2-3).

a) **Table initiale d'observations**



	A	B	C
Gène1	0	180	130
Gène2	110	156	220
Gène3	32	60	64
⋮	⋮	⋮	⋮
Total	8k	15k	20k

b) **Table après raréfaction**



	A	B	C
Gène1	0	96	52
Gène2	110	83,2	88
Gène3	32	32	25,6
⋮	⋮	⋮	⋮
Total	8k	8k	8k

Figure 2-5-Exemple de raréfaction de la table de comptages des gènes effectué au sein de MGP

Normalisation :

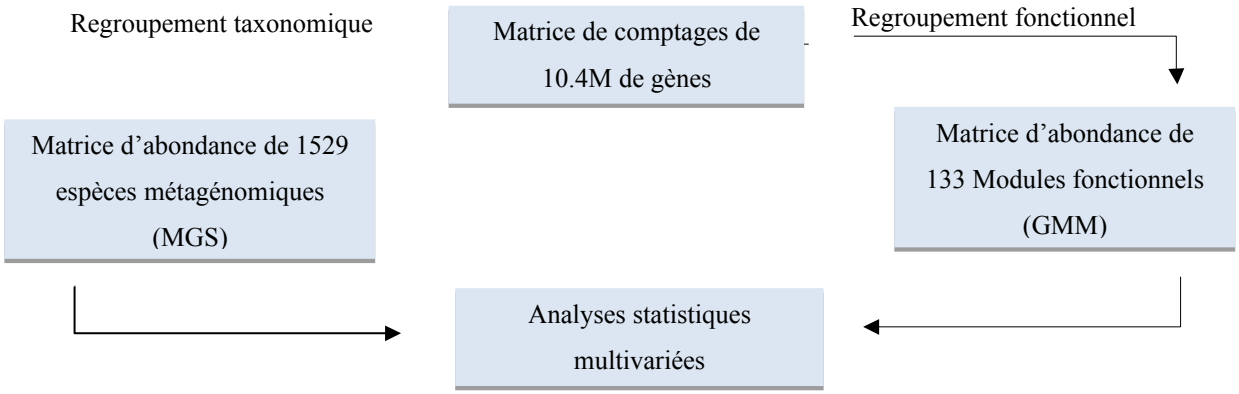
Après le downsizing, la matrice de comptage des gènes est normalisée pour que les abondances soient comparables entre les gènes pour un même individu. Les comptages de chaque gène sont divisés par la taille du gène, afin de prendre en compte le fait que plus un gène est grand, plus il y aura de lectures qui lui seront attribuées. Ensuite, le vecteur de comptage est divisé par la somme des comptages pour obtenir des fréquences. La matrice issue de cette normalisation est une matrice de fréquence des gènes où la somme de l'abondance des gènes vaut 1 pour chaque individu.

Dans notre étude, nous allons appliquer sur la matrice de gènes downsizée d'autres méthodes de normalisation que nous allons décrire dans le chapitre suivant. Enfin, plutôt que de travailler directement sur la matrice de 10.4M de gènes, nous allons travailler sur version résumée de cette matrice, où les gènes ont été regroupés en groupes taxonomiques ou fonctionnels en utilisant des à priori biologiques. La partie suivante de ce chapitre décrit la méthodologie de construction de ces données résumées.

2.2 Méthodologie pour le traitement des données de comptages des gènes

Plusieurs traitements bio-informatiques sont réalisés sur les données de comptages des gènes. Ces traitements consistent à faire des regroupements ou « clustering » taxonomiques et fonctionnels afin d'obtenir de nouvelles matrices avec un nombre de variables nettement plus faible.

Import
Preprocess
Inference
Validation
Graphics



2.2.1 Regroupement taxonomique

Le catalogue de référence de 10.4 millions de gènes du microbiote intestinal humain contient 1529 clusters correspondant à des espèces bactériennes métagénomiques ou « MGS ». La méthode utilisée pour obtenir ces clusters, consiste à regrouper les gènes dont l'abondance covarie au sein d'un grand nombre d'échantillons métagénomiques. Parmi ces groupes de gènes coabondants, seuls les plus grands (≥ 500 gènes) sont considérés comme des MGS. Les autres correspondent à des petits éléments micro-organiques (plasmides, phages, etc.) ou artéfactuels qui ne sont pas pris en compte dans la suite des analyses. Dans chaque MGS, l'ensemble des gènes « essentiels », ou gènes marqueurs, est défini comme celui des 50 gènes les plus corrélés entre eux.

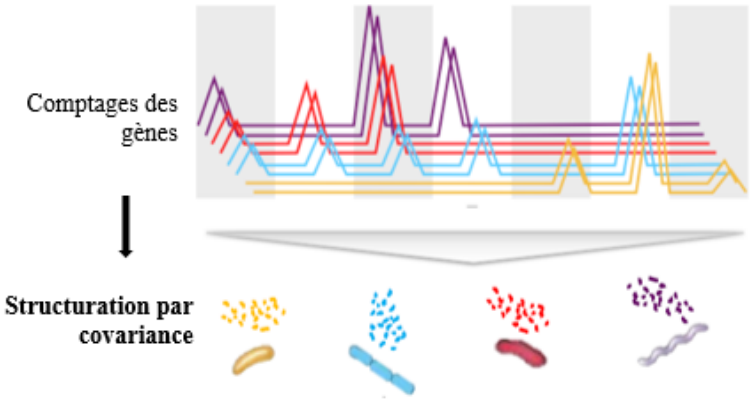


Figure 2-6- Regroupement taxonomique

Au sein d'un échantillon, l'abondance des MGS est calculée en faisant la moyenne des comptages des 50 gènes marqueurs qui la composent. Si moins de 10 % de ces gènes sont présents, le comptage de la MGS est mis à 0. On obtient ainsi une nouvelle matrice de comptages avec les 1529 MGS en ligne et les échantillons en colonne. Cette matrice peut contenir un très grand nombre de zéros (entre 70% et 90), à interpréter comme l'absence d'une espèce bactérienne.

2.2.2 Regroupement fonctionnel

Lorsque cela est possible, les gènes du catalogue de référence sont assignés à des groupes fonctionnels (KO) selon leur similarité avec les gènes qui composent ces groupes. L'abondance des KO est calculée comme la somme des abondances des gènes associés à ce KO.

Les KO intervenant dans une même réaction métabolique sont regroupés au sein d'un GMM (Gut Metabolic Module) L'abondance d'un GMM est calculée à partir de l'abondance des différents KO qui le compose. Dans le microbiote intestinal humain, 133 GMM ont été décrits. La nouvelle matrice de comptages est composée des 133 GMM en ligne et des échantillons en colonne.

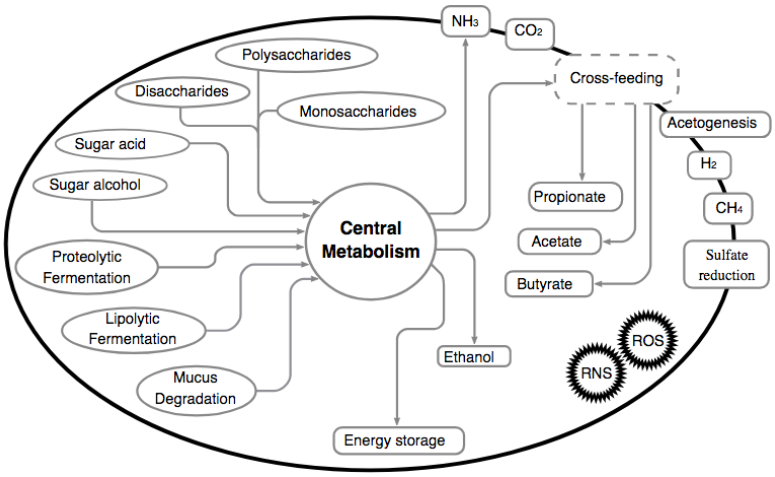


Figure 2-7- Regroupement fonctionnel

2.3 Théorie statistique

Les méthodes multivariées permettent d'analyser un profil métagénomique dans sa globalité. En effet, si on veut comparer des profils métagénomiques dans leur ensemble, une analyse multivariée apportera une réponse objective. Au-delà de la comparaison globale des profils, les analyses multivariées permettent aussi de montrer les associations entre les espèces, utiles pour envisager les hypothèses sur la mécanique de mise en place de certaines maladies. Enfin les analyses multivariées permettent de visualiser la répartition et la variabilité des échantillons (variabilité intra-groupe versus inter-groupes). Elles peuvent s'appliquer sur tous les types de données (binaire, abondance ou pourcentage) et sur des études supervisées ou non supervisées.

Dans notre étude nous avons utilisé principalement des méthodes d'analyses multivariées non supervisées afin de trouver les structures latentes dans les données d'abondance du microbiote intestinal humain. Nous avons également complété nos analyses par une approche supervisée qui va chercher à décrire et prédire l'appartenance de nos échantillons « individus » à des groupes prédéfinis (par exemple le groupe des individus sains et le groupe des individus malades). Commençons par expliquer la différence entre apprentissage supervisée et non supervisée.

2.3.1 Méthodes supervisées vs non supervisées

En apprentissage supervisé, on cherche à produire des règles à partir d'une base de données d'apprentissage contenant des variables d'entrée et des valeurs cibles. On dit que les variables explicatives sont annotées par leurs sorties. La cible est la réponse que l'algorithme devrait produire à partir de l'entrée. Dans notre exemple, les données d'entrée sont les comptages d'espèces bactériennes (MGS) ou des modules fonctionnels (GMM) pour différents individus et la cible est un caractère phénotypique, par exemple le statut (sain/malade).

Soit X = ($X^{(1)}$, $X^{(2)}$, … $X^{(n)}$) des variables dites explicatives ou prédictives observées sur n individus et Y une variable à expliquer. Dans le cas d'apprentissage supervisé, on cherche à trouver une fonction f susceptible de prédire Y à partir de X :

$$Y = f(X) + \varepsilon$$

Où ε est un bruit ou une erreur de mesure.

Si la variable Y à expliquer est qualitative, on parle de classification tandis que si Y est quantitative on parle de régression. La sortie Y dans notre exemple est qualitative (statut sain ou malade des échantillons de notre cohorte).

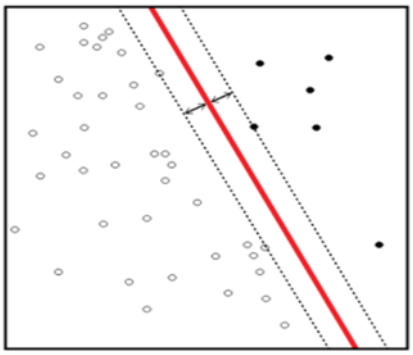


Figure 2-8- Modèle de classification

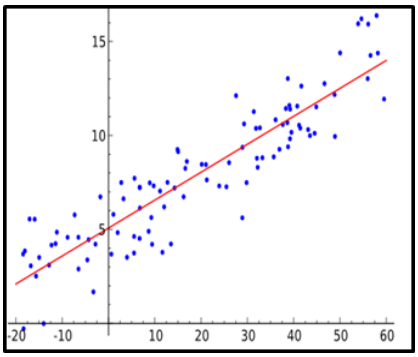


Figure 2-9- Modèle de régression

Dans le cas où les données d'entrée ne sont pas annotées, on parle alors d'apprentissage non supervisé. La phase d'apprentissage s'applique dans ce cas pour trouver **les similarités et les**

différences au sein de ces données, et à regrouper celles-ci en classes homogènes partageant des caractéristiques communes. C'est un problème de recherche de classe (aussi appelé clustering).

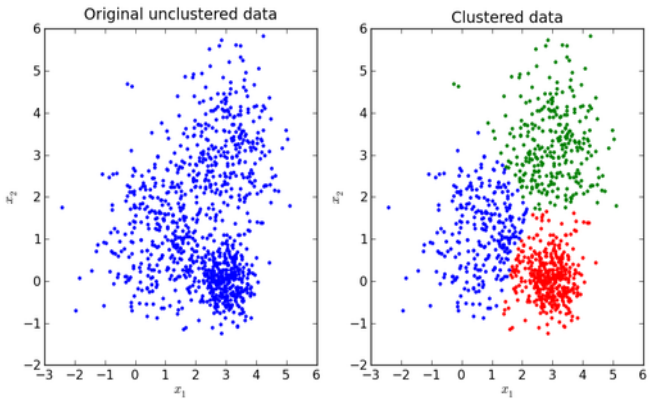


Figure 2-10- Modèle de clustering

Voici une brève description des méthodes non supervisées que nous avons utilisées dans notre étude :

2.3.2 Méthodes non supervisées

L'apprentissage non supervisé permet d'extraire des connaissances à partir d'échantillons non étiquetés. Il existe de nombreuses méthodes et aucune n'est systématiquement meilleure que les autres. En effet, le choix de la méthode se fait en fonction du problème et du type de données.

La première méthode d'analyse multivariée descriptive non supervisée utilisée pour le traitement de nos données d'abondance est la méthode d'ordination.

2.3.2.1 Ordination

"**Gauch (1982):** Ordination primarily endeavors to represent sample and species relationships as faithfully as possible in a low dimensional space"

L'ordination est le nom donné par les écologues des communautés aux différentes méthodes de réduction de dimension utilisées dans cette discipline. Il s'agit de représenter la

(dis)similarité entre les échantillons, construite à partir des valeurs de plusieurs variables (les abondances) qui leurs sont associées, de sorte que les échantillons similaires soient représentés proches les uns des autres et les échantillons dissemblables soient plus éloignés les uns des autres.

En effet, les méthodes d'ordination donnent un aperçu des relations de similarité entre échantillons en termes d'abondance des espèces, la distance relative d'une paire d'échantillons reflétant leur dissemblance relative. Clark et Green (1988) les définissent comme une analyse d'une matrice de données de n échantillons par p espèces, grâce à laquelle on obtient une nouvelle série de variables qui prédisent de façon optimale la structure des relations entre les variables originales p. Les méthodes diffèrent selon l'approche mathématique utilisé pour calculer la similitude/(dis)similitude entre les espèces et les échantillons, et selon la façon dont l'algorithme d'ordination détermine les nouveaux axes représentant les nouvelles variables. Actuellement, il existe plusieurs techniques d'ordination. Voici une petite description des méthodes d'ordination utilisées dans notre projet : l'analyse en composantes principales (ACP) et le positionnement multidimensionnel (MDS).

2.3.2.1.1 L'analyse en composantes principales (ACP)

L'analyse en composantes principales est la méthode d'ordination la plus fondamentale. Elle est à la base des méthodes d'ordination parce que les autres méthodes tenteront de produire le même type de graphiques d'ordination que l'ACP en analysant des données qui ne sont pas tout à fait appropriées ou qui sont inappropriées pour l'ACP.

C'est une technique qui permet de trouver **des espaces de dimensions plus petites** dans lesquels il est possible d'observer au mieux les individus. Sa démarche essentielle consiste à transformer les variables quantitatives initiales, plus ou moins corrélées entre elles, en des variables quantitatives, non corrélées, qui s'expriment comme combinaisons linéaires des variables initiales et sont appelées composantes principales.

Globalement l'ACP consiste à rechercher la direction suivant laquelle le nuage de points des observations s'étire au maximum. A cette direction correspond la première composante principale. La seconde composante principale est déterminée de telle sorte à être indépendante de la première, et donc lui être perpendiculaire, à la première et à capturer de nouveau la

direction d'étirement maximum. Ces deux composantes forment le premier plan principal. Cette opération est réitérée séquentiellement de manière à trouver toutes les composantes principales expliquant le maximum de variance.

L'ACP prend uniquement en compte les dépendances linéaires entre les variables et ne peut donc pas fournir une projection pertinente pour une distribution non-gaussiennes de points. En effet, la situation idéale pour appliquer cette méthode est **une distribution multivariée normale**, ce qui n'est pas le cas avec nos matrices de comptages métagénomiques creuses. Ceci nous a amené à penser à une transformation des données qui peut réduire l'asymétrie des distributions d'abondances d'espèces, et ainsi être appropriée pour l'étude de la diversité entre les différentes observations. Pour les données métagénomiques de comptages à inflation de zéros, la transformation la plus utilisée est : y' = log (y + 1) parce que la borne inférieure des valeurs est 0. Log (0) = –Inf, mais log (0 + 1) = 0. C'est la transformation que nous avons utilisée dans notre étude avant d'appliquer l'ACP.

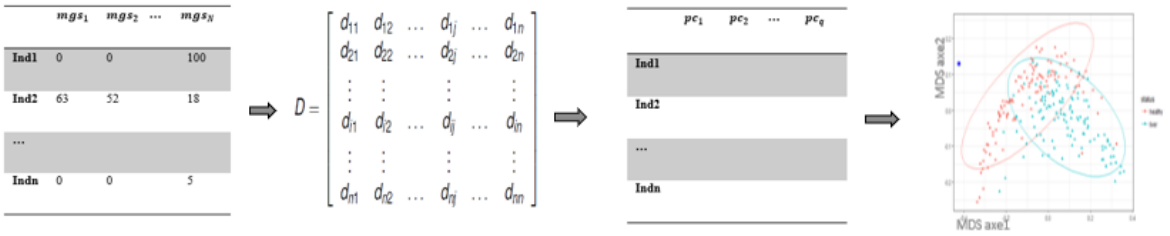
2.3.2.1.2 Le positionnement multidimensionnel (MDS) :

La deuxième méthode d'ordination utilisée dans notre étude est le positionnement multidimensionnel (Messick et Abelson [1956]), ou MDS pour « Multi Dimensional Scaling » en anglais, qui permet de construire une représentation en faible dimension des points de l'espace. Son objectif est de construire, à partir d'une matrice de distances ou de mesures de similarité calculées sur chaque paire de points, une représentation euclidienne des individus dans un espace de dimension réduite qui préserve "au mieux" ces distances.

Supposons que nous ayons un ensemble de données X = ($X^{(1)}$, $X^{(2)}$, … $X^{(n)}$) composé de n observations où chaque observation $x_i = (x_{i1}, x_{i2},.., x_{i})$ est composée de N caractéristiques

Soit D une matrice symétrique de taille n×n où chaque élément d_{ij} représente la distance entre x_i et x_j et $d_{ii}= 0$. L'idée du MDS est de trouver une configuration de points y_i, i= 1...n dans un espace euclidien de dimension plus réduite qui conserverait les distances entre les points initiaux x_i. Autrement dit, il cherche les points y_i, dans un espace de dimension q < N tels que :

$$d(y_i, y_i) \approx d_{ij} = d(x_i, x_j)$$



Attention, la distance d entre les y dans l'espace de départ et celle entre les x dans l'espace d'arrivée peuvent être de nature très différente. En particulier, la distance dans l'espace de départ peut n'être qu'une (dis)similarité. Voici quelques propriétés et définitions élémentaires à propos de la notion de **(dis)similarité** :

Une similarité ou dissimilarité est toute application à valeurs numériques qui permet de mesurer le lien entre les individus d'un même ensemble ou entre les variables. Pour une similarité le lien est d'autant plus fort que sa valeur est grande.

Un indice de dissimilarité d vérifie : (k, l, m sont trois individus)

1. La dissimilarité d'un individu avec lui-même est nulle : d (k, k) = 0 ;
2. La dissimilarité entre deux individus différents est positive : d (k, l) ≥ 0 ; La dissimilarité est symétrique : d (k, l) =d (l, k) ;

Une distance vérifie en plus :

1. La distance entre deux individus différents est strictement positive : d (k, l) = 0 ⇒ k = l ;
2. L'inégalité triangulaire : d (k, m) ≤ d (k, l) + d (l, m). De nombreux indices de dissimilarité ne vérifient pas cette dernière propriété.

Nous présentons ensuite les mesures de dissimilarité utilisées dans notre projet.

Mesure basée sur les abondances des espèces :

Ce type de distance permet d'évaluer la dissimilarité entre deux échantillons, en termes d'abondance d'espèces présentes dans chacun de ces échantillons.

- **Dissimilarité de Bray-Curtis :**

La dissimilarité de Bray-Curtis prend le nom de ses auteurs, J. Roger Bray et John T. Curtis et est définie comme suit :

$$d_{BC} = \sum_s ¿ n_s^1 - n_s^2 \vee / \sum_s \left(n_s^1 + n_s^2 \right)$$

Où, n_s^1 et n_s^2 sont les comptages de l'espèce « s » dans le premier et le deuxième échantillon.

L'indice de dissimilarité de Bray-Curtis est compris entre 0 (les deux échantillons ont la même composition en espèces avec les mêmes abondances) et 1 (les échantillons sont totalement dissemblables).

Mesure basée sur les présences/absences des espèces :

- **Dissimilarité de Jaccard :**

$$d_{BC} = \sum_s 1_{\{n_s^1 > 0, n_s^2 = 0\}} + 1_{\{n_s^2 > 0, n_s^1 = 0\}} / \sum_s 1_{\{n_s^1 + n_s^2 > 0\}}$$

L'indice de dissimilarité de Jaccard est aussi compris entre 0 (les deux échantillons ont la même composition en espèces mais pas nécessairement avec les mêmes abondances) et 1 (aucune espèce n'est partagée entre les deux échantillons).

Différence entre dissimilarité de Bray et dissimilarité de Jaccard :

Le graphe ci-dessous explique la différence entre les deux mesures de dissimilarités Bray et Jaccard.

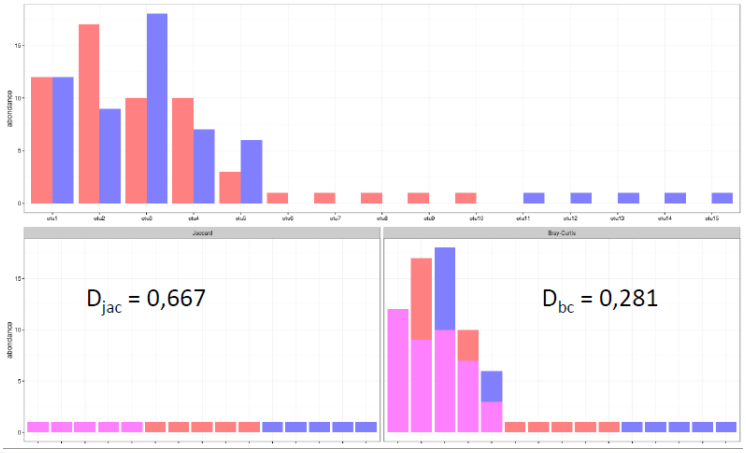


Figure 2-11- Explication de la différence entre la dissimilarité de Bray et Jaccard

C'est la représentation des abondances de plusieurs OTU_s (MGS_s) au sein de deux échantillons (échantillon 1 en rouge et échantillon 2 en bleu). Le violet représente les abondances des MGS partagées entre les deux échantillons.

A travers ce graphe, on peut remarquer que la dissimilarité de Jaccard donne le même poids à toutes les MGS_s, alors que pour la dissimilarité de Bray-Curtis, les poids sont proportionnels à l'abondance de l'espèce.

Le choix de la dissimilarité doit être basé sur la question des caractéristiques des données que nous souhaitons décrire et sur la manière dont la mesure résume le signal dans nos données. En effet, que ce soit appropriée ou non, dépend largement de ce que nous considérons comme des informations pertinentes dans l'ensemble de données.

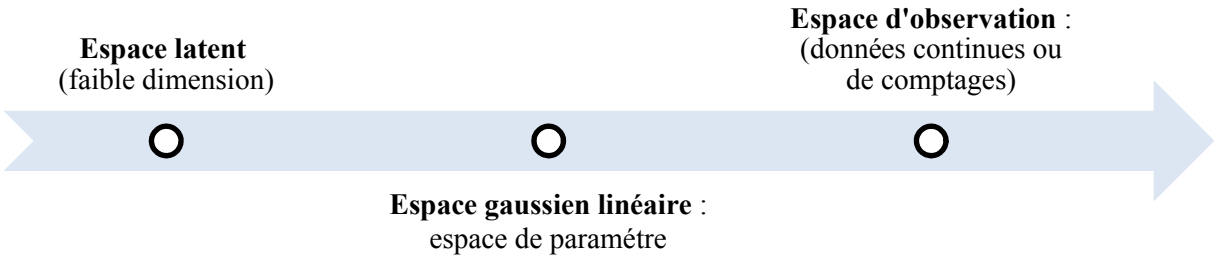
2.3.2.2 ACP probabiliste (pPCA)

La meilleure situation pour appliquer les méthodes d'analyse multivariée (ordination ou clustering) est celle d'une distribution multivariée gaussienne. Or, notre étude s'appuie sur l'exploration des données non normales « données de comptages contenant énormément de zéros ».

Une nouvelle méthode probabiliste a été développée par un groupe de mathématiciens, biostatisticiens au sein de l'unité MAIaGE de l'INRA de Jouy-en-Josas dont le but d'explorer des données de comptages à inflation de zéros. C'est la méthode d'ACP probabiliste pour données de type poisson log-normales.

L'un des objectifs de notre étude est de tester l'efficacité de cette méthode probabiliste dans un **contexte métagénomique**. Nous allons appliquer cette méthode sur nos matrices de comptages d'espèces bactériennes ou de modules fonctionnels.

Dans cette partie, nous allons introduire cette méthode et expliquer ces étapes. Commençons tout d'abord par introduire l'ACP probabiliste gaussienne :



L'ACP probabiliste gaussienne :

C'est la version probabiliste de l'analyse en composantes principales ou **pPCA** [Minka, 2000 ; Mohamed et al., 2009 ; Tipping et Bishop, 1999]. Contrairement à l'ACP standard, l'ACP probabiliste suppose que les données suivent une loi normale : dans un premier temps, on échantillonne un vecteur gaussien Wi dans un espace latent de faible dimension q, avec coordonnées indépendantes. Dans un deuxième temps, on applique une transformation linéaire B à ce vecteur pour l'injecter dans un espace de grande dimension. Enfin, on rajoute un bruit gaussien isotrope à cette transformation linéaire pour obtenir les observations. Formellement, l'ACP probabiliste gaussienne peut s'écrire de la façon suivante :

$$Y_i = \mu + W_i B + \varepsilon_i, \ \varepsilon_i \ N(\cdot)$$

Où, μ : la moyenne

B : représente la matrice de dépendance entre les variables latentes et les observations.

$W_i \quad N(0_q, I_q)$: les variables latentes

Le modèle peut aussi se réécrire indépendamment sous la forme : $Y_i \quad N(\mu, BB^T + \sigma^2 I_p)$

Une formulation différente, qui reprend le cadre hiérarchique décrit plus haut, est la suivante :

- Espace latent R^q W_i i.i.d. $W_i \quad N$ ¿)
- Espace de paramètre R^p $Z_i = \mu + W_i B$ $Z_i \quad N$ ¿)
- Espace d'observation R^p $Y_{ij} \vee Z_{ij}$ indep. $Y_{ij} \vee Z_{ij} \quad N(Z_{ij}, \sigma^2)$

Cette formulation hiérarchique se prête bien à des extensions. Plutôt que d'utiliser le même vecteur de moyenne μ pour tout le monde, on peut le faire dépendre de covariables Xi et/ou d'un offset O_i et écrire ainsi $\mu_i = O_i + X_i \Theta + W_i B$. La prise en compte de covariables permet de corriger les dépendances induites par ces covariables et de trouver d'autres types de structure. Voici également une représentation graphique (extrait de l'article : Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic Poisson PCA.), illustrant ces 3 étapes :

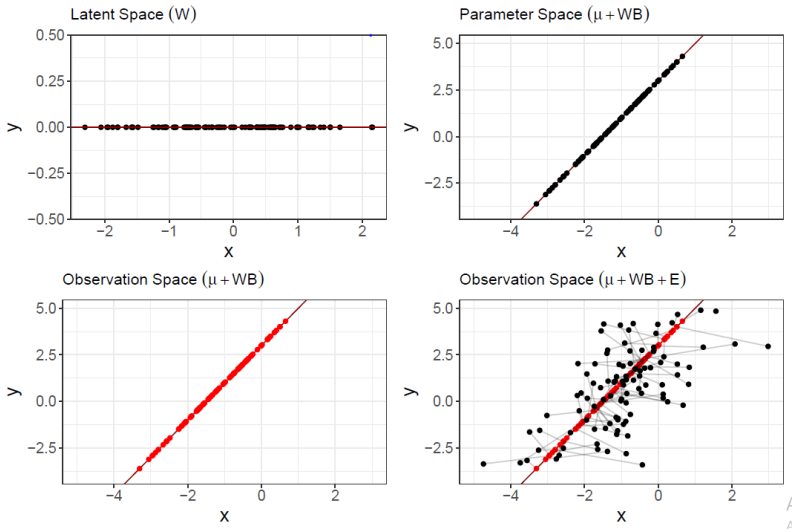


Figure 2-12- ACP probabiliste gaussienne

L'ACP probabiliste de poisson log-normal :

Cette méthode est une généralisation de l'ACP probabiliste gaussienne à des données de comptage, dans laquelle on remplace la dernière étape de bruit gaussien par une étape de bruit poissonnien. Le modèle est appelé Poisson Log-Normal (PLN) parce que la distribution marginale de chacun des comptages suit cette distribution. De la même façon que dans la version gaussienne, on peut intégrer une matrice d'offsets et une matrice de covariables au modèle. Dans ce contexte précis, l'offset joue le rôle de facteur de normalisation.

Soit Y_{ij}le comptage de l'espèces j dans un échantillon i

Y_{ij} ne suit pas une distribution normale comme dans le cas précédant, mais plutôt une distribution de Poisson de paramètre Z_{ij}. Y_{ij} P ¿ ¿**),** $Z_i = \mu_{ij} + W_i B_j$où $\mu_{ij} = o_{ij} + x_i B_j$

o_{ij}: représente la matrice des offsets : profondeur de séquençage de MGS j dans l'échantillon i * tailles des MGS j. Cette matrice permet de normaliser les données hétérogènes de comptages directement au niveau du modèle sans passer par du downsizing (sur les reads) et sans diviser par la longueur des MGS. En effet, ces deux mesures : profondeur de séquençage et longueur de MGS sont proportionnelles aux comptages.

x_i: représente les covariables cliniques caractérisant chaque échantillon i.

W_i : représente les variables latentes.

De même que précédemment, on peut écrire le modèle sous forme hiérarchique : Pour plus de détails, le lecteur pourra consulter l'article : (J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic Poisson PCA.).

- Espace latent R^q W_i i.i.d. W_i N ¿)
- Espace de paramètre R^p $Z_i = \mu_i + W_i$B Z_i N ¿)
- Espace d'observation N^p Y_{ij} { Z ¿$_{ij}$ Indep. Y_{ij} { Z ¿$_{ij}$ P ¿ ¿)

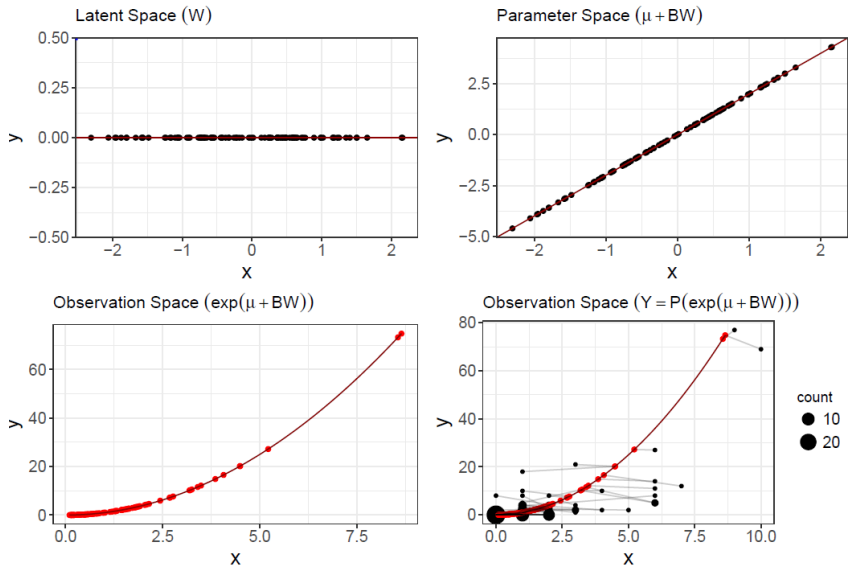


Figure 2-13- ACP probabiliste de poisson log-normal

2.3.3 Méthodes supervisées

2.3.3.1 Analyse linéaire discriminante (LDA)

Chercher à mesurer ce qui sépare des groupes connus est ce qu'on appelle discriminer. L'analyse linéaire discriminante, appelée aussi analyse discriminante linéaire de Fisher, est une méthode de réduction du nombre de dimensions proposée par Fisher en 1936. Cette méthode s'applique lorsque les classes des individus sont connues. L'idée a été de créer une méthode pour choisir entre les différentes combinaisons linéaires des variables celles qui maximisent l'homogénéité de chaque classe et maximise leur séparation. En d'autres termes, cette méthode consiste à chercher un espace vectoriel de faible dimension qui maximise la variance inter-classe.

Cette méthode essaye explicitement de modéliser la différence entre les classes de données contrairement à l'ACP, qui n'a pas d'information de classe et ne peut pas en tenir compte.

PCA **LDA**

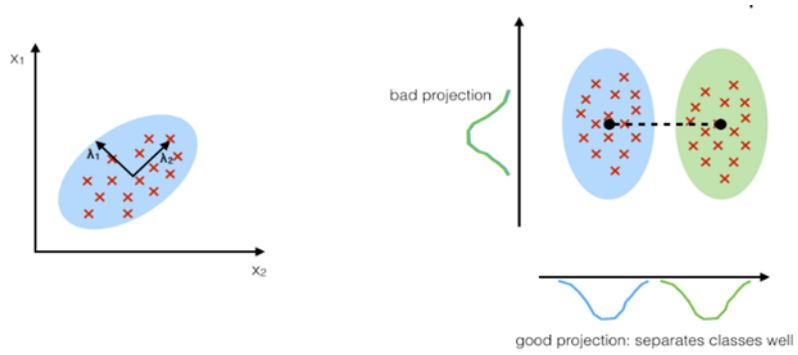


Figure 2-14- Différence entre l'ACP et l'analyse discriminante linéaire

2.4 Principaux packages utilisés sous Rstudio

Les deux librairies sous R phyloseq [McMurdie and Holmes, 2013] et PLNmodels [https://github.com/jchiquet/PLNmodels] regroupent un grand nombre des méthodes d'analyses multivariées qu'on a utilisées dans notre étude.

2.4.1 Phyloseq

Le package R phyloseq est disponible gratuitement sur le Web à partir de GitHub et de Bioconductor. Phyloseq fournit un ensemble d'outils pour faciliter l'importation, le stockage, l'analyse et la visualisation des données du microbiome.

Structure des données phyloseq :

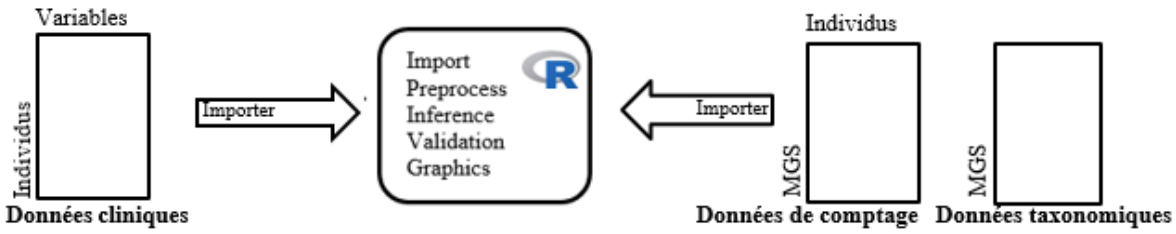


Figure 2-15- Structure des donnés dans le package phyloseq sous R

2.4.2 PLNmodels

PLNmodels implémente une série de méthodes basées sur les modèles Poisson log-normaux pour l'analyse des données de comptage multivariées. En particulier, les observations sont tirées d'une **distribution de Poisson Log-Normale.**

Conclusion

Tout au long de ce chapitre, nous avons décrit la méthodologie utilisée pour explorer les données de comptages métagénomiques. Dans le prochain chapitre, nous appliquerons ces méthodes sur différents jeux de données de comptages et commenterons leurs pertinences.

Chapitre III

Résultats et discussions

3 Réalisation : Résultats et discussions

3.1 Analyse descriptive des données cliniques

Nous avons commencé à travailler sur une cohorte humaine de 237 individus chinois [Qin Nature 2014] sains et malades afin d'explorer le microbiote intestinal lié à la cirrhose du foie.

La cirrhose est une maladie hépatique relativement grave dans la mesure où elle peut endommager le foie de façon irréversible. La principale cause de la cirrhose du foie est la consommation excessive d'alcool pendant une longue période.

Une analyse descriptive des données cliniques aide à construire une base de connaissance sur notre cohorte. En effet, pour chaque patient, plusieurs caractéristiques cliniques ont été relevées à savoir : le statut (sain/malade), l'âge, le sexe, l'indice de masse corporelle, les indices de gravité de la maladie, les tests du bon fonctionnement des reins, etc. (Tableau 3-1)

SAMPLE	STATUS	GENDER	AGE	BMI	ALCOHOL_RELATED	CREATININE
H1	healthy	female	40	20.03	N	53
H10	healthy	female	40	20.2	N	65
H11	healthy	female	38	19.95	N	57
H12	healthy	male	40	22.55	N	72
H13	healthy	female	56	21.48	N	47

Tableau 3-1- Extrait du tableau des données cliniques du projet «cirrhose du foie»

Le but de cette partie est de mesurer et visualiser la répartition des différentes variables (sexe, âge, …) entre les différents status, en utilisant des violin-plots et des barplots, pour vérifier s'il existe des différences systématiques entre patients sains et malades au niveau des variables cliniques

Cette étude peut nous aider par la suite dans la partie exploration et modélisation afin de tester l'influence de certains paramètres cliniques sur la variation du microbiote.

La variable « sexe » :

Nous avons visualisé à l'aide d'un barplot la répartition du sexe « male / female » au sein de nos groupes d'intérêt « sains/malades » : (figure 3-1)

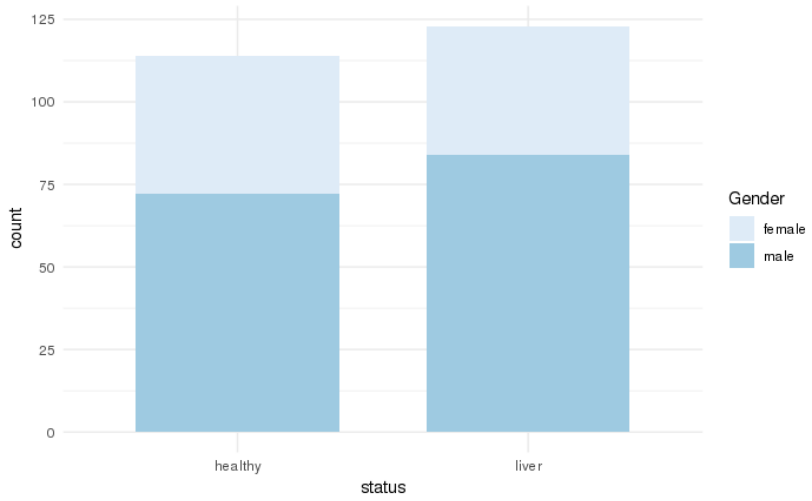


Figure 3-16- Répartition des groupes sains/malades selon le sexe

Les patients malades de la cirrhose du foie « liver » sont légèrement plus nombreux que les sains « healthy ». Mais que la balance des genres a l'air équilibrée entre les groupes

La variable « âge » :

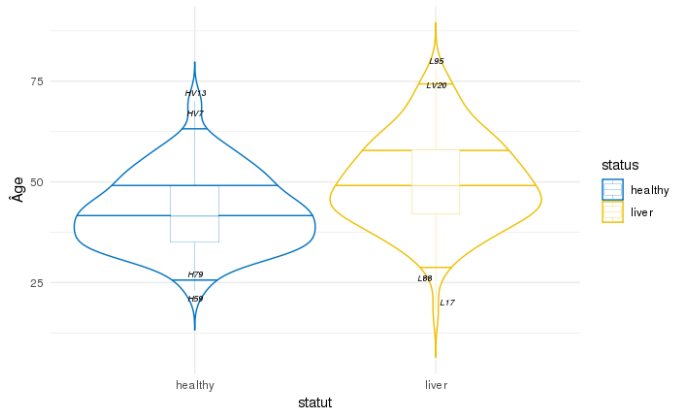


Figure 3-17- Répartition des groupes sains/malades selon l'âge

On constate que la moyenne d'âge des individus malades est plus élevée que celle des individus sains.

La variable « richesse » :

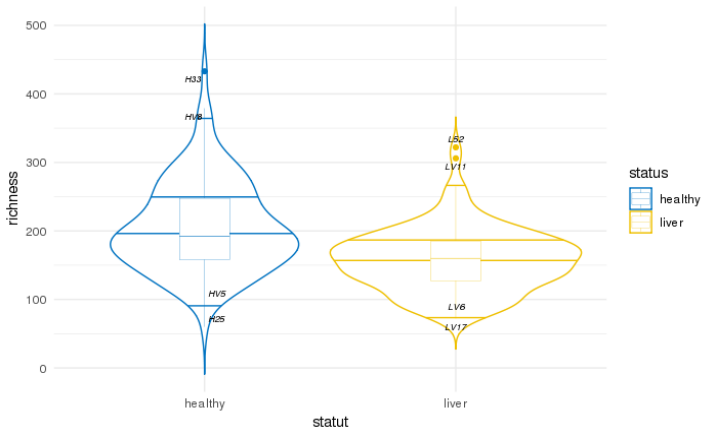


Figure 3-18- Répartition des groupes sains/malades selon la richesse

La variable « richesse » représente le nombre d'espèces bactériennes distinctes présentes chez chaque individu. Ce graphe montre que les individus sains sont plus riches en espèces bactériennes que ceux atteints de la cirrhose du foie.

La variable « créatinine » :

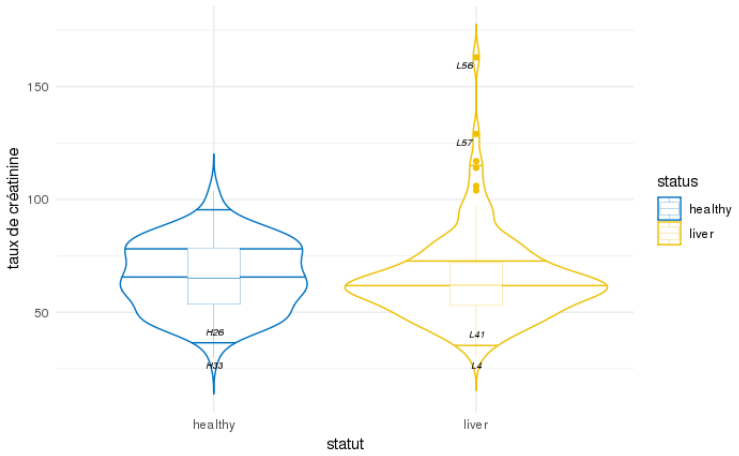


Figure 3-19- Répartition des groupes sains/malades selon le taux de la créatinine

Le taux de la créatinine dans le sang est un indicateur précieux permettant d'évaluer l'activité rénale chez un individu. La créatinine est issue de la dégradation de la créatine, qui est synthétisée par le foie et stockée dans les muscles où elle joue un rôle important dans la production d'énergie. En effet, Une insuffisance rénale peut-être induite par la sévérité d'une cirrhose du foie.

D'après le graphe ci-dessus, nous remarquons que la moyenne du taux de créatinine des individus sains est plus élevée que celle des individus malades.

3.2 Analyses multivariées des données de comptages

Les méthodes statistiques exploratoires multivariées sont une famille de méthodes permettant de faciliter la visualisation des données et de révéler leur structure sous-jacente.

Dans notre étude, deux types de méthodes d'analyse multivariée sont appliquées dans un cadre non supervisé : les méthodes d'ordination, telle que l'analyse en composantes principales (ACP), et le positionnement multidimensionnel (MDS), et les méthodes probabilistes (pPCA). Pour évaluer ces méthodes dans un contexte métagénomique, deux cohortes de deux maladies différentes (la cirrhose du foie et la spondylarthrite ankylosante) ont été intégrées et analysées à l'aide de différents outils de visualisation sous Rstudio. Ce document présente seulement les résultats de l'étude faite sur la première cohorte (la cirrhose du foie).

Dans cette partie, nous considérons un ensemble d'échantillons $(x_i)_{i=1,\dots,n}$, pour lequel on dispose d'un tableau de comptages représenté par une matrice, notée Y. Cette matrice est de taille n × p, où n correspond au nombre d'échantillons et p au nombre d'espèces observées dans l'ensemble des échantillons. Les entrées de Y sont des valeurs de comptages, notées y_{ij}, correspondant au nombre d'organismes de l'espèce j observés dans l'échantillon x_i.

Les méthodes d'analyses multivariées que nous allons appliquer s'appuient sur une matrice de comptages des MGS dans un premier temps et une matrice des modules fonctionnels dans un second temps.

3.2.1 Données de comptages d'espèces bactériennes (MGS)

Pour la cohorte de la cirrhose du foie, notre matrice de comptages des MGS est de taille : n= 237 et p=1529. Il y a une proportion très élevée de zéros dans cette matrice (88 %). (Tableau 3-2)

		MGS				
		CAG00001_1_hs_10.4	CAG00001_hs_9.9	CAG00002_1_hs_10.4	CAG00002_2_hs_10.4	….
INDIVIDUS	H25	0	0	0	0	….
	H34	0	0	0	0	….
	H66	0	0	0	0	….
	L82	0	0	14	0	….
	H13	0	0	0	0	….

Tableau 3-2-Tableau de comptages extrait des données du projet «cirrhose du foie»

Les lignes correspondent à cinq identifiants d'échantillons et les colonnes à quatre identifiants de MGSs différentes.

3.2.1.1 Méthodes non supervisées

Dans cette partie du rapport, nous souhaitons répondre à la question suivante : Parmi les 237 échantillons sains ou atteins de la cirrhose du foie, existe-t-il des échantillons (individus) qui se ressemblent en termes de composition en espèces bactériennes **?**

Commençons par l'application des méthodes d'ordination.

3.2.1.1.1 Méthodes d'ordination :

L'analyse en composantes principales :

Pour avoir un premier aperçu de nos données de comptages non gaussiennes, nous avons appliqué une ACP.

Voici les graphes représentant la projection des individus sur le premier plan factoriel :

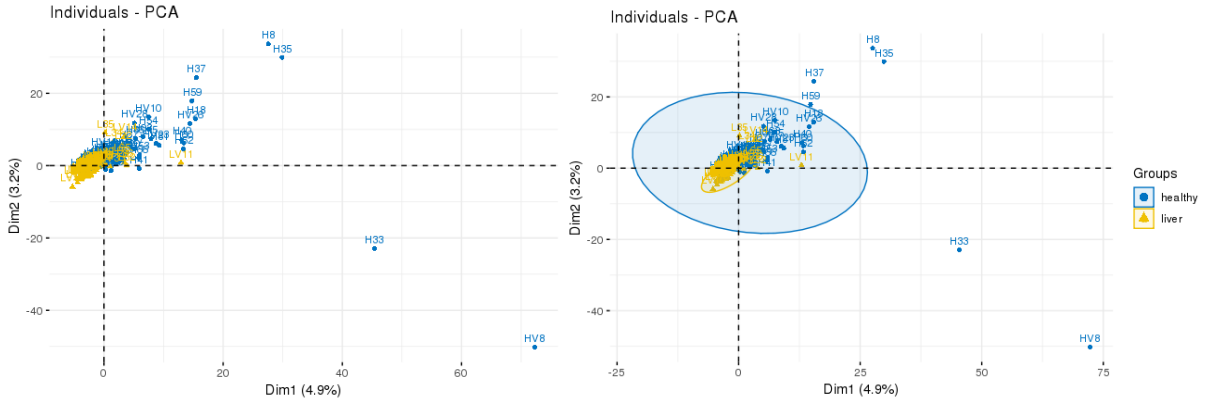


Figure 3-20- Graphes des individus après l'application d'une ACP

Les résultats de l'ACP pour la cohorte de la cirrhose du foie sont pauvres, c'est-à-dire cette méthode n'est pas pertinente en termes de recherche de structure du microbiote intestinal humain, parce qu'une faible variabilité des individus est mise en évidence. Après la coloration des points selon la variable « status » nous pouvons constater que bien qu'il n'y a pas de séparation marquée entre les deux groupes « healthy » et « liver », les deux nuages de points ne sont pas superposés. Ce résultat est bien prévu, vu la nature des données traitées. En effet, l'ACP prend uniquement en compte les dépendances linéaires entre les variables et ne peut donc pas fournir une projection pertinente pour une distribution non-linéaire de points.

Dans le but de réduire cette asymétrie des distributions de nos données d'abondances d'espèces, nous avons appliqué une transformation logarithmique ($y' = \log(y+1)$) avant d'appliquer une ACP. Voici la projection des individus sur le premier plan factoriel :

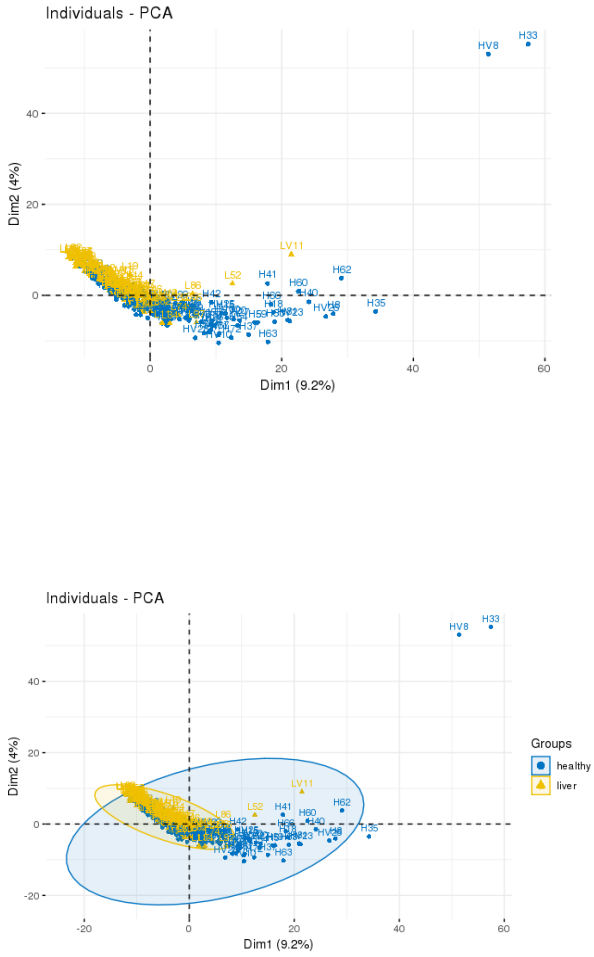


Figure 3-21- Graphes des individus après une transformation logarithmique

Nous remarquons une petite amélioration au niveau de la dispersion des individus sains et malades. Le pourcentage de variance expliqué sur le premier plan factoriel est passé de 8.1% à 13.2%. La transformation logarithmique nous a pas permis d'avoir une séparation des deux groupes « sain/ malade ».

Il est intéressant de remarquer qu'il y a deux outliers (H33, HV8) qui étirent la représentation. En effet, en revenant à la partie analyse descriptive des données cliniques, nous pouvons constater que « H33 » et « HV8 » sont les deux individus les plus riches en espèces bactériennes dans notre cohorte (figure 3-3). Nous avons également remarqué que l'individu

« H33 » possède le taux de créatinine le plus faible dans le groupe des sains (figure 3-4), qui est égale à 30 µmol/l, alors que le taux normal doit se situer entre 50 et 100 µmol/l.

Dans le but d'améliorer le résultat de l'ACP, nous avons pensé : à supprimer de ces deux outliers. Voici alors les graphes d'individus dans le premier plan factoriel :

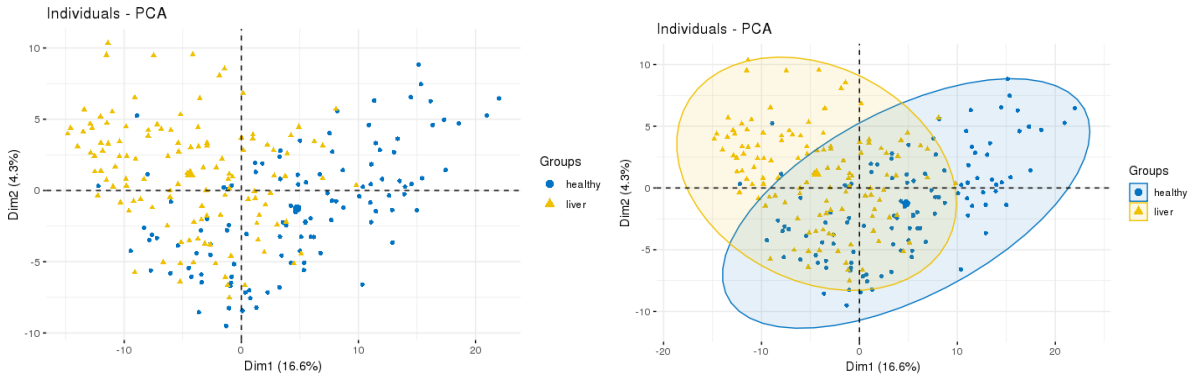


Figure 3-22- Graphes de l'ACP après suppression des outliers

La séparation des deux groupes apparaît plus clairement sur cette représentation, même si la distinction entre les deux groupes n'est pas parfaite. Le pourcentage de variance expliqué sur le premier plan factoriel est passé de 13.2 % à 21 %.

La séparation des deux groupes d'intérêt : sains/malades sur le deuxième et le troisième plan factoriel est moins franche que celle observée sur le premier plan. Nous avons également constaté une baisse du pourcentage de variance expliqué sur ces deux plans factoriels.

Voici par exemple le graphe de la projection des individus sur le troisième plan factoriel :

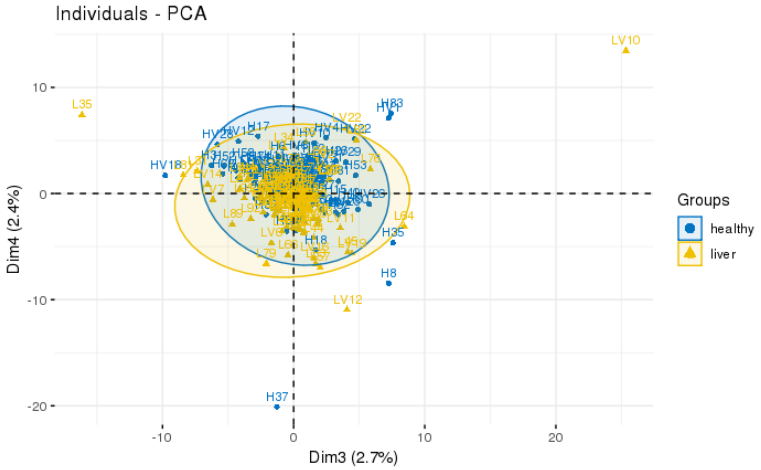


Figure 3-23- Projection des individus sur le troisième plan factoriel

Le positionnement multidimensionnel (MDS) :

À partir d'une matrice de dissimilarité, nous avons essayé de former des groupes sur la base d'une classification hiérarchique ascendante afin de voir si ces groupes collent avec le statut (sain/malade). En effet, à l'aide de la fonction « ordination-plot » du package « phyloseq » de R, Nous avons appliqué la méthode MDS en utilisant les deux dissimilarités : Bray-Curtis et Jaccard, et comparé leurs efficacités dans la recherche de ces groupes.

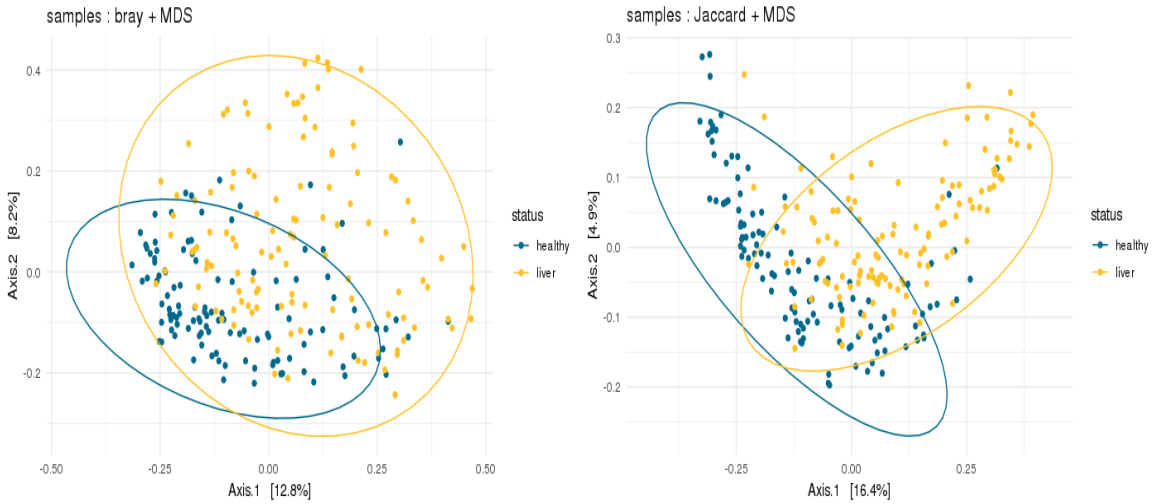


Figure 3-24- Graphes d'individus pour la méthode d'ordination MDS

Les deux graphes d'ordination MDS montrent qu'il n'y a pas de séparation claire et marquée entre les deux groupes sains/malades. Cependant, nous pouvons constater que la dissimilarité de Jaccard (graphe de droite) est légèrement meilleure que celle de Bray-Curtis en termes de séparation. Il est cependant intéressant de noter qu'avec la représentation de Bray-Curtis, la variabilité - en termes de composition microbienne - des individus malades est plus grande que celle des individus sains. La différence entre les deux mesures de dissimilarité utilisées met en évidence que la séparation entre sains et malades est plus forte en termes de « répertoire d'espèces » qu'en terme de « composition des communautés ».

Dans le but d'améliorer nos résultats trouvés en utilisant les méthodes d'ordination, nous avons appliqué la méthode d'ACP probabiliste de poisson log-normal spécifiquement conçue pour des données de comptages.

3.2.1.1.2 L'ACP probabiliste de poisson log-Normal :

Dans cette partie, nous avons normalisé notre matrice de comptages en utilisant une matrice d'offset (pour prendre en compte la profondeur de séquençage et la taille des gènes, cf. chapitre 2).

Nous avons appliqué cette méthode en utilisant la fonction « PLNPCA » du package « PLNmodels » de R.

Figure 3-25- PPCA de poisson log-normal sur les données de comptages des MGS

Les deux graphes ci-dessus montrent que l'application de la méthode d'ACP probabiliste de poisson log-normal sur la cohorte de la cirrhose du foie n'a pas amélioré les résultats trouvés avec les méthodes d'ordination. En effet, il n'y a pas une séparation claire entre les deux groupes d'intérêt (sains/malades). Il est à noter que la densité le long du premier axe montre une petite différence entre les individus sains et les individus malades : les individus malades (points jaunes) sont plus dispersés que les individus sains (points bleus). La variabilité, en termes de composition d'espèces, est donc plus importante chez les malades.

La matrice de comptage est extrêmement creuse car un grand nombre de MGS n'est présent que chez un très faible nombre d'individu. Nous avons réduit le nombre de variables en ne conservant que les MGS les plus abondantes (celles présentes dans au moins de 10% des échantillons). Les nouvelles dimensions de notre jeu de données sont : 237 échantillons et 392 MGS.

PPCA de poisson log normal sur la matrice de comptage contenant les MGS les plus abondantes :

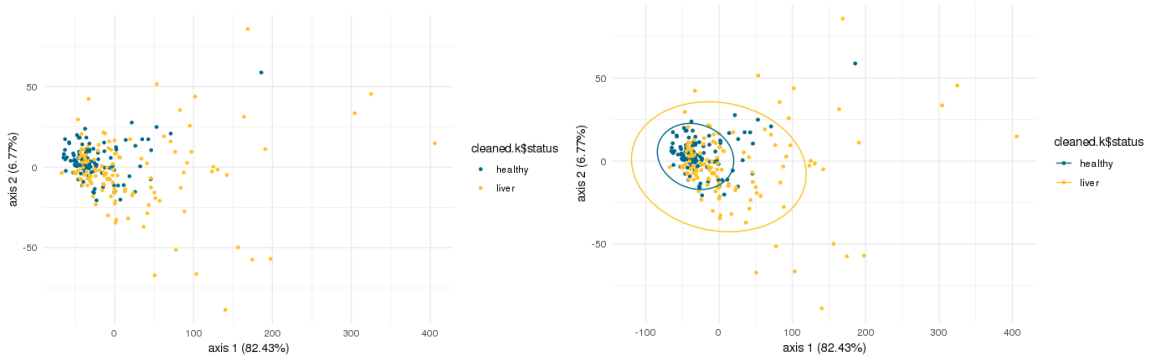


Figure 3-26- PPCA de poisson log-normal sur les données de comptages contenant les MGS les plus abondants

Nous remarquons une petite amélioration au niveau de la dispersion des individus sains et malades.

L'étude de l'influence des covariables « sexe et âge » sur la variabilité des données de comptages des MGS les plus abondantes :

L'ACP probabiliste permettant d'ajouter des covariables afin de de corriger les dépendances induites par ces covariables et de trouver d'autres types de structure, nous avons choisi de prendre en compte le sexe et l'âge des individus qui sont de potentiel facteurs de confusion au vu de leurs distributions différentes entre les sains et les malades.

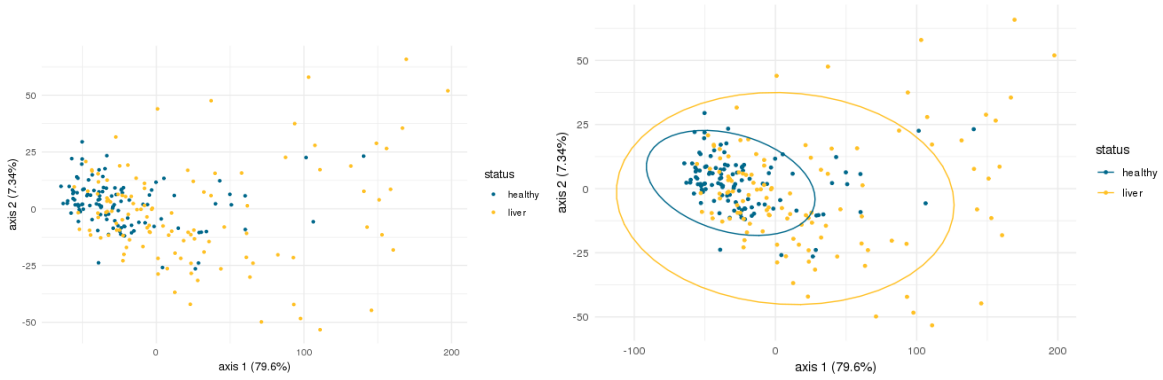


Figure 3-27- Graphes d'individus après l'ajout des covariables

L'ajout des covariables « Âge et sexe » a permis d'améliorer légèrement la dispersion des données, mais pas de séparer totalement les deux groupes sain/malade.

Le faible effet de ces deux covariables choisies, peut être aussi constaté à partir des deux graphes (violon-plots) représentant la répartition de la variable « statut » selon l'âge et le sexe dans la partie d'analyse descriptive.

La méthode de l'ACP probabiliste du poisson log-normal n'a pas réussit à trouver les structures liées au statut : sain/malade dans la cohorte de la cirrhose du foie. Cependant, il est connu [Qin Nature 2014] que la richesse en MGS est un facteur structurant du microbiote intestinal. Nous avons construit à partir de la matrice de comptages Y, une nouvelle variable continue representant la richesse en espèces bactériennes.

Etude de la richesse :

Les points sont ici colorés selon « la richesse » en MGS. Voici le graphe de l'ACP probabiliste du poisson log-normal :

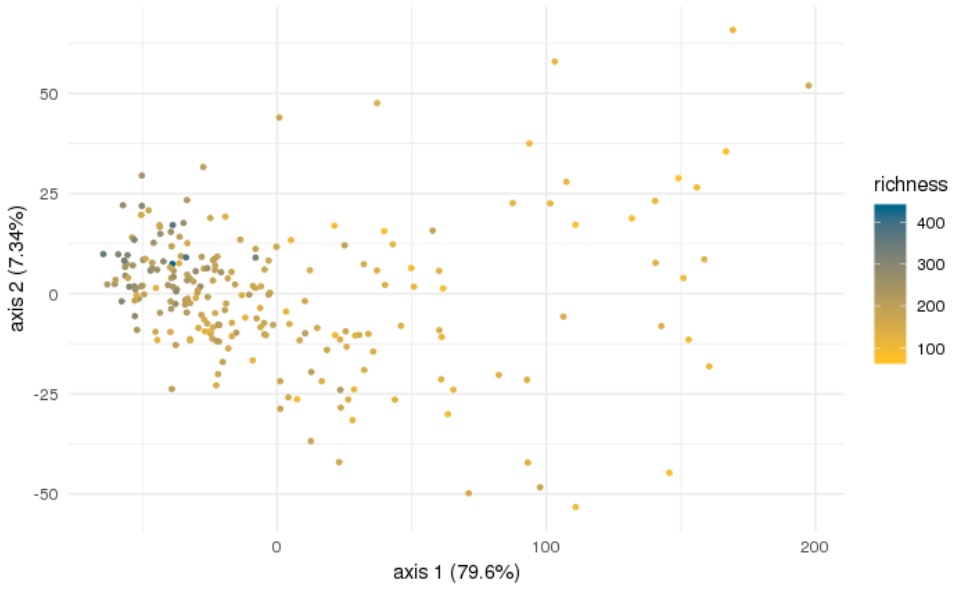


Figure 3-28- Graphe des individus coloré selon la richesse en espèces bactériennes

Nous remarquons que les individus les plus riches sont à gauche (bleu foncé) et les individus les moins riches sont à droite (jaune). On peut alors dire que cette méthode nous a permis de trouver une séparation entre les deux communautés « riche/ pauvre » en espèces bactériennes dans la cohorte de la cirrhose du foie.

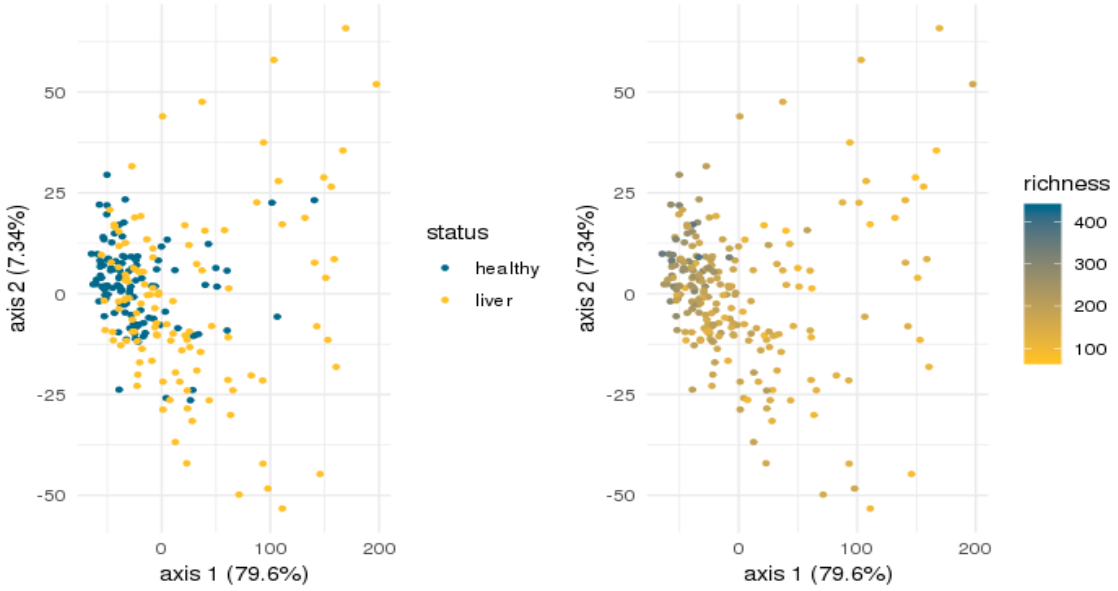


Figure 3-29- Comparaison entre le graphe d'individus coloré selon la variable «status» et celui coloré selon «la richesse»

À travers ce graphe, on peut constater qu'il existe une relation entre la richsse et le statut clinique. Cette relation avait déjà été mise en évidence dans la partie analyse descriptive des données cliniques. En effet, on peut remarqué que les individus sains sont plus riches en especes bacteriennes que les individus atteints de la cirrhose du foie.

Afin de confirmer l'effet richesse dans la structuration de nos données de comptages, nous avons pensé à la prise en compte de la richesse comme covariable dans notre modèle. Ceci est representé dans le graphe ci-dessous :

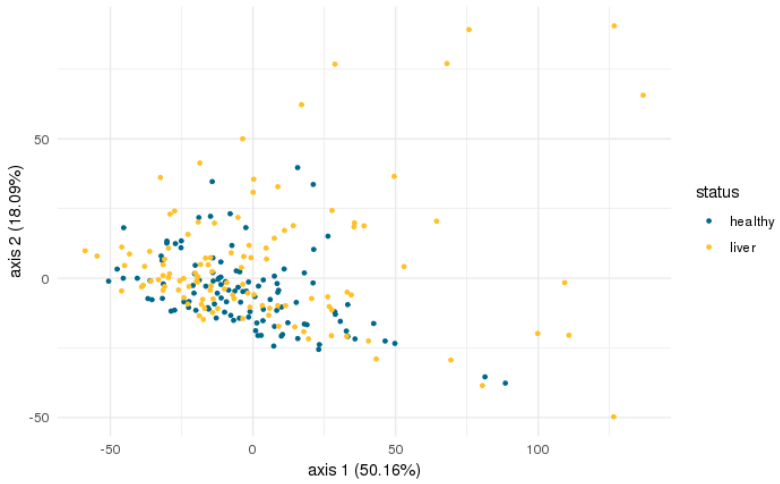


Figure 3-30- PPCA en prenant en compte la richesse comme covariable

La différence entre les deux communautés « sains/malades » ne se voit plus au niveau global. En effet, la variabilité de nos données ne colle pas avec le statut. Nous pouvons alors confirmer l'effet de la richesse dans la structuration de nos données de comptages des MGS .

3.2.1.2 Méthode supervisée

Dans le cas non supervisé, nous avons recherché les structures des communautés bactériennes du microbiote intestinal pour voir si elles correspondaient au statut clinique. Dans le cas supervisé, nous avons un échantillon d'apprentissage contenant des variables d'entrée (les comptages des 1529 MGS pour les 237 individus) et la cible (le statut : sain/malade).

Nous avons appliqué la méthode supervisée « analyse linéaire discriminante » à l'aide de la fonction « PLNLDA » du package PLNmodels de R.

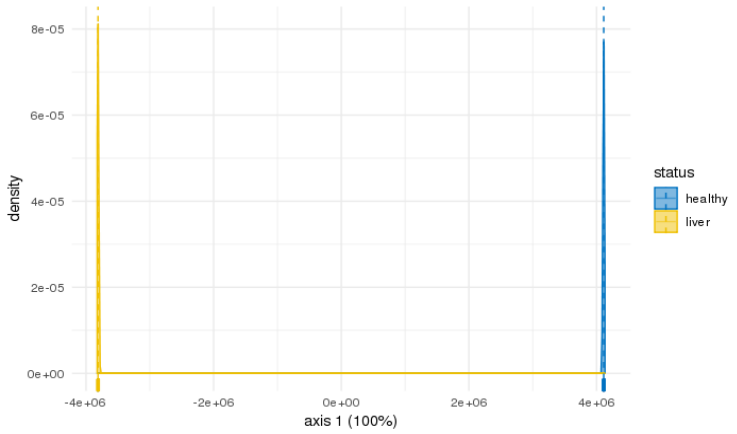


Figure 3-31- Positions des échantillons le long du premier axe de l'analyse discriminante et densité de ces positions, obtenue via une analyse PLNLDA de la matrice de comptage de tous les MGS

Nous constatons une séparation totale entre les deux groupes d'individus « sains et malades ». Ce résultat est attendu. En effet, avec un nombre de variable assez élevé (1529 >> nombre d'observation), Nous pouvons facilement trouver des combinaisons linéaires d'abondance des espèces pour prédire le statut clinique.

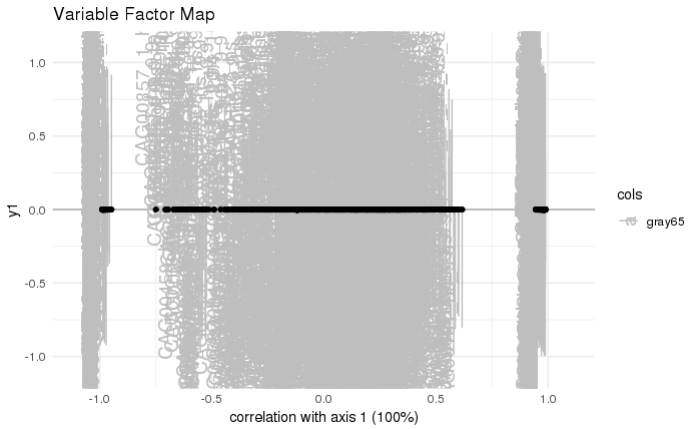


Figure 3-32- Contribution de chacune des MGS au premier axe de la LDA avant suppression des espèces les moins occurrentes.

On peut noter que les nombreuses MGS positionnées en –1 ou +1 sont souvent spécifique d'un unique échantillon.

Pour pouvoir mieux comprendre et interpréter ce résultat dans un contexte biologique, nous avons pensé à appliquer cette même méthode sur la matrice contenant uniquement les MGS les plus occurrentes construite précédemment.

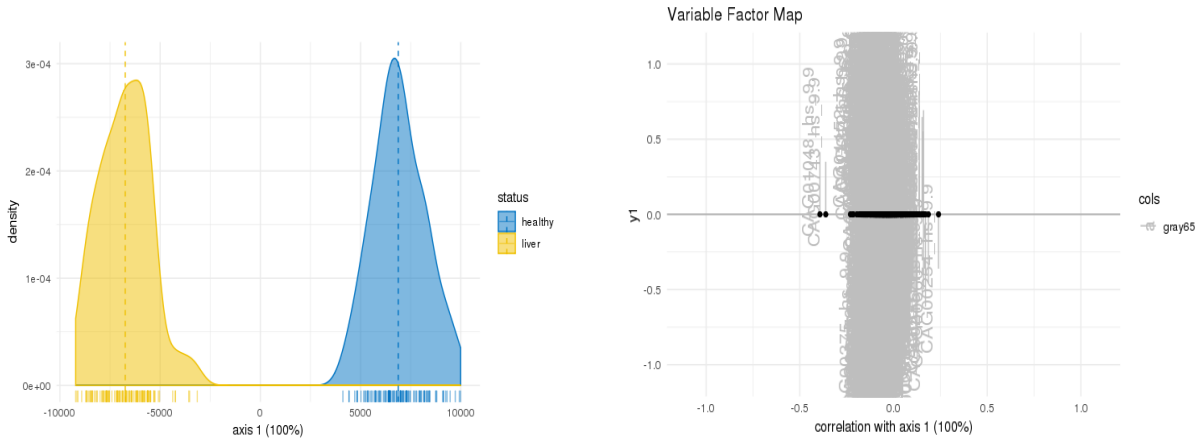


Figure 3-33- Idem Figs. 3-16 et 3-17 mais en se restreignant aux MGS les plus occurrentes

Nous constatons ici une séparation plus faible de deux groupes sains/malades. Ce résultat peut être expliqué par le fait que les espèces rares sont les responsables à la séparation totale de deux communités sains/ malades ou parce qu'elles contribuaient fortement au surapprentissage.

3.2.1.3 Conclusion sur les méthodes non supervisées utilisées :

Les résultats d'application de la méthode d'analyse multivariée probabiliste « pPCA » sur les données de comptages des MGS n'est pas pertinente en termes de recherche des structures qui collent avec le phénotype « sain/malade ». Mais, nous avons réussi à trouver les structures cohérentes avec la variable « richesse » en espèces bactériennes. En effet, nous avons trouvé que les sains sont plus riches en MGS que individus atteints d'une cirrhose du foie. L'application de la méthode d'ordination « MDS », nous a aussi aidé à conclure que la mesure de dissimilarité de Jaccard est meilleure que la dissimilarité de Bray et que la distance

euclidienne (ACP) dans la séparation entre sains et malades en termes de « répertoire d'espèces ».

Dans la prochaine partie, nous allons appliquer les mêmes méthodes sur des données de comptages des fonctions des gènes. Notre but est de tester la robustesse de ces trois méthodes d'analyses multivariées non supervisées dans un contexte métagénomique fonctionnel : l'exploration des interactions entre les modules fonctionnels.

3.2.2 Données de comptages des modules fonctionnels

Nous considérons, dans cette partie, le même ensemble d'échantillons (237 individus sains ou malades de la cirrhose du foie), pour lequel on dispose d'un tableau de comptages représenté par une matrice et contenant les valeurs de comptages des modules fonctionnels (fonctions des gènes), observés au sein de chaque individu.

Les méthodes d'analyses multivariées que nous allons appliquer s'appuient sur cette matrice de comptages normalisée.

Pour la cohorte de la cirrhose du foie, notre matrice de comptages des modules fonctionnels est de taille : n= 237 et p=133. Voici les différents graphes d'individus après l'application des différentes méthodes non supervisées (ACP, MDS et pPCA) sur cette matrice :

Graphes d'ACP après une transformation logarithmique :

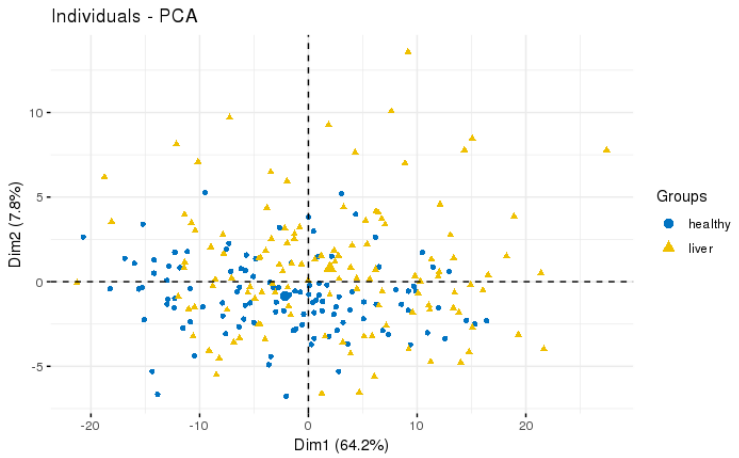


Figure 3-34- Graphes d'ACP sur des modules fonctionnels

Graphes du positionnement multidimensionnel (MDS) :

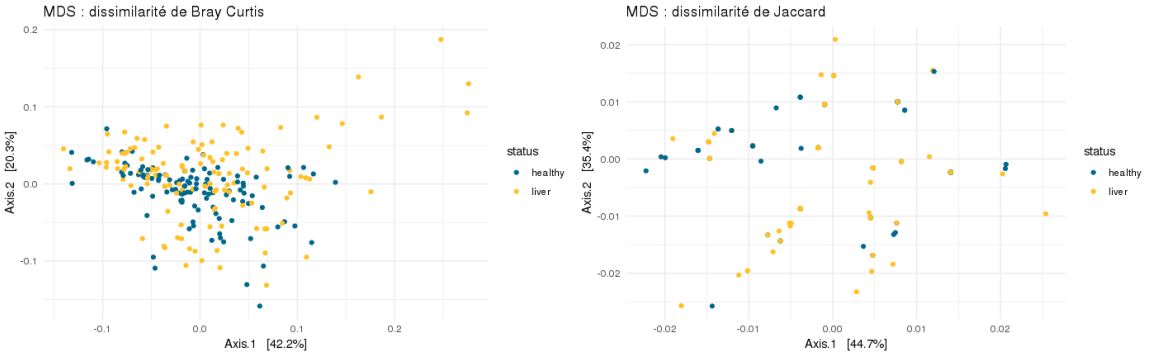


Figure 3-35- Graphes de MDS en utilisant les deux mesures de Bray-Curtis et Jaccard sur des modules fonctionnels

Graphes d'ACP probabiliste de poisson log-normal :

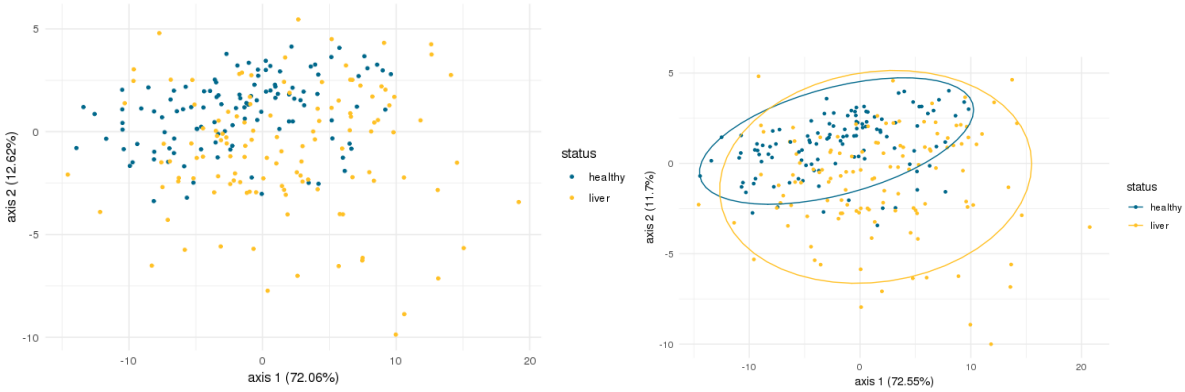


Figure 3-36- Graphes de pPCA de poisson log-normal sur des modules fonctionnels

Conclusion sur les 3 méthodes utilisées :

A travers les graphes représentés ci-dessus, on peut conclure qu'il n'y a pas une séparation claire entre les deux groupes d'intérêt « sains/malades » dans un contexte métagénomique fonctionnel. Cependant, il est à noter que la densité le long du premier axe montre une petite différence entre les individus sains et les individus malades : les individus malades (points jaunes) sont plus dispersés que les individus sains (points bleus). La variabilité, en termes de fonctions des gènes, est donc plus importante chez les malades. Ceci est bien remarquable

surtout dans les deux graphes d'individus après l'application d'une MDS en utilisant la mesure de Bray et après l'application de la méthode probabiliste pPCA.

Chapitre IV
Visualisation des résultats: R Shiny

4 Visualisation des résultats : R Shiny

Les analyses biostatistiques de routine réalisées sur les matrices de comptage de MGS sont réalisées à MGP à l'aide une plateforme d'analyse développée en RShiny, « ShRCAn ». L'application ShRCAn permet d'importer les données de comptage de gènes brutes et les données cliniques, d'automatiser le prétraitement des données, de construire la matrice de comptage d'espèces bactériennes (MGS), de clusteriser les échantillons, d'étudier la diversité locale ou « alpha-diversité » (la richesse en espèces bactériennes), d'analyser la répartition taxonomique des MGS et de réaliser l'analyse différentielle des abondances de MGS en fonction d'un critère phénotypique.

Présentation de ma contribution à l'application ShRCAn :

J'ai ajouté un nouvel onglet qui permet à l'utilisateur de visualiser les graphiques d'ordination en fonction de la distance ou de l'indice de dissimilarité (Bray, Jaccard, Jensen–Shannon divergence « jsd »...) choisi, ainsi que les graphiques de l'ACP probabiliste en ajoutant des covariables cliniques souhaitées. J'ai aussi implémenté la possibilité de colorer ces graphiques en fonction d'une variable cliniques d'intérêt, quantitative ou qualitative, par exemple l'état de l'individu « sain/malade » comme représenté dans les différentes figures du présent rapport.

J'ai également embelli les graphes en utilisant différents thèmes et plusieurs palettes de couleurs. Voici deux captures d'écran de l'application (les autres sont présentées en annexe 1) :

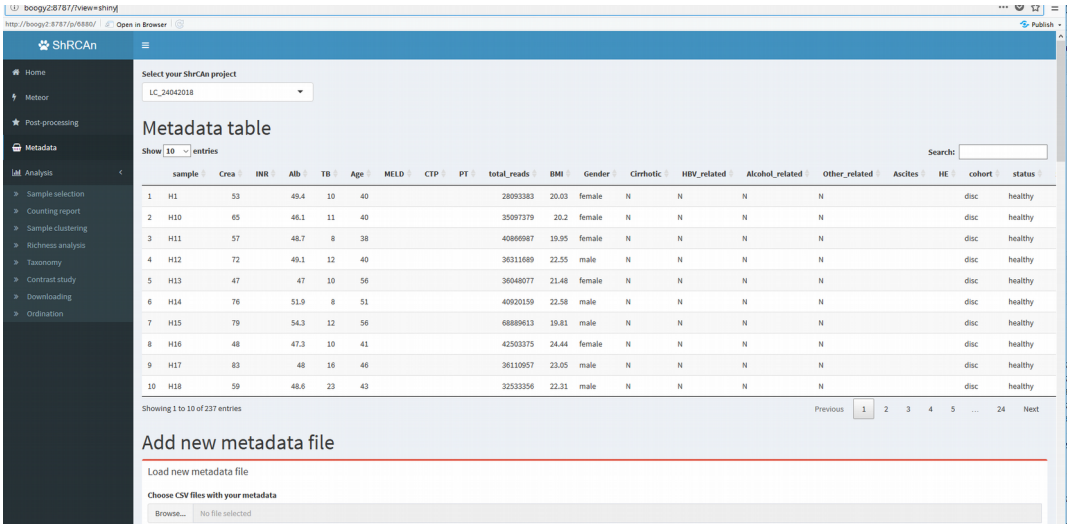


Figure 4-37- Import des données cliniques dans l'application **ShRCAn**

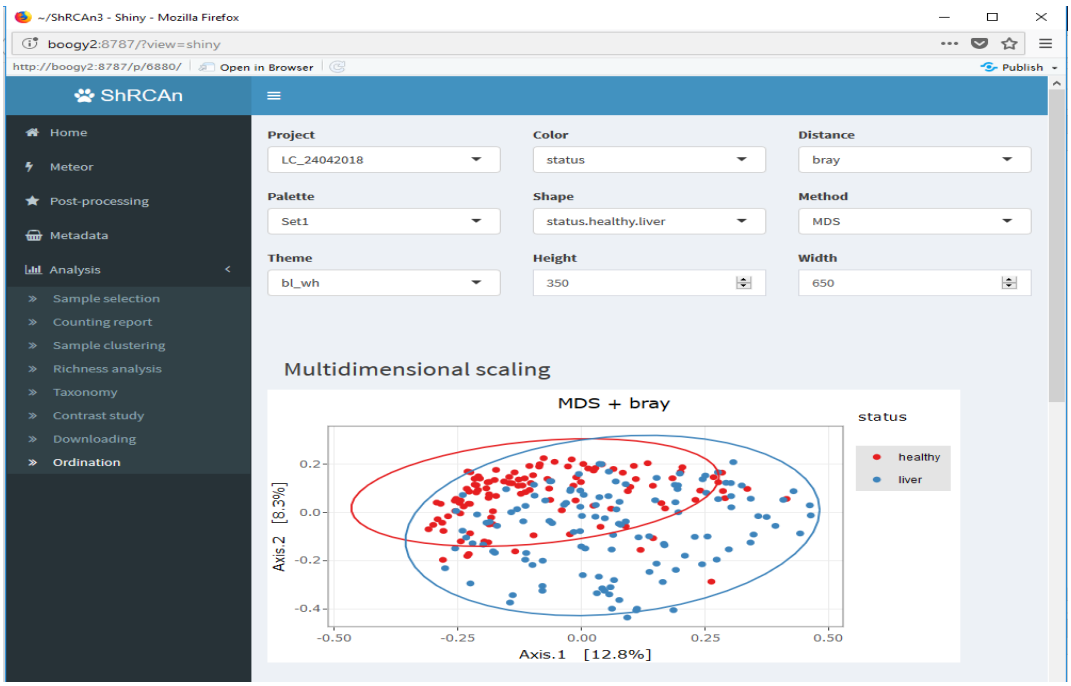


Figure 4-38- Visualisation des graphes d'ordination dans **ShRCAn**

Conclusion et perspectives

Les travaux de ce rapport se placent dans le cadre d'une activité de recherche dans le domaine de la biostatistique sur des problématiques de fouille des données de comptages métagénomiques, en grande dimension, bruitées et à inflation de zéros, afin d'explorer le lien entre le microbiote intestinal et la santé humaine.

L'objectif global de ce projet a été de proposer ou trouver les méthodes d'analyses statistiques multivariées non supervisées les plus adaptées et d'évaluer leur pertinence dans l'extraction de la structure existante entre les communautés microbiennes.

Dans un premier temps, nous nous sommes intéressés à l'application des méthodes d'ordination permettant de donner un aperçu des relations de similarité entre les individus en termes d'abondance des espèces bactériennes. L'objectif visé par cette méthode est de comparer entre les résultats de différentes mesures d'ordination appliquées à deux indices de dissimilarité : Bray-Curtis et Jaccard.

Dans un second temps, nous nous sommes intéressés à tester l'efficacité d'une méthode probabiliste spécifiquement conçue pour des données de comptages. C'est une ACP

probabiliste basée sur un modèle de poisson log-normal, avec l'intégration des covariables afin de tester leurs influences sur la structure des données de comptages métagénomiques.

Nous avons finalement développé une application Rshiny pour diffuser et rendre disponible ces méthodes pour l'équipe d'accueil « InfoBioStat ».

Ces deux méthodes n'ont été pas trop pertinentes dans l'dentification des structures de microbiote intestinal liées au statut « sain /malade » dans les différents jeux de données utilisés dans notre étude. Mais, Nous avons réussi à trouver les structures qui collent avec la richesse en espèces bactériennes.

Nous avons bien conscience que notre travail ne constitue qu'un début des réflexions pour l'exploration des données de comptages métagénomiques au sein de l'unité MétaGénoPolis de l'INRA de Jouy-en-Josas. Il serait alors judicieux d'appliquer ces méthodes sur d'autres jeux de données de comptages afin de tester leur robustesse dans un contexte métagénomique, d'utiliser d'autres méthodes non supervisées (Exemple : la méthode de PLS…), de tester la robustesse de la méthode supervisée utilisée dans notre étude en utilisant de la validation croisée par exemple, ou encore d'étudier les MGS une par une afin de déterminer les structures d'intérêt du microbiote intestinal humain.

Bibliographie

Ouvrages imprimés

[1] CHATFIELD, Chris, ZIDEK, Jim, et LINDSEY, Jim. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2010.

[2] GREENACRE, Michael et PRIMICERIO, Raul. *Multivariate analysis of ecological data*. Fundacion BBVA, 2014.

[3] KINDT, Roeland et COE, Richard. *Tree diversity analysis: a manual and software for common statistical methods for ecological and biodiversity studies*. World Agroforestry Centre, 2005.

Ouvrages électroniques

[1] EHRLICH, Stanislav Dusko. *The human gut microbiome impacts health and disease. Comptes rendus biologies*, 2016, vol. 339, no 7-8, p. 319-323. Disponible sur : https://www.sciencedirect.com/science/article/pii/S1631069116300312.

[2] NIELSEN, H. Bjørn, ALMEIDA, Mathieu, JUNCKER, Agnieszka Sierakowska, *et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes*. *Nature biotechnology*, 2014, vol. 32, no 8, p. 822. Disponible sur : https://www.ncbi.nlm.nih.gov/pubmed/24997787.

[3] KANEHISA, Minoru, SATO, Yoko, KAWASHIMA, Masayuki, et al. *KEGG as a reference resource for gene and protein annotation.* Nucleic acids research, 2015, vol. 44, no D1, p. D457-D462. Disponible sur : https://www.genome.jp/kegg/ko.html.

[4] VIEIRA-SILVA, Sara, FALONY, Gwen, DARZI, Youssef, et al. Species–function relationships shape ecological properties of the human gut microbiome. Nature microbiology, 2016, vol. 1, no 8, p. 16088. Disponible sur https://www.nature.com/articles/nmicrobiol201688.

[5] GUAN, Yue et DY, Jennifer. *Sparse probabilistic principal component analysis. In : Artificial Intelligence and Statistics*. 2009. p. 185-192. Disponible sur : https://www.researchgate.net/publication/220320523_Sparse_Probabilistic_Principal_Component_Analysis.

[6] CHIQUET, Julien, MARIADASSOU, Mahendra, et ROBIN, Stéphane. *Variational inference for probabilistic Poisson PCA*. 2017. arXiv preprint arXiv:1703.06633. Disponible sur : https://arxiv.org/abs/1703.06633.

[7] LACHENBRUCH, Peter A. et GOLDSTEIN, M. *Discriminant analysis. Biometrics*, 1979, p. 69-85. Disponible sur : https://www.jstor.org/stable/252993.

[8] MCMURDIE, Paul J. et HOLMES, Susan. *phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data*. PloS one, 2013, vol. 8, no 4, p. e61217. Disponible sur : https://github.com/joey711/phyloseq.

[9] QIN, Nan, YANG, Fengling, LI, Ang, et al. *Alterations of the human gut microbiome in liver cirrhosis*. Nature, 2014, vol. 513, no 7516, p. 59. Disponible sur : https://www.ncbi.nlm.nih.gov/pubmed/25079328

[10] AITCHISON, John et HO, C. H. *The multivariate Poisson-log normal distribution.* Biometrika, 1989, vol. 76, no 4, p. 643-653. Disponible sur : https://academic.oup.com/biomet/article-abstract/76/4/643/254386.

[11] COLLINS, Michael, DASGUPTA, Sanjoy, et SCHAPIRE, Robert E. *A generalization of principal components analysis to the exponential family.* In : Advances in neural information processing systems. 2002. p. 617-624.

[12] INOUYE, David I., YANG, Eunho, ALLEN, Genevera I., et al. *A review of multivariate distributions for count data derived from the Poisson distribution.* Wiley Interdisciplinary Reviews: Computational Statistics, 2017, vol. 9, no 3, p. e1398. Disponible sur : https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1398.

[13] TIPPING, Michael E. et BISHOP, Christopher M. *Probabilistic principal component analysis.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1999,

vol. 61, no 3, p. 611-622. Disponible sur : http://www.robots.ox.ac.uk/~cvrg/hilary2006/ppca.pdf.

[14] Wikipédia. *Bray-Curtis dissimilarity*. wikipédia, l'encyclopédie libre, 2018. URL http://en.wikipedia.org/wiki/Bray%E2%80%93Curtis_dissimilarity. [En ligne ; Page disponible le 8-avril-2018].

[15] BRAY, J. Roger et CURTIS, John T. *An ordination of the upland forest communities of southern Wisconsin.* Ecological monographs, 1957, vol. 27, no 4, p. 325-349. Disponible sur : https://esajournals.onlinelibrary.wiley.com/doi/10.2307/1942268.

[16] PARK, Eun et LORD, Dominique. *Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity.* Transportation Research Record : Journal of the Transportation Research Board, 2007, no 2019, p. 1-6.

[17] DORE, Joël, EHRLICH, Dusko, MONNET, Véronique, et al. *microbiote, la révolution intestinale*. *Dossier de presse INRA 2017,*2017.

[18] COTILLARD, Aurélie, KENNEDY, Sean P., KONG, Ling Chun, et al. *Dietary intervention impact on gut microbial gene richness.* Nature, 2013, vol. 500, no 7464, p. 585.

[19] LE CHATELIER, Emmanuelle, NIELSEN, Trine, QIN, Junjie, et al. *Richness of human gut microbiome correlates with metabolic markers*. Nature, 2013, vol. 500, no 7464, p. 541.

[20] ODINTSOVA, Vera, TYAKHT, Alexander, et ALEXEEV, Dmitry. *Guidelines to Statistical Analysis of Microbial Composition Data Inferred from Metagenomic Sequencing*. Current issues in molecular biology, 2017, vol. 24, p. 17-36.

[21] MANDAL, Siddhartha, VAN TREUREN, Will, WHITE, Richard A., et al. *Analysis of composition of microbiomes: a novel method for studying microbial composition*. Microbial ecology in health and disease, 2015, vol. 26, no 1, p. 27663.

[22] MCMURDIE, Paul J. et HOLMES, Susan. *phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data*. PloS one, 2013, vol. 8, no 4, p. e61217.

Travaux universitaires

[1] PAVOINE, Sandrine. *Méthodes statistiques pour la mesure de la biodiversité. Thèse de doctorat en biostatistiques.* Université Claude Bernard - LYON I. Soutenu le 06 décembre 2005.

[2] Frédéric CAILLEUX. *Impact du microbiote intestinal dans la maladie alcoolique du foie*. Thèse de doctorat. Université Paris Sud XI. Soutenue le 07 avril 2014.

[3] ALMEIDA, Mathieu. *Caractérisation de flores microbiennes intestinale humaine et fromagère par méthode de métagénomique quantitative*. Thèse de doctorat en Sciences de la Vie. Université Paris Sud XI Orsay. Soutenu le 07 juin 2013.

Annexe 1: Captures d'écran de l'application R Shiny

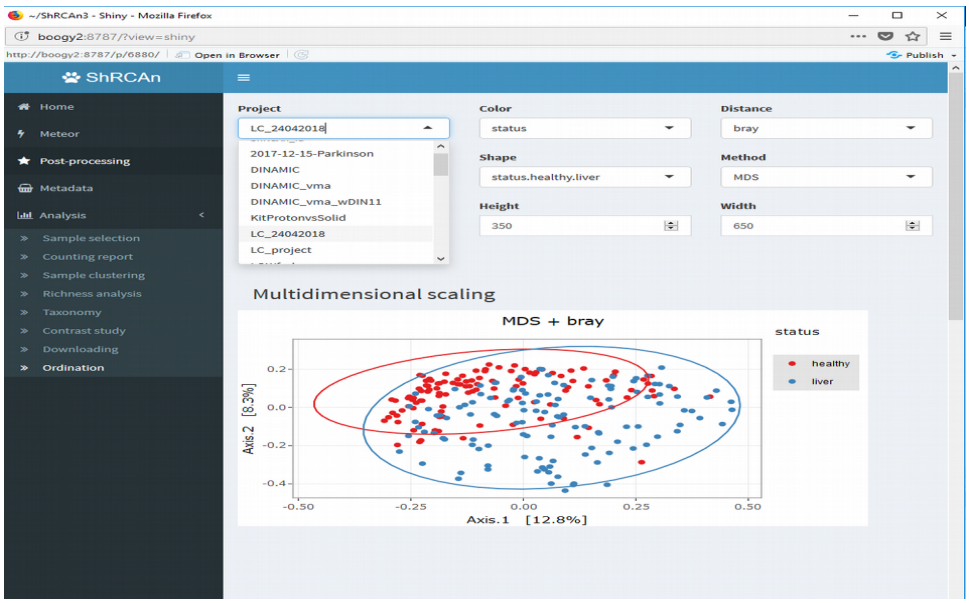


Figure 4-39- Choix du projet dans ShRCAn

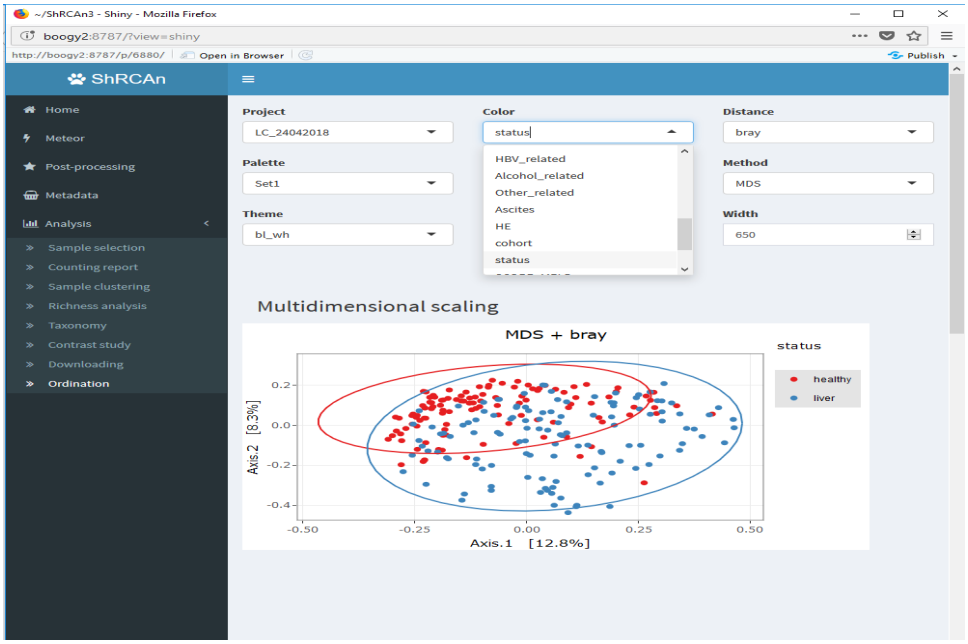
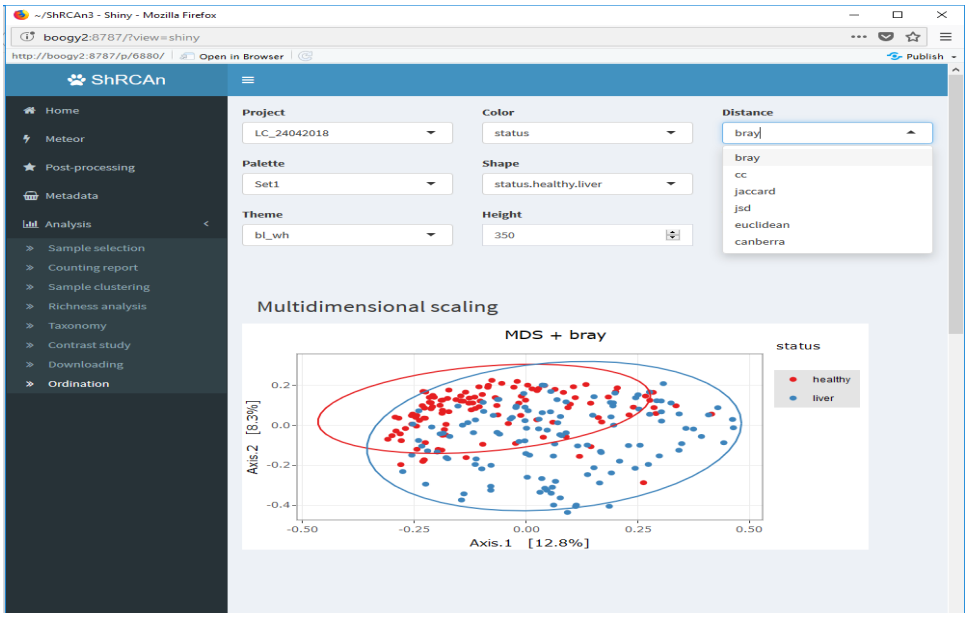


Figure 4-40- Coloration des graphiques d'ordination en fonction d'une variable cliniques d'intérêt, quantitative ou qualitative



Annexe 4-1- Choix de distance ou (dis)similarité pour appliquer une méthode d'ordination

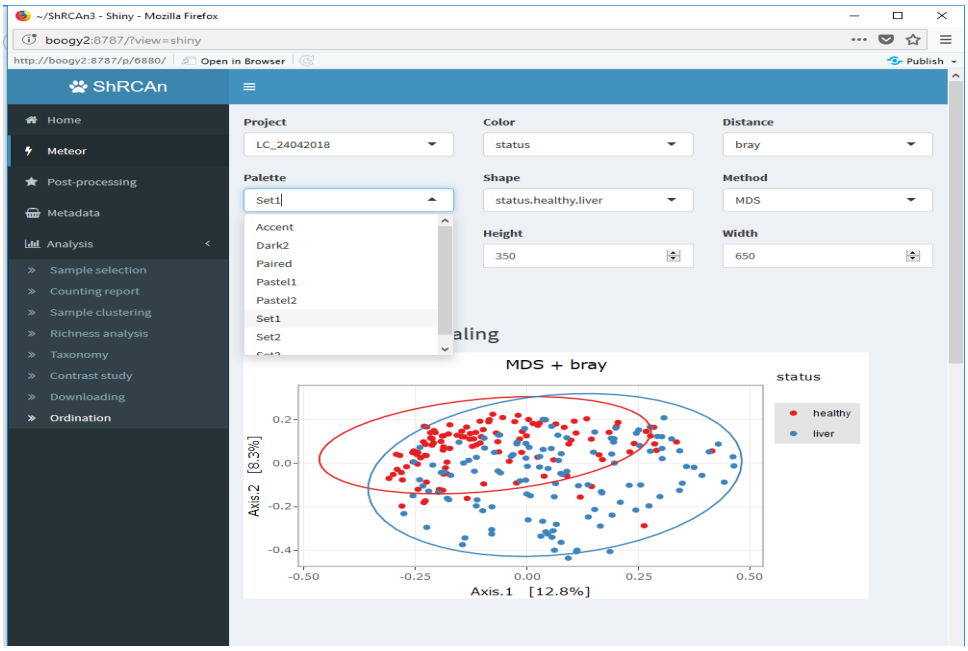


Figure 4-41- Choix de palette des couleurs

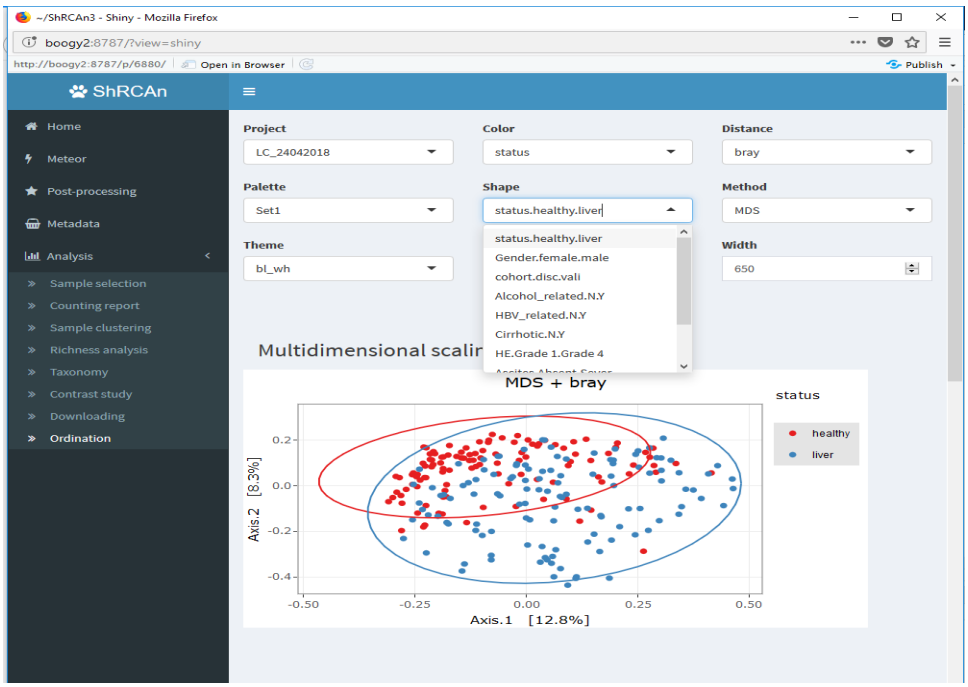


Figure 4-42- Graphe de l'ACP probabiliste de poisson log-normal

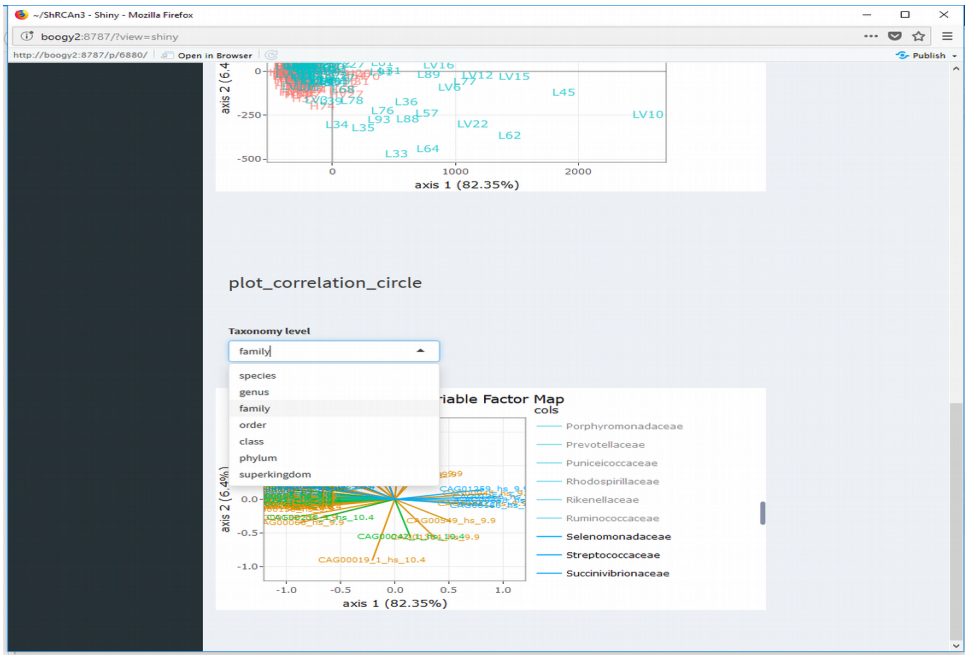


Figure 4-43- Cercle de corrélation d'une ACP probabiliste de poisson log-normal

Annexe 2: Code R