Étude de données de microbiote intestinal issues de la cohorte de jumeaux "TwinsUK" à l'aide de méthodes de régression incorporant la notion de distance

Jimmy Mullaert

UNIVERSITÉ PARIS DESCARTES

AVERTISSEMENT

Cette thèse d'exercice est le fruit d'un travail approuvé par le jury de soutenance et réalisé dans le but d'obtenir le diplôme d'Etat de docteur en médecine. Ce document est mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt toute poursuite pénale.

Code de la Propriété Intellectuelle. Articles L 122.4

Code de la Propriété Intellectuelle. Articles L 335.2-L 335.10

USPC
Université Sorbonne Paris Cité

UNIVERSITÉ PARIS DESCARTES

Faculté de Médecine PARIS DESCARTES

Année 2019 N° 70

THÈSE

POUR LE DIPLÔME D'ÉTAT

DE

DOCTEUR EN MÉDECINE

Étude de données de microbiote intestinal issues de la cohorte de jumeaux "TwinsUK" à l'aide de méthodes de régression incorporant la notion de distance

Présentée et soutenue publiquement
le 14 mai 2019

Par

Jimmy MULLAERT
Né le 15 août 1985 à Reims (51)

Dirigée par Mme Le Professeur France Mentré, PU-PH

Jury :

M. Le Professeur Erick Denamur, PU-PH ..... Président
Mme Le Professeur Sylvie Chevret, PU-PH
Mme Le Docteur Aurélie Bourmaud, MCU-PH
M. Le Docteur Olivier Tenaillon, Directeur de Recherche INSERM

USPC Université Sorbonne Paris Cité

Remerciements

Je remercie tout d'abord les membres du jury pour avoir accepté d'évaluer ce travail et pour le temps consacré à sa relecture.

Je remercie tout particulièrement le Professeur Erick Denamur, président du jury, pour ses commentaires et ajouts décisifs dans le manuscrit et la motivation qu'il insuffle chaque jour aux membres de l'unité de recherche UMR1137 IAME.

Je remercie également le Professeur France Mentré, ma directrice de thèse, pour sa patience à toute épreuve, sa constante bienveillance depuis que j'ai rejoins son équipe de recherche, ainsi que pour sa rigueur scientifique incontestable.

Merci également au Professeur Sylvie Chevret et à Aurélie Bourmaud. Votre participation à ce jury me touche beaucoup et nous aurons l'occasion, je l'espère, de travailler ensemble par la suite, sur des projets d'enseignement comme de recherche.

Merci également à tous ceux qui ont rendu ce travail possible : Olivier Tenaillon et son équipe en particulier. Les données qui ont servies pour ce travail d'analyse ont demandé des heures d'efforts à beaucoup de personnes et je leur en suis reconnaissant.

Merci aux membres de l'équipe de recherche BIPID à laquelle j'appartiens, ainsi qu'aux membres du Département d'Epidémiologie, Biostatistiques et Recherche Clinique du groupe hospitalier HUPNVS, au sein duquel j'effectue mon activité hospitalière, pour leurs précieux commentaires lors de la préparation de la soutenance orale.

Merci également à ma femme, mes enfants et amis pour m'avoir soutenu dans cette longue épreuve et jusqu'à la soutenance finale.

Sommaire

Introduction

Le terme "microbiote intestinal" désigne la composition du tube digestif en microorganismes. Il peut s'agir soit de la composition bactérienne, la plus étudiée [22, 54, 56], mais aussi de la présence de virus (appelée virome ou phageome) [12, 26, 38, 40, 54] ou de champignons (mycobiome) [10, 12, 32, 55]. On estime entre 10^{13} et 10^{14} le nombre de bactéries présente dans le tube digestif humain, soit entre 1 et 10 fois le nombre de cellule composant le corps humain [59,68]. Cette importance en nombre en fait le reservoir bactérien le plus étudié et et on suspecte depuis au moins un siècle que ces bactéries jouent un rôle dans des fonctions essentielles à la vie [56,65].

L'étude du microbiote intestinal est récemment passé dans une nouvelle ère avec l'émergence des méthodes de séquençage haut-débit, qui ont rendu quasiment obsolète les anciennes méthodes basées sur la culture cellulaire [22,27,66] en apportant deux avantages décisifs : d'une part leur sensibilité à mettre en évidence des bactéries difficilement cultivables, mais aussi par leur coût réduit relativement au débit de données qu'elles sont capables de fournir. A titre d'illustration, il est possible aujourd'hui d'obtenir, à partir d'un échantillon de selle, des informations riches sur la composition bactérienne en une journée et pour un coût de l'ordre de la centaine d'euros.

Ces dernières recherches ont permis de mettre en évidence l'altération du microbiote intestinal (appelée dysbiose) dans de nombreuses pathologies. Ainsi, l'étude du microbiote intestinal représente aujourd'hui un enjeu de santé publique. Sans surprise, ce sont les maladies intestinales qui ont, les premières, concentré les recherches [46]. L'inventaire commence par les maladies inflammatoires chroniques de l'intestin (MICI), comprenant la maladie de Crohn et la rectocolite hémorragique. Ce sont des maladies très invalidantes avec une prévalence en France d'environ 20 cas pour 10^5 habitants. Des modifications du contenu microbien intestinal ont été décrites chez des patients souffrant de MICI [72] : ces derniers hébergent dans leur tube digestif une proportion plus importante de certaines espèces (Enterobactéries, Fusobactéries) et moins importante d'autres espèces (Clostridia, Faecalibacterium). Un lien avec le patrimoine génétique de l'hôte a même été décrit : des gènes de susceptibilité ont été identifiés comme NOD2, ATG16L1 ou bien Card9, qui permet aux lymphocytes T de reconnaitre les micro-organismes [30, 33]. Ces résultats reposent uniquement sur des études observationnelles et ne permettent pas, à ce stade, de conclure à une éventuelle causalité. Cependant, un premier pas dans ce sens a été fait dans une étude récente [31], qui montre même qu'il est possible d'induire une inflammation de l'épithélium intestinal en greffant un microbiote particulier à des souris axéniques (c'est-à-dire nées et élevés dans des conditions stériles, sans aucune colonisation bactérienne intestinale).

Les bactéries du tube digestif participant aux échanges de nutriments, on retrouve une association entre composition du microbiote et des maladies métaboliques comme le diabète, l'obésité et la malnutrition [20, 21, 51]. Plus surprenant, il a été montré un lien avec des maladies neurologiques ou psychiatriques comme les troubles du spectre

autistique, troubles de l'humeur [39] et la schizophrénie [14, 48], à tel point que l'on parle maintenant d'un véritable axe intestin-cerveau [60]. On peut également citer dans ce cadre les travaux récents sur des maladies neurologiques liées à l'âge : les maladies d'Alzheimer [29] et de Parkinson [47].

Le domaine de la résistance aux antibiotiques est également un thème de recherche pour lequel le rôle du microbiote intestinal est probablement important. On assiste à une baisse du rythme de développement de nouvelles molécules et une hausse simultanée de la prévalence de la résistance bactérienne aux antibiotique [67]. Les politiques récentes de limitation de prescription tentent de pallier ce phénomène, mais leur efficacité n'est que relative. Dans ce contexte, l'intestin humain représente un réservoir de bactéries, dont certaines peuvent être porteuse de gènes de résistance [64], susceptibles de se transmettre entre bactéries et se propager ensuite à l'environnement. Ces bactéries résistantes peuvent également être sélectionnées par les antibiothérapies (et particulierement par β lactamines et carbapénèmes) prises par l'individu en ville ou lors d'une hostpitalisation [61, 62]. L'étude de l'effet des antibiotiques sur le microbiote intestinal peut ainsi permettre de comprendre et de mettre en évidence des facteurs favorisant l'émergence de résistance.

Un autre domaine dans lequel l'étude de l'évolution du microbiote intestinal sous l'effet d'antibiotique peut rendre un grand service pour la pratique clinique est la prévention des colites à C. *difficile*, dont les facteurs de risque connus sont l'âge avancé, l'hospitalisation et la prise d'antibiotiques. Cette bactérie peut être portée de manière asymptomatique (on estime la prévalence de ce portage à 5%) et on comprend aujourd'hui encore mal les raisons qui favorise sa croissance et sa virulence. Il est vraisemblable que la modification de l'environnement bactérien induite par l'antibiothérapie soit une condition favorisant la croissance de C. *difficile* [53], mais la caractérisation de ces modifications dans la composition microbienne de l'intestin n'est à ce jour pas précisément connue. Un travail récent [8] montre, sur un modèle animal de colite à C. *difficile*, que la baisse de la diversité consécutive à une antibiothérapie est fortement associée à la mortalité. Ici encore, il n'est pas clair si la dysbiose joue le rôle de marqueurs d'un risque infectieux ou bien si elle est un réel facteur de risque. Des essais cliniques concluants de transplantation fécale semblent pourtant plaider en faveur de l'hypothèse d'une réelle causalité [23, 41], même si les effectifs de ces études restent modestes.

Parmi toutes les bactéries qui composent le microbiote humain, E. *coli* est probablement la plus étudiée, bien connue pour son implication dans des toxi-infection alimentaires ou bien les infections génitales basses [17, 42]. Dans des conditions particulières d'immunodépression ou de stress, elle peut devenir virulente et entraîner une septicémie gravissime, dont la mortalité associée est d'environ 50%. Parallèlement à ces formes pathogènes, E. *coli* est également un hôte commensal du tube digestif humain et représente entre $0,1$ et 1% du microbiote intestinal [16, 19]. Elle est, en revanche, dominante si l'on s'intéresse uniquement aux bactérie aérobies et possède la particularité d'être la première à coloniser le tube digestif après la naissance [6, 28]. Son abondance relative peut augmenter au décours de certains états pathologiques comme cela a été décrit dans le cas

des MICI [69]. La bactérie E. *coli* représente également un enjeu de santé publique par sa capacité à porter et transmettre des gènes de résistance aux antibiotiques. La prévalence de portage d'une souche multi-résistante (porteuse d'une betalactamase à spectre élargi) est d'environ 6% en région Parisienne [50] et cette prévalence est en augmentation. Les individus de retour d'un voyage à l'étranger sont également à haut risque d'être porteur d'une souche multi-résistante [3].

Dans cette thèse, nous allons nous intéresser plus en détail à la composition en E. *coli* du microbiote intestinal et sur les associations statistiques qui peuvent exister entre le profil de colonisation à E. *coli* et la composition du microbiote bactérien complet. L'hypothèse principale de recherche est que E. *coli* pourrait influencer la composition de tout le microbiote, puisqu'elle créée les conditions anaérobies nécessaires à la colonisation par d'autres espèces. On utilisera pour cela des données de la cohorte TwinsUK [44, 45, 63], dont des résultats concernant le microbiote des jumeaux homozygotes comparativement aux dizygotes ont été publiés [25]. Dans cette publication, les auteurs ont pu obtenir une description du microbiote intestinal de 977 jumeaux (171 paires de jumeaux monozygotes et 245 paires de jumeaux dizygotes) en utilisant une méthode basée sur l'amplification d'une fraction du gène codant pour l'ARN ribosomal 16S bactérien. Pour chaque échantillon, le matériel génétique amplifié est séquencé, ce qui fournit une importante liste d'environ 80 millions de fragments (aussi appelés reads), qui proviennent chacun d'une espèce bactérienne présente dans l'échantillon. On obtient ainsi une table de compte qui décrit, pour chaque espèce identifiée et chaque échantillon, le nombre de reads détectés.

Cette thèse propose de poursuivre l'exploitation de ces données en poursuivant trois objectifs :

1. On dispose de données sur l'âge, le sexe et l'indice de masse corporelle des participants à la cohortece qui permet de tester, à l'aide de méthodes statistiques appropriées, l'association entre la composition du microbiote et ces variables.
2. Les espèces bactériennes identifiées peuvent être représentées sous forme d'un arbre phylogénétique qui matérialise les distances génétiques entre elles. Un enjeu important pour les tests d'association est de prendre en compte cette information afin d'exploiter l'ensemble des données disponibles.
3. Enfin, le protocole d'amplification de l'ARN 16S utilisé ne permet pas de détecter la présence de E. *coli*, ni de décrire les souches présentes, car elle ne dispose pas d'une résolution suffisante. On propose donc de coupler les données de microbiote de la cohorte twinsUK avec des données obtenues dans l'unité INSERM 1137 (IAME) à partir des mêmes échantillons, afin d'explorer le lien qui pourrait exister entre le microbiote et le profil de colonisation à E. *coli*.

Les méthodes statistiques pour exploiter une table de compte sont multiples, mais peu d'entre-elles permettent de prendre en compte l'information a priori contenue dans l'arbre phylogénétique qui structure les espèces composant le microbiote intestinal. Les difficultés statistiques qui apparaissent sont de deux ordres. La première concerne le nombre d'espèces important détecté qui est de l'ordre de plusieurs centaines (768 dans le cas de

nos données, après sélection des espèces présentes dans plus de 50% des échantillons). Dans ces conditions, la comparaison de deux groupes pour chacune de ces espèces mène à un nombre de test important et il est important de prévoir une correction pour la multiplicité des tests, afin de correctement contrôler l'erreur de type I (ou le taux de fausses découverte en fonction de l'application). Ce problème de test multiples s'amplifie encore si l'on considère pas uniquement les espèces, mais aussi des regroupements d'espèces (par genre, famille,...) comme c'est l'usage dans de nombreuses publications. La deuxième difficulté est de préserver un maximum de puissance statistique, puisque les nombreux tests que l'on peut réaliser à partir d'une table de compte ne sont pas indépendants et qu'une correction du seuil de significativité naive de type Bonferonni serait probablement très conservatrice. D'autres stratégies de correction existent, mais nécessitent des hypothèses sur la structure de corrélation des statistiques de tests qui sont difficiles à établir sur ces données.

Les méthodes de regression "distance-based" permettent à la fois de contourner la difficulté de la dimensionnalité des données et d'incorporer de l'information *a priori* sur la structure phylogénétique sous-jacente, par la spécificationd d'une fonction de distance adaptée [43, 73]. Elles consistent en une première étape de réduction de la dimensionnalité de la table de compte par positionnement multidimensionnel (multidimentional scaling), qui permet d'incorporer une fonction de distance qui rend compte de la proximité génétique entre espèces. Une fois la dimension des données réduite, des régressions multivariées standard sont utilisées.

La suite de ce manuscrit est organisé de la manière suivante. La section 1 présentera la cohorte TwinsUK ainsi que les manipulations et étapes bioinformatiques menant aux tables de compte. Dans la section 2, on présentera les méthodes de régression dites "distance-based", qui permettent de résoudre simultanément les deux difficultés méthodologiques décrites ci-dessus. Enfin, la section 3 sera consacrée aux résultats des analyses, qui seront discutés dans la conclusion.

1 Les données de microbiote intestinal

Les données qui servent de support à cette thèse sont issues de participant volontaires appartenant à la cohorte TwinsUK [44, 45, 63]. On commencera donc par décrire cette cohorte ainsi que les multiples projets qui y ont été associés. Puis on décrira les méthodes de préparation des échantillons pour l'analyse du microbiote (analyse 16S) et l'analyse de la composition en E. coli, qui ont servi de base à l'analyse statistique.

1.1 La cohorte twinsUK et ses projets associés

La cohorte TwinsUK a été initiée au Royaume Uni en 1992 et rassemble environ 12000 jumeaux adultes volontaires. L'objectif principal était de décrire l'héritabilité de l'ostéoporose, de l'arthrose et des infections osteo-articulaires chez la femme. A partir de l'année 1995, les jumeaux de sexe masculins ont été recruté, de sorte que la cohorte se compose d'une majorité (83%) de femmes. L'âge des participants varie de 16 à 100 ans. Les principales caractéristiques des participants sont résumés dans la table 1.1.

Les participants subissent une visite médicale d'inclusion qui s'est déroulée entre 1992 et 2004, puis deux visites de suivi (l'une en 2004-2007 et l'autre en 2007-2010). Dans l'intervalle entre les visites, les participants sont invités à répondre à des questionnaires de santé annuels. De courtes visites de suivi sont également prévues. Au total, de très nombreuses informations sont recueillies, la liste des phénotypes est disponible à l'adresse suivante : http ://www.twinsuk.ac.uk/wp-content/uploads/2012/06/pheno_phenotypes.xls

On dispose également de données de type "omics" sur les participants de la cohorte twinsUK. 5710 jumeaux ont été génotypés pour un maximm de 610000 SNP (single nucleotide polymorphism) à l'aide des puces Illumina HumanHap300 Bead Chip et Illumina HumanHap610 Quad Chip. Ces données ont permis la publication de nombreuses études de type GWAS (genome-wide association sudy), dont l'objectif est de montrer une association entre un phénotype binaire ou quantitatif et des marqueurs génétiques fréquents.

Si l'on s'intéresse au variants génétiques rares, c'est-à-dire ceux dont la fréqence allélique dans la population générale est inférieure à 5%, d'autres données sont nécessaires car les puces à ADN contiennent très peu de variants rares. Ainsi, le projet UK10K mené en collaboration avec le Wellcome Trust Sanger Institute prévoit le séquençage du génome complet de 2000 participants de la cohorte TwinsUK. D'autres projets ont utilisé les participants de la cohorte TwinsUK pour produire des données omiques. C'est le cas par exemple du projet EpiTwin, en collaboration avec l'institut de génomique de Pékin, qui prévoit d'étudier comparativement l'état de méthylation de l'ADN entre jumeaux. On peut également citer le projet MuTHER (Multiple Tissue Human Expression Resource, [49]), financé par le Wellcome Trust, dont l'objectif est d'étudier le profil d'expression des gènes dans des échantillons de peau, de tissus adipeux et de lignée lymphoblastoïde (*i.e.* des cellules B périphériques immortalisées *in vitro*).

Enfin, des échantillons de selle pour une étude métagénomique du microbiote intestinal ont été recueillis. Leur traitement préalable à l'analyse bioinformatique et statistique est décrit dans le paragraphe suivant.

	Visite d'inclusion (1992–2004)	Première visite de suivi (2004–07)	Seconde visite de suivi (2007–10)
N	5725	3725	3125
Âge (années)	46.8 ± 12.6	52.5 ± 13.4	59.6 ± 9.3
Sexe (% homme/% femme)	7.2 / 92.8	10.8/89.2	0/100
Zygosité (% DZ/% MZ)	64.3/35.7	45/55	52/48
Déjà marié (%)	81.1	88.5	90.1
Âge de sortie de l'école (années)	17.3 ± 3.6	17.7 ± 3.6	17.4 ± 3.6
Niveau socioéconomique bas (%)	21.4	19.8	17.5
Etat de santé perçu (bon/mauvais) (%)	7.2	8.0	12.3
Age des premières règles (années)	13.0 ± 1.6	12.9 ± 1.5	12.9 ± 1.5
Fumeur (%)	18.4	15.1	9.3
Ancien fumeur (%)	28.4	27.9	37.3
Consommation d'alcool (nombre par semaine)	4 (0–9)	4 (1–10)	2 (0–5)

TABLE 1.1 – Caractéristiques des individus de la cohorte TwinsUK qui ont participé au visites cliniques succesives (adapté de [45]).
DZ : Dizygote, MZ : monozygote, UZ : zygosité inconnue
Niveau socioéconomique bas : deux premiers quintiles du score IMD (Index for multiple deprivation) dans la population générale.

1.1.1 Préparation des échantillon pour l'analyse 16S

Les échantillons de selle nécessaires à l'analyse ont été prélevés à domicile et stockés à 2°C pendant au maximum deux jours avant d'être congelés à −80°C. Ces échantillons ont subi une extraction d'ADN selon les procédures standard, puis une amplification par PCR de la fraction V4 de l'ARN16S bactérien à l'aide d'amorces adaptées. La région V4 (représentée schématiquement figure 1.1) a une longueur d'environ 1500 paires de base. Cette région est très utilisée en métagénomique car elle satisfait simultanément deux conditions. Elle est à la fois suffisamment variable pour que son séquençage permette de déduire l'espèce bactérienne dont elle provient, mais également suffisamment conservée, par endroits, pour que des amorces universelles permettent de l'amplifier de manière efficace, quelle que soit l'espèce concernée.

Le séquençage proprement dit s'est fait sur une plateforme Illumina MiSeq avec des reads paired-end de 250 paires de base de long. Pour aboutir au séquençage, le protocole prévoit l'amplification de la région à séquencer par PCR, puis la constitution d'une collection fragments d'environ 800 paires de base maximum. Des oligonucléotides particuliers appelés adaptateurs sont alors fixés aux deux extrémités des fragments. Ces derniers comprennent des séquences qui permettront de fixer le fragment lors du séquençage ainsi qu'une séquence spécifique à chaque patient, appelée barcode, qui permet de séquencer la matériel génétique de plusieurs patients en même temps.

A la sortie, le séquenceur produit une liste de reads, c'est-à-dire de courtes séquences qu'il s'agit d'assembler en exploitant les chevauchements afin de reconstituer les séquences de la région V4 de toutes les espèces bactériennes présentes dans l'échantillon initial. Ce travail bioinformatique est décrit dans la section 1.2.

1.1.2 Préparation des échantillon pour l'analyse E. coli

Une des difficultés lorsqu'il s'agit d'explorer le profil de colonisation à E. coli est que cette espèce est largement minoritaire dans le tube digestif humain, ce qui exclu d'utiliser la même préparation que pour l'analyse 16S. Dans le cas de l'analyse pour E. coli, nous avons utilisé des échantillons de selles généreusement transmis par l'équipe de Ruth Ley et réalisé une culture cellulaire sur un milieu favorable à la croissance des bactéries E. coli. Pour chaque échantillon, 100 clones ont été isolés et l'ensemble de leur matériel génétique a été séquencé.

Au total, il y a donc deux différences majeures entre les pipeline 16S et le pipeline E. coli. La première provient du fait qu'il y a une étape de culture et de repiquage dans la préparation des échantillons pour l'analyse E. coli. La deuxième est que l'ensemble du matériel génétique des bactéries E. coli sont amplifiés et séquencés, alors que dans le cas du pipeline 16S, seul une portion (la partie V4) de l'ARN 16 des bactéries sert de base à la détermination des espèces présentes et de leur abondance. Il est logique que pour distinguer deux souches de E. coli, le niveau de définition atteint par l'analyse de la seule région V4 est insuffisant, ce qui justifie que, pour l'analyse E.coli, l'ensemble du génome

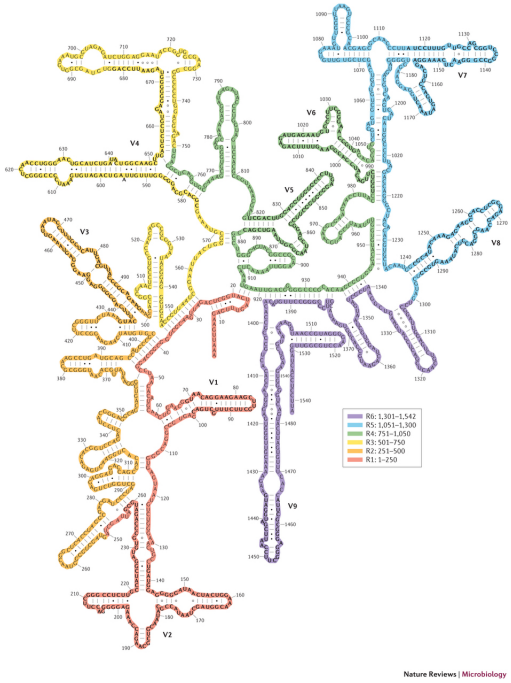

FIGURE 1.1 – Représentation schématique de l'ARN 16S bactérien (tiré de [71]). La région V4 se situe en haut à gauche.

soit séquencé.

Le génome de la bactérie E. coli est particulier par rapport au génome humain : deux bactéries de la même espèce ne possèdent pas nécessairement les mêmes gènes. Seul environ la moitié du génome (appelé le core genome) est présent dans toutes les souches [18]. L'autre moitié, appelée génome accessoire, pouvant être soit présent, soit absent en fonction de la souche étudiée. Cela rend difficile la définition d'un génome de référence pour l'espèce E.coli. Ainsi, nous avons sélectionné 56 souches qui représentent la diversité de l'espèce E.coli et constitué 56 génomes de référence, qui sont spécifiques à ces souches et non à l'espèce entière. Le traitement bioinformatique qui suit le séquençage consiste donc, dans le cas de l'analyse E. coli, à affecter chaque read à un ou plusieurs des 56 génomes de référence.

1.2 Traitement bioinformatique

La partie bioinformatique dans le circuit des données arrive à la suite de la production par le séquenceur des fichiers au format fastq contenant les séquences des reads lus par la machine. On distingue alors plusieurs étapes qui séparent ce fichier de la production de la table de compte qui servira par la suite à l'analyse statistique.

1.2.1 Contrôle qualité et filtrage des reads

Tout d'abord, il est nécessaire d'opérer un contrôle de qualité afin d'éliminer les reads de mauvaise qualité qui pourraient impacter l'analyse statistique à venir. La stratégie exacte à adopter pour ce contrôle qualité n'est pas consensuelle et fait l'objet de recherches [75].

Le processus d'amplification de l'ADN qui sert de base au séquençage à haut débit est particulièrement sensible à des contamination par de l'ADN exogène. On utilise le barcode qui a été greffé sur les fragments d'ADN à séquencer afin de démultiplexer les reads, c'est-à-dire identifier de quel échantillon provient chaque read et d'éliminer les reads provenant d'une contamination.

Le séquenceur produit, en plus des séquences des reads, une indication sur la confiance que l'on en en chaque bases de la séquence. Il est ainsi possible d'élimier les reads qui sont de mauvaise qualité. Le processus complet de contrôle qualité appliqué est précisément décrit dans [7]. Il s'appuie, entre autres, sur le score Phred fourni par le séquenceur. Ce score Q est relié à la probabilité d'erreur p d'identification d'une base par la formule

$$Q = -10 \log_{10} p$$

Ainsi, un score de 50 pour une certaine base de la séquence signifie que la probabilité d'erreur de séquençage sur cette base est de 10^{-5}. Le principal risque qui motive le choix d'un score seuil élevé est celui d'obtenir une diversité plus importante que la réelle diversité de l'échantillon, aussi appelée α-diversité [7].

De manière générale, la probabilité d'une erreur de séquençage sur une base donnée augmente entre le début à la fin du read. L'algorithme de filtrage employé par la suite logicielle qiime qui a été utilisée [9,57] consiste donc à tronquer les reads lorsque le score de qualité passe en dessous d'un seuil de tolérance qui est généralement de 30, soit une erreur pour mille bases). Le read tronqué est ensuite complètement éliminé si sa taille est inférieure à un autre seuil prédéfini.

1.2.2 Alignement et production de la table de compte pour les données 16S

L'étape d'alignement est probablement la plus coûteuse en terme de ressources informatiques nécessaires. Il s'agit en effet de tester, pour chaque read, s'il est susceptible de provenir d'une ou plusieurs espèces. Les stratégies pour identifier des OTU à partir d'une liste de reads provenant peuvent formellement se classer en trois catégories :

— la stratégie "de novo" consiste à réaliser des clusters de séquences homologues à 97% à partir des reads disponibles. Ainsi, la règle est de considérer que des séquençes identiques à 97% proviennent de la même OTU. Cette méthode présente l'avantage de ne pas nécessiter de séquence de référence, mais ne permet pas de nommer les OTU. Elle est également très demandeuse en ressources de calcul.

— la stratégie "closed reference" consiste à utiliser des séquences de références pour des OTU déjà connus et à tenter d'aligner chaque read avec l'un de ces OTU. Ici, on peut véritablement nommer les espèces identifiées et la complexité de l'algorithme d'alignement est bien moindre que dans le cas de la stratégie "de novo". En revanche, comme le microbiote contient de nombreuses espèces inconnues, un grand nombre de reads ne pourront être classés en utilisant cette stratégie.
— la stratégie "open reference" réalise un compromis entre les deux stratégies précédentes. On réalise d'abord la méthode "closed reference" à partir d'une base de données publique, puis les reads inclassables sont regroupés en OTU par une stratégie "de novo". Ainsi, tous les reads sont utilisés, mais seule une partie des OTU identifiés peuvent être nommés. C'est cette dernière stratégie qui est la plus utilisée dans la littérature et que l'on a utilisé pour obtenir la table de compte 16S.

Ces méthodes sont toutes implémentées dans la suite logicielle qiime, qui a été également utilisée pour réaliser l'identification des OTU. Cette étape fournit en sortie une table de compte qui indique, pour chaque OTU identifié et chaque échantillon, combien de reads ont été trouvés. Elle fournit également un arbre phylogénétique qui indique la hiérarchie entre les OTU. Comme on le verra dans la section 2, cette information sera très utile à l'analyse statistique.

Le pipeline bioinformatique utilisé pour l'analyse des données 16S est susceptible de générer un nombre très important d'OTU, du fait de l'utilisation de la stratégie "open reference". Cela représente un problème puisque la dimensionnalité des données peut dépasser le nombre d'échantillons disponibles. Il est donc nécessaire de réaliser une sélection sur les OTU à inclure dans l'analyse statistique, par exemple en ignorant les OTU faiblement représentés. Pour ce travail, on a choisi de ne garder que les OTU identifiés dans au moins 50% des échantillons.

1.2.3 Alignement et production de la table de compte pour les données E. coli

Dans le cas de l'alignement des reads du jeu de données Coli, nous disposons de 56 génomes de référence et la stratégie "closed reference" a été utilisée pour affecter chaque read à une souche. L'idée de cette stratégie est d'inférer, à partir d'un ensemble de reads, la composition de l'échantillon à analyser. Il s'agit d'une tâche difficile dans la mesure où :

— Les souches présentes dans l'échantillon peuvent tout à fait ne pas faire partie de la liste des 56 souches de référence.
— Lors de l'alignement, il arrive fréquemment que des reads mappent sur plusieurs génomes de références. Ce comportement n'est pas surprenant dans la mesure où ces génomes de référence sont très proches génétiquement (ils concernent tous la même espèce) et l'aligneur doit nécessairement tolérer les erreurs qui peuvent être des mutations ponctuelles ou des erreurs de séquençage.

Il faut donc ici choisir une règle d'affectation pour traiter ces cas ambigus. Nous avons pour cela testé trois stratégies :

— La plus naïve consiste à éliminer les reads qui s'alignent sur plus d'un génome de référence. Cette méthode n'est pas réellement satisfaisante puisqu'elle revient à éliminer de nombreux reads, ce qui peut avoir des conséquences sur l'analyse statistique réalisée en aval. En effet, seule une faible proportion des reads ne s'alignent que sur un seul génome de référence.
— Une stratégie plus élaborée consiste à affecter de manière équitable chaque read à l'ensemble des génomes de référence sur lesquels il s'aligne. Si, par exemple, un read s'aligne sur deux génomes, on affecte 1/2 read à chaque génome.
— Enfin, il est possible d'être encore plus précis dans la répartition des reads multiples en procédant dans un premier temps à l'affectation des reads uniques. Puis, les reads multiples sont affecté au prorata des proportions observées des reads uniques (et non uniformément comme dans la stratégie précédente). Ainsi, on utilise les proportions observées de reads unique comme estimateur de la proportion vraie.

Pour ce travail de thèse, nous nous sommes limités à une comparaison entre ces trois stratégies, après en avoir réalisé l'implémentation en langage python. D'autres sont explorées dans la littérature. On peut citer, à titre d'exemple, les approches basées sur les fréquences d'observation de k-mer [52], l'utilisation de gènes spécifiques marqueurs de certaines branches [58], ou des méthodes probabilistes [1].

Pour les données de métagénomique E. coli, nous disposons d'une table de compte similaire à celle des données 16S, à ceci près que la liste des OTU est remplacée par la liste des 56 souches de référence. Dans la mesure où l'on dispose des génomes complets de ces 56 souches, il est aisé de construire un arbre phylogénétique qui décrit la proximité génétique entre ces souches. Ainsi, d'un point de vue formel, le type de donnée qui servira de base à l'analyse statistique est identique entre les données issues de l'analyse 16S et celles issues de l'analyse de la métagénomique E. coli. La description de méthodes applicables pour décrire ce type de données et étudier des associations avec des variables d'intérêt fait l'objet de la section 2.

2 Les méthodes statistiques "distance-based"

Dans cette section, on va s'intéresser au traitement statistique que l'on peut réaliser à partir des données générées par l'analyse 16S ou l'analyse métagénomique E. coli. Ces données présentent plusieurs caractéristiques qui vont influencer le choix des méthodes statistiques utilisées pour les analyser. Il s'agit de données

- *de comptage* : le nombre de reads est un nombre entier. Il est parfois très grand (de l'ordre de plusieurs milliers), mais peut également valoir 0. Ces données sont fréquemment surdispersées par rapport à une loi de Poisson et nécessitent des méthodes d'inférence adaptées
- *de grande dimension* : pour chaque échantillon, on observe des comptes pour chaque OTU (environ 800) ou chaque souche de référence E. coli (il y en a 56). Ici encore, des méthodes spécifiques de réduction de la dimensionnalité sont nécessaires.
- *hiérarchisées* : l'arbre phylogénétique donne une information sur la distance génétique entre deux OTU (ou deux souches pour l'analyse métagénomique). Cette information ne doit pas être ignorée. En effet, la partie expérimentale et bioinformatique qui a conduit à la table de compte peut générer des erreurs de classements entre reads. Ces erreurs sont plus probables entre souches génétiquement proches de sorte qu'il existe une incertitude sur les valeurs des comptes. La prise en compte de la structure phylogénétique est donc susceptible d'apporter de la robustesse vis-à-vis de ces erreurs et ainsi d'améliorer la puissance statistique des tests d'associations par la réduction d'un biais de classement.

Dans les parties suivantes, on va décrire plus précisément la notion de distance entre échantillons (section 2.1) et décrire les méthodes statistiques qui exploitent les distances entre échantillons (section 2.2-2.4).

2.1 Les distances entre échantillon

Une fois normalisées, les données de microbiote intestinal se résument, pour chaque patient, à un vecteur dont les composantes donnent l'abondance relative de chaque OTU dans l'échantillon considéré. Afin de décrire ces données, il est utile de définir une distance entre deux échantillons qui permet de mesurer si ces échantillons sont semblables dans leur composition microbienne ou bien très différents.

En mathématiques, on appelle distance une fonction de deux paramètres et à valeurs réelles d qui vérifie les quatre propriétés suivantes :

- La positivité : $\forall(x,y), d(x,y) \geq 0$
- La séparabilité : $\forall(x,y), d(x,y) = 0 \Rightarrow x = y$
- La symétrie : $\forall(x,y), d(x,y) = d(y,x)$
- L'inégalité triangulaire $\forall(x,y,z), d(x,y) \leq d(x,z) + d(z,y)$

Cette dernière propriété traduit simplement le fait que la longueur du chemin direct entre

deux échantillons est toujours plus petite que celle d'un chemin "détourné" qui passe par un troisième échantillon. Lorsque cette dernière propriété n'est pas satisfaite, mais que toutes les autres le sont, on parle alors de dissimilarité et non de distance. Ce sera, par exemple, le cas de la dissimilarité de Bray-Curtis, présentée à la section 2.1.2.

2.1.1 Les distances usuelles

Un première manière de définir la distance entre deux échantillon est d'étendre la distance naturelle (celle que l'on peut mesurer avec une règle graduée) aux espaces de dimensions plus grande que 2 ou 3. Plus précisément, si on note p le nombre d'espèces microbiennes étudiées (ou d'OTU en cas de données provenant de séquençage de l'ARN 16S), si l'on dispose de deux échantillons $x = (x_1, ..., x_p)$ et $y = (y_1, ..., y_p)$, on définit la distance Euclidienne entre les échantillons x et y, notée $d_2(x, y)$ par la formule suivante :

$$d_2(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Il existe d'autres distances couramment employées sur l'espace $\mathbb{R}^p$ sont les distances, notées d_1 et d_∞ dont la définition est :

$$d_1(x, y) = \sum_{i=1}^{p} |x_i - y_i|$$

$$d_\infty(x, y) = \sup_{i \in [\![1,p]\!]} |x_i - y_i|$$

Ces distances usuelles sont toutes calculées à partir du vecteur des différences entre compte pour chaque espèce. La différence entre elles provient de la contribution relative des grandes différences par rapport aux petites. Pour la distance d_∞, seul compte l'espèce pour laquelle la différence entre comptes est la plus grande. A l'inverse, pour la distance d_1, toutes les différences entre comptes sont additionnées pour contribuer à la distance globale. Pour la distance d_2 enfin, le carré dans la formule favorise les fortes différences entre compte, ce qui en fait un compromis entre les deux dernières distance dans le poids donné aux grandes différences relativement aux petites. Pour être totalement exhaustif dans la description des distances usuelles sur $\mathbb{R}^p$, il faudrait sans doute mentionner les distances de Hölder, dont les trois exemple ci-dessous ne sont que des cas particulier, mais cela sortirait du cadre de cette thèse.

Il est possible d'apporter des modifications aux distances usuelles, en particulier de pondérer les p espèces bactériennes a priori. Cela permet, pourvu que les poids soient positifs, de privilégier certaines espèces par rapport à d'autres tout en conservant les propriétés d'une distance. Concrètement, si on se donne un vecteur de poids $w = (w_1, ..., w_p) \in \mathbb{R}^p$ tel que, pour tout i, $w_i \geq 0$ et $\sum_i w_i = 1$, la distance Euclidienne

"pondérée" s'écrit alors

$$d_{2,w}(x,y) = \sqrt{\sum_{i=1}^{p} w_i (x_i - y_i)^2}$$

2.1.2 Les distances prenant en compte la phylogénie

Si on utilise l'une des distances précédemment décrites dans le cadre de l'analyse des table de compte bactériens, on ignore totalement l'information concernant la phylogénie sous-jacente. Deux échantillons comportant chacun une espèce différente seront jugés aussi "éloignés" l'un de l'autre si les deux espèces sont proches génétiquement que si elle sont très différentes. Cette perte d'information est préjudiciable et peut se traduire par une perte de puissance statistique. Ainsi, de nouvelles distances prenant en compte la phylogénie qui structure la liste des OTU ont été proposées.

La plus populaire de ces distance est sans doute la distance Unifrac [35–37], qui existe sous de nombreuses variantes : non-pondérée, pondérée, à variance corrigée, généralisée. Cette distance utilise l'information contenue dans un arbre phylogénétique dont le nombre de feuille correspond au nombre d'OTU p dans la table de compte. Contrairement aux distances usuelles décrites dans la section précédente, la distance Unifrac est une somme portant sur les branches de l'arbre (qui sont au nombre de $2p-1$ pour un arbre enraciné). Chaque branche, indexée par l'indice i, possède une longueur notée b_i. On note $L = \sum_i b_i$ la longueur totale de l'arbre. On note respectivement p_i^x et p_i^y les proportions de comptes de l'échantillon x (resp. l'échantillon y) situés en aval de la branche considérée.

La première des distance Unifrac a être proposée fut la distance dite "non-pondérée" :

$$d_U(x,y) = \frac{1}{L} \sum_{i=1}^{2p-1} b_i |\mathbb{1}_{p_i^x > 0} - \mathbb{1}_{p_i^y > 0}|$$

La distance entre les échantillons x et y correspond, selon cette formule, à la proportion de longueur de l'arbre spécifique à un des deux échantillons. Comme d_U représente une proportion, on a toujours $d_U(x,y) \leq 1$, ce qui n'est pas une propriété nécessaire à la définition d'une distance. On remarquera d'ailleurs que la terminologie "distance" est inappropriée pour cette fonction car la propriété de séparation n'est pas vérifiée. Pour être parfaitement précis, il faudrait mentionner que, si l'on transforme la table des comptes en matrice de 0 et 1 selon la règle suivante :

— les comptes à 0 restent à 0

— les comptes strictement positifs sont fixés à 1,

alors la distance Unifrac non pondérée est bien une distance, au sens de la définition donnée ci-dessus. Cette méthode de seuillage forte permet de s'affranchir des problèmes de normalisation de la table de compte : seuls comptent les entrées pour lesquelles le compte est strictement positif. Ce comportement peut malgré tout poser problème pour les arbres relativement petits ou bien pour des échantillons particulièrement riches : dans

ces conditions, tous les échantillons seront à distance 0 les uns des autres et on a perdu la totalité de l'information contenue dans la table de compte.

Pour contourner cette difficulté et incorporer dans la distance l'information relative à l'abondance des différentes espèces, la distance Unifrac pondérée a été proposée :

$$d_W(x,y) = \frac{1}{Z_{xy}} \sum_{i=1}^{2p-1} b_i |p_i^x - p_i^y|,$$

où Z_{xy} est un facteur de normalisation qui assure que $d_W(x,y) \leq 1$, soit

$$Z_{xy} = \sum_{i=1}^{2p-1} b_i (p_i^x + p_i^y)$$

Remarquons que ce facteur, tel qu'il est défini dans [36], n'est pas le seul choix possible. On aurait pu utiliser L ou bien la somme des $\min(1, (p_i^x + p_i^y))$, le premier choix étant le plus simple du point de vue de l'effort de calcul, puisqu'il ne dépend pas du couple d'échantillon (x,y) dont on souhaite calculer la distance. Le second donne une constante de normalisation plus petite que celle proposée, ce qui augmente l'étendue des valeurs possibles pour $d_W(x,y)$. Dans la suite de cette thèse, on adoptera la normalisation standard par Z_{xy}.

Il est important de noter que la distance d_W fait intervenir la quantité $|p_i^x - p_i^y|$, qui peut varier continument entre 0 et 1. Ainsi, si deux échantillons diffèrent de 1% dans le compte d'un OTU, le poids que cette différence représenta dans la distance totale est plus ou moins important important en fonction de la fréquence moyenne de l'OTU considéré. Plus précisément, la même différence relative aura un poids plus important si elle concerne un OTU fréquent. Dans le cas de la distance non pondérée d_U, les OTU rares avaient le même poids que les OTU fréquents puisque c'est la quantité $|\mathbb{1}_{p_i^x>0} - \mathbb{1}_{p_i^y>0}|$ qui était utilisée. Si l'on souhaite qu'un écart relatif entre comptes de deux échantillons pour un OTU particulier ait un poids qui ne dépende pas de la fréquence de cet OTU, on utilisera la distance d_0 qui fait intervenir, pour chaque nœud de l'arbre, une différence relative (et non plus absolue). Cette distance est définie par :

$$d_0(x,y) = \frac{1}{L} \sum_{i=1}^{2p-1} b_i \frac{|p_i^x - p_i^y|}{p_i^x + p_i^y}$$

A partir des deux distances d_0 et d_W, il est possible d'introduire toute une famille de distances paramétrées par un réel $\alpha \in [0,1]$. On appellera donc distance Unifrac

généralisée la distance d_α définie par :

$$d_\alpha(x,y) = \frac{1}{Z'_{xy}} \sum_{i=1}^{2p-1} b_i (p_i^x + p_i^y)^\alpha \frac{|p_i^x - p_i^y|}{p_i^x + p_i^y}$$

avec une constante de normalisation

$$Z'_{xy} = \sum_{i=1}^{2p-1} b_i (p_i^x + p_i^y)^\alpha$$

Le terme distance Unifrac *généralisée* est justifié dès lors que l'on remarque que, pour $\alpha = 0$, on retrouve la distance d_0 avec la même normalisation. On remarquera également que d_W est retrouvée avec le choix $\alpha = 1$.

D'autres distances basées sur un arbre phylogénétiques ont été proposées. Dans [11], les auteurs proposent de résoudre le problème des poids relatifs des OTU rares et fréquents dans le calcul de la distance par une autre approche. Il définissent la distance Unifrac corrigée ou "variance-adjusted" de la manière suivante :

$$d_{VA}(x,y) = \frac{1}{Z''_{xy}} \sum_{i=1}^{2p-1} b_i \frac{|p_i^x - p_i^y|}{\sqrt{m(m-m_i)}},$$

où m_i et m désignent, respectivement, la somme des comptes pour les deux échantillons au nœud i et la somme totale $\sum_i m_i$. Le facteur de normalisation Z est construit de la même manière que pour les distances Unifrac généralisées, soit

$$Z''_{xy} = \sum_{i=1}^{2p-1} \frac{b_i}{\sqrt{m(m-m_i)}}$$

Pour cette dernière distance de la famille des distances Unifrac, l'idée est de normaliser l'écart $|p_i^x - p_i^y|$ par son écart-type sous l'hypothèse d'une répartition aléatoire et uniforme des comptes deux échantillons sur l'ensemble de l'arbre.

2.2 Les méthodes statistiques applicables au tables de compte

Dans cette thèse, on se propose d'analyser des données de microbiote intestinal par une méthode en deux étapes. La première consiste, à partir de la matrice des distances Unifrac entre échantillons, à reconstruire une table de "pseudo-comptes", de manière à ce que les distances Euclidiennes entre échantillon de "pseudo-comptes" soient les plus proches possibles des distances Unifrac entre échantillons. Une fois cette reconstruction faite, on insère les pseudo-comptes comme prédicteurs dans une régression linéaire ou logistique standard.

Il s'agit donc d'utiliser des méthodes de régression standard (linéaire ou logistique), mais en leur fournissant en entrée non pas la table de compte brute, mais une table transformée de façon à rendre semblables (au sens de la distance Euclidienne) les colonnes correspondant échantillons semblables (au sens de la distance Unifrac). La méthode est semblable à celle qui permet de produire des cartes dites "anamorphose" qui représentent une région ou un territoire, mais dans lesquelles la distance usuelle (en km) est remplacée par une autre distance (par exemple le temps de transport). La figure 2.1 montre l'exemple d'une telle carte de France.

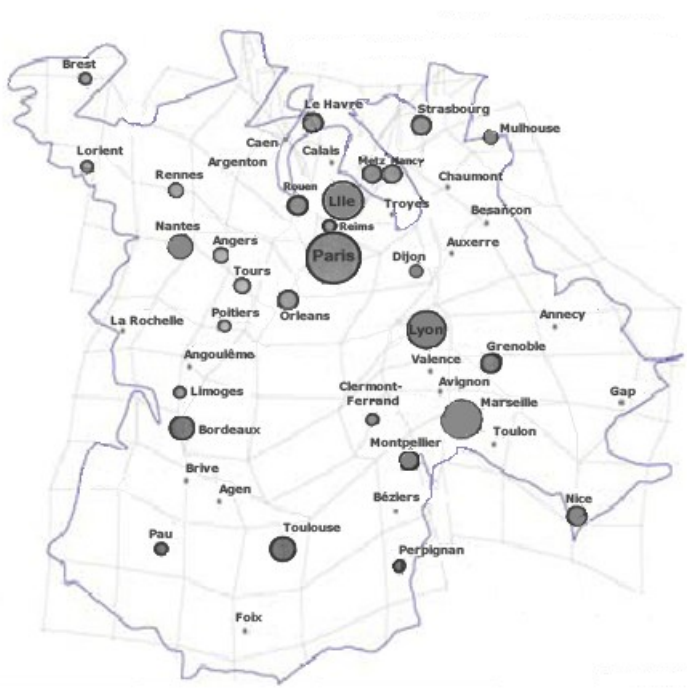

FIGURE 2.1 – Représentation anamorphique de la France selon la distance "temps de parcours en train" (source SNCF)

Dans cette section, on commencera par un rappel sur la méthode de positionnement multidimensionnel (MDS) qui permet de fournir une transformation anamorphique. On traitera d'abord le cas Euclidien puis plus général au paragraphe 2.3. On décrira ensuite comment incorporer les pseudo-comptes obtenus dans un modèle linéaire ou logistique au paragraphe 2.4 et on finira par évoquer les autres méthodes disponibles dans la littérature qui utilisent les distances Unifrac pour tester l'association entre microbiote intestinal et des variables d'intérêt dans le paragraphe 2.5.

2.3 Le positionnement multidimensionnel (MDS)

Le positionnement multidimensionnel ou (MDS pour multidimensional scaling) est une méthode statistique descriptive qui appartient à la famille des méthodes de réduction de dimensionalité, tout comme l'analyse en composantes principales ou la famille des analyses factorielles. Brièvement, l'analyse en composantes principales est une méthode qui travaille à partir d'une matrice de données X de taille (n, p), où n représente le nombre d'échantillons et p le nombre de dimensions mesurées pour chaque échantillon. L'idée est de travailler sur la matrice de Gram $G = X^t X$, qui est une matrice carré de taille (n, n) qui peut s'interpréter comme une matrice de variance-covariance pour peu que les données de départ soient centrées. Les valeurs propres de G sont nécessairement positives et ses vecteurs propres définissent alors, si on les ordonne par ordre décroissant des valeurs propres, des axes orthogonaux qui permettent de projeter les données initiales (ou des données complémentaire). Le premier axe a la propriété que la projection du nuage de point initial sur celui-ci maximise la dispersion le long de cet axe et ainsi de suite pour les axes successifs.

Par rapport à l'analyse en composantes principales, la méthode MDS travaille à partir d'une matrice de distance entre échantillons, donc une matrice carré de taille (n, n). Lorsque la distance considérée est la distance Euclidienne sur $\mathbb{R}^p$, la méthode coïncide avec l'ACP standard. Mais l'intérêt véritable du MDS est de pouvoir utiliser des distances non standard et adaptées au problème statistique posé

2.3.1 Le cas Euclidien

Pour fixer les idées, on commencera par décrire la méthode MDS dans le cas de la distance Euclidienne. A partir d'une matrice de données $X \in M_{n,p}(\mathbb{R})$, on construit la matrice carrée $D(X) \in M_n(\mathbb{R})$ des distance entre échantillons. On remarque alors trois propriétés de la matrice D qui découlent de cette construction :

- Les coefficients de la matrice D sont tous positifs
- Les coefficients diagonaux sont tous nuls
- La matrice D est symétrique
- Si on choisis trois indices i, j, k, alors on a l'inégalité triangulaire $D_{ij} \leq D_{ik} + D_{kj}$

On peut à présent se poser la question de savoir si, à partir de la matrice des distances $D(X)$, il est possible de reconstruire les données qui ont servi à la calculer. La réponse est en fait négative car la matrice $D(X)$ est invariante si le nuage de point subit une transformation rigide (translation, réflexion ou rotation). En revanche, il est possible de reconstruire un nuage de points Y tel que X et Y soient images d'un de l'autre par une transformation rigide.

La procédure à suivre pour cela est la suivante : on commence par reconstruire la matrice de Gramm $G = X^t X$ à partir de la matrice $D(X)$. Ce passage est possible en raison d'une identité qui n'est valable que dans le cas où c'est la distance Euclidienne qui a servi à construire la matrice de distance $D(X)$. On introduit pour cela les notations

suivantes :

$$v = (1, ..., 1) \in \mathbb{R}^n, \qquad H = I_n - \frac{1}{n} v^t v$$

Géométriquement, H est une matrice de projection orthogonale sur l'hyperplan orthogonal à la droite $\mathbb{R}(1, ..., 1)$. On a alors l'identité

$$G = -\frac{1}{2} H D(X)^{*2} H$$

où la notation $D(X)^{*2}$ désigne la matrice dont les coefficients sont les carrés des coefficients de $D(X)$. Une propriété intéressante de la matrice de Gramm est que son rang est égal à celui de X, c'est-à-dire p si l'on considère que la matrice X est de rang maximal.

On achève la reconstruction à l'aide d'une décomposition spectrale standard. La matrice de Gram possède p valeurs propres strictement positives que l'on note $(\lambda_1, ..., \lambda_p)$ avec $\lambda_1 \geq ... \geq \lambda_p > 0$. Si on note $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p, 0, ..., 0)$ et P, la matrice des vecteurs propres associés, on a alors

$$G = P \Lambda^t P$$

Il s'agit maintenant d'écrire $G = Y_p^t Y_p$, avec $Y_p \in M_{n,p}(\mathbb{R})$. Pour cela, on se donne maintenant un entier positif $r \leq p$ et on introduit la matrice

$$J_r = \begin{pmatrix} I_r \\ 0_{n-r,p} \end{pmatrix} \in M_{n,r}(\mathbb{R})$$

de sorte que $Y_p = P \Lambda^{1/2} J_p$ vérifie bien $G = Y_p^t Y_p$.

Enfin, l'égalité $G = X^t X = Y_p^t Y_p$ assure qu'il existe une transformation rigide qui envoie X sur Y_p. On a donc bien reconstruit la matrice initiale X, à transformation rigide près.

Avec p dimensions, on obtient une reconstruction parfaite. Si on se donne un entier positif $r \leq p$, on peut poser, de la même manière $Y_r = P \Lambda^{1/2} J_r$. On a alors la remarquable propriété que la matrice de distance $D(Y_r)$ est la meilleure approximation de $D(X)$ réalisable avec une matrice de taille n, r. En d'autres termes :

$$Y_r = \underset{Y \in M_{n,r}(\mathbb{R})}{\text{Arg min}} \, \|D(X) - D(Y)\|_2^2,$$

où $\|M\|_2^2$ désigne la somme des carrés des coefficients de M. Dans la littérature autour du MDS, on utilise plutôt le stress $S(Y)$ défini comme

$$S(Y) = \frac{\|D(X) - D(Y)\|_2^2}{\|D(X)\|_2^2}$$

On a, bien entendu, également

$$Y_r = \underset{Y \in M_{n,r}(\mathbb{R})}{\text{Arg min}}\, S(Y).$$

De plus, la quantité $S(Y_r)$ décroit à mesure que r augmente, ce qui signifie que la précision de la reconstruction s'améliore si on s'autorise un nombre de dimension supérieur.

Ainsi, on dispose, dans le cas Euclidien, d'une méthode pour reconstruire, à transformation rigide près, un nuage de point à partir de la matrice des distances deux à deux. On peut également approximer ce nuage en dimension inférieure et mesurer la qualité de la reconstruction. Malheureusement, lorsque la matrice $D(X)$ est construite à partir d'une distance autre que la distance Euclidienne, un certain nombre de propriétés deviennent fausses.

2.3.2 Le cas général

Dans cette partie, on s'intéresse au cas plus complexe où la matrice $D(X)$ est construite à partir d'une distance quelconque et non nécessairement Euclidienne. C'est le cas, par exemple, de la famille des distances Unifrac. Le problème du MDS se formule alors de la manière suivante : étant donné une matrice D et un entier positif r, existe-t-il une configuration de points $Y_r \in M_{n,r}(\mathbb{R})$ telle que la matrice des distances Euclidiennes entre ces points $D(Y_r)$ approxime le mieux la matrice des distances initiales D ? La méthode décrite dans le cas Euclidien permet de répondre approximativement à cette question.

Les différences avec le cas Euclidien sont de trois ordres :

— La matrice $G = -1/2HD^{*2}H$ (appelée matrice de Gower dans ce cas) n'est plus nécessairement positive. Ainsi, on ne pourra reconstruire $Y_r = P\Lambda^{1/2}J_r$ que pour $r \leq n_+$, où n_+ désigne le nombre de valeurs propres positives de G.

— Par conséquent, une reconstruction parfaite n'est pas toujours possible. Cela n'est en réalité pas un inconvénient dans le contexte d'une analyse statistique puisque, en pratique, c'est surtout la possibilité de reconstruire une approximation dans un espace de faible dimension qui est utile.

— Enfin, la qualité de la reconstruction, mesurée par le stress $S(Y_r)$, n'est plus nécessairement monotone quand on augmente le nombre de dimensions r. Le plus souvent, cette qualité commence par s'améliorer lorsque r est petit, passe ensuite par un minimum avant de se dégrader lorsque r devient plus grand.

Ainsi, on dispose d'une méthode qui permet, à partir de la table de comptes et d'une distance Unifrac, d'obtenir une table de pseudo-comptes dont les distances Euclidiennes approximent les distances Unifrac entre échantillons. On va maintenant décrire comment inclure ces pseudo-comptes dans un modèle linéaire ou logistique standard.

2.4 Les "pseudo-comptes" commes prédicteurs dans un modèle linéaire ou logistique

En statistiques, le modèle linéaire et ses généralisations occupent une place importante dans l'arsenal des méthodes permettant de décrire une association entre variables explicatives et une variable d'intérêt. Concrètement, si on note Y le vecteur de la variable dépendante de taille n et $X \in M_{n,p}(\mathbb{R})$ la matrice des variables explicatives (fréquemment augmentée d'une colonne constante égale à un), l'idée est de construire à partir de X un prédicteur linéaire $X\beta$ en multipliant par un vecteur β de coefficients (appelés effets fixes). Ce prédicteur représente, par l'intermédiaire d'une fonction de lien f, l'espérance de la variable dépendante Y.

$$\mathbb{E}(Y) = f(X\beta)$$

Dans le cadre d'une variable Y continue, on choisit $f = Id$ et on obtient le modèle linéaire. Si Y est une variable binaire, on choisit le lien logit et on obtient le modèle logistique. Il existe d'autre possibilités de couples loi de Y/fonction de lien, mais ces autres cas sortent du cadre de cette thèse.

En pratique, on a pas accès à l'espérance de la variable Y, mais uniquement à certaines réalisations, dont les observations représentent un vecteur également noté Y, observé simultanément avec les covariables X. Si l'on suppose qu'il existe une variable aléatoire ϵ, de loi normale centrée, de variance σ^2 et indépendante de X telle que

$$Y = f(X\beta + \epsilon), \tag{1}$$

alors on peut estimer, par la méthode du maximum de vraisemblance, les paramètres $\hat{\beta}$ et $\hat{\sigma}^2$ qui rendent les observation (X, Y) les plus vraisemblables, à supposer que le modèle (1) est le bon.

2.4.1 Test d'hypothèses

A ce stade, le problème est alors de tester la significativité statistique de ces coefficients. L'absence d'influence d'une composante X_k parmi les covariables sur l'espérance de Y se traduit par la relation $\beta_k = 0$. En pratique, on se pose donc la question de savoir si l'estimateur $\hat{\beta}_k$ est suffisamment éloigné de 0 pour rejeter l'hypothèse nulle $\beta_k = 0$ à un niveau de confiance donné. Plus généralement, on pourrait vouloir tester simultanément l'influence de plusieurs covariables $(X_{k_1}, ..., X_{k_r})$. Pour cela, nous utiliserons dans cette thèse deux statistiques de test :

— La statistique de Wald

— La statistique du rapport de vraisemblance

La statistique de Wald consiste à normaliser l'estimateur par son écart-type. A noter qu'un estimateur de la variance de $\hat{\beta}_k$ est fourni par la méthode du maximum de

vraisemblance. Ainsi, on utilise la statistique de test

$$z_k = \frac{|\hat{\beta}_k|}{\sqrt{\mathrm{Var}\hat{\beta}_k}},$$

que l'on compare à 1.96 si on souhaite un risque alpha standard à 5%. Une possibilité équivalente est de comparer z_k^2 au seuil 3.84, qui correspond au 95ième percentile d'un chi-deux à un degré de liberté. L'avantage de cette dernière formulation est qu'elle peut être généralisée au cas où on souhaite tester simultanément plusieurs covariables. Si on choisit des indices $(k_1, ..., k_r)$ et que l'on note V la matrice de variance-covariance des estimateurs $\hat{\beta}_k = (\hat{\beta}_{k_1}, ..., \hat{\beta}_{k_r})$, alors la statistique

$$z_k^2 =^t \hat{\beta}_k V^{-1} \hat{\beta}_k$$

est distribuée, sous l'hypothèse nulle et pour une taille d'échantillon très grande, selon un chi-deux à r degrés de liberté, ce qui permet d'obtenir une valeur seuil pour rejeter l'hypothèse nulle $\beta_k = 0$.

Dans cette thèse, on utilisera cette version multivariée du test de Wald sur le vecteur des coefficients associés aux pseudo-comptes. Comme la matrice des pseudo-comptes est construite à une rotation/réflexion près, il serait intéressant que la statistique de test z_k^2 soit elle aussi invariante par rotation/réflexion de la matrice des covariables. C'est heureusement le cas.

Cette statistique de Wald multivariée peut également se construire d'une autre manière. Si $\hat{\beta}_k$ est un vecteur de coefficients à tester et admet V comme matrice de variance-covariance. Si on se donne un vecteur de contraste c, alors la loi de ${}^t c\hat{\beta}_k$ est, sous H_0, une loi normale centrée dont la variance est ${}^t cVc$. On peut alors tenter de trouver une direction c_0 optimale au sens où

$$c_0 = \mathrm{Arg}\max_c \frac{{}^t c\hat{\beta}_k}{\sqrt{{}^t cVc}}$$

Lorsque l'on résout ce problème d'optimisation, on trouve $c_0 = V^{-1}\hat{\beta}_k$ comme direction optimale et on retrouve la statistique de test de Wald z_k. Finalement, le test de Wald multivarié est un test de Wald univarié appliqué à la direction la plus favorable parmi les covariables testées.

Une autre manière de tester la nullité de certains coefficients est d'utiliser le test du rapport de vraisemblance (LRT). Ce test utilise la propriété que, sous l'hypothèse nulle $\beta_k = 0$, la quantité

$$D = 2\log\left(\frac{L_{\mathrm{full}}}{L_{\mathrm{red}}}\right)$$

est distribuée, sous l'hypothèse nulle et pour une taille d'échantillon très grande, selon

un chi-deux à r degrés de liberté, si L_{full} désigne la vraisemblance du modèle complet et L_{red} celle du modèle ou on a enlevé les covariables testées. Tout comme la statistique de Wald, la statistique du rapport de vraisemblance est invariante si les covariables testées sont transformées par rotation et/ou réflexion.

Au total, ces deux statistiques de test (Wald et LRT) permettent d'évaluer l'influence d'une table de pseudo-comptes sur l'espérance d'une variable d'intérêt tout en respectant le fait que la table de pseudo-comptes n'est connue qu'à une transformation rigide près.

2.5 Autres méthodes statistiques prenant en compte la phylogénie

Récemment, d'autres méthodes que celle présentée ci-dessus ont été proposées dans la littérature pour utiliser l'information que représente la phylogénie pour analyser une table de compte. On en liste ici deux à titre d'exemple.

2.5.1 PERMANOVA

La méthode PERMANOVA [2] généralise l'analyse de variance ANOVA classique, qui permet de comparer deux ou plusieurs groupes quant à la valeur d'une variable quantitative. Au lieu de comparer les carrés des distances intra-groupe aux carrés des distances inter-groupe dans une décomposition classique de la variance globale, on compare les distance inter-échantillons au sein d'un groupe aux distances inter-échantillons globales. On calcule ainsi un rapport qui peut être comparé à une statistique de Fisher.

Lorsque la distance est Euclidienne, la somme des distances entre échantillons à l'intérieur d'un groupe est égale à la somme des distances au centroïde de ce groupe et la statistique suit exactement à une loi de Fisher. En revanche, l'utilisation de distances non Euclidienne fait que la loi de la statistique de test sous l'hypothèse nulle n'est pas connue. La p-value est donc obtenue en permutant les labels de groupes aléatoirement et en recalculant à chaque fois la statistique de test. On obtient ainsi une distribution empirique, dont on dérive une p-value empirique.

2.5.2 MiRKAT

La méthode MirKAT [74], acronyme pour Microbiome Regression-Based Kernel Association Test, est une méthode dite "à noyau", construite sur le modèle de SKAT en génétique humaine. Le point de départ est le modèle de régression logistique suivant,

$$\text{logit}(P(Y_i = 1)) = \alpha + \beta' Z + \beta X_i$$

dans lequel Z désigne les éventuelles covariables et X_i la i-ième colonne de la matrice de comptes X. Dans ce modèle, on suppose que les composantes du vecteur β sont des effets aléatoires de moyenne nulle et de variance τw_j. Dans cette situation, le test $\beta = 0$ correspondant à l'hypothèse nulle d'absence d'association entre la variable d'intérêt Y

et la composition du microbiote revient finalement à tester $\tau = 0$. Ce genre de test est appelé "variance-component test". La statistique est la suivante :

$$Z = \frac{1}{2}\, {}^t(y - \hat{y})G(y - \hat{y})$$

où y et $\hat{y}$ désignent respectivement la valeur et l'estimation sous H_0 de la variable expliquée. Sous l'hypothèse que la matrice G soit positive (on a vu que cela pouvait nécessiter de tronquer les valeurs propres négatives), Z suit asymptotiquement une loi qui est un mélange de chi-deux, ce qui permet de dériver une p-value de manière quasiment analytique (voire [13, 15, 34] pour les méthodes applicables à ce cas), ce qui représente un avantage en termes de temps de calcul par rapport à la méthode PERMANOVA dont la p-value est obtenue par permutations.

Afin de contourner le problème du choix de la distance (et donc du noyau G) à utiliser, les auteurs proposent un test ad-hoc, où plusieurs distances sont testées et seule la distance donnant la meilleure p-value est retenue. Dans ce cas en revanche, une méthode de permutation est bien sûr nécessaire afin de contrôler l'erreur de type I. Cette méthode donne une meilleure puissance lorsqu'aucune hypothèse a priori ne peut être faite sur la distance la plus judicieuse.

3 Résultats

Cette section est dédiée à la description des tables de comptes issues du pipeline 16S et E. coli, ainsi qu'aux résultats de l'analyse de l'association entre table de compte et trois phénotypes disponibles : âge, sexe et indice de masse corporelle.

3.1 Sélection des individus et des OTU

Le pipeline bioinformatique "open reference" permet d'identifier 111273 OTU. En réalité, beaucoup de ces OTU ne sont observées que dans un petit nombre d'échantillons ($N = 43372$ pour les OTU observés que dans un échantillon, $N = 17614$ pour deux échantillons). La table de compte brute est donc extrêmement éparse avec 97.5% de comptes nuls. On a donc sélectionné pour l'analyse les 842 OTU détectés avec un compte strictement supérieur à 0 dans au moins 50% des échantillons et ignoré les autres. Ainsi, l'arbre phylogénétique associé à l'analyse 16S comporte 842 feuilles correspondant à ces 842 OTU. Il est représenté dans la figure 3.1 et fait apparaitre les deux principaux phyla bactériens (Firmicutes et Bacteroidetes), ainsi que d'autres phyla minoritaires en nombre d'OTU.

Pour l'analyse de la composition en E. coli, la stratégie "closed reference" a été utilisée et les 56 génomes de référence sélectionnés sont observés avec des comptes strictement positifs dans l'ensemble des échantillons. La topologie de l'arbre phylogénétique correspondant est représenté dans la figure 3.2. On remarque que le panel ne contient qu'une seule souche appartenant au sous-type génétique F, alors que les autres sous-types génétiques sont plus équitablement représentés.

Sur les 12000 jumeaux composant la cohorte twinsUK, on dispose d'informations phénotypique chez 512 individus. Les femmes sont majoritaires ($N = 445$, 87%). L'âge moyen est de 60 ans avec un écart-type de 11 ans. L'indice de masse corporelle (BMI) moyen est de 26 kg/m^2 avec un écart-type de 5,2 kg/m^2.

Les échantillons de selles utilisés pour l'analyse de la composition en E. coli du microbiote sont disponibles pour 439 individus, alors que le jeu de données 16S concerne 411 individus. On voit donc que l'on a au départ deux jeux de données très différents : la table de compte 16S comporte un grand nombre de variables et un nombre plus faible d'échantillons tandis que la situation est inversée pour la table E. Coli.

Une approche naïve de modélisation directe de la table des comptes consisterait à supposer que ceux-ci suivent une loi de Poisson dont le paramètre serait à estimer. Cette hypothèse impose une relation entre la moyenne des comptes et leur dispersion, qui doivent être égales. Or, il est fréquent d'observer des phénomènes de comptage pour lesquels cette hypothèse n'est pas réaliste. Les figures 3.3 et 3.4 montrent, pour la table de compte de E. coli et de l'analyse 16S que nos données souffrent de ce phénomène de sur-dispersion puisque les points se trouvent largement au dessus de la première bissectrice. En l'état, une modélisation à partir d'une loi de Poisson ne serait pas appropriée

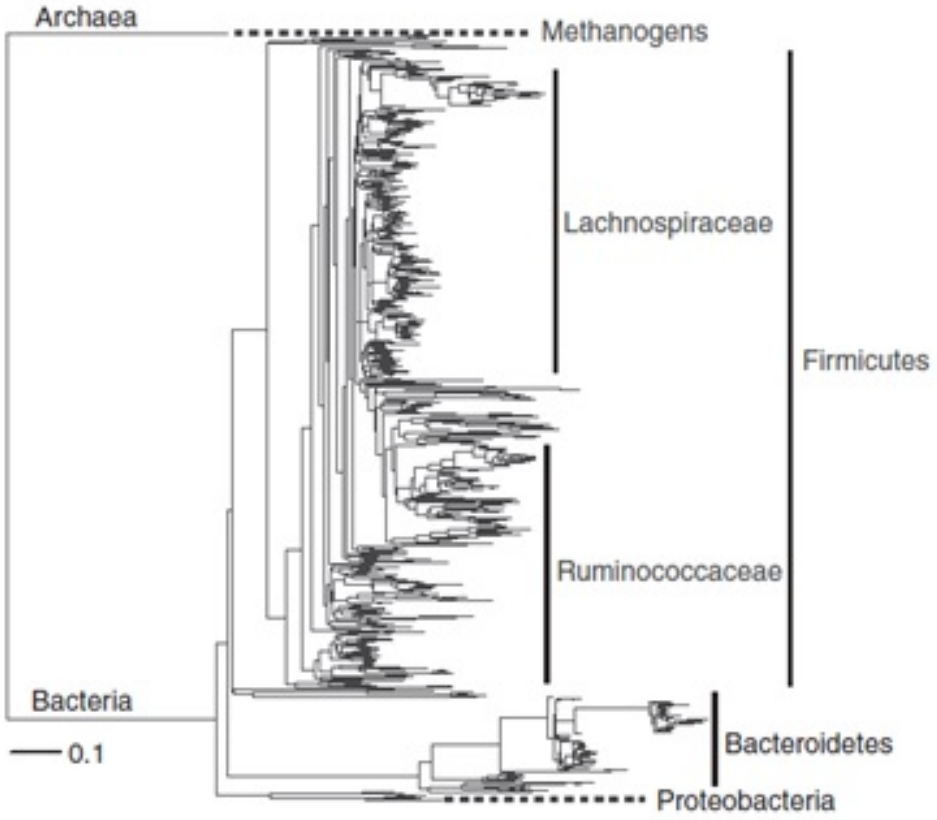

FIGURE 3.1 – Structure de l'arbre phylogénétique de l'analyse 16S (adapté de [25])

puisqu'elle mènerait à une forte inflation de l'erreur de type I des tests asymptotiques qui seraient réalisés. Pour contourner cette difficulté, d'autres approches plus complexes de modélisation des comptes ont été proposées dans [70], en s'appuyant sur un modèle multinomial par exemple.

3.2 Description des distances entre échantillons

On va maintenant décrire les distances entre échantillons pour les deux jeux de données. La figure 3.5 représente les distances Unifrac pondérées entre échantillons issues des données de la table de compte de E. coli. La figure 3.6 représente ces mêmes distances, mais issues de la table de compte 16S. On remarque que l'histogramme contient quelques distances nulles qui correspondent à la diagonale de la matrice des distances (un échantillon est à distance nulle de lui même).

Dans le cas du jeu de données 16S, la distribution des distance a une allure gaussienne

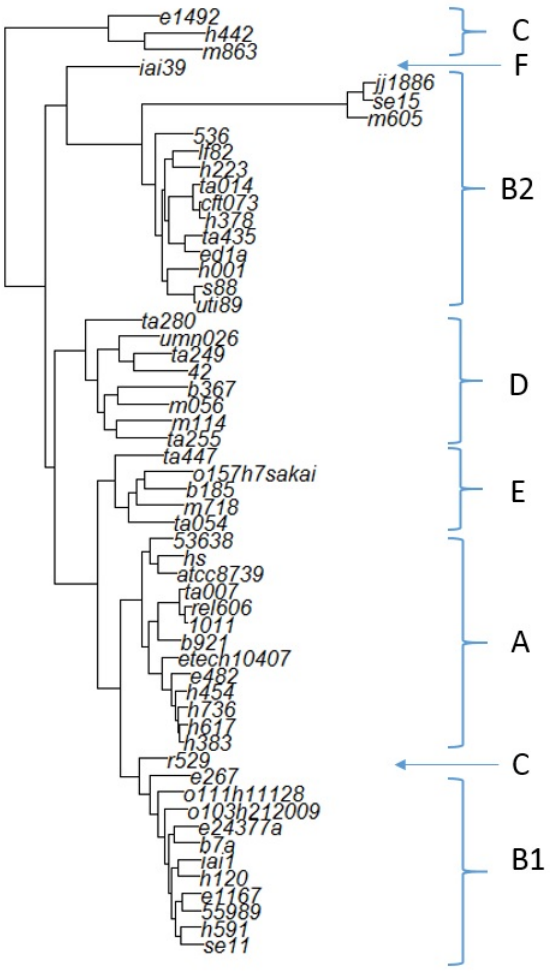

FIGURE 3.2 – Structure de l'arbre phylogénétique E. coli. Les lettres indiques le sous-type génétique de la souche

avec une faible dispersion autour de sa moyenne. Cette allure est caractéristique des données de grande dimension (on rappelle qu'il y a 842 OTU dans la table 16S). Le fait que la loi des distances entre échantillons soit proche d'une loi normale découle du théorème central limite et du fait que la distance Unifrac s'écrit comme une somme qui porte sur les branches de l'arbre phylogénétique.

3.3 Qualité de la reconstruction MDS

On s'intéresse à présent à la reconstruction MDS de pseudo-compte selon la méthode décrite au paragraphe 2.3.2. Les figures 3.7 et 3.8 représentent,pour les jeux de données E. coli et 16S respectivement, la variation de l'erreur entre la matrice de distance initiale et

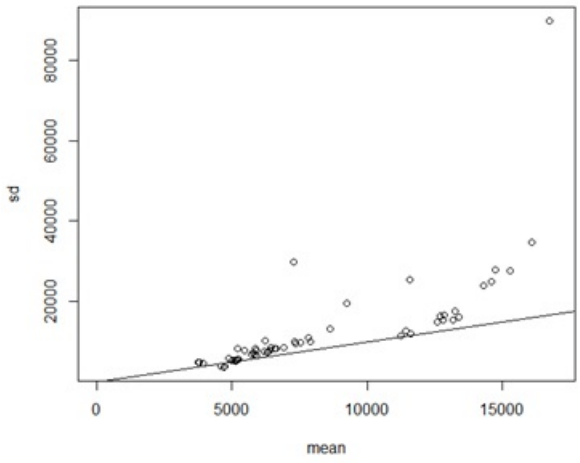

FIGURE 3.3 – Dispersion des comptes de E. coli en fonction de la moyenne observée ($n = 56$). La droite représente la relation attendue sous l'hypothèse que les comptes suivent une loi de Poisson

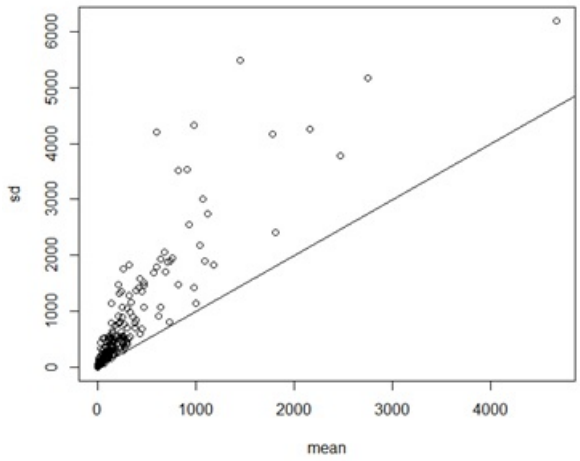

FIGURE 3.4 – Dispersion des comptes de l'analyse 16S en fonction de la moyenne observée ($n = 842$). La droite représente la relation attendue sous l'hypothèse que les comptes suivent une loi de Poisson

celle reconstruite en fonction du nombre de valeurs propres conservées. Pour des raisons de lisibilités, on a représenté que les 100 premières valeurs propres dans les deux figures, même si le nombre de valeurs propres positives était supérieur.

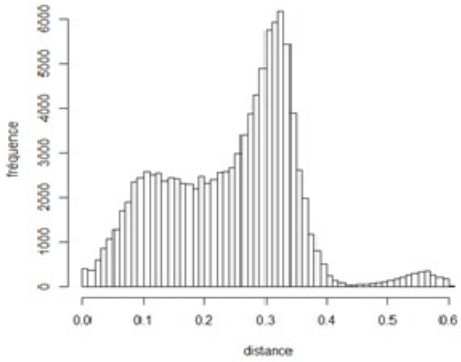

FIGURE 3.5 – Histogramme des distances Unifrac pondérées entre échantillons de E. coli

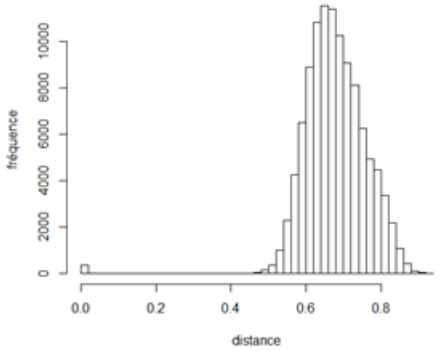

FIGURE 3.6 – Histogramme des distances Unifrac pondérées entre échantillons de l'analyse 16S

Pour la table E. coli, on constate que l'erreur commence par diminuer lorsque la dimensionnalité de la reconstruction augmente, puis finit par remonter au delà du seuil $p = 7$. Le même comportement est observé pour la table 16S, mais la reconstruction optimale est atteinte pour $p = 29$ dans ce cas.

Le fait que la qualité de la reconstruction ne soit pas monotone n'est pas paradoxal si on se rappelle que la distance Unifrac utilisée n'est pas Euclidienne. Ainsi, la matrice de Gower n'est pas positive et les résultats d'optimalité présentés au paragraphe 2.3.1 ne s'appliquent pas. Dans la suite, on retiendra les deux reconstruction optimales, c'est-à-dire avec un nombre de dimension de 7 et 29 pour les données E. coli et 16S respectivement.

Une autre manière d'illustrer cette optimalité est de former les diagrammes de She-

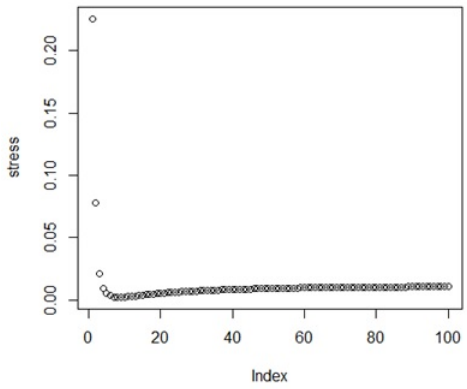

FIGURE 3.7 – Stress de la reconstruction des pseudo-comptes de E. coli

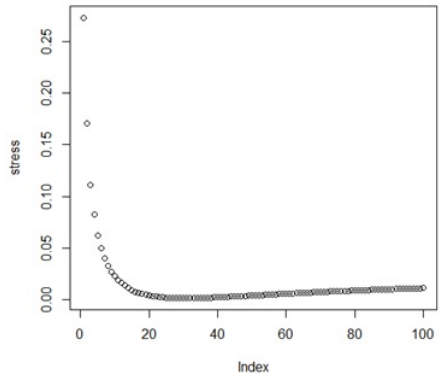

FIGURE 3.8 – Stress de la reconstruction des pseudo-comptes de 16S

pard (figure 3.9). Ces diagrammes représentent, pour une dimensionnalité fixée, les distances reconstruites en fonction des distances originales. Idéalement, les points devraient donc s'aligner le long de la première bissectrice. En pratique, on évalue visuellement la qualité de la reconstruction en estimant la dispersion du nuage de points autour de la première bissectrice. Une reconstruction parfaite correspond au cas où tous les points sont alignés sur la première bissectrice.

On constate que la dispersion est très importante lorsque l'on contrait la reconstruction sur une droite ($p = 1$), et qu'elle s'améliore pour $p = 7$. Dans le cas des données issues de la table E. coli, on constate également que prendre $p = 27$ n'améliore pas significativement la qualité de la reconstruction (en réalité, cela la dégrade légèrement). En revanche, pour les données 16S, le choix $p = 27$ donne un résultat qui, visuellement,

parait bien meilleur que le choix $p = 7$.

3.4 Tests d'association avec les variables phénotypiques

L'association entre microbiote et l'indice de masse corporelle (BMI) a été testée en adoptant un codage quantitatif pour le BMI et un test LRT. On retrouve une association significative pour la table 16S (p=0.0017) et non significative pour les données E. coli (p=0.22). Pour l'âge et toujours avec un codage quantitatif, l'association est également significative pour la table 16S (p=0.0011) et non significative pour les données E. coli (p=0.86). Enfin, l'association avec le sexe a le même profil avec p=0.026 pour les données 16S et p=0.44 pour les données E. coli.

Dans les figures 3.10, 3.11 et 3.12, on examine la sensibilité de ces conclusions lorsque l'on fait varier la dimensionnalité p. On a représenté en trait plein la médiane d'un chi-deux à p degrés de libertés et en traits pointillés la borne supérieure d'intervalle de confiance à 95%. On peut donc constater graphiquement si le test est significatif en regardant si la statistique de test est au dessus des traits pointillés. On constate que le jugement de significativité est relativement peu sensible au choix de la dimensionnalité, excepté pour des très faibles valeurs de p (pour lesquels la reconstruction MDS n'est pas d'une qualité satisfaisante) et pour les très grandes valeurs, pour lesquelles la puissance statistique diminue.

On remarque également que la statistique de test LRT est systématiquement plus grande que cette du test de Wald. On constate même des valeurs très conservatrices pour la statistique de Wald de l'association entre le sexe et les données E. coli qui s'expliquent par le fait que l'échantillon contient très majoritairement des femmes. De manière surprenante, l'ajout d'un degré de liberté peut parfois permettre de faire un saut important à la statistique de test, alors que la matrice des distances n'en est que très peu modifiée. Il est probable que l'ajout d'un degré de liberté autorise une configuration de points significativement plus fidèle bien que la matrice des distance ne varie presque pas.

Enfin, il est intéressant de noter que, malgré une dimensionnalité plus importante, la méthode utilisée donne un résultat significatif pour le jeu 16S et donc qu'elle arrive a extraire de manière efficace l'information contenu dans la matrice des distance. Pourtant, à la lecture des histogrammes 3.5 et 3.6, il aurait été légitime de penser que le jeu de données E. coli était le plus prometteur des deux.

3.5 Tests d'association croisée entre les données Coli et 16S

L'hypothèse de recherche qui a initialement motivé ce travail était de tester si la composition du microbiote en E. coli déterminait celle à l'échelle de toutes les bactéries. Pour tester cette hypothèse, on a utilisé la même méthode que dans la section précédente, mais en changeant la variable expliquée. La figure 3.13 (gauche) représente les résultats d'un test d'association entre la composition en E. coli et les trois premières composantes principales de reconstruction des pseudo comptes 16S. De manière symétrique, on a testé

(figure 3.13-droite) l'association entre la composition bactérienne et les trois premières composantes principales de reconstruction des pseudo comptes issus de la table E. coli. Il semble que tous ces tests sont non significatifs ce qui peut être expliqué par deux hypothèses :

— Les compositions microbiennes et le pattern de colonisation à E. coli suivent deux mécanismes indépendants
— ou bien la méthode utilisée ne permet pas de mettre en évidence ce phénomène (manque de puissance).

La première explication joue, en quelque sorte, le rôle de l'hypothèse nulle. On s'efforcera donc, dans la suite, de tester d'autres méthodes statistiques, comme celles présentées au paragraphe 2.5.1. Il est également possible de tenter de gagner en puissance en considérant des classes de E. coli moins nombreuses que les 56 génomes de référence utilisés. Sur ce point, la catégorisation des souches de E. coli en type A, B1 et B2 pourrait être prometteuse. Concernant la table issue de l'analyse 16S, il est possible d'en dériver des paramètres tels que la diversité ou la proportion de Firmicutes par exemple. Ces approches permettent de simplifier la modélisation et de gagner en puissance si cette hypothèse de simplification est fondée.

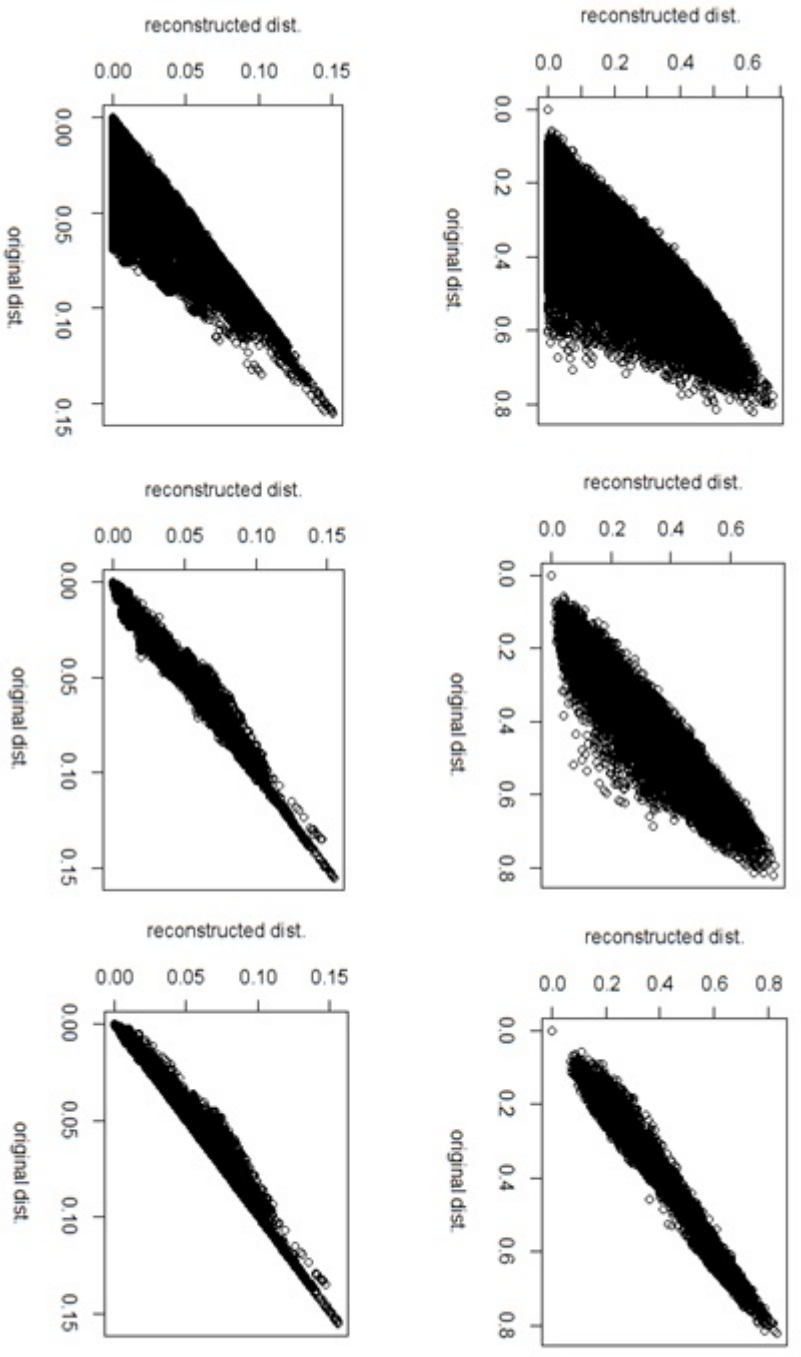

FIGURE 3.9 – Diagramme de Shepard représentant les distances originales contre les distances reconstruites pour les données E. coli (gauche) et 16S (droite) pour $r = 1$ (haut), $r = 7$ (milieu) et $r = 29$ (bas)

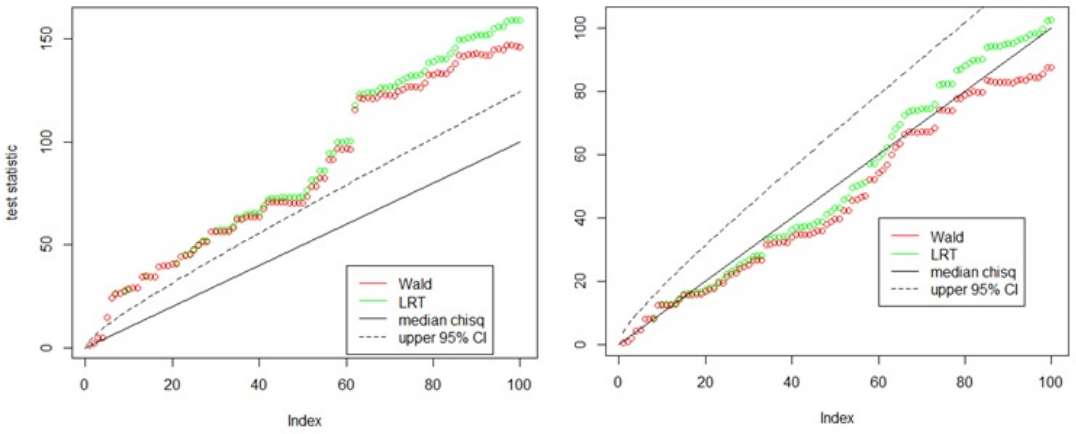

FIGURE 3.10 – Statistique de test en fonction de la dimensionnalité pour l'association entre microbiote intestinal et BMI
(gauche : 16S, droite : E. coli)

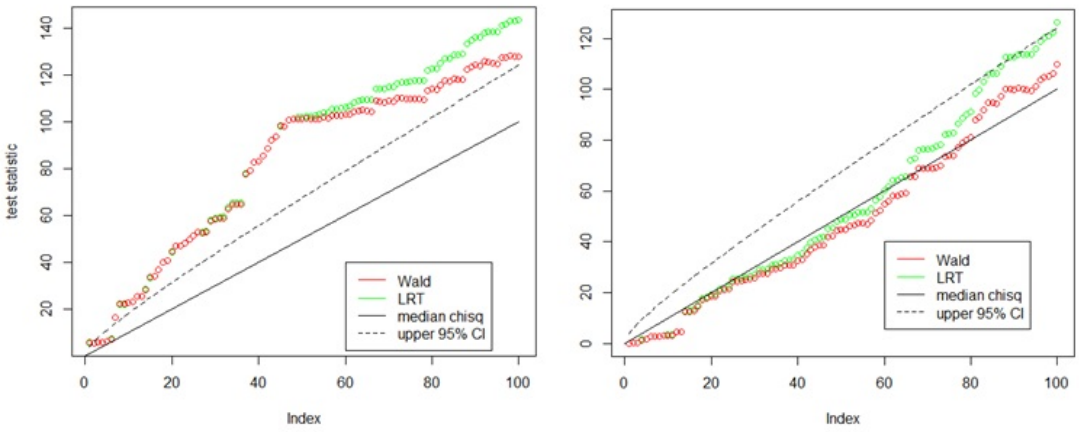

FIGURE 3.11 – Statistique de test en fonction de la dimensionnalité pour l'association entre microbiote intestinal et âge
(gauche : 16S, droite : E. coli)

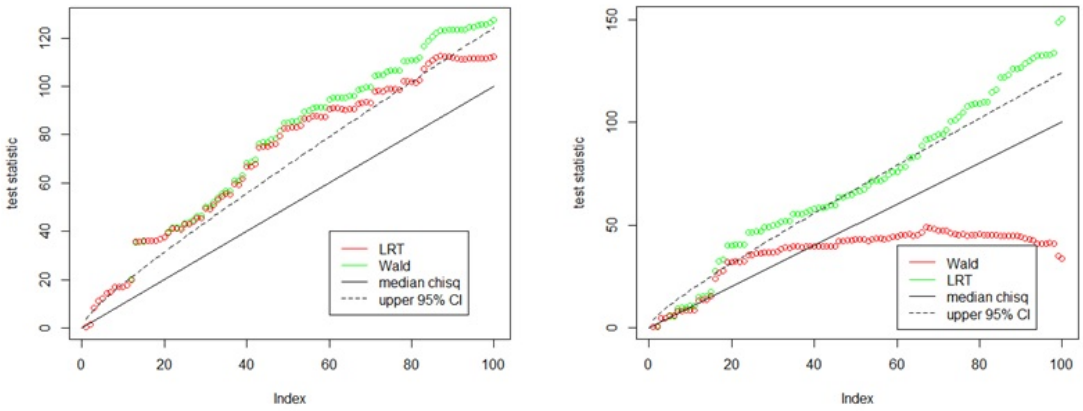

FIGURE 3.12 – Statistique de test en fonction de la dimensionnalité pour l'association entre microbiote intestinal et sexe
(gauche : 16S, droite : E. coli)

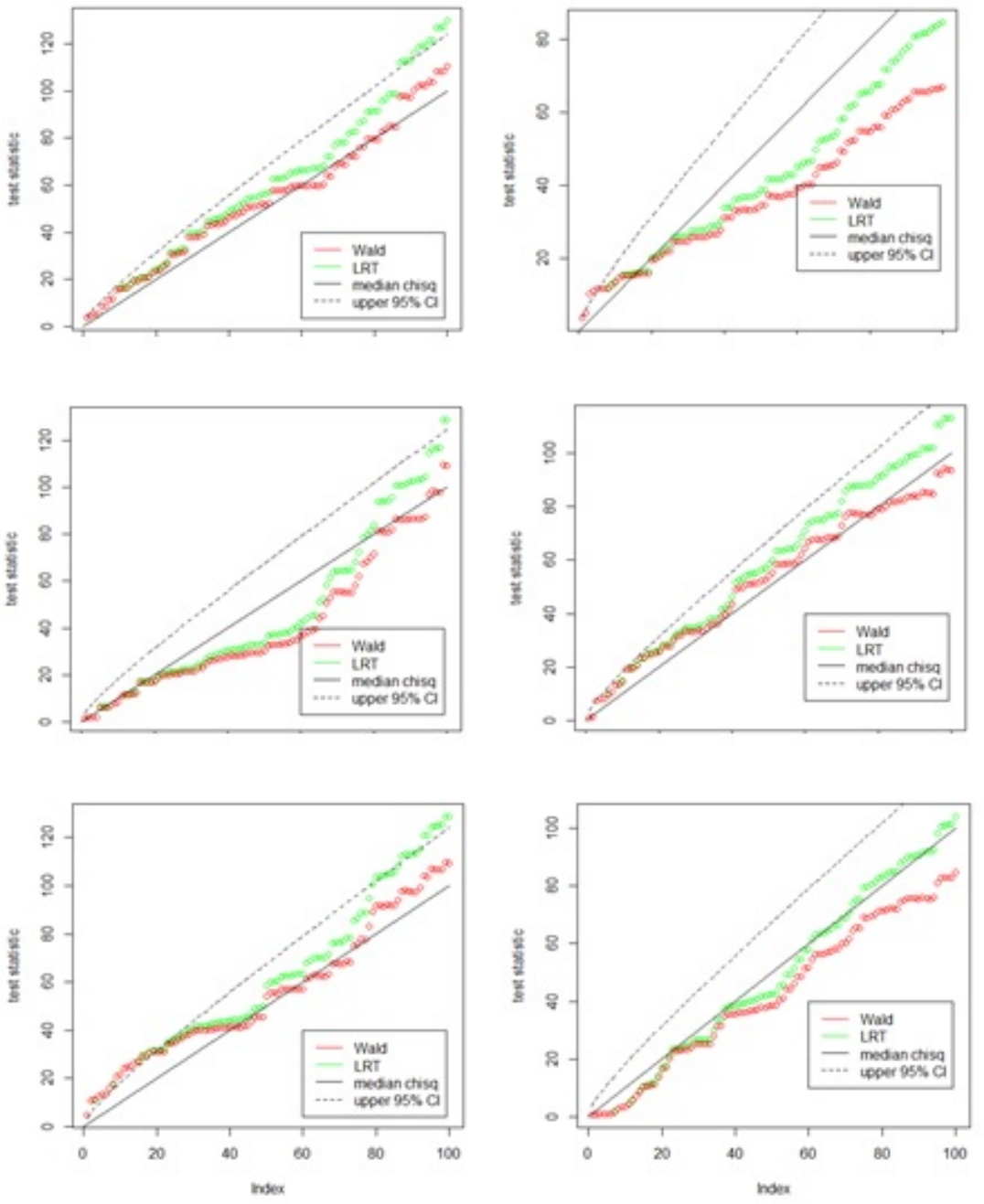

FIGURE 3.13 – Statistique de test en fonction de la dimensionnalité pour l'association entre trois premières composantes de la reconstruction MDS et microbiote (variable explicative : gauche E. coli, droite 16S)

Conclusion et perspectives

Cette thèse est consacrée à l'application de méthodes statistiques adaptées aux données de microbiote intestinal issues d'analyses basées sur l'ARN 16S ou metagénomiques. Sa principale contribution est la mise en évidence d'une association statistique entre la composition microbienne du tube digestif et l'âge, du sexe et du BMI, sur la cohorte de patients TwinsUK.

La méthode statistique employée est particulière et tiens compte de la structure des données. Elle consiste à utiliser l'information de l'arbre phylogénétique qui structure les OTU (dans le cas de l'analyse 16S) ou les souches de références (dans le cas de l'analyse métagénomique E. coli) afin de construire une matrice de distance Unifrac entre échantillon. Cette matrice sert dans un second temps à altérer la table de compte afin de rendre plus proche au sens de la distance Euclidienne les échantillons proche au sens de la distance Unifrac. Les variables explicatives ainsi construites entrent, après une étape de sélection de dimensionnalité dans un modèle multivarié standard.

Dans ce travail, la méthode proposée a été appliquée à deux types de données très différentes :

— les données issues du pipeline 16S sont de grande dimension (842 OTU) et couvrent l'ensemble du microbiote intestinal. C'est sur ces données qu'un signal a été mis en évidence.
— les données issues de l'analyse métagénomique de la population de E. coli, qui sont de dimension plus raisonnable ($p = 56$) et qui décrivent la composition de la population de E. coli dans les échantillons de selles.

Par rapport à une stratégie naive qui consisterait à tester les 842 associations possibles entre abondance relative d'un OTU et une variable dépendante, les méthodes distance-based présentent l'avantage de ne faire qu'un unique test statistique. Il est important de noter que la combinatoire des tests possibles est possiblement bien plus grande (elle est en réalité presque doublée) si l'on inclut les tests d'association avec des comptes cumulés d'OTU à un niveau plus élevé de l'arbre pylogénétique (espèce, genre, famille, phylum,..), comme cela est souvent réalisé dans la littérature. Ainsi, la stratégie présentée évite la nécessité d'ajuster les p-value obtenues pour prendre en compte la multiplicité des tests et assurer un contrôle de l'erreur de type I ou du taux de fausse découvertes (FDR). Ces méthodes d'ajustement comportent des limites (perte de puissance, hypothèses nécessaires sur la structure de corrélation des différentes statistiques de test [4, 5]).

Sur les exemples étudiés dans ce travail, l'association qui est mise en évidence entre la composition microbienne du tube digestif et l'âge, le sexe et l'indice de masse corporelle se montre très robuste vis-à-vis du choix de la dimensionnalité de la reconstruction. Ce résultat est également important dans la mesure où le critère de minimisation du stress utilisé n'est pas le seul critère possible et que le nombre de dimensions retenues avec un autre critère pourrait être différent.

Contrairement à notre hypothèse de recherche selon laquelle le profil de colonisation à E. coli serait susceptible d'être associé aux variables âge, sexe, BMI, ou d'influencer la composition microbienne globale, la méthode employée ne parvient pas à mettre en évidence une telle association. Plusieurs explications sont possibles à ce phénomène. On peut imaginer, par exemple que la relation fait intervenir un niveau hiérarchique supérieur de la classifaction taxonomique comme le groupe génétique de E. coli (A, B1, B2,...). On peut également envisager que la composition en E. coli influence le microbiote intestinal par l'intermédiaire d'une variable comme la diversité, et que la méthode de regression distance-based n'est pas très puissante pour mettre en évidence cette association. Enfin, il n'es pas possible d'exclure que l'hypothèse de travail initale soit fausse, malgré les arguments physiologiques en présence, et que le profil de colonisation à E. coli n'a en réalité aucun lien avec la composition microbienne globale, ces deux communautées vivant en totale indépendance l'une de l'autre.

Plusieurs éléments de ce travail sont susceptibles d'être approfondis à l'occasion de recherches ultérieures (dont certaines sont d'ailleurs déjà en cours). Le premier concerne l'identification des OTU associés au signal mis en évidence sur les données 16S. En effet, la méthode basé sur les distance élimine complètement l'information sur les OTU dès la première étape de construction de la matrice de distance. Lorsqu'une association significative, il serait intéressant de pouvoir identifier les OTU qui y contribuent. Deux stratégies peuvent être imaginées pour y parvenir :

— Des tests individuels protégés en cas de significativité de l'association globale.
— La caractérisation des échantillons dont la position dans l'espace de reconstruction MDS est "extrême" par rapport à la variable expliquée.

Un deuxième point concerne l'analyse "closed reference" des données de métagénomique E. coli. On a utilisé sur ce point une stratégie empririque d'affectation des reads aux 56 souches de référence alors que d'autres stratégies sont décrites dans la littérature [1,52,58]. Cependant, leur applicabilité à notre cas particulier d'un panel de génomes de références de la même espèce (et sont donc très semblables) est incertaine. De plus, l'évaluation qui a été menée sur ces méthodes par leurs auteurs couvrent uniquement l'étape d'affectation des reads mais pas l'impact sur la stratégie d'analyse statistique qui suit. Les questions en suspens sont donc ici nombreuses :

— Quel est la robustesse de la méthode globale (de l'échantillon jusqu'au test statistique) vis-à-vis du panel de génomes de référence ?
— Comment évolue la complexité de l'algorithme d'affectation des reads si la taille du panel augmente ? Actuellement, des milliers de génomes de E. coli sont disponibles sur des bases de données publiques. Est-il faisable ou souhaitable d'en inclure beaucoup plus dans le panel ?
— Existe-t-il des critères pour qualifier un bon panel ? On imagine aisément que le panel doit représenter toute la diversité rencontrée dans les échantillons. La question est cependant plus complexe concernant sa taille. Une taille importante permet d'être plus précis dans la description de l'échantillon, mais présente l'inconvénient d'augmenter la dimensionnalité des données, ce qui peut être un handicap pour

l'analyse statistique.

Enfin, on pourra explorer d'avantage l'hypothèse d'une association entre le profil de colonisation à E. coli exploré par métagénomique et le microbiote intestinal. L'approche réalisée dans ce travail, qui consiste à utiliser les axes de reconstruction MDS des données 16S comme des variables expliquée dans une régression distance-based avec les données E. coli, présente l'inconvénient de rompre la symétrie qui existe a priori sur les données. On a d'ailleurs également réalisé l'analyse inverse en considérant les axes de la reconstruction MDS des données E. coli comme variable expliquée. Une autre stratégie serait d'utiliser d'autres méthodes qui préservent la symétrie entre les données 16S et E. coli. L'analyse canonique des corrélations ([24]) est une piste intéressante. Elle permet d'étudier les corrélations entre deux groupes de variables quantitatives. En réalisant des combinaisons linéaires optimales des comptes 16S et E. coli, on peut faire en sorte d'obtenir deux variables dont le coeficient de corrélation est maximal. D'autres combinaisons peuvent ensuite être trouvées et l'algorithme fournit une suite de coefficients de corrélation, dont la décroissance peut être comparée à celle que l'on observerait en permutant les comptes des OTU de chaque échantillon. Des recherches sur ce point sont en cours afin d'incorporer la structure de l'arbre phylogénétique dans l'algorithme, ce que la méthode actuelle ne permet pas.

Au total, ce travail de thèse aura permis de valoriser des données riches de microbiote intestinal, comportant à la fois une description du microbiote sous la forme d'une table d'OTU et d'une analyse métagénomique sur la population E. coli. L'utilisation de méthodes adaptées à ce type de données et prenant en compte la notion de distance génétique entre OTU (ou entre souches) permet de mettre en évidence des associations sans tomber dans l'eceuil de la multicplicité des test qui représente une réelle menace pour la qualité de la recherche dans ce domaine. Ce travail participera, on l'espère, à ouvrir la voie vers une standardisation des procédures pour ce type de données, couvrant l'ensemble du processus de traitement depuis l'analyse bioinformatique des données brutes en sortie de séquenceur jusqu'à l'analyse statistique.

Table des figures

Références

[1] T.-H. Ahn, J. Chai, and C. Pan. Sigma : Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics*, 31(2) :170–177, Jan. 2015.

[2] M. J. Anderson. Permutational Multivariate Analysis of Variance (PERMANOVA). In *Wiley StatsRef : Statistics Reference Online*, pages 1–15. American Cancer Society, Nov. 2017.

[3] L. Armand-Lefèvre, A. Andremont, and E. Ruppé. Travel and acquisition of multidrug-resistant Enterobacteriaceae. *Medecine Et Maladies Infectieuses*, pages 431–441, Mar. 2018.

[4] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1) :289–300, 1995.

[5] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4) :1165–1188, Aug. 2001.

[6] K. A. Bettelheim and S. M. Lennox-King. The acquisition of Escherichia coli by new-born babies. *Infection*, 4(3) :174–179, 1976.

[7] N. A. Bokulich, S. Subramanian, J. J. Faith, D. Gevers, J. I. Gordon, R. Knight, D. A. Mills, and J. G. Caporaso. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, 10(1) :57–59, Jan. 2013.

[8] C. Burdet, S. Sayah-Jeanne, T. T. Nguyen, P. Hugon, F. Sablier-Gallis, N. Saint-Lu, T. Corbel, S. Ferreira, M. Pulse, W. Weiss, A. Andremont, F. Mentré, and J. d. Gunzburg. Antibiotic-induced dysbiosis predicts mortality in an animal model of Clostridium difficile infection. *Antimicrobial Agents and Chemotherapy*, pages 00925–18, July 2018.

[9] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5) :335–336, May 2010.

[10] M. R. Chacón, J. Lozano-Bartolomé, M. Portero-Otín, M. M. Rodríguez, G. Xifra, J. Puig, G. Blasco, W. Ricart, F. J. Chaves, and J. M. Fernández-Real. The gut mycobiome composition is linked to carotid atherosclerosis. *Beneficial Microbes*, 9(2) :185–198, Feb. 2018.

[11] Q. Chang, Y. Luan, and F. Sun. Variance adjusted weighted UniFrac : a powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*, 12 :12–118, Apr. 2011.

[12] Y. Chu, M. Z. Jiang, B. Xu, W. J. Wang, D. Chen, X. W. Li, Y. J. Zhang, and J. Liang. Specific changes of enteric mycobiota and virome in inflammatory bowel disease. *Journal of Digestive Diseases*, 19(1) :2–7, Jan. 2018.

[13] R. B. Davies. Algorithm AS 155 : The Distribution of a Linear Combination of chi2 Random Variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3) :323–333, 1980.

[14] T. G. Dinan, Y. E. Borre, and J. F. Cryan. Genomics of schizophrenia : time to consider the gut microbiome ? *Molecular Psychiatry*, 19(12) :1252–1257, Dec. 2014.

[15] P. Duchesne and P. Lafaye De Micheaux. Computing the distribution of quadratic forms : Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4) :858–862, 2010.

[16] P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman. Diversity of the human intestinal microbial flora. *Science*, 308(5728) :1635–1638, June 2005.

[17] W. A. Ferens and C. J. Hovde. Escherichia coli O157 :H7 : Animal Reservoir and Sources of Human Infection. *Foodborne Pathogens and Disease*, 8(4) :465–487, Apr. 2011.

[18] S. Fukiya, H. Mizoguchi, T. Tobe, and H. Mori. Extensive genomic diversity in pathogenic Escherichia coli and Shigella Strains revealed by comparative genomic hybridization microarray. *Journal of Bacteriology*, 186(12) :3911–3921, June 2004.

[19] Y.-D. Gao, Y. Zhao, and J. Huang. Metabolic Modeling of Common Escherichia coli Strains in Human Gut Microbiome. *BioMed Research International*, pages 1–11, 2014.

[20] L. Garidou, C. Pomié, P. Klopp, A. Waget, J. Charpentier, M. Aloulou, A. Giry, M. Serino, L. Stenman, S. Lahtinen, C. Dray, J. S. Iacovoni, M. Courtney, X. Collet, J. Amar, F. Servant, B. Lelouvier, P. Valet, G. Eberl, N. Fazilleau, V. Douin-Echinard, C. Heymes, and R. Burcelin. The Gut Microbiota Regulates Intestinal CD4 T Cells Expressing RORgt and Controls Metabolic Disease. *Cell Metabolism*, 22(1) :100–112, July 2015.

[21] E. Gianchecchi and A. Fierabracci. On the pathogenesis of insulin-dependent diabetes mellitus : the role of microbiota. *Immunologic Research*, pages 242–256, July 2016.

[22] S. R. Gill, M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778) :1355–1359, June 2006.

[23] M. Girotra, S. Garg, R. Anand, Y. Song, and S. K. Dutta. Fecal Microbiota Transplantation for Recurrent Clostridium difficile Infection in the Elderly : Long-Term Outcomes and Microbiota Changes. *Digestive Diseases and Sciences*, pages 3007–3015, July 2016.

[24] I. González, S. Déjean, P. Martin, and A. Baccini. Cca : An r package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12) :1–14, 2008.

[25] J. K. Goodrich, J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, J. T. Bell, T. D. Spector, A. G. Clark, and R. E. Ley. Human genetics shape the gut microbiome. *Cell*, 159(4) :789–799, Nov. 2014.

[26] A. Górska, S. Peter, M. Willmann, I. Autenrieth, R. Schlaberg, and D. H. Huson. Dynamics of the human gut phageome during antibiotic treatment. *Computational Biology and Chemistry*, pages 420–427, Mar. 2018.

[27] A. Hiergeist, J. Gläsner, U. Reischl, and A. Gessner. Analyses of Intestinal Microbiota : Culture versus Sequencing. *ILAR journal*, 56(2) :228–240, 2015.

[28] P. D. Houghteling and W. A. Walker. Why is initial bacterial colonization of the intestine important to the infant's and child's health ? *Journal of Pediatric Gastroenterology and Nutrition*, 60(3) :294–307, Mar. 2015.

[29] C. Jiang, G. Li, P. Huang, Z. Liu, and B. Zhao. The Gut Microbiota and Alzheimer's Disease. *Journal of Alzheimer's Disease*, 58(1) :1–15, 2017.

[30] B. Kelsall. Getting to the guts of NOD2. *Nature Medicine*, 11(4) :383–384, 2005.

[31] D. Kim, Y.-G. Kim, S.-U. Seo, D.-J. Kim, N. Kamada, D. Prescott, Mathias Chamaillard, D. J. Philpott, P. Rosenstiel, N. Inohara, and G. Núñez. Nod2-mediated recognition of the microbiota is critical for mucosal adjuvant activity of cholera toxin. *Nature Medicine*, 22(5) :524–530, 2016.

[32] C. A. Kumamoto. The Fungal Mycobiota : Small Numbers, Large Impacts. *Cell Host & Microbe*, 19(6) :750–751, 2016.

[33] B. Lamas, M. L. Richard, V. Leducq, H.-P. Pham, M.-L. Michel, G. Da Costa, C. Bridonneau, S. Jegou, T. W. Hoffmann, J. M. Natividad, L. Brot, S. Taleb, A. Couturier-Maillard, I. Nion-Larmurier, F. Merabtene, P. Seksik, A. Bourrier, J. Cosnes, B. Ryffel, L. Beaugerie, J.-M. Launay, P. Langella, R. J. Xavier, and H. Sokol. CARD9 impacts colitis by altering gut microbiota metabolism of tryptophan into aryl hydrocarbon receptor ligands. *Nature Medicine*, 22(6) :598–605, 2016.

[34] H. Liu, Y. Tang, and H. H. Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4) :853–856, Feb. 2009.

[35] C. Lozupone, M. Hamady, and R. Knight. UniFrac – An online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, 7 :371, Aug. 2006.

[36] C. Lozupone and R. Knight. UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71(12) :8228–8235, Dec. 2005.

[37] C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh, and R. Knight. UniFrac : an effective distance metric for microbial community comparison. *The ISME journal*, 5(2) :169–172, Feb. 2011.

[38] Y. Ma, X. You, G. Mai, T. Tokuyasu, and C. Liu. A human gut phage catalog correlates the gut phageome with type 2 diabetes. *Microbiome*, 6(1) :24, Feb. 2018.

[39] F. Mangiola, G. Ianiro, F. Franceschi, S. Fagiuoli, G. Gasbarrini, and A. Gasbarrini. Gut microbiota in autism and mood disorders. *World Journal of Gastroenterology*, 22(1) :361–368, Jan. 2016.

[40] P. Manrique, B. Bolduc, S. T. Walk, J. van der Oost, W. M. de Vos, and M. J. Young. Healthy human gut phageome. *Proceedings of the National Academy of Sciences of the United States of America*, 113(37) :10400–10405, 2016.

[41] S. L. Martz, M. Guzman-Rodriguez, S.-M. He, C. Noordhof, D. J. Hurlbut, G. B. Gloor, C. Carlucci, S. Weese, E. Allen-Vercoe, J. Sun, E. C. Claud, and E. O. Petrof. A human gut ecosystem protects against C. difficile disease by targeting TcdA. *Journal of Gastroenterology*, pages 452–465, June 2016.

[42] A. Mazzariol, A. Bazaj, and G. Cornaglia. Multi-drug-resistant Gram-negative bacteria causing urinary tract infections : a review. *Journal of Chemotherapy (Florence, Italy)*, 29(sup1) :2–9, Dec. 2017.

[43] D. B. McArtor, G. H. Lubke, and C. S. Bergeman. extending multivariate distance matrix regression with an effect size measure and the asymptotic null distribution of the test statistic. *Psychometrika*, 82(4) :1052–1077, Dec. 2017.

[44] A. Moayyeri, C. J. Hammond, D. J. Hart, and T. D. Spector. The UK Adult Twin Registry (TwinsUK Resource). *Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies*, 16(1) :144–149, Feb. 2013.

[45] A. Moayyeri, C. J. Hammond, A. M. Valdes, and T. D. Spector. Cohort Profile : TwinsUK and healthy ageing twin study. *International Journal of Epidemiology*, 42(1) :76–85, Feb. 2013.

[46] P. K. Mukherjee, B. Sendid, G. Hoarau, J.-F. Colombel, D. Poulain, and M. A. Ghannoum. Mycobiota in gastrointestinal diseases. *Nature Reviews Gastroenterology & Hepatology*, 12(2) :77–87, Feb. 2015.

[47] A. Mulak and B. Bonaz. Brain-gut-microbiota axis in Parkinson's disease. *World Journal of Gastroenterology*, 21(37) :10609–10620, Oct. 2015.

[48] K. Nemani, R. Hosseini Ghomi, B. McCormick, and X. Fan. Schizophrenia and the gut-brain axis. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 56 :155–160, Jan. 2015.

[49] A. C. Nica, L. Parts, D. Glass, J. Nisbet, A. Barrett, M. Sekowska, M. Travers, S. Potter, E. Grundberg, K. Small, s. K. Hedman, V. Bataille, J. T. Bell, G. Surdulescu, A. S. Dimas, C. Ingle, F. O. Nestle, P. d. Meglio, J. L. Min, A. Wilk, C. J. Hammond, N. Hassanali, T.-P. Yang, S. B. Montgomery, S. O'Rahilly, C. M. Lindgren, K. T. Zondervan, N. Soranzo, I. Barroso, R. Durbin, K. Ahmadi, P. Deloukas,

M. I. McCarthy, E. T. Dermitzakis, T. D. Spector, and T. M. Consortium. The Architecture of Gene Regulatory Variation across Multiple Human Tissues : The MuTHER Study. *PLOS Genetics*, 7(2) :e1002003, 2011.

[50] M.-H. Nicolas-Chanoine, C. Gruson, S. Bialek-Davenet, X. Bertrand, F. Thomas-Jean, F. Bert, M. Moyat, E. Meiller, E. Marcon, N. Danchin, L. Noussair, R. Moreau, and V. Leflon-Guibout. 10-Fold increase (2006-11) in the rate of healthy subjects with extended-spectrum g-lactamase-producing Escherichia coli faecal carriage in a Parisian check-up centre. *The Journal of Antimicrobial Chemotherapy*, 68(3) :562–568, Mar. 2013.

[51] A. Paun and J. S. Danska. Modulation of type 1 and type 2 diabetes risk by the intestinal microbiome. *Pediatric Diabetes*, pages 469–477, Aug. 2016.

[52] G. Rosen, E. Garbarine, D. Caseiro, R. Polikar, and B. Sokhansanj. Metagenome fragment classification using N-mer frequency profiles. *Advances in Bioinformatics*, 2008 :205969, 2008.

[53] C. L. Ross, J. K. Spinler, and T. C. Savidge. Structural and functional changes within the gut microbiota and susceptibility to Clostridium difficile infection. *Anaerobe*, pages 37–43, May 2016.

[54] E. Scarpellini, G. Ianiro, F. Attili, C. Bassanelli, A. De Santis, and A. Gasbarrini. The human gut microbiota and virome : Potential therapeutic implications. *Digestive and Liver Disease : Official Journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver*, 47(12) :1007–1012, Dec. 2015.

[55] K. Schei, E. Avershina, T. Øien, K. Rudi, T. Follestad, S. Salamati, and R. A. Ødegård. Early gut mycobiota and mother-offspring transfer. *Microbiome*, 5(1) :107, Aug. 2017.

[56] S. Schippa and M. P. Conte. Dysbiotic events in gut microbiota : impact on human health. *Nutrients*, 6(12) :5786–5805, Dec. 2014.

[57] P. D. Schloss. Evaluating different approaches that test whether microbial communities have the same structure. *The ISME journal*, 2(3) :265–275, Mar. 2008.

[58] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8) :811–814, June 2012.

[59] R. Sender, S. Fuchs, and R. Milo. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, 14(8) :e1002533, Aug. 2016.

[60] E. Sherwin, K. V. Sandhu, T. G. Dinan, and J. F. Cryan. May the Force Be With You : The Light and Dark Sides of the Microbiota-Gut-Brain Axis in Neuropsychiatry. *CNS drugs*, pages 1019–1041, July 2016.

[61] M. O. A. Sommer, G. M. Church, and G. Dantas. The human microbiome harbors a diverse reservoir of antibiotic resistance genes. *Virulence*, 1(4) :299–303, Aug. 2010.

[62] M. O. A. Sommer, G. Dantas, and G. M. Church. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*, 325(5944) :1128–1131, Aug. 2009.

[63] T. D. Spector and F. M. K. Williams. The UK Adult Twin Registry (TwinsUK). *Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies*, 9(6) :899–906, Dec. 2006.

[64] A. Sullivan, C. Edlund, and C. E. Nord. Effect of antimicrobial agents on the ecological balance of human microflora. *The Lancet Infectious Diseases*, 1(2) :101–114, Sept. 2001.

[65] E. Thursby and N. Juge. Introduction to the human gut microbiota. *Biochemical Journal*, 474(11) :1823–1836, June 2017.

[66] S. G. Tringe, C. v. Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. Comparative Metagenomics of Microbial Communities. *Science*, 308(5721) :554–557, Apr. 2005.

[67] H. H. Wang and D. W. Schaffner. Antibiotic Resistance : How Much Do We Know and Where Do We Go from Here ? *Applied and Environmental Microbiology*, 77(20) :7093–7095, Oct. 2011.

[68] W. B. Whitman, D. C. Coleman, and W. J. Wiebe. Prokaryotes : The unseen majority. *Proceedings of the National Academy of Sciences*, 95(12) :6578–6583, June 1998.

[69] S. E. Winter, M. G. Winter, M. N. Xavier, P. Thiennimitr, V. Poon, A. M. Keestra, R. C. Laughlin, G. Gomez, J. Wu, S. D. Lawhon, I. E. Popova, S. J. Parikh, L. G. Adams, R. M. Tsolis, V. J. Stewart, and A. J. Bäumler. Host-derived nitrate boosts growth of E. coli in the inflamed gut. *Science*, 339(6120) :708–711, Feb. 2013.

[70] F. Xia, J. Chen, W. K. Fung, and H. Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4) :1053–1063, Dec. 2013.

[71] P. Yarza, P. Yilmaz, E. Pruesse, F. O. Glöckner, W. Ludwig, K.-H. Schleifer, W. B. Whitman, J. Euzéby, R. Amann, and R. Rosselló-Móra. Uniting the classification of cultured and uncultured bacteria and archaea using 16s rRNA gene sequences. *Nature Reviews Microbiology*, 12(9) :635–645, 2014.

[72] S. F. Øyri, G. Műzes, and F. Sipos. Dysbiotic gut microbiome : A key element of Crohn's disease. *Comparative Immunology, Microbiology and Infectious Diseases*, 43 :36–49, Dec. 2015.

[73] M. A. Zapala and N. J. Schork. Statistical Properties of Multivariate Distance Matrix Regression for High-Dimensional Data Analysis. *Frontiers in Genetics*, 3 :3–190, Sept. 2012.

[74] N. Zhao, J. Chen, I. Carroll, T. Ringel-Kulka, M. Epstein, H. Zhou, J. Zhou, Y. Ringel, H. Li, and M. Wu. Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *American Journal of Human Genetics*, 96(5) :797–807, May 2015.

[75] Q. Zhou, X. Su, and K. Ning. Assessment of quality control approaches for metagenomic data analysis. *Scientific Reports*, 4 :6957, Nov. 2014.

Titre : Étude de données de microbiote intestinal issues de la cohorte de jumeaux « TwinsUK » à l'aide de méthodes de régression incorporant la notion de distance.

Objectif : Montrer une association entre la composition bactérienne du microbiote intestinal et des phénotypes comme l'âge, le sexe ou l'IMC
Méthodes : Les données sont des tables de comptes issues du séquençage de l'ARN 16S bactérien. On dispose également de données de métagénomique de la population d'E. coli. Le traitement statistique de ces tables de compte consiste à procéder à une transformation MDS qui incorpore la notion de distance génétique entre Operational Taxonomic Unit (OTU) ou entre génomes de référence, avant d'utiliser la table transformée comme covariable dans un modèle de régression usuel.
Résultats : Des informations phénotypiques sont disponibles chez 512 individus qui sont majoritairement des femmes (n=445, 87%). L'âge moyen est de 60 ans (sd 11 ans) et l'IMC moyen est à 26kg/m² (sd 5,2kg/m²). Après filtrage et contrôle qualité, on retient 842 OTU et 411 échantillons pour l'analyse issue de l'ARN 16S et 56 génomes de référence et 439 échantillons pour l'analyse de la population d'E. coli. L'analyse s'est faite avec une dimensionnalité de 29 pour l'analyse 16S et 7 pour la population d'E. coli. On montre une association significative entre la composition microbienne mesurée par l'analyse 16S et l'âge, le sexe et l'IMC (p=0,0011, p=0,026 et p=0,0017 respectivement), mais on ne montre pas d'association entre la population d'E. coli et ces phénotypes (p=0,86, p=0,44 et p=0,22 respectivement).
Conclusion : L'utilisation de l'information sur la structure génétique qui sous-tend la table de compte dans l'analyse statistique permet de mettre en évidence des associations tout en évitant l'écueil des tests multiples.
Mots clés : Microbiote ; Analyse de régression ; ARN ribosomique 16S ; metagenomique

Title : Distance-based methods for the study of the gut microbiota from participants of the TwinsUK cohort

Objective : To show an association between gut microbiota composition and phenotypes as age, sex and BMI. Methods : Data were count tables from NGS sequencing of 16S bacterial RNA. We also analyzed data from metagenomics analysis of E. coli population. Statistical analysis is performed in two steps. First, a MDS transform is apply on the count tables to embed the notion of genetic distance between Operational Taxonomic Unit (OTU) or reference genome. Then standard multivariate regression model are used on transformed data.
Results : Phenotypic data were available for 512 individuals who were mainly women (n=445, 87%). The mean age was 60y (sd 11y) and mean BMI was 26kg/m² (sd 5,2kg/m²). After filtering and quality control, we analyzed 842 OTU and 411 samples for the 16S analysis and 56 reference genome and 439 sample for the E.coli metapopulation analysis. The dimensionality for the MDS reconstruction was 29 for the 16S analysis and 7 for the E.coli metapopulation analysis. We showed a significant association between the gut bacterial composition measured by the 16S pipeline and age, sex and BMI (p=0.0011, p=0.026 et p=0.0017 respectively). But we failed to show the same association from E.coli metapopulation data (p=0.86, p=0.44 et p=0.22 respectively).
Conclusion : Taking into account a priori information on the genetic structure underlying the count table seems to allow showing statistical associations between gut microbiota and phenotypes while avoiding the pitfalls of multiple testing
Keywords : Microbiota ; Regression Analysis ; RNA, Ribosomal, 16S ; metagenomic

Université Paris Descartes
Faculté de Médecine Paris Descartes
15, rue de l'Ecole de Médecine
75270 Paris cedex 06