Année 2017 N°.

Thèse de
DOCTORAT en MÉDECINE

DIPLÔME D'ÉTAT

Par
Sacha SCHUTZ

Née le 18 Mai 1984

Présentée et soutenue publiquement le 23 Mai 2017

Description du microbiote pulmonaire chez les patients atteints de mucoviscidose

Président :
Professeur Claude FEREC

Directeur de Thèse :
Docteur Geneviève HERY-ARNAUD

Membres du jury :
Professeur Cédric LE MARECHAL
Docteur Emmanuelle GENIN
Professeur Philippe LANOTTE

Engagement de non-plagiat

Je, soussigné Sacha SCHUTZ, interne en biologie moléculaire au CHRU de Brest, déclare être pleinement informé que le plagiat de documents ou de parties de documents publiés sur toute forme de support, y compris l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Date : 17/05/2017

Signature :

Licence

UNIVERSITE DE BRETAGNE OCCIDENTALE

FACULTE DE MEDECINE ET DES SCIENCES DE LA SANTE DE BREST

DOYENS HONORAIRES :

Professeur H. FLOCH

Professeur G. LE MENN (†)

Professeur B. SENECAIL

Professeur J. M. BOLES

Professeur Y. BIZAIS (†)

Professeur M. DE BRAEKELEER (†)

DOYEN — **Professeur C. BERTHOU**

PROFESSEURS ÉMÉRITES

CENAC Arnaud	Médecine interne
LEHN Pierre	Biologie Cellulaire

PROFESSEURS DES UNIVERSITÉS EN SURNOMBRE

COLLET Michel	Gynécologie - Obstétrique
MOTTIER Dominique	Thérapeutique
RICHE Christian	Pharmacologie fondamentale
LEFEVRE Christian	Anatomie

PROFESSEURS DES UNIVERSITÉS - PRATICIENS HOSPITALIERS DE CLASSE EXCEPTIONNELLE

BOLES Jean-Michel	Réanimation Médicale
COCHENER - LAMARD Béatrice	Ophtalmologie
DEWITTE Jean-Dominique	Médecine & Santé au Travail
FEREC Claude	Génétique
GILARD Martine	Cardiologie
JOUQUAN Jean	Médecine Interne
OZIER Yves	Anesthésiologie et Réanimation Chirurgicale
ROBASZKIEWICZ Michel	Gastroentérologie - Hépatologie

PROFESSEURS DES UNIVERSITÉS - PRATICIENS HOSPITALIERS DE 1ÈRE CLASSE

BAIL Jean-Pierre	Chirurgie Digestive
BERTHOU Christian	Hématologie – Transfusion
BLONDEL Marc	Biologie cellulaire
BRESSOLLETTE Luc	Médecine Vasculaire
DE PARSCAU DU PLESSIX Loïc	Pédiatrie
DE BRAEKELEER Marc	Génétique
DELARUE Jacques	Nutrition
DUBRANA Frédéric	Chirurgie Orthopédique et Traumatologique
FENOLL Bertrand	Chirurgie Infantile
FOURNIER Georges	Urologie
GOUNY Pierre	Chirurgie Vasculaire
HU Weiguo	Chirurgie plastique, reconstructrice & esthétique ; brûlologie
KERLAN Véronique	Endocrinologie, Diabète & maladies métaboliques
LACUT Karine	Thérapeutique
LEROYER Christophe	Pneumologie
LE MEUR Yannick	Néphrologie
LE NEN Dominique	Chirurgie Orthopédique et Traumatologique
LOZAC'H Patrick	Chirurgie Digestive
MANSOURATI Jacques	Cardiologie
MARIANOWSKI Rémi	Oto. Rhino. Laryngologie
MISERY Laurent	Dermatologie – Vénérologie
MERVIEL Philippe	Gynécologie médicale : option gynécologie obstétrique
NEVEZ Gilles	Parasitologie et Mycologie
NONENT Michel	Radiologie & Imagerie médicale
PAYAN Christopher	Bactériologie – Virologie; Hygiène
REMY-NERIS Olivier	Médecine Physique et Réadaptation
SALAUN Pierre-Yves	Biophysique et Médecine Nucléaire
SARAUX Alain	Rhumatologie
SIZUN Jacques	Pédiatrie
STINDEL Éric	Biostatistiques, Informatique Médicale & technologies de communication

TILLY - GENTRIC Armelle	Gériatrie & biologie du vieillissement
TIMSIT Serge	Neurologie
VALERI Antoine	Urologie
WALTER Michel	Psychiatrie d'Adultes

PROFESSEURS DES UNIVERSITÉS - PRATICIENS HOSPITALIERS DE 2ÈME CLASSE

ANSART Séverine	Maladies infectieuses, maladies tropicales
AUBRON Cécile	Réanimation ; médecine d'urgence
BEN SALEM Douraied	Radiologie & Imagerie médicale
BERNARD-MARCORELLES Pascale	Anatomie et cytologie pathologiques
BEZON Eric	Chirurgie thoracique et cardiovasculaire
BOTBOL Michel	Psychiatrie Infantile
BROCHARD Sylvain	Médecine Physique et Réadaptation
CARRE Jean-Luc	Biochimie et Biologie moléculaire
COUTURAUD Francis	Pneumologie
DAM HIEU Phong	Neurochirurgie
DELLUC Aurélien	Médecine interne
DEVAUCHELLE-PENSEC Valérie	Rhumatologie
GIROUX-METGES Marie-Agnès	Physiologie
HUET Olivier	Anesthésiologie - Réanimation Chirurgicale/Médecine d'urgences
LIPPERT Éric	Hématologie ; transfusion : option hématologie
LE MARECHAL Cédric	Génétique
L'HER Erwan	Réanimation Médicale
MONTIER Tristan	Biologie Cellulaire
NOUSBAUM Jean-Baptiste	Gastroentérologie - Hépatologie
PRADIER Olivier	Cancérologie - Radiothérapie
RENAUDINEAU Yves	Immunologie
SEIZEUR Romuald	Anatomie-Neurochirurgie

PROFESSEUR DES UNIVERSITÉS - PRATICIEN LIBÉRAL

LE RESTE Jean Yves	Médecine Générale
LE FLOC'H Bernard	Médecine Générale

PROFESSEUR DES UNIVERSITÉS ASSOCIÉS À MI-TEMPS

BARRAINE Pierre	Médecine Générale

PROFESSEUR DES UNIVERSITÉS - LRU

BORDRON Anne	Biochimie et Biologie moléculaire

MAÎTRES DE CONFÉRENCES DES UNIVERSITÉS – PRATICIENS HOSPITALIERS DE HORS CLASSE

LE MEVEL Jean Claude	Physiologie
PERSON Hervé	Anatomie

MAÎTRES DE CONFÉRENCES DES UNIVERSITÉS – PRATICIENS HOSPITALIERS DE 1ÈRE CLASSE

ABGRAL Ronan	Biophysique et Médecine nucléaire
CORNEC Divi	Rhumatologie
DE VRIES Philine	Chirurgie infantile
DOUET-GUILBERT Nathalie	Génétique
HERY-ARNAUD Geneviève	Bactériologie – Virologie; Hygiène
HILLION Sophie	Immunologie
JAMIN Christophe	Immunologie
LE BERRE Rozenn	Maladies infectieuses-Maladies tropicales
LE GAC Gérald	Génétique
LE ROUX Pierre-Yves	Biophysique et Médecine nucléaire
LODDE Brice	Médecine et santé au travail
MIALON Philippe	Physiologie
MOREL Frédéric	Médecine & biologie du développement & de la reproduction
PLEE-GAUTIER Emmanuelle	Biochimie et Biologie Moléculaire
QUERELLOU Solène	Biophysique et Médecine nucléaire
VALLET Sophie	Bactériologie – Virologie ; Hygiène

MAÎTRES DE CONFÉRENCES DES UNIVERSITÉS – PRATICIENS HOSPITALIERS DE 2ÈME CLASSE

LE GAL Solène	Parasitologie et Mycologie
LE VEN Florent	Cardiologie
PERRIN Aurore	Biologie et médecine du développement & de la reproduction
TALAGAS Matthieu	Cytologie et histologie

MAÎTRES DE CONFÉRENCES DES UNIVERSITÉS – PRATICIENS HOSPITALIERS STAGIAIRES

UGUEN Arnaud	Anatomie et Cytologie Pathologiques

MAITRE DE CONFÉRENCES – PRATICIEN LIBERAL

NABBE Patrice	Médecine Générale

MAÎTRES DE CONFÉRENCES ASSOCIÉS DES UNIVERSITÉ MI-TEMPS

BARAIS Marie	Médecine Générale
CHIRON Benoît	Médecine Générale

MAÎTRES DE CONFÉRENCES DES UNIVERSITÉS

BERNARD Delphine	Biochimie et biologie moléculaire
FAYAD Hadi	Génie informatique, automatique et traitement du signal
HAXAIRE Claudie	Sociologie - Démographie
KARCHER Brigitte	Psychologie clinique
LANCIEN Frédéric	Physiologie
LE CORRE Rozenn	Biologie cellulaire
MIGNEN Olivier	Physiologie
MORIN Vincent	Électronique et Informatique

MAÎTRES DE CONFÉRENCES ASSOCIÉS DES UNIVERSITÉS A TEMPS COMPLET

MERCADIE Lolita	Rhumatologie

MAÎTRES DE CONFÉRENCES ASSOCIÉS DES UNIVERSITÉS A MI - TEMPS

SCHICK Ulrike	Cancérologie, radiothérapie : option radiothérapie

AGRÉGÉS / CERTIFIÉS DU SECOND DEGRÉ

MONOT Alain	Français
RIOU Morgan	Anglais

Remerciements

Au Professeur Claude FERREC, qui me fait le grand honneur de présider le jury de cette thèse. Qu'il en soit chaleureusement remercié.

Au Docteur Geneviève HERY-ARNAUD, directrice de thèse, qui m'a fait découvrir toute la magie et la complexité du microbiote.

Au Professeur Cedric LE MARECHAL ansi que le Docteur Paul GUEUGEN pour m'avoir accueilli dans le service de biologie moléculaire et m'avoir tant apprit.

Au Docteur Emmanuelle GENIN, qui me fait le plaisir d'être membre de ce jury. Et qui sans son autorisation, aucun des calculs bio-informatiques n'auraient pu être réalisé.

Au Docteur Philippe LANOTTE, pour avoir accepté d'être présent par les voies du numériques

À Perrine qui m'a toujours soutenu,

À ma mère,
À mon père et tous ceux qui ne dédicacent pas leurs thèses à eux même.
À mes grands-parents Denise et Armand.

À mes compagnons de Fac,
Raphaël, avec qui jamais je n'aurais autant dépensé d'argent au Relai H. Florian, pour ses cours en sciences politiques et son sens du partage. Guillaume D, qui s'amuse aussi avec le microbiote. André pour m'avoir apprit pendant un cours soporifique que les variables membres d'une classe sont contiguës en mémoire. Andreas, pour être mon modèle de la perfection intellectuelle. Thibaud pour m'avoir plongé dans la dépendance des bandes dessinées. Mon harem de cointerne Élodie, Marie, Margot, Charlotte, Varoona est (Kevin : je ne savais pas où te mettre) pour m'avoir boosté dans les derniers jours de rédactions.

À mes amis de Caen,
Sassan, Paul, Lorène, Simon, Beinjamin, Jeff ...
Je remercie également mes amis qui ont relu ma thèse. Jeremie, Anne-sophie, Andreas, Thibaud et Olivier mon compagnon baroudeur.

À tous mes amis du Master de Bioinformatique de Rennes,
qui m'ont redonné une nouvelle jeunesse d'anarchiste libriste. Natir, Aluriak, Chloé, David, Jérôme, Pierre ...

Je n'oublie pas de remercier toute l'équipe de biologie moléculaire qui a supporté ma maladresse.

Enfin, je remercie particulièrement mon ordinateur et toute la communauté open source qui m'a permis de produire cette thèse.

Table des matières

Avant-propos

Depuis Louis Pasteur, les bactéries ont toujours été associées aux maladies. La mise en évidence des agents pathogènes comme la syphilis ou la peste n'a pas aidé à tordre le cou de ce stéréotype. La médecine les a donc longtemps considérées comme les ennemies à combattre plutôt que des alliées. Aujourd'hui, personne ne peut nier que les traitements anti-infectieux ont diminué la morbidité des maladies infectieuses. Les avancées majeures dans les domaines de l'hygiène, la vaccination et les antibiotiques ont conduit à faire diminuer la prévalence des maladies infectieuses voire à les faire disparaître. Cependant, la destruction systématique et massive des micro-organismes qui font partie de nous depuis des milliers d'années pourrait bien être la cause de l'émergence de nouvelles maladies[2] (Figure 1).

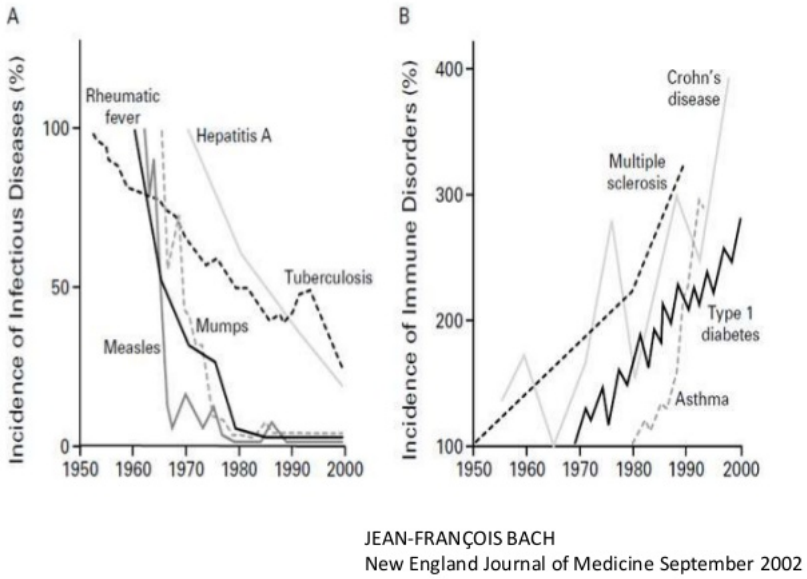

FIGURE 1 – Incidence des maladies infectieuses et auto-immunes en Europe au cours de 50 ans.

Les méthodes récentes d'exploration de ce monde microscopique comme le séquençage haut-débit ont permis aux bactéries de retrouver leurs lettres de noblesse. Elles sont présentes partout et jouent le premier rôle dans le fonctionnement des écosystèmes. Elles sont, par exemple, impliquées dans le cycle de l'azote et permettent à la biomasse d'absorber le diazote atmosphérique. Les bactéries sont ainsi la source primaire dans laquelle les organismes puisent pour construire leurs protéines et leurs acides nucléiques. Elles peuvent survivre dans les milieux les plus inhospitaliers. Les archées (anciennement archéobactéries) peuvent vivre dans des environnements d'acidité et de température exceptionnellement élevées ou faibles. On les retrouve dans les fonds océaniques où, privées

de lumière, elles sont la seule source d'énergie pour la faune grâce à la chimiosynthèse à l'instar de la photosynthèse. Le corps humain est lui aussi un environnement riche en bactéries. Pour la majorité d'entre elles, les méthodes de cultures classiques ne permettaient pas de les détecter. Mais à présent, nous savons que les régions anatomiques autrefois considérées stériles contiennent beaucoup de bactéries. On les retrouve sur tous les muqueuses et épithéliums du corps où elles se regroupent en communautés. La peau est colonisée par *Propionibacterium*, *Corynebacterium* et *Staphylococcus*[52]. Le vagin est colonisé essentiellement par *Lactobacillus* et la bouche principalement par *Streptococcus*[52]. La muqueuse intestinale est colonisée par une « flore » bactérienne ou microbiote dominé par les anaérobies pouvant contribuer jusqu'à 0,2 kg[i] au poids corporel[43]. En échange de l'hospitalité, le microbiote participe au fonctionnement physiologique de l'organisme. Il aide à la digestion en dégradant par exemple les sucres du lait maternel chez le nouveau-né[5;52]. Il participe à la synthèse des vitamines essentielles (K, B12, B9)[26;52]. Il éduque notre système immunitaire et fait barrière à l'implantation de tout nouvel agent pathogène. Tout déséquilibre du microbiote intestinal ou dysbiose peut avoir des conséquences pour notre santé. La liste des affections associées est longue[52]. On y retrouve la maladie de Crohn, la maladie cœliaque, le cancer de l'intestin, le syndrome du côlon irritable, l'obésité, le diabète de type 1, l'asthme, l'eczéma, la sclérose en plaques, la polyarthrite rhumatoïde, la maladie d'Alzheimer et même l'autisme.
Un exemple avec une application clinique est l'infection à *Clostridium difficile*. La destruction du microbiote intestinal par des antibiotiques donne l'opportunité à la souche toxinogène de *C. difficile* de s'installer et provoquer une colite pseudo-membraneuse. Un des traitements proposés après plusieurs échecs de traitement antibiotique est la transplantation fécale visant à régénérer le microbiote du patient.
Le microbiote nous amène donc à reconsidérer notre individualité. Nous ne sommes plus seulement un organisme multicellulaire composé d'un seul génome ; mais plutôt un écosystème où cellules eucaryotes et micro-organismes vivent en symbiose. Cette relation n'est pas figée dans le temps et peut varier entre le commensalisme, le parasitisme et le mutualisme. Des études récentes ont permis d'estimer que pour un être humain le nombre de cellules microbiennes est de 39 milliards comparé aux 30 milliards de cellules humaines[43]. En associant les gènes bactériens, le génome d'un être humain passe de 23 000 à 3,3 millions de gènes[38] avec toute la complexité des interactions que cela engendre. Les scientifiques ont nommé *holobionte* cet écosystème vivant. L'ensemble des génomes est appelé *hologénome*.
Il faut toutefois rester prudent quant au rôle donné aux microbiotes et éviter de tomber dans l'excès. Nombreuses sont les publications scientifiques qui se contredisent ou qui confondent corrélation et causalité. Ces publications ont d'ailleurs conduit à la créa-

i. Les anciennes études estimaient le poid du microbiote digestive à 2 kg

tion du hashtag humoristique sur Twitter, *#GutMicrobiomeAndRandomSomething* qui consister à générer des titres d'études aléatoires. Les études sur le microbiote doivent être étayées par des études fonctionnelles afin de trouver des relations de cause à effet. Les corrélations doivent être réalisées sur des populations plus grandes avec un suivi dans le temps plus conséquent. Les nouvelles technologies de séquençage haut-débit vont dans ce sens en collectant toujours plus de données.
Il est encore trop tôt pour dire si cette science va révolutionner la médecine ou s'il s'agit d'un effet de mode. Au regard de l'évolution biologique et de l'importance du microbiote chez d'autres espèces, les paris sont ouverts. Ne l'oublions pas, ce sont d'anciennes bactéries, que nous appelons aujourd'hui des mitochondries (Figure 2) qui constituent avec nos cellules eucaryotes le paroxysme de la relation symbiotique.

FIGURE 2 – La mitochondrie est l'exemple de symbiose ultime entre eucaryote et procaryote.

Définitions

Termes utilisés en écologie

Le microbiote[15] est l'ensemble des micro-organismes (bactéries, levures, champignons, virus) vivant dans un environnement donné.

Le microbiome[15] est l'ensemble des génomes du microbiote. Le plus souvent le microbiome fait référence aux génomes bactériens seuls. On parle alors de virome et de mycobiome pour les virus et champignons respectivement.

La biocénose est le terme écologique dans un sens large désignant l'ensemble des organismes vivants dans un environnement appelé **biotope**. Biocénose et biotope forment ensemble un **écosystème**.

Une symbiose est une association durable entre deux organismes. Leurs relations peuvent être mutualiste, parasitaire ou commensale.

L'OTU (*Operational Taxonomic Unit*) est un terme utilisé en phylogénie pour désigner un groupe d'individus semblables. Dans la majorité des cas, il s'agit de l'espèce dans la classification de Linné. En microbiologie, un OTU est défini par un groupe de bactéries ayant une similarité dans leurs séquences d'ADNr 16S supérieur ou égal à 97 %.

L'abondance absolue est le nombre d'individu appartement au même OTU dans un échantillon. L'abondance relative est l'expression en pourcentage de l'abondance absolue. L'abondance est évaluée en calculant le nombre de séquence (reads) d'un OTU.

La dominance caractérise un OTU dont l'abondance est supérieure à 90 % dans un échantillon.

La table des OTUs est à la base de toutes analyses en écologie. Elle correspond à un tableau à double entrée contenant l'abondance par échantillon et par OTU. Dans le tableau 3, l'abondance relative pour *Staphylococcus aureus* est de 68 %.

	échantillon 1	échantillon 2	échantillon 3
OTU 1 (*S. aureus*)	68 %	12 %	25 %
OTU 2 (*E. coli*)	40 %	24 %	25 %
OTU 3 (*P. aeruginosa*)	28 %	64 %	50 %

FIGURE 3 – La table des OTUs : l'abondance relative de 3 espèces bactériennes pour 3 échantillons

La diversité alpha est une mesure de la biodiversité au sein d'un échantillon. Elle correspond à l'analyse d'une colonne de la table des OTUs. La richesse, l'indice de Chao1, de Shannon et de Simpson sont des indices de diversité alpha.

La diversité bêta est une analyse descriptive de la biodiversité entre plusieurs échantillons. Elle correspond à l'analyse de l'ensemble de la table des OTUs. L'approche la plus courante est de réaliser une analyse multivariée par des méthodes d'ordination. Il s'agit de représenter un graphique à n dimensions, impossible à représenter dans toutes ses dimensions, en le projetant dans un espace à deux ou trois dimensions.

La richesse (*richness*) est le nombre d'OTUs présents dans un échantillon. Les deux échantillons suivants ont la même richesse (2), mais pas les mêmes abondances.

échantillon 1 : 4 *Streptoccus*, 4 *Escherichia*

échantillon 2 : 432 *Streptoccus*, 12 *Escherichia*

L'équitabilité (*evenness*) indique si les OTUs d'un échantillon sont répartis uniformément. L'équitabilité du premier échantillon est plus grande que celle du second.

échantillon 1 : 50 *Streptococcus*, 50 *Escherichia*

échantillon 2 : 432 *Streptococcus*, 12 *Escherichia*

L'indice Chao1 est une estimation de la richesse réelle (*in vivo*) par rapport à la richesse observée dans l'échantillon (*in vitro*). Cet indice part du principe que si l'échantillon contient beaucoup de singletons (OTU détecté une seule fois), il est fort probable que la richesse réelle soit plus grande que la richesse de l'échantillon. La formule est la suivante :

$$Chao_1 = S_{observed} + \frac{a^2}{2b} \tag{1}$$

avec **S** la richesse observée, **a** le nombre de singletons et **b** le nombre de doubletons.

L'indice de Shannon est un indicateur évaluant à la fois la richesse et l'équitabilité dans un échantillon. Il se calcule de la même façon que l'entropie de Shannon.

$$Shannon = \sum_{i=1}^{S} p_i ln(p_i) \tag{2}$$

avec **p** la fréquence d'un OTU parmi les **S** OTUs présents dans l'échantillon.

L'indice de Simpson est un indicateur évaluant la probabilité que deux individus sélectionnés aléatoirement dans un échantillon donné soient de la même espèce. L'indice est de sens contraire aux précédents. La formule est la suivante :

$$Simpson = 1 - \sum_{i=1}^{S} p_i \tag{3}$$

avec **p** la fréquence d'un OTU parmi les **S** OTUs présents dans l'échantillon

La courbe de raréfaction est utilisée pour déterminer si la profondeur de séquençage est suffisante pour caractériser la diversité d'un échantillon. Pour comprendre, imaginons un sac noir contenant des billes (reads) de différentes couleurs (espèces). Avec deux couleurs dans le sac, une petite poignée (un petit échantillon) de billes sera suffisante pour évaluer la diversité. Vous ne trouverez pas de nouvelle couleur avec plus de billes dans votre main. En revanche, avec d'avantage de couleurs dans le sac, il faudra une poignée plus importante pour évaluer la diversité. La courbe de raréfaction permet d'évaluer si cette poignée de billes est suffisamment représentative. Pour générer cette courbe, des groupes de reads de taille croissante (1..n) sont tirés aléatoirement depuis l'échantillon. Le groupe est reporté sur l'axe X et le nombre d'OTUs trouvé dans ce groupe est reporté sur l'axe Y. Une courbe s'aplatissant indique une bonne profondeur de séquençage (Figure 4).

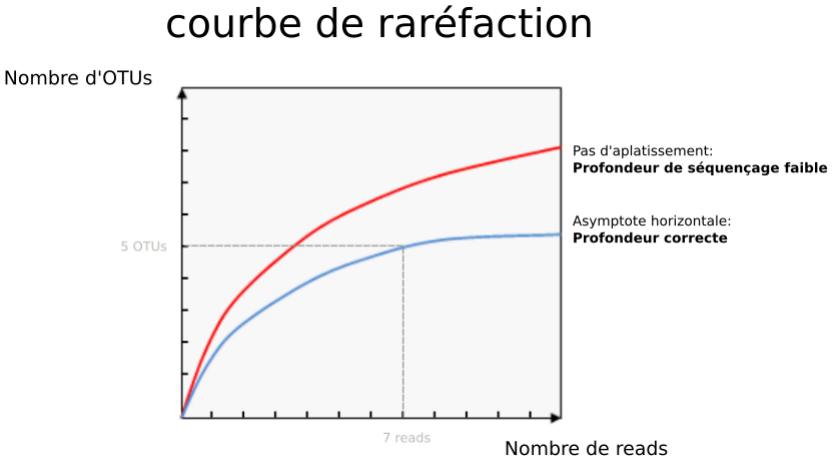

FIGURE 4 – Exemple de courbes de raréfactions. La courbe bleue témoigne d'un bon échantillonnage en s'aplatissant précocement. Plus de reads ne permettraient pas de découvrir de nouveaux OTUs contrairement à la courbe rouge.

Termes utilisés en bio-informatique

Un pipeline est un ensemble d'étapes de calcul informatique. Chaque étape prend en entrée des fichiers pour en produire de nouveaux dans sa sortie. On peut comparer cela aux étapes d'une recette de cuisine. Sans parallélisation, un cuisinier (le processeur) doit attendre de faire fondre le beurre avant de battre les œufs en neige (exécution synchrone). En parallélisant, le cuisinier peut réaliser plusieurs étapes en même temps. Battre les œufs pendant que le beurre fond (exécution asynchrone). Maintenant, si l'objectif est de produire 188 gâteaux (188 analyses) et que l'on dispose de 64 cuisiniers (64 processeurs), l'organisation des tâches devient complexe si l'on veut maximiser le rendement. Pour cela, on dispose d'outils comme *Snakemake*[23], qui permettent de trouver la meilleure façon d'optimiser les tâches en construisant ce qu'on appelle un graphe orienté acyclique.

NGS (Next Generation Sequencing) désigne l'ensemble des techniques de séquençage de nouvelle génération permettant de séquencer un nombre très important de fragments d'ADN.

La métagénomique est une méthode d'étude du contenu en matériel génétique présent dans un milieu grâce aux techniques de séquençage haut débit. Contrairement à la génomique qui s'intéresse au génome d'un individu, la métagénomique s'intéresse aux génomes d'une population d'individus. Dans son sens strict, la métagénomique correspond à l'étude de l'ensemble des séquences d'ADN. L'analyse d'un seul gène, comme celui de l'ARNr 16S est associé à tort au terme métagénomique, mais son usage reste courant. On lui préférera le terme de **métagénétique** ou **métataxonomique**.

Un *read* est un terme bio-informatique désignant une séquence d'ADN issue d'un séquençage haut débit. Selon les technologies, la longueur des reads varie le plus souvent entre 100 et 300 paires de bases.

Une librairie est l'ensemble des fragments d'ADN préparés pour être séquencés.

Un score de qualité Phred ou QScore exprime la confiance que l'on porte au séquençage. Il est lié de façon logarithmique à la probabilité d'erreur d'identification d'un nucléotide. Par exemple un score de 10 équivaut à une 1 erreur sur 10. Un score de 40 à 1 erreur sur 10000.

$$Q = -10 \ \log_{10} P \tag{4}$$

Le score Q est associé de façon logarithmique à la probabilité d'erreur P de s'être trompé en séquençant un nucléotide.

Une ordination est une analyse multivariée visant à décrire des objets caractérisés par plusieurs variables. Par exemple un échantillon est caractérisé par différentes abondances de bactéries. Il existe plusieurs méthodes d'ordination. La plus connue étant l'Analyse en Composantes Principales (PCA) basée sur une matrice de covariances. En écologie, l'Analyse en Coordonnées Principales (PCoA) est préférée. Elle se base sur une matrice de distances. Elle consiste à projeter des points d'un espace à n dimensions vers un espace visible en 2 dimensions. La figure 5 illustre ce principe avec un exemple à deux dimensions seulement.

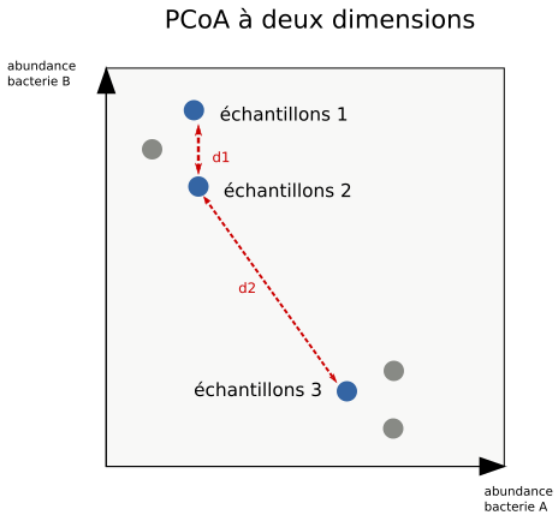

FIGURE 5 – Représentation de plusieurs échantillons ne contenant que deux espèces A et B. Chaque échantillon est assimilable à un point de coordonnées (abondance A ; abondance B). Il est alors possible de calculer des distances entre les échantillons. La plus connue étant la distance euclidienne. Mais d'autres existent comme la distance de Bray-Curtis. Sur cette figure, les échantillons 1 et 2 ont des microbiotes proches, car la distance d1 qui les sépare est petite. Dans la réalité, il y a plus d'espèces donc plus de dimensions. Une PCoA consiste à réduire cet espace multidimensionnelle en un espace 2D qui respecte les distances entre les points.

La distance de Bray-Curtis est un indice de dissimilarité entre deux échantillons qui s'assimile à une distance entre 2 points dans un espace de taille n (Figure 5). Plus la distance est grande et plus les deux échantillons sont dissemblables. La formule est la suivante :

$$BC_{jk} = 1 - \frac{2\sum_{i=1}^{p} min(N_{ij}, N_{ik})}{\sum_{i=1}^{p}(N_{ij} + N_{ik})} \quad (5)$$

Où N_{ij} est l'abondance d'une espèce i dans l'échantillon j et N_{ik} l'abondance de la même espèce i dans l'échantillon k. Le terme min(.,.) correspond au minimum obtenu pour deux comptes sur les mêmes échantillons. Les sommes situées au numérateur et dénominateur sont réalisées sur l'ensemble des espèces présentes dans les échantillons.

Un fichier FASTQ contient les séquences des reads et leurs qualités par nucléotides exprimés en Phred Score. Chaque read est représenté par 4 lignes dans un fichier textuel (Figure 6). Ces fichiers sont produit par le séquençage haut-débit.

FIGURE 6 – Exemple d'un read contenu dans un fichier FASTQ. Les scores de qualités sont encodés avec des symboles ASCII. Par exemple le quatorzième T est associé au symbole « : » correspondant au score phred de 25, soit une probabilité de 0,3 % d'erreur.

Le multiplexage est utilisé pour pouvoir séquencer plusieurs échantillons différents dans une même librairie. Pour cela, les ADNs de chaque échantillon sont marqués avec une courte séquence nucléotidique appelée barcode ou index suivant la technologie. Lors de l'analyse bio-informatique, les données sont démultiplexées afin d'associer chaque read à son échantillon.

Un graphe est un objet mathématique défini par des nœuds (nodes) reliés par des arrêtes (edges). Les graphes permettent de répondre à différents problèmes mathématiques. Par exemple, le problème des sept ponts de Königsberg, qui consiste à savoir s'il existe une promenade passant par 7 ponts une seule fois. La théorie des graphes est très utilisée en bio-informatique.

Introduction

La mucoviscidose

Une maladie génétique

La mucoviscidose est une maladie génétique autosomique récessive grave qui touche en France 1 naissance sur 5400[40]. La Bretagne est la région la plus touchée avec une prévalence de 1/3000[40]. La loi de Hardy Weinberg estime qu'en Bretagne, 1 personne sur 27 est porteuse d'une mutation dans *CFTR* à l'état hétérozygote. Cette haute prévalence s'explique probablement par un effet fondateur associé à un avantage sélectif pour les individus porteurs de l'allèle muté[ii]. Le gène *CFTR* impacté se situe sur le chromosome 7 en position q31.2. Il est constitué de 27 exons pour 250 188[35] paires de bases. Il code un canal chlore AMP dépendant permettant les échanges des ions chlorures au niveau des membranes cellulaires. Il est également impliqué dans le transport du thiocynate (SCN-) et des bicarbonates (HCO3-)[39]. On dénombre à ce jour 2017 mutations[17] mises en cause dans la mucoviscidose. La perte d'une phénylalanine en position 508 par délétion de trois nucléotides c.1521-1523delCTT (anciennement ΔF508) cause à elle seule 80 % des mucoviscidoses[17]. Ces mutations sont responsables d'une protéine défectueuse ou d'une absence de canaux sur les membranes cellulaires.

Cliniquement, la mutation entraîne une insuffisance pancréatique exocrine et une infertilité par disparition des canaux déférents. Des signes digestifs, hépatiques et articulaires sont également présents. L'atteinte de la fonction respiratoire est la plus bruyante cliniquement. En effet, au niveau de l'épithélium broncho-pulmonaire, l'absence d'une protéine CFTR fonctionnelle est responsable d'une déshydratation du mucus, le rendant plus visqueux et empêche les cils bronchiques de jouer leurs rôles.

Du fait de la forte prévalence de la maladie, un dépistage précoce est réalisé systématiquement chez tous les nouveau-nés (test de Guthrie) afin d'adapter au plus tôt la prise en charge. Seul le test de la sueur permet de poser le diagnostic. Le dépistage prénatal basé sur l'ADN circulant est actuellement à l'étude[18].

Le traitement repose avant tout sur une prise en charge respiratoire (kinésithérapie, dornase[iii], antibiothérapie). Les recherches en thérapies génétiques sont encourageantes[33]. Les traitements correcteurs/potentiateurs de la protéine CFTR, comme l'*ivacaftor* sont les seuls traitements curateurs, mais concernent uniquement certaines mutations rares comme la G551D. La greffe pulmonaire est le dernier recours.

ii. Plusieurs hypothèses ont été proposées, notamment celle de la diminution des pertes hydriques chez les sujets hétérozygotes lors des grandes épidémies de choléra.

iii. Une DNase qui fluidifie les sécrétions trop riches en ADN provenant de l'afflux des neutrophiles in situ et de leur apoptose prématurée.

Une maladie infectieuse et inflammatoire

L'atteinte pulmonaire dans la mucoviscidose est caractérisée par des réactions inflammatoires qui dégradent progressivement la fonction respiratoire. Cette inflammation est la conséquence d'un afflux excessif de polynucléaire avec libération de médiateurs[20] probablement en réponse à une agression microbienne[iv]. Plusieurs pathogènes sont impliqués. Chez les jeunes enfants[9], *Haemophilus influenzae* et *Staphylococcus aureus* sont le plus souvent responsable. *Achromobacter xylosoxidans*, *Burkholderia cepacia complex* et *Stenotrophomonas maltophilia* sont retrouvés parmi les sujets plus âgés[9].

Mais c'est *Pseudomonas aeruginosa* qui caractérise l'atteinte pulmonaire dans la mucoviscidose en marquant un tournant décisif dans l'évolution de la maladie. Ce bacille aérobie strict est une bactérie de l'environnement rarement retrouvée parmi les sujets sains[37]. En revanche, dans la mucoviscidose, il est mis en évidence[24] chez 60 % des patients jeunes et plus de 90 % des patients adultes.

La primocolonisation à *P. aeruginosa* semble avoir lieu tôt dans l'enfance[37]. Il y a ensuite une phase de latence (Figure 7), variable entre les individus, marquée par des épisodes de recolonisation intermittentes. À ce moment, l'éradication par des antibiotiques reste possible. Puis survient le passage à la chronicité. *P. aeruginosa* s'adapte à son milieu et s'installe à long terme. Il perd certains caractères de virulence, mais devient résistant aux antibiotiques[24]. Son phénotype change. Il se transforme pour devenir mucoïde en sécrétant un film d'alginate qui le protège du système immunitaire. Les mécanismes sous-jacents à cette adaptation sont ingénieux. La forte densité en bactéries est responsable de l'activation de certains gènes par un processus appelé *quorum sensing*[41]. Dans ce processus, chaque bactérie communique avec ses voisines par des signaux chimiques. Le génome de *P. aeruginosa* devient également hyperpermutable[9] pour présenter une plus grande diversité génétique au regard de la sélection naturelle[v] et donc évoluer plus rapidement.

À ce stade, le traitement antibiotique n'est plus curatif mais symptomatique et l'évolution tend inexorablement vers un déclin de la fonction respiratoire.

L'approche clinique est donc préventive afin de retarder le passage à la chronicité. Elle vise à éliminer *P. aeruginosa* dès qu'il est détecté en culture grâce à une surveillance rapprochée[24]. La culture étant peu sensible, d'autres méthodes d'identification peuvent être employées. La détection des anticorps anti-pyocyaniques par ELISA a montré peu de sensibilité et n'est plus recommandée à l'heure actuelle[?] . La PCR ciblée associant les gènes des protéines bactériennes *oprL*, *gyrB* et *ecfX* s'est montrée plus sensible et plus spécifique que la culture[?] .

En pratique, la colonisation chronique est définie lorsque 3 expectorations sont rendues

iv. D'autres hypothèses suggèrent que l'inflammation précède la colonisation[20]

v. En biologie évolutive, il s'agit d'évolvabilité

positives en culture, successivement au cours d'un suivi mensuel ou bimensuel[24]. Une autre classification, celle de Lee[27] définit 4 groupes (Tableau 1) :

TABLE 1 – Classification de Lee[27]

chronique	$>$ 50 % des cultures sont positives sur 12 mois
intermédiaire	$\leq$ 50 % des cultures sont positives sur 12 mois
Free	Cultures nég. sur 12 mois avec des antécédents de positivité à *P. aeruginosa*
Never	Cultures nég. sur 12 mois sans antécédents de positivité à *P. aeruginosa*

À ce jour, la physiopathologie de l'infection broncho-pulmonaire chez les patients atteints de mucoviscidose est encore mal comprise ; Par exemple, on ne sait pas clairement pourquoi *P. aeruginosa* s'installe préférentiellement chez les patients atteints de mucoviscidose. Plusieurs hypothèses ont été proposées :

Déshydratation du mucus[9] : La mutation *CFTR* serait responsable d'un mucus plus visqueux et moins épais associé à la perte de fonction des cils bronchiques. Les bactéries seraient alors piégées dans ce mucus pour être exposées plus longtemps à l'épithélium respiratoire et ainsi déclencher une réponse inflammatoire.

Augmentation de salinité[9] : L'augmentation de la salinité du mucus aurait pour conséquence de désactiver des peptites antimicrobiens présents comme les bêta-defensines, lysozymes ou les lactoferrines.

Adhérence aux cellules[9] : *P. aeruginosa* serait un ligand naturel des cellules épithéliales portant la protéine CFTR non mutée. Après internalisation, la bactérie serait détruite.

Création d'un milieu favorable : L'inflammation précède la colonisation à *P. aeruginosa* et crée un milieu favorable. Par exemple la production d'alanine et de lactate[6] est une source de carbone métabolisable par *P. aeruginosa*.

Stimulation du système immunitaire : *P. aeruginosa* stimulerait le système immunitaire de façon à éliminer tous ses concurrents et ainsi se développer seul dans sa niche écologique[?] .

Une anomalie du microbiote[? 21;34] : A l'instar de ce qui est bien démontré au niveau intestinal, le microbiote respiratoire jouerait le rôle de barrière contre l'implantation d'agents pathogènes exogènes. La dysbiose induite par la mucoviscidose permettrait à *P. aeruginosa* de s'installer

Le microbiote pulmonaire

Bien qu'il soit en contact avec le milieu extérieur, l'arbre respiratoire (comprenant la trachée, les bronches et les alvéoles) a longtemps été considéré stérile avec les méthodes de culture classique. Il a fallu attendre l'avènement du séquençage haut débit

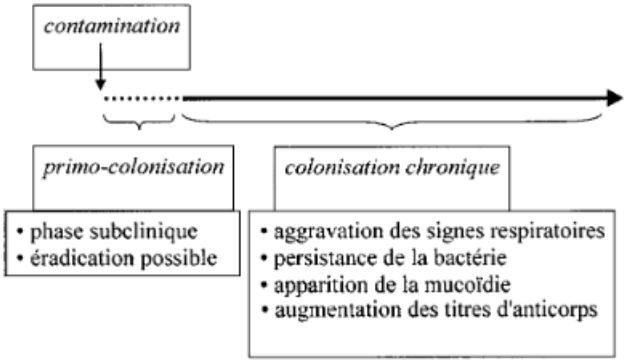

FIGURE 7 – Histoire naturelle de l'infection pulmonaire à *P. aeruginosa* dans la mucoviscidose (P.Plésiat 2003[37]).

pour mettre en évidence le microbiote pulmonaire[21;3;13].

L'arbre respiratoire chez le fœtus [vi] est stérile. Il se colonise comme l'intestin, à la naissance lors du passage de la filière génitale puis se différencie au cours du temps. Il est beaucoup moins riche que le microbiote digestif, mais plus dynamique en raison d'un flux aérien bidirectionnel. Les bactéries du microbiote respiratoire proviennent de l'air ambiant, des voies supérieures, mais aussi du tube digestif par un phénomène de micro aspirations[11]. Une étude a d'ailleurs montré une concordance entre le microbiote digestif et respiratoire[11].

Le microbiote pulmonaire est dominé par le phylum des *Firmicutes* (*Streptococcus*) et des *Bacteroidetes* (*Prevotella*). Les genres retrouvés majoritairement sont *Streptococcus*, *Prevotella*, *Fusobacteria*, *Veillonella*, *Haemophilus*, *Neisseria* et *Porphyromonas*. L'arbre respiratoire étant en continuité directe avec les voies aériennes supérieures il y a un gradient microbien de l'oropharynx jusqu'aux alvéoles. Certains genres bactériens sont communs, comme *Streptococcus*, *Staphylococcus*, *Haemophilus* et *Moraxella*. Tandis que d'autres genres comme *Corynebacterium* et *Dolosigranulum* ne sont retrouvés essentiellement au niveau du nez et de l'oropharynx.

Le microbiote pulmonaire est dynamique. Il peut varier dans l'espace à cause de la structure irrégulière de l'arbre bronchique. Un prélèvement au niveau d'un foyer infectieux, par exemple, se distinguera nécessairement d'un foyer sain. Il existe également des gradients physico-chimiques qui selon la région peuvent sélectionner des espèces. Les cavernes tuberculeuses par exemple se trouvent essentiellement dans le lobe supérieur en raison d'une concentration en oxygène plus élevée qui favorise ce bacille aérobie strict. D'autres études[16;8] ont montré que la diversité diminue avec l'âge ou encore que la structure du microbiote change avec une exposition tabagique. Enfin, les pathologies respiratoires[21] sont une autre source de variabilité du microbiote. Il est par exemple

vi. Un article récent critique l'idée d'un fœtus stérile[36]

démontré dans la mucoviscidose qu'il y une augmentation de richesse et une diminution de diversité. L'asthme et la BPCO sont également associés à des anomalies du microbiote pulmonaire[12;34].
Il semble cependant que ce soit l'individu lui-même qui soit la principale source de variabilité avant tout autre facteur[50].

Exploration du microbiote pulmonaire

Méthodes de prélèvement

Le microbiote respiratoire est exploré en séquençant l'ADN des micro-organismes présents dans un échantillon. Toutes les méthodes de recueil sont possibles, mais les prélèvements protégés (combicath [vii], Liquide Broncho Alvéolaire (LBA)) sont recommandés afin d'éviter une contamination par les voies aériennes supérieures. En pratique, les patients atteints de mucoviscidose sont suivis par le recueil des prélèvements moins invasifs comme les prélèvements pharyngés ou les expectorations spontanées ou induites (examen cytobactériologique du crachat (ECBC)). Dans ces dernières, la qualité du prélèvement peut être évaluée en comptant le nombre de cellules épithélium (normalement < 25 par champ à l'objectif x10) et de polynucléaire (normalement > 10 par champ à l'objectif x10) selon le score de Bartlett adapté par Murray et Washington[?] . Des prélèvements sont également réalisables *in situ* sur les explants lors des greffes pulmonaires.

Séquençage haut débit

Le séquençage de nouvelle génération permet de séquencer de grandes quantités d'ADN dans un échantillon et ainsi de déterminer sa composition en bactéries ou autres micro-organismes. À titre d'exemple, un séquenceur Sanger classique permet de lire des fragments d'ADN d'environ 800 pb parallélisable jusqu'à 96 fois en augmentant le nombre de capillaires. À l'inverse, un séquenceur de nouvelle génération lit des fragments plus courts de l'ordre de 150 pb. Mais la parallélisation peut aller jusqu'à 20 milliards de fois en un seul run sur un Illumina Novaseq.
Deux stratégies de séquençages sont utilisées en écologie microbienne :
La stratégie shotgun consiste à séquencer l'ensemble des ADN présents dans l'échantillon sans discernement, qu'ils soient humains ou bactériens. Les séquences sont filtrées puis les génomes bactériens sont reconstruits par des méthodes bio-informatiques complexes.
La stratégie Amplicon est moins coûteuse sur le plan de l'analyse. Elle consiste à séquencer un gène spécifiquement bactérien et suffisamment variable pour discriminer une espèce. Il s'agit du gène de l'ARNr 16S.

vii. Cette brosse télescopique est glissée au travers d'un fibroscope et dirigée sous contrôle de la vue dans une petite bronche. Le cathéter interne est alors poussé, expulsant le bouchon et permettant d'avancer la brosse de quelques centimètres, pour réaliser le prélèvement bactériologique.

L'ARNr 16S est un ARN non codant participant à la structure de la petite sous unité des ribosomes bactériens. Il est composé de 1542 nucléotides et forme plusieurs boucles dans sa structure secondaire (Figure 8). L'alignement des séquences d'ADNr 16S entre plusieurs espèces met en évidence des régions constantes et 9 régions variables (Figure 9). Les régions constantes permettent de concevoir des amorces s'hybridant sur tous les ADNs bactériens. Les régions variables apportent la spécificité taxonomique permettant d'identifier l'espèce bactérienne.

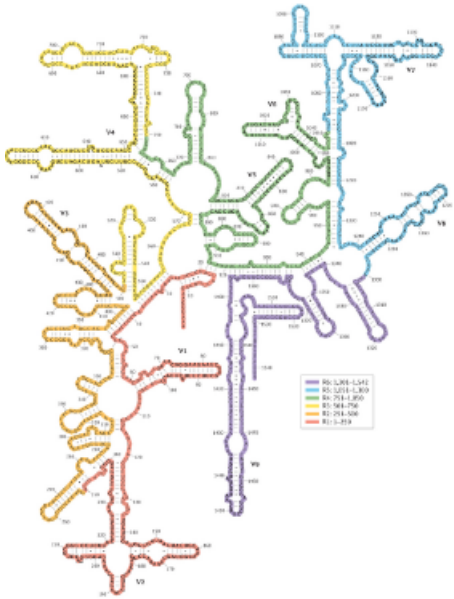

FIGURE 8 – Structure secondaire de l'ARNr 16S

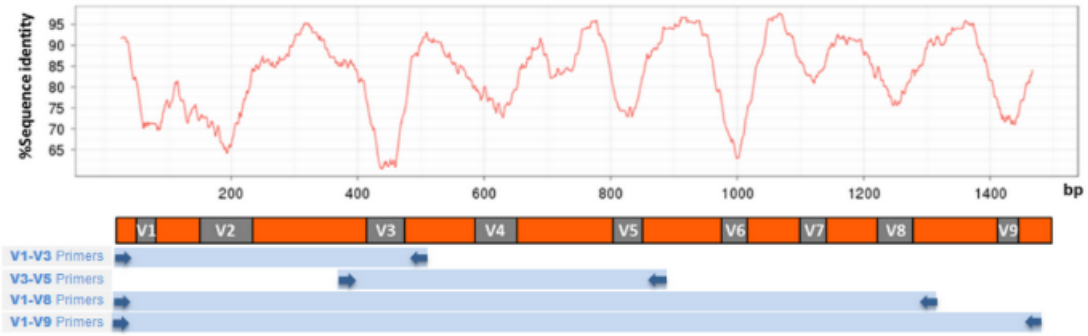

FIGURE 9 – Régions constantes et variables de l'ADNr 16S

En séquençant l'ensemble des génomes bactériens, la stratégie *shotgun* est plus informative, car elle permet de prédire la fonction d'un microbiote. En effet, les transferts génétiques horizontaux amènent à dissocier l'espèce de sa fonction. Deux bactéries d'une

même espèce peuvent avoir des fonctions différentes. L'inférence fonctionnelle à partir de la deuxième stratégie est possible, mais déconseillée. La stratégie 16S reste toutefois une méthode simple pour décrire les populations bactériennes présentes. C'est cette stratégie qui a été utilisée dans notre étude.

Objectif de l'étude Mucobiome

L'objectif principal de l'étude MUCOBIOME est de déterminer si dans la mucoviscidose, le microbiote respiratoire influence la colonisation à *P. aeruginosa*. Pour cela, nous avons suivi pendant trois ans, une cohorte de 96 patients atteints de mucoviscidose exempts de *P. aeruginosa* depuis au moins 1 an. L'exploration de leur microbiote par la stratégie 16S a été réalisée à partir d'expectorations bronchiques obtenues dans le cadre de leur suivi longitudinal. À partir des données générées par le séquençage, nous avons effectué l'étude descriptive et analytique de leur microbiote en créant un pipeline bio-informatique dédié. L'objectif principal de ce travail de thèse a été la mise au point et l'évaluation de ce pipeline créé à façon pour l'étude MUCOBIOME.

Matériel et Méthodes

Recueil des données

Quatre-vingt-seize patients atteints de mucoviscidose ont été suivis sur 3 ans (2008-2011) dans une étude prospective multicentrique (Nantes, Brest, Roscoff) appelée *Mucobiome*. Le comité de protection des personnes (CPP VI-Ouest) et le comité d'éthique du CHRU de Brest ont approuvé le protocole. Tous les patients (ou les parents pour les mineurs) ont signé un consentement éclairé. Le protocole a fait l'objet d'une déclaration de biocollection à l'ARS et au MESR (n° DC-2008-214).

Les expectorations des patients ont été recueillies lors des séances de kinésithérapie respiratoire, tous les 3 mois selon le calendrier des recommandations officielles. En pratique, sur l'ensemble de la cohorte suivie, l'intervalle médian entre 2 consultations a été de 3,4 mois. Les patients devaient avoir un génotypage *CFTR* et un test à la sueur positif. Les transplantés pulmonaires ont été exclus de l'étude. Pour être inclus, le patient devait être exempt de *P. aeruginosa* depuis au moins un an d'après les résultats de la culture bactérienne des expectorations. Le patient devait être en capacité d'expectorer spontanément (exclusion des prélèvements pharyngés). La première culture bactérienne positive à *P. aeruginosa* était le « endpoint » de l'étude. À cette étape, le patient était sorti de l'étude pour être réinclus un an plus tard en cas de maintien de la négativité des cultures bactériennes obtenue par l'antibiothérapie d'éradication anti-*P. aeruginosa*. D'autres données clinico-biologiques (Age, Sexe, IMC, type de mutation, Présence ou absence de P. aeruginosa en culture et en qPCR, score cytologique de Bartlett, Categorie Free ou Never selon Lee[27]) ont également été recueillies. Au total, 707

expectorations ont été recueillies à l'issue de l'étude MUCOBIOME. Pour ce travail portant sur l'étude du microbiote broncho-pulmonaire, 188 ont été sélectionnées dont 22 expectorations positives à *P. aeruginosa*.

La procédure d'extraction des ADN des échantillons est détaillée dans la publication de F. Gall[25]. Brièvement les échantillons ont été liquéfiés avec du Dithiotréitol (DTT). Les protéines ont été dégradées avec de la protéinase K. Les parois bactériennes ont été fragmentées par sonication (DTT par sonication, Elamsonic S10, Singen, Germany). Après 10 min de centrifugation, L'ADN a été extrait à partir du culot à l'aide du kit QUIAamp DNA Minikit (Quagen). Les extraits d'ADN ont été envoyés pour séquençage par le prestataire GATC.

Séquençage

La librairie a été générée en amplifiant la région V3-V5 à l'aide du couple d'amorces forward CCTACGGGAGGCAGCAG et reverse CCGTCAATTCMTTTRAGT et du kit MiSeq Reagent Kits v3.
Le séquençage a été réalisé sur Illumina MiSeq pour produire un couple de séquences chevauchantes de 300 pb (Figure 10). Après fusion du couple, le séquençage permet de lire 535 pb correspondant à la région V3-V5.
Environ 25 millions de reads sont produits par run MiSeq. En multiplexant à l'aide de 94 indexes, les 188 échantillons ont été séquencés sur 2 runs pour produire 188 x 2 fichiers fastq.

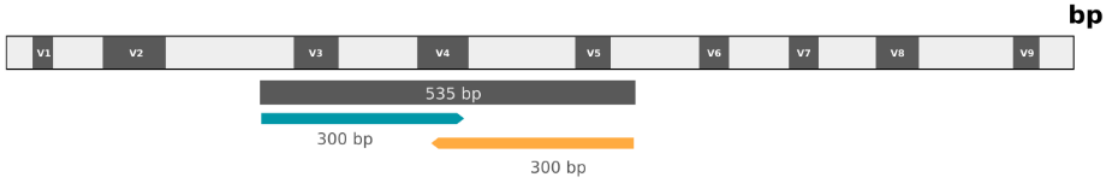

FIGURE 10 – le couple de séquence de 300 pb permet de recouvrir l'ensemble de la région V3-V5 de l'ARN 16S

Analyse bio-informatique

L'analyse des 188 paires de fichiers fastq a été réalisée grâce à un pipeline bio-informatique, appelé *mucobiome*, conçu et testé dans le cadre de cette étude. Cet outil modélise l'ensemble du pipeline sous forme d'un graphe direct acyclique (DAG) et le résout afin d'optimiser la parallélisation.
Le pipeline mucobiome prend en entrée, les 188 couples de fichiers FASTQ provenant du séquençage et produit un fichier BIOM contenant la table des OTUs. Les figures 11 et 12 illustrent toutes les étapes du pipeline qui sont décrites ci-dessous.

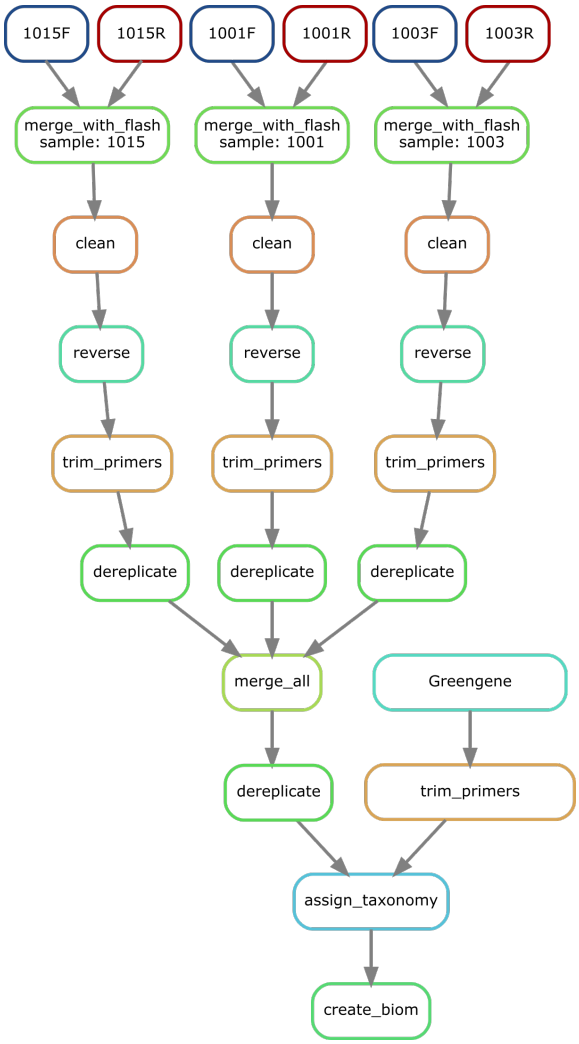

FIGURE 11 – Graphe du pipeline simplifié sur 3 échantillons (1015, 1001 et 1003). **Merging** : Les reads pairs de 300 pb sont fusionnés pour produire un fichier fastq contenant des reads de 535 pb. **Cleaning** : Les reads de mauvaise qualité sont supprimés. **Reversing** : Les reads sont transformés en leurs séquences complémentaires pour pouvoir être alignés. **Trimming** : Seule la séquence entre les primers V3-V5 est conservée. **Dereplicating** : Les séquences dupliquées sont retirées. **Merging** : L'ensemble des séquences est regroupé dans un seul fichier. **Taxonomy assignement** : Les séquences sont alignées sur la base de données *Greengene*. **create_biom** : la table des OTUs est créée

Étape du pipeline

Merging : Fusions des reads

2 fastq en entrée 1 fastq en sortie.

La première étape du pipeline consiste à fusionner le couple de fichiers fastq afin de produire un seul fichier contenant des plus longues séquences de 535 pb correspondant à la région V3-V5 de l'ADNr 16S. Le programme **Flash**[29] a été utilisé avec les paramètres

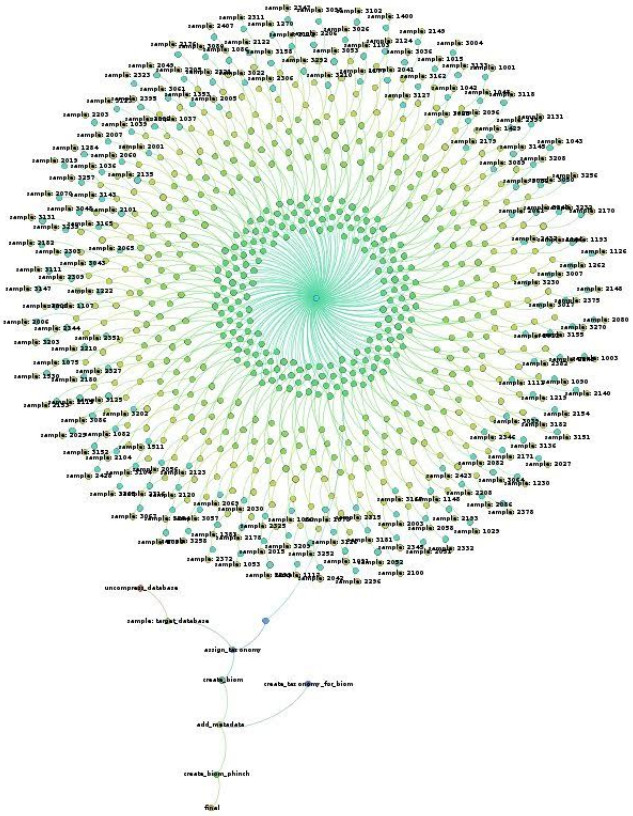

FIGURE 12 – le graphe du pipeline sur l'ensemble des échantillons. Le tournesol montre le niveau de parallélisation nécessaire.

par défaut. À partir de deux fichiers fastq, ce dernier recherche le meilleur alignement entre deux reads et produit un fichier fastq contenant les reads fusionnés. Une analyse qualitative des reads a été réalisée avec **FastQt**[44] avant et après fusion.

Cleaning : Filtrage des qualités

1 fastq en entrée 1 fastq en sortie.

Les données de séquençage haut débit peuvent contenir beaucoup d'erreurs. Il est important de supprimer les reads de mauvaise qualité pour gagner en spécificité. Le filtrage des reads de mauvaise qualité est réalisé avec le programme **sickle**[22]. Son algorithme repose sur l'utilisation d'une fenêtre glissante de taille définie (par défaut : 20 pb). Cette fenêtre glisse le long de la séquence et pour chaque position calcule la moyenne des scores de qualité dans cette fenêtre. Si successivement le score moyen passe sous un certain seuil, le read est supprimé. Les paramètres utilisés sont ceux par défaut : un score de 20 avec une fenêtre glissante de 20 pb. Une analyse qualitative des reads a été réalisée avec **FastQt**[44] après le filtrage.

Reversing : Séquence complémentaire

1 fastq en entrée 1 fasta en sortie.

Les reads produits par le séquenceur ne sont pas orientés dans le même sens que la base de données. *Greengene*[10]. Pour permettre l'alignement, les séquences ont été remplacées par leurs séquences complémentaires grâce au programme **seqtk**[19]. Par la même occasion les scores de qualités devenus inutiles sont supprimés. Les séquences sont sauvegardées dans un fichier fasta.

Trimming : Suppression des primers

1 fasta en entrée 1 fasta en sortie.

Pour permettre un alignement parfait entre les reads et la base de données, les primers sont retirés et seule la séquence V3-V5 est conservée. Cette étape est réalisée aussi bien pour les données du séquençage que la base de données *Greengene*. Le programme **cutadapt**[31] est utilisé avec une tolérance de 0,1 par défaut.

Dereplicating : Suppression des doublons

1 fasta en entrée 1 fasta en sortie.

Cette étape consiste à supprimer toutes les séquences doublons. En procédant ainsi, on s'assure de ne pas répéter l'assignation taxonomique plusieurs fois sur des reads identiques. Le nombre de reads dupliqués est conservé pour être pris en considération lors du calcul des abondances. Il s'agit d'une étape d'optimisation permettant d'économiser en temps de calcul. La déréplication a été réalisée avec **vsearch**[46] et sa fonction *–derep_fulllength*

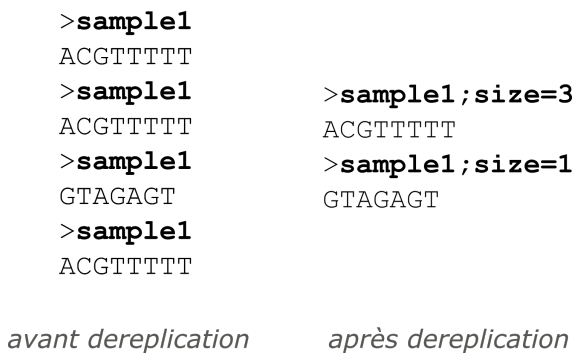

FIGURE 13 – Exemple de déréplication d'un fichier fasta

Assignation taxonomique

2 fasta en entrée, 1 fichier biom en sortie.

L'assignation taxonomique consiste à associer chaque read à son taxon. Nous avons utilisé la stratégie *close référence* dont le rôle est de comparer chaque read à une base de données avec un seuil de 97 % de similarité. Cet algorithme est de complexité N.

C'est-à-dire que le temps de calcul est directement proportionnel au nombre de reads testé. La base de données *Greengene*[10] version mai 2013 a été utilisée. C'est un fichier fasta contenant 1 262 986 séquences et 203 452 OTUs.
La stratégie d'assignation *de novo* n'a pas été utilisée. Cette dernière, de complexité N^2, consiste à comparer les reads entre eux pour former des groupes, sans utilisation d'une base de donnée. L'assignation taxonomique a été réalisé avec **vsearch** et sa fonction *–usearch_ global*.

Analyse de la table des OTUs

L'analyse de la table des OTUs a été réalisée avec R et le package **phyloseq**[32]. Les OTUs ont été regroupés par genre bactérien. La normalisation de la table des OTUs s'est faite en pondérant le nombre de reads au total. Les abondances absolue et relative, la prévalence, la dominance et la courbe de raréfaction ont été mesurées pour chaque échantillon. Les diversités alpha de type Chao1 et Shannon ont été mesurées sur tous les échantillons. L'évolution de ces indices a été représentée graphiquement dans le temps et par patient. La diversité bêta a été évaluée grâce une Analyse en Coordonnées Principales utilisant la distance de Bray-Curtis. Le core microbiota[48] a été calculé. Il est défini comme l'ensemble des taxons retrouvé dans plus de 50 % des échantillons avec une abondance supérieure à 0,1 %. La corrélation entre les genres a été recherchée. Les catégories Free et Never ont été comparées d'après leurs diversités alpha et bêta.

Résultats

Séquençage et pipeline

Données de séquençage

Après démultiplexage, 188 x 2 fichiers Fastq ont été générés soit deux fichiers par échantillons. La longueur des reads dans l'ensemble est de 301 paires de bases. Au total, 115 002 297 reads ont été produits sur deux runs MiSeq (Figure **??**). Avec en moyenne 616 900 reads par échantillon. Un minimum de 61 422 reads pour l'échantillon 2154 et un maximum de 1 071 188 pour l'échantillon 3165.

Qualité des reads

La figure 15 montre la qualité typique d'un fichier fastq produit par le séquenceur. Une baisse de qualité importante est observée à hauteur du 250e nucléotide. Tous les fichiers fastq présentaient le même profil. À cause de cette baisse de qualité, en moyenne 37,8 % des reads pairs n'ont pas pu être fusionnés. L'étape de filtrage a permis de ramener la qualité médiane au-dessus de 20 (Figure 16). Au total, seulement 49.24 % des reads ont été conservés pour l'analyse (Figure **??**) avec des bornes allant de 37,30 % à 61,13 %.

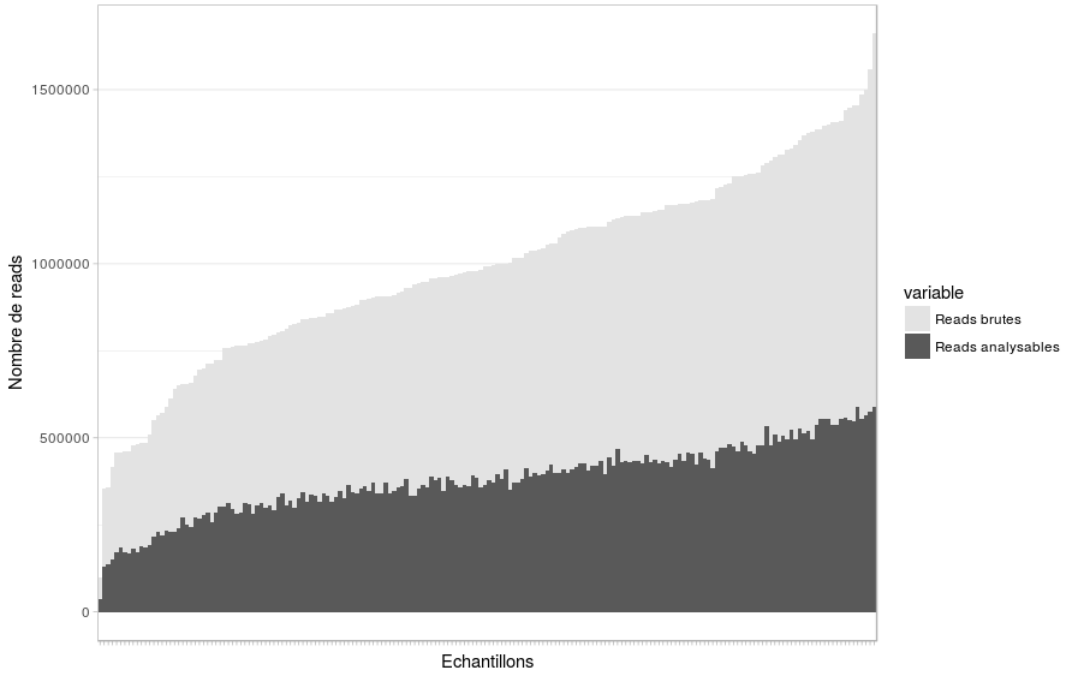

FIGURE 14 – Nombre de reads par échantillon. En clair le nombre de reads bruts produits par le séquenceur. En sombre les reads analysables.

Assignation taxonomique

99.88 % (4615960) des reads analysables ont reçu une assignation taxonomique soit 10517 OTUs détectés [viii] correspondant à 54 genres bactériens (Figure 17).

Profondeur de séquençage

La figure 18 montre les courbes de raréfaction pour les 188 échantillons. Dans l'ensemble elles s'aplatissent précocement, témoignant d'un bon niveau d'échantillonnage. Les quelques échantillons n'ayant pas attend l'asymptote horizontale ont tout de même été conservés.

Résultats descriptifs

Composition du microbiote

Cinquante-quatre genres (Figure 17) et sept phyla (Figure 19) bactériens sont retrouvés dans l'ensemble des échantillons analysés. Les deux phyla majoritaires sont *Firmicute* (42.59 %) et *Proteobacteria* (31.48 %). Parmi les *Firmicutes* majoritaires, on trouve *Streptococcus* et *Staphylococcus*. Au sein des *Proteobacteria*, *Neisseria* et *Haemophilus* sont les genres les plus abondants. Le tableau 2 résume l'ensemble des résultats en y associant la prévalence des genres bactériens parmi les 188 échantillons. Par exemple *Streptococcus*, *Neisseria*, *Prevotella*, *Granulicatella*, *Gemella*, *Veillonella* et *Fusobacte-*

viii. Pour chaque espèce, il y a plusieurs OTUs définis dans *Greengene*

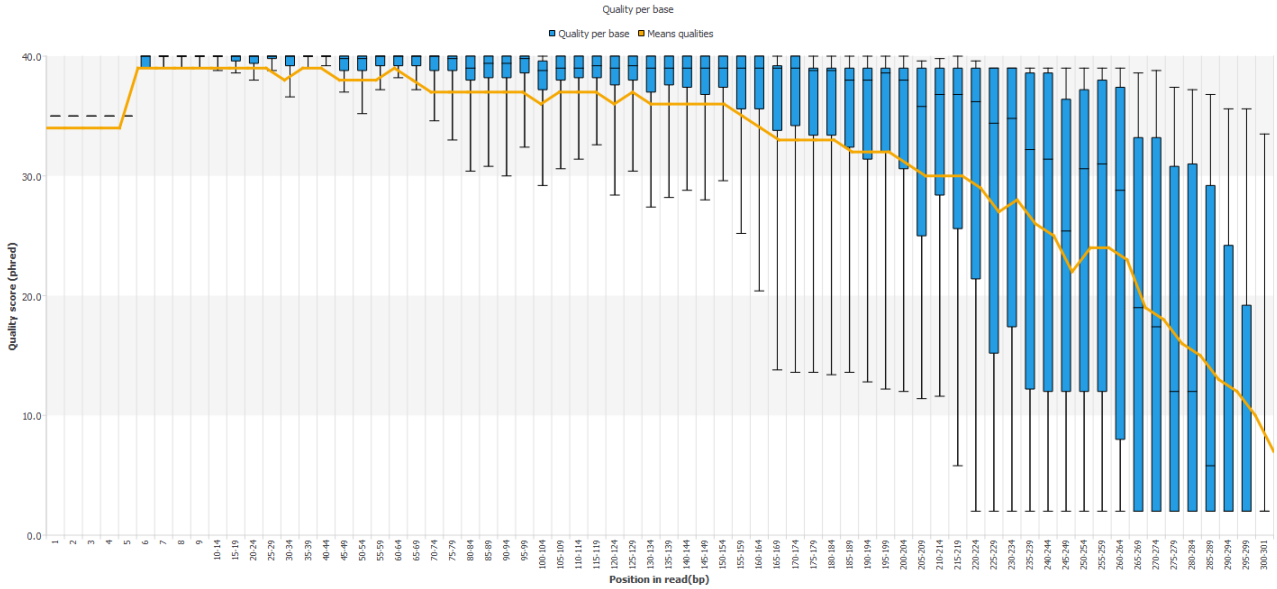

FIGURE 15 – Qualité par nucléotide des reads forward de l'échantillon 1003. **Axe X** : la position sur le read. **Axe Y** : La distribution des qualités

rium sont très prévalents, car présents dans plus de 185 échantillons. D'autres genres sont dominants. Il s'agit de *Streptococcus*, *Neisseria*, *Haemophilus* et *Staphyloccoccus*. *Stenotrophomonas* et *Achromobacter* sont retrouvées dans 64 et 8 échantillons respectivement. *Pseudomonas* est retrouvé dans 53 échantillons, dont un, au moins avec une abondance de 25.94 %. *Burkholderia* est retrouvé seulement dans deux échantillons à moins de 1 %.

Le core microbiota est constitué de 14 genres bactériens (Figure 20). Sa distribution est illustrée dans la figure 21. *Streptococcus* respecte grossièrement une distribution normale variant de la quasi-absence à la dominance avec une moyenne de 30 % par échantillon. *Neisseria* est le deuxième genre le plus abondant avec une moyenne de 18 %. Les abondances de *Staphylococcus* et *Haemophilus* sont faibles dans la plupart des échantillons. Mais pour quelques échantillons, ces genres sont dominants. Les autres genres ont une abondance faible qui varie faiblement. Ils ne sont jamais retrouvés comme dominants.

Évolution de l'alpha diversité

Les figures 31,32 et 33 en annexe, montrent l'évolution des diversités alpha par patient au cours du temps en utilisant les indices Chao1, Observed et Shannon.

La richesse (Observed , Chao1) par patient varie de 10 à 40 genres bactériens. La richesse du patient 256 est faible. Celle du patient 26 élevée. Certains patients ont des richesses stables au court du temps. Les patients 25, 26 et 74 conservent leurs richesses sur plus de 5 prélèvements successifs. Les patients 20 et 69 présentent une stabilité globale entrecoupée par des pertes de biodiversité. La richesse du patient 8 diminue

	Reads count (% of all reads)	Mean of Relative abundance %	Standard deviation of Relative abundance %	Minimum relative abundance %	Maximumn relative abundance %	Present N (% of all samples)	Most abundant N (% of all samples)
Streptococcus	1358573 (29.83)	31.02	20.38	0.02	90.50	188 (100)	95 (175.93)
Neisseria	807931 (17.74)	17.29	18.04	0.00	93.53	185 (98.4)	40 (74.07)
Haemophilus	686186 (15.07)	13.43	24.26	0.00	99.77	183 (97.3)	29 (53.7)
Staphylococcus	377739 (8.3)	7.62	16.13	0.00	98.27	175 (93.1)	12 (22.22)
Prevotella	227445 (4.99)	5.37	5.86	0.00	29.91	187 (99.5)	0 (0)
Granulicatella	199458 (4.38)	4.45	4.11	0.00	28.42	185 (98.4)	0 (0)
Gemella	161797 (3.55)	3.74	3.67	0.00	17.48	182 (96.8)	0 (0)
Stenotrophomonas	129326 (2.84)	2.96	14.61	0.00	94.17	53 (28.2)	7 (12.96)
Fusobacterium	91380 (2.01)	2.22	2.77	0.00	16.19	182 (96.8)	0 (0)
Veillonella	88875 (1.95)	2.16	2.66	0.00	14.16	187 (99.5)	0 (0)
Porphyromonas	81748 (1.8)	1.84	3.29	0.00	25.90	162 (86.2)	0 (0)
[Prevotella]	50207 (1.1)	1.24	2.09	0.00	17.22	165 (87.8)	0 (0)
Leptotrichia	37561 (0.82)	0.89	1.53	0.00	13.48	170 (90.4)	0 (0)
Pseudomonas	37178 (0.82)	0.85	4.02	0.00	27.45	48 (25.5)	1 (1.85)
Moraxella	45865 (1.01)	0.79	6.37	0.00	76.71	46 (24.5)	2 (3.7)
Achromobacter	17729 (0.39)	0.52	7.08	0.00	97.14	7 (3.7)	1 (1.85)
Oribacterium	22236 (0.49)	0.51	0.76	0.00	5.70	178 (94.7)	0 (0)
Capnocytophaga	20073 (0.44)	0.50	0.96	0.00	7.71	174 (92.6)	0 (0)
Lautropia	20785 (0.46)	0.45	1.22	0.00	11.23	136 (72.3)	0 (0)
Actinobacillus	17430 (0.38)	0.42	2.57	0.00	34.08	93 (49.5)	1 (1.85)
Parvimonas	10957 (0.24)	0.26	0.71	0.00	5.96	147 (78.2)	0 (0)
Kingella	8289 (0.18)	0.19	0.52	0.00	5.08	144 (76.6)	0 (0)
Atopobium	7356 (0.16)	0.18	0.34	0.00	2.43	158 (84)	0 (0)
Moryella	7731 (0.17)	0.17	0.42	0.00	3.84	140 (74.5)	0 (0)
Peptostreptococcus	5793 (0.13)	0.15	0.31	0.00	1.82	120 (63.8)	0 (0)
Catonella	5851 (0.13)	0.14	0.19	0.00	1.20	157 (83.5)	0 (0)
Actinomyces	5196 (0.11)	0.13	0.23	0.00	1.93	161 (85.6)	0 (0)
Aggregatibacter	5657 (0.12)	0.13	0.39	0.00	3.12	83 (44.1)	0 (0)
Eikenella	4327 (0.1)	0.10	0.33	0.00	2.40	138 (73.4)	0 (0)
Rothia	3333 (0.07)	0.07	0.16	0.00	1.26	117 (62.2)	0 (0)
Morganella	3317 (0.07)	0.06	0.78	0.00	10.67	5 (2.7)	0 (0)
Megasphaera	2425 (0.05)	0.06	0.19	0.00	2.00	87 (46.3)	0 (0)
Selenomonas	800 (0.02)	0.02	0.07	0.00	0.58	108 (57.4)	0 (0)
Tannerella	895 (0.02)	0.02	0.04	0.00	0.38	116 (61.7)	0 (0)
Treponema	229 (0.01)	0.01	0.03	0.00	0.27	34 (18.1)	0 (0)
Burkholderia	217 (<0.01)	0.01	0.09	0.00	1.17	2 (1.1)	0 (0)
Corynebacterium	317 (0.01)	0.01	0.02	0.00	0.16	77 (41)	0 (0)
Campylobacter	344 (0.01)	0.01	0.02	0.00	0.14	88 (46.8)	0 (0)
Peptoniphilus	31 (<0.01)	0.00	0.00	0.00	0.03	11 (5.9)	0 (0)
Vagococcus	13 (<0.01)	0.00	0.00	0.00	0.02	10 (5.3)	0 (0)
Curvibacter	12 (<0.01)	0.00	0.00	0.00	0.01	11 (5.9)	0 (0)
Mycoplasma	49 (<0.01)	0.00	0.00	0.00	0.02	18 (9.6)	0 (0)
Lactococcus	18 (<0.01)	0.00	0.00	0.00	0.02	7 (3.7)	0 (0)
Paludibacter	128 (<0.01)	0.00	0.02	0.00	0.22	18 (9.6)	0 (0)
Anaerococcus	137 (<0.01)	0.00	0.05	0.00	0.64	2 (1.1)	0 (0)
Schwartzia	41 (<0.01)	0.00	0.01	0.00	0.13	5 (2.7)	0 (0)
Filifactor	85 (<0.01)	0.00	0.02	0.00	0.17	5 (2.7)	0 (0)
Peptococcus	37 (<0.01)	0.00	0.00	0.00	0.02	16 (8.5)	0 (0)
Lactobacillus	198 (<0.01)	0.00	0.02	0.00	0.29	26 (13.8)	0 (0)
Enhydrobacter	11 (<0.01)	0.00	0.00	0.00	0.01	7 (3.7)	0 (0)
Dialister	195 (<0.01)	0.00	0.02	0.00	0.13	39 (20.7)	0 (0)
Proteus	10 (<0.01)	0.00	0.00	0.00	0.02	7 (3.7)	0 (0)
Enterococcus	45 (<0.01)	0.00	0.01	0.00	0.11	22 (11.7)	0 (0)
Butyrivibrio	103 (<0.01)	0.00	0.02	0.00	0.21	10 (5.3)	0 (0)

TABLE 2 – Résumé détaillé de l'assignation taxonomique des reads. **Legende :** Nombre de reads - Moyenne des abondances relatives - écart-type des abondances relatives - plus petite abondance relative - plus grande abondance relative - Nombre d'échantillons où le genre est présent - Nombre d'échantillon où le genre est dominant

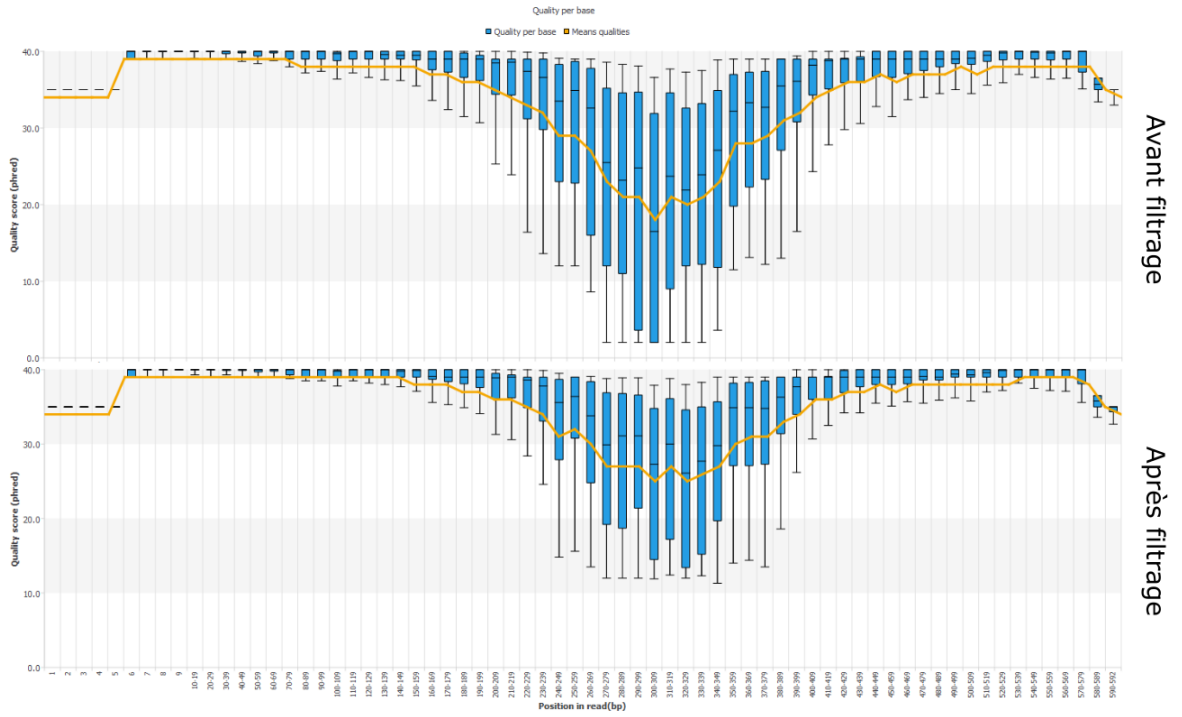

FIGURE 16 – Qualité par nucléotide des reads fusionnés pour l'échantillon 1003

progressivement. Les patients 23, 232 et 248 ont des fluctuations plus chaotiques. L'indice de Shannon montre que pour un richesse constante dans le temps, l'équitabilité est différente. Par exemple, la richesse du patient 223 est stable sur la figure, 31 mais son indice de shannon varie sur la figure 33 témoignant d'une distribution différente de ses bactéries. Seul le patient 26 semble conserver à la fois une richesse et une équitabilité stables au court du temps.

Évolution des abondances

Les figures 34, 35 et 36 montrent l'évolution des abondances au cours du temps pour chaque patient. Ces graphiques nous permettent d'interpréter plus finement les graphiques d'alpha diversité précédents. Par exemple, la diminution de richesse du patient 8 est liée à la présence de *Stenotrophomonas* sur les 4 échantillons. La perte de diversité du 3e prélèvement du patient 223 est causée par l'apparition de *Neisseria* qui devient dominant. Le patient 3 montre une diminution progressive *d'Haemophilus* parallèlement à une augmentation de *Streptococcus*.
Le patient 256 présente une dominance à *Stenotrophomonas* sur l'ensemble de ses échantillons. *Achromobacter* est dominant dans le premier échantillon du patient 211. D'autres montrent un phénomène de résilience. Par exemple, le patient 223 récupère un microbiote identique après une colonisation quasi omnisciente à *Haemophilus*.
Dans l'ensemble, il existe deux types de population bactérienne : une population de bac-

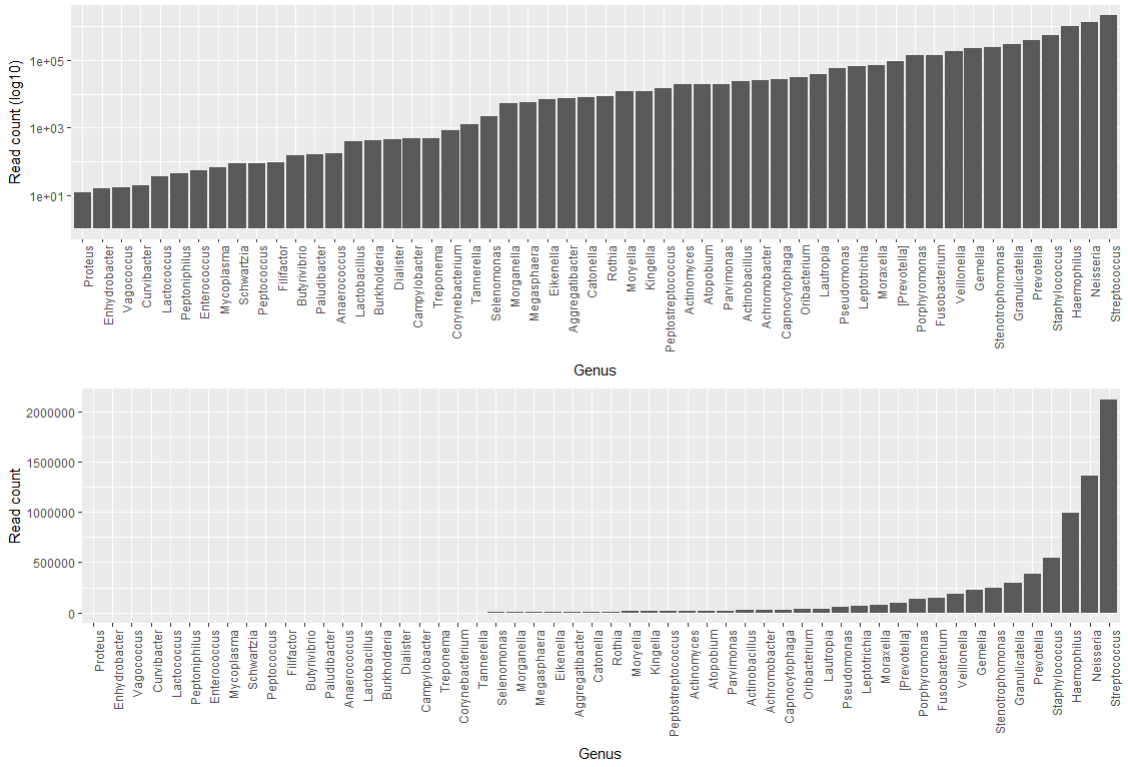

FIGURE 17 – Nombre de reads par genre bactérien

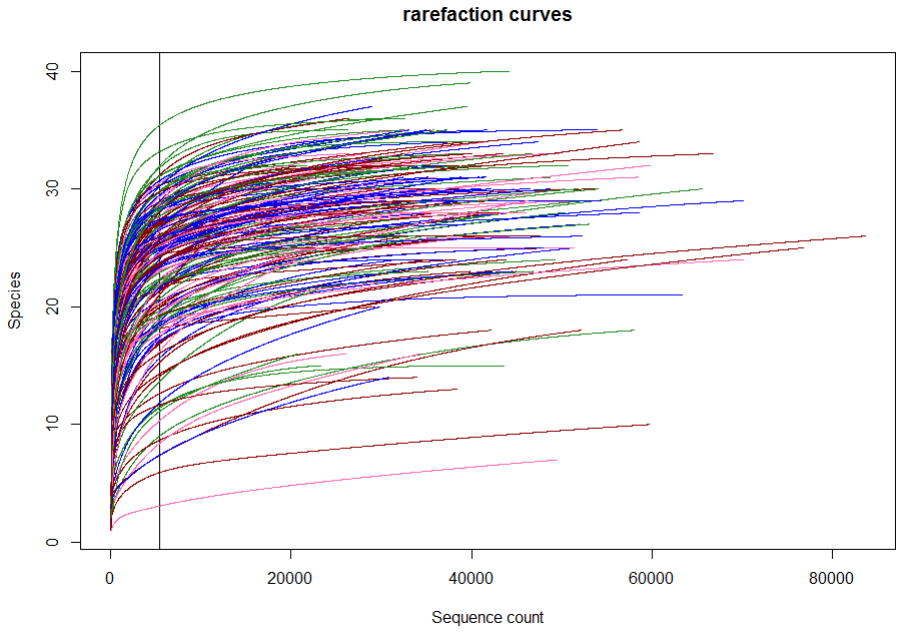

FIGURE 18 – Courbe de raréfaction pour les 188 échantillons.

téries constantes et minoritaires (*Fusobacterium*, *Granulicatella*, *Gemella*, *Veillonella*, *Parvimonas*, *Leptotrichia*,*Oribacterium*, *Capnoctophaga*,*Catonella*) ; une autre popula-

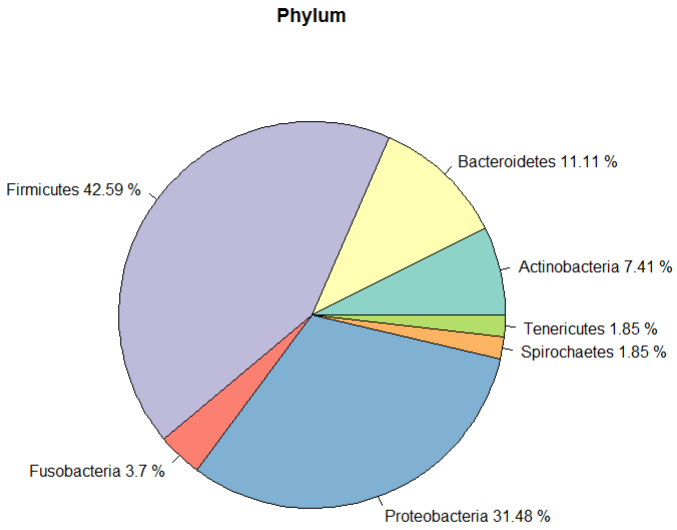

FIGURE 19 – Phyla retrouvés dans les 188 échantillons

tion de bactéries très fluctuantes dans le temps jouant alternativement le rôle de genre dominant ou au contraire de genre absent. (*Haemophilus*, *Streptococcus*, *Neisseria* et *Prevotella*). La figure 22 zoome sur le patient 43 pour montrer ces deux populations. La figure du patient 23 est mise également à titre exemple, car elle montre l'apparition de *P. aeruginosa*.

Bêta Diversité

La bêta diversité sur l'ensemble des échantillons a été réalisée par une méthode d'ordination de type PCoA en utilisant les distances de Bray-Curtis (Figure 24,25). Les deux axes principaux expliquent respectivement 28.5 % et 17.4 % de la variabilité. Certains échantillons d'un même patient sont très proches sur le graphique d'ordination. Par exemple le patient 69 et 003 ont des échantillons dont les points se confondent. Aucune des analyses tenant compte des paramètres biocliniques colligés pendant l'étude MUCOBIOME comme l'âge, le sexe, le poids, la prise d'antibiotique et le type de mutation du gène *CFTR* n'a mis en évidence des groupes distincts de microbiote. Le score cytologique n'a pas non plus montré de différence. La variabilité est expliquée principalement par la dominance des genres *Neisseria*, *Streptococcus* et *Haemophilus* comme le montre la figure 24.

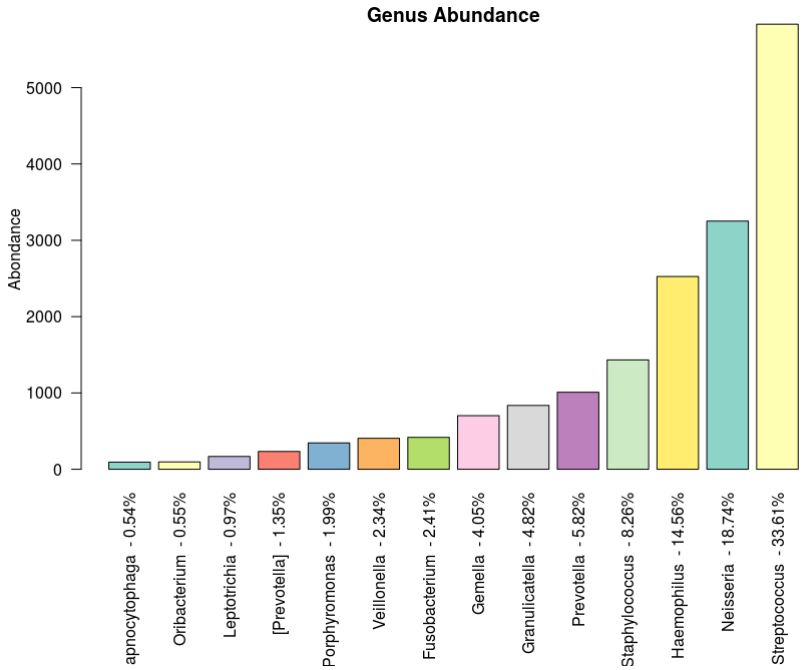

FIGURE 20 – Distribution du core microbiota dans l'ensemble des échantillons

Présence de P. aeruginosa

Sur les 188 échantillons, 22 étaient positifs à *P.aeruginosa* en culture. L'ADNr 16S de Pseudomonas est détecté dans 48 échantillons. Dans la majorité des cas, l'abondance relative de P. aeruginosa est faible. Sur les 48 échantillons, 39 ont une abondance relative en dessous de 3 %. L'abondance la plus forte est de 27,4 % et correspond au « end point » positif en culture du patient 10. Les tableaux 3 et 4 comparent le nombre d'échantillons arborant des reads ADNr 16S de *P.aeruginosa* avec ceux détectés par qPCR (données de l'étude de Héry-Arnaud et al., 2017[?]) et par culture. En considérant arbitrairement la méthode 16S comme un gold standard, ces deux tableaux nous permettent de mesurer la sensibilité et la spécificité de la détection par qPCR ou par culture. Ainsi la qPCR oprL est très sensible et peu spécifique avec un seul vrai négatif pour 31 faux positifs. La culture est en revanche très spécifique avec 9 faux positifs.

	qPCR négative	qPCR positive
16S absent	141	31
16S présent	1	15

TABLE 3 – Concordance entre la présence d'ADNr 16S de Pseudomonas et sa détection en qPCR oprL

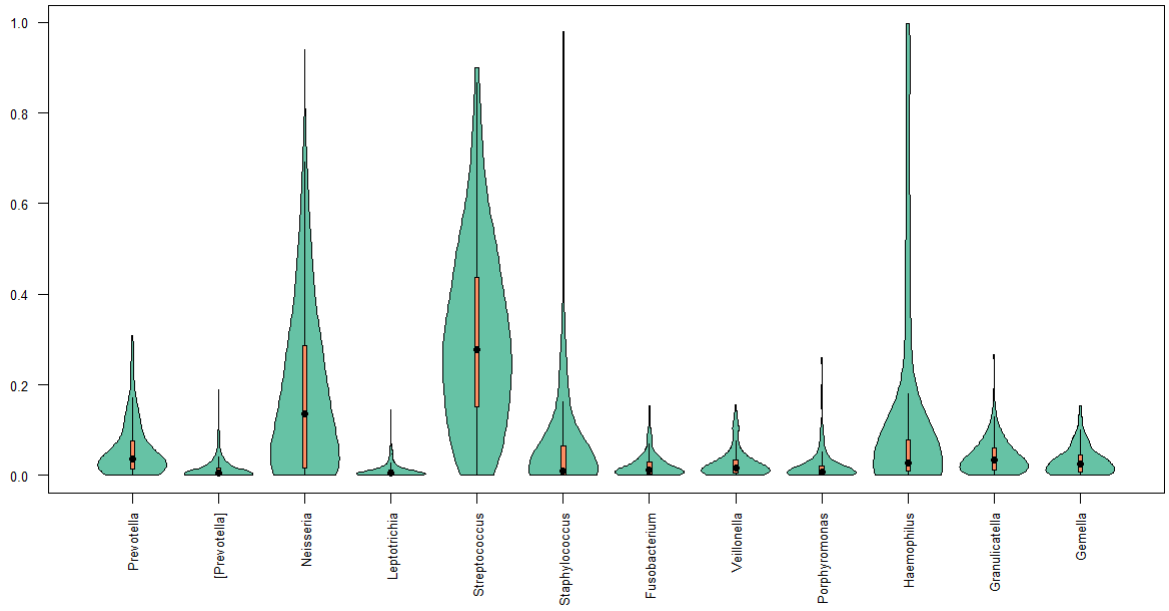

FIGURE 21 – Variation du core microbiota dans l'ensemble des échantillons

	Culture négative	Culture positive
16S absent	163	9
16S present	3	13

TABLE 4 – Concordance entre la présence d'ADNr 16S de Pseudomonas et sa détection en culture

Résultats analytiques

Corrélation entre les genres bactériens

La figure 26 montre les corrélations linéaires réalisées entre les genres du core microbiota. La corrélation la plus forte est entre *Prevotella* et *Veillonella* avec un coefficient de Pearson à 0.60. *Streptococcus* et *Haemophilus* évoluent dans le sens inverse avec un coefficient à -0.41.

Comparaison entre échantillons Free et Never

La figure 27 compare les indices de Shannon entre les échantillons Free et Never. Le T-test / ANOVA n'a pas montré de différence significative (p-value > 0.05).
La figure 28 est une autre PCoA réalisée avec microbiomeanalyst [ix] tenant de discriminer les groupes Free et Never. Les échantillons Never ont plus de différences entre eux que les échantillons Free qui semblent converger. Les deux groupes sont significativement différents avec une PERMANOVA p-value < 0.001.

ix. http ://www.microbiomeanalyst.ca

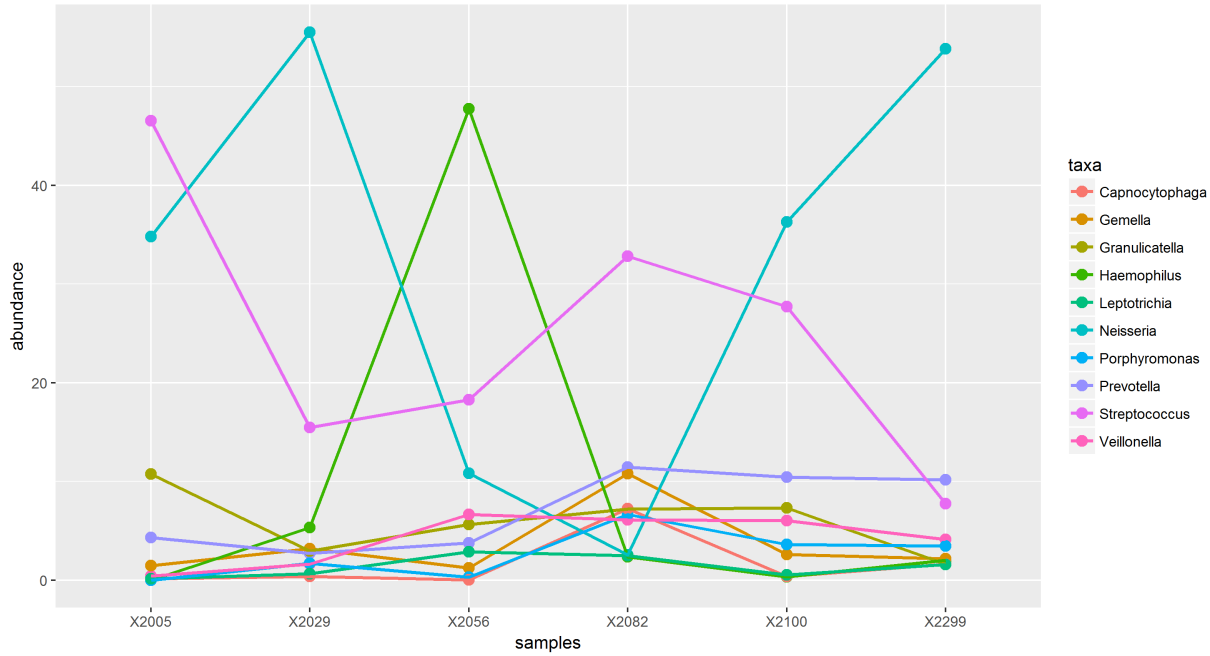

FIGURE 22 – Évolution des abondances pour le patient 43. Notez la population stable et la population fluctuante

.

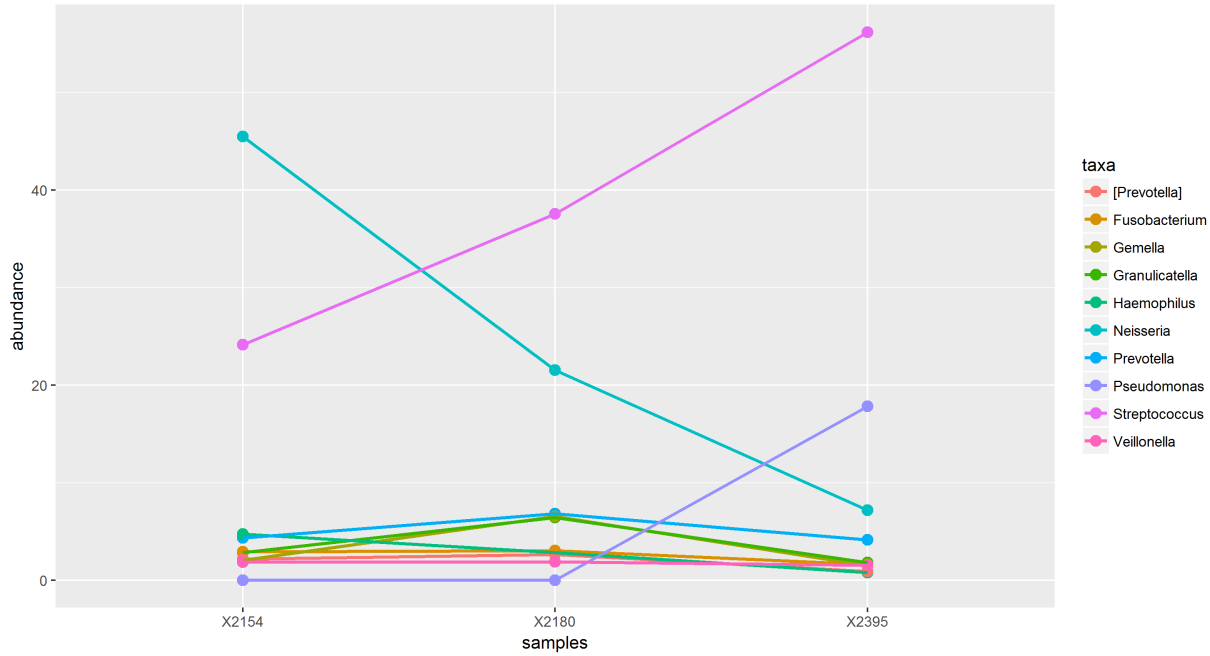

FIGURE 23 – Évolution des abondances pour le patient 54. Notez l'apparition de *Pseudomonas* dans le dernier échantillon

.

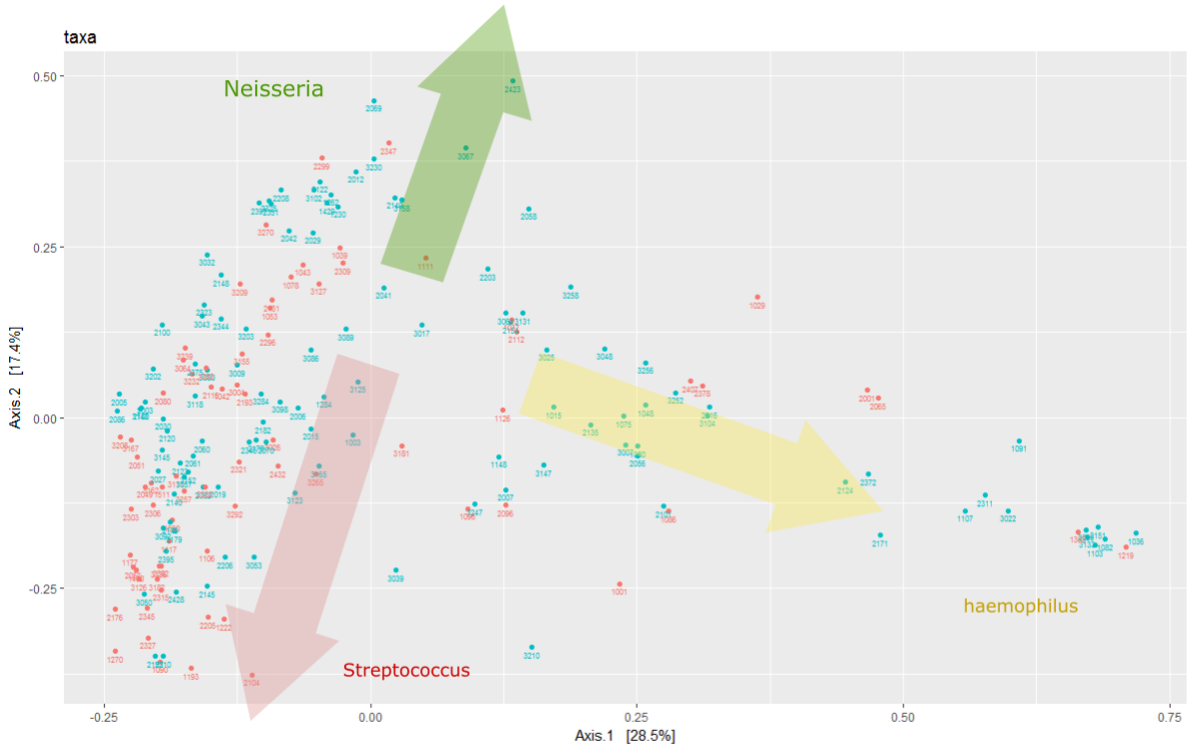

FIGURE 24 – Analyse en coordonnées principales sur les 188 échantillons en utilisant les distances de Bray-Curtis. Chaque point est un échantillon labellisé par l'identifiant du patient. La variabilité s'explique principalement par *Haemophilus*, *Streptococcus* et *Neisseria*

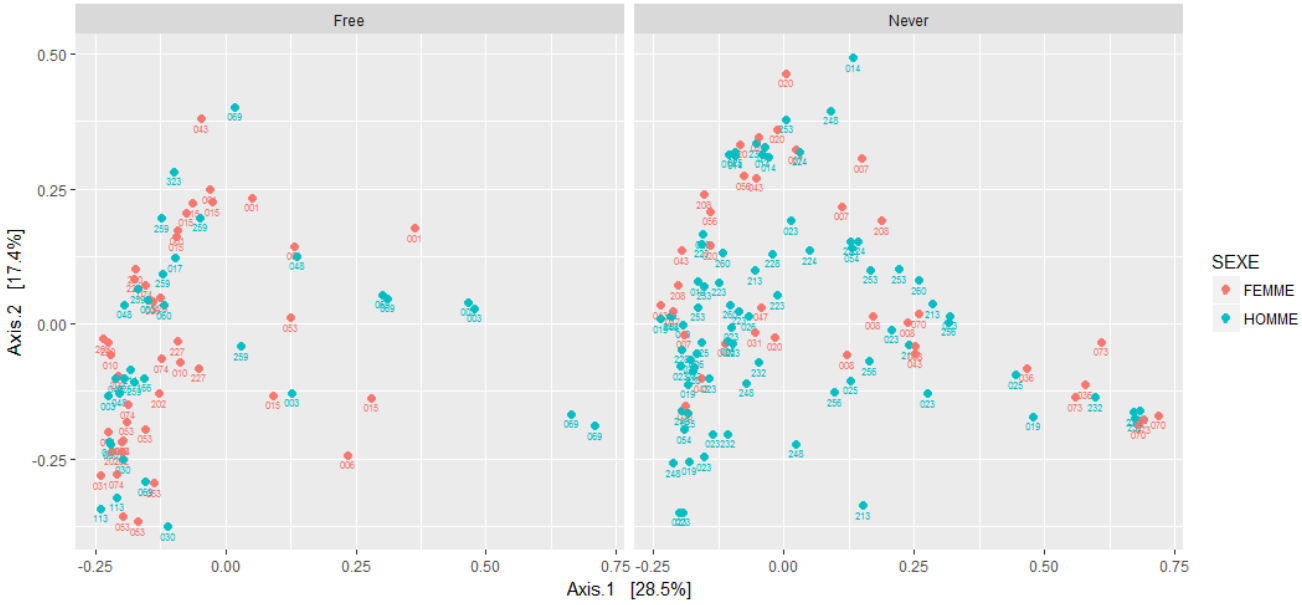

FIGURE 25 – Analyse en coordonnées principales sur les 188 échantillons en utilisant les distances de Bray-Curtis. A gauche les échantillons Free, à droite les échantillons Never. Le sexe est mis en couleur à titre indicatif.

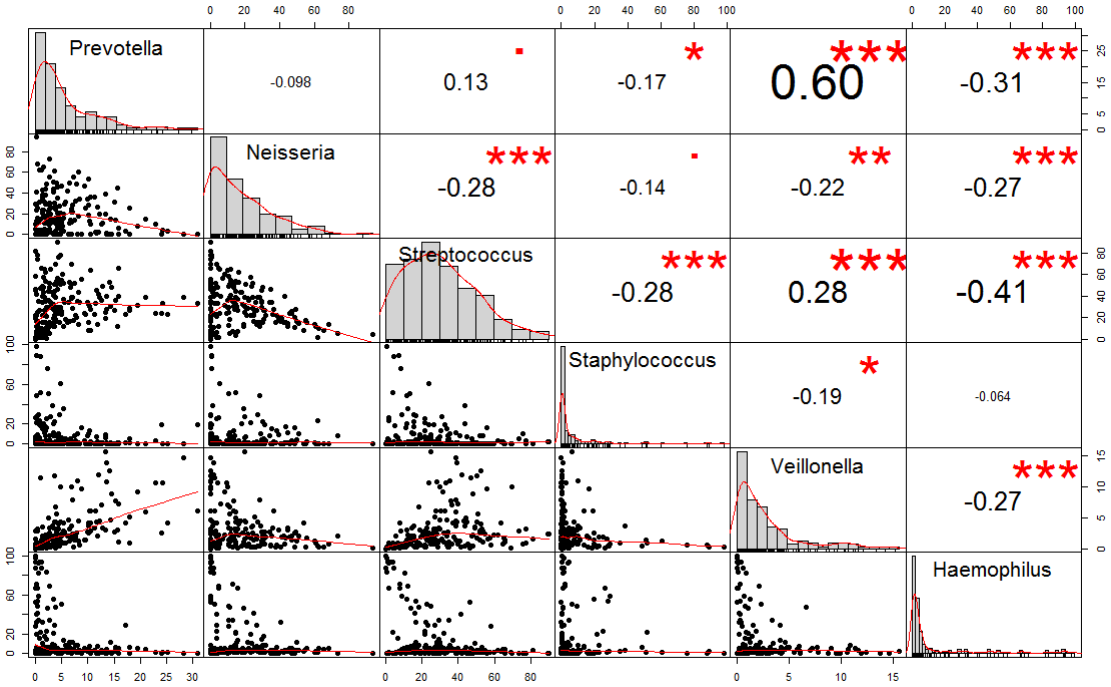

FIGURE 26 – Corrélation des abondances entre les genres du core microbiota. En diagonale, la distribution des genres par échantillon. A gauche, les corrélations entre deux genres. A droite, le score de Pearson entre deux genres. *Prevotella* et *Veillonella* sont fortement corrélés.

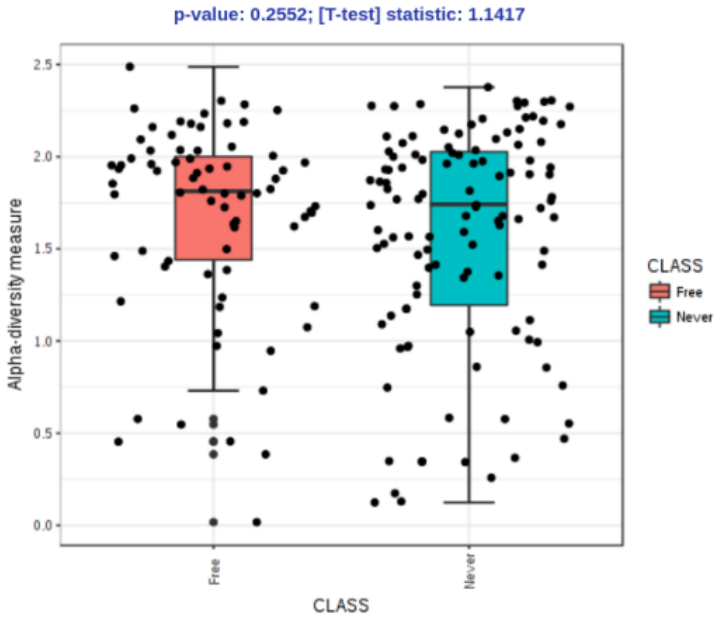

FIGURE 27 – Comparaison des diversités de Shannon entre les échantillons Free et Never

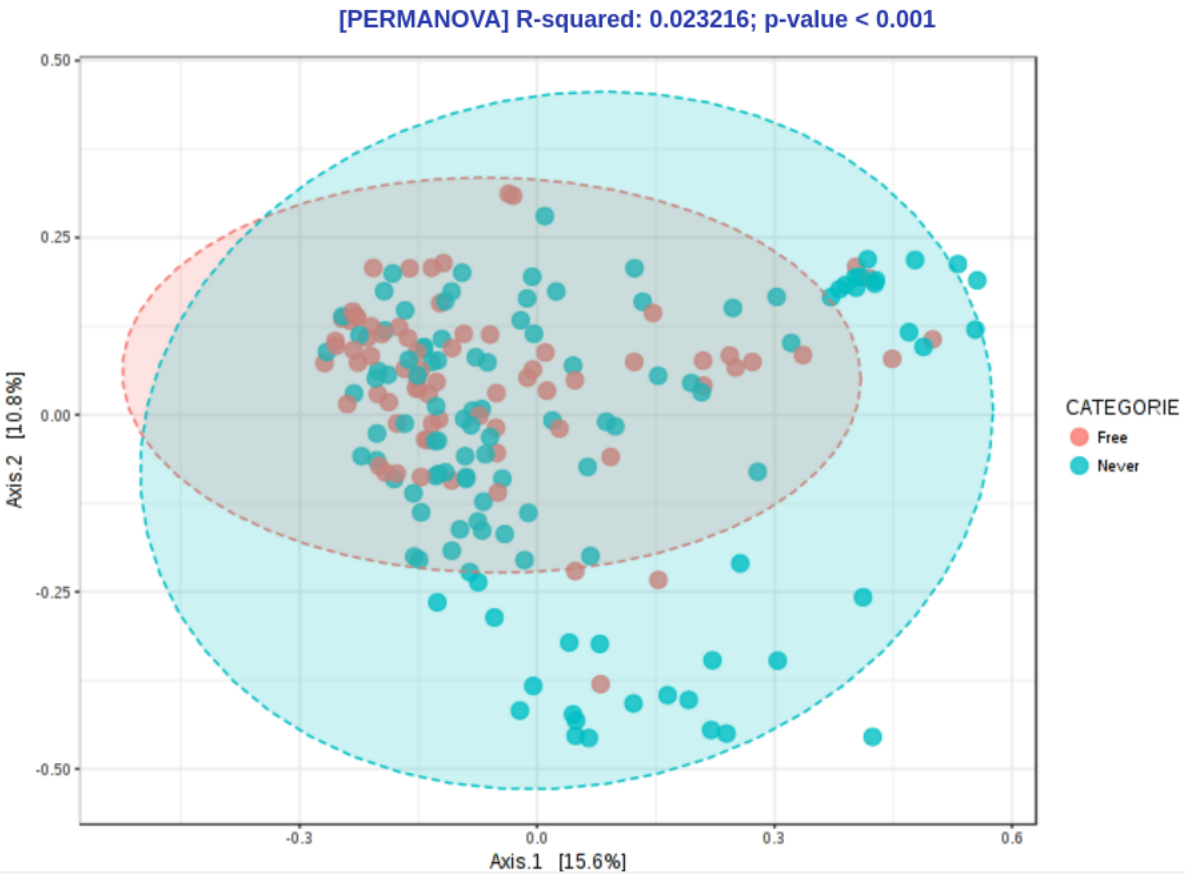

FIGURE 28 – Permanova entre les échantillons Free et Never sur une PCoA en utilisant la distance de Bray-Curtis

Discussion

Production des données

Séquençage

La version 3 du kit MiSeq Reagent Kit produit des reads plus longs (300 pb) que dans la version précédente (250 pb) afin d'augmenter la précision de l'assignation taxonomique (Figure 29). En contrepartie, la qualité du séquençage est médiocre aux extrémités avec des scores qui chutent en dessous de 20. À cause de cela, seulement 43 % de reads sont exploitables. C'est un problème connu lié à la chimie du kit que Illumina n'a pas corrigé à ce jour. La profondeur de séquençage reste toutefois suffisante comme l'attestent les courbes de raréfaction. En utilisant d'autres amorces ciblant des régions de l'ADNr 16S différentes, nous pourrions augmenter la taille du chevauchement pour gagner en profondeur, mais en diminuant la longueur des reads.
Les séquenceurs de 3e générations comme le Mini Ion (Oxford Nanopore) ou le PacBio (Pacific Bioscience) apporteraient un avantage certain. Ces derniers sont capables de

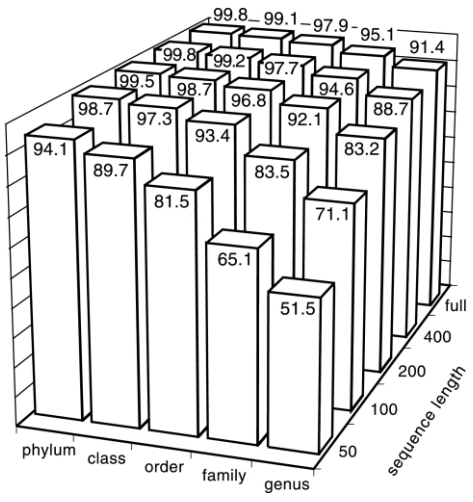

FIGURE 29 – Sensibilité de détection de la stratégie 16S en fonction de la longueur des reads. (Wang et al. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy)

produire des reads très longs (10 000 pb en moyenne) et pourraient facilement séquencer l'intégralité de l'ADNr 16S. De plus, contrairement à Illumina, les fragments d'ADN sont directement lus (Technologie SMS : Single Molecule Sequencing) évitant ainsi tout biais d'amplification pouvant surévaluer les abondances. La qualité des reads de ces séquenceurs de 3e génération est en revanche de mauvaise qualité. Malgré cela, deux études[49;4] ont déjà montré que leurs résolutions taxonomiques atteignent facilement celui de l'espèce.

Pipeline

Par rapport aux autres logiciels comme *QIIME*[7] ou *MOTHUR*[42], le pipeline mucobiome est plus rapide car il distribue la charge de calcul de chaque échantillon dans un *thread* dédié grâce à *Snakemake*[23]. Il est par ailleurs économe en temps de calcul grâce la dé-réplication qui teste une seul fois un même read contre la base de données. Sans cette option, sur la même machine, le pipeline mucobiome a mis 21h de calcul contre 1h29 avec l'option. La stratégie close-reference utilisée dans le pipeline mucobiome a permis d'assigner jusqu'au niveau du genre, plus de 99 % des reads analysés. La base de données *Greengene* est donc assez exhaustive pour analyser le microbiote respiratoire, car tous les genres sont connus. En revanche, la résolution taxonomique n'atteint pas le rang de l'espèce. Il a en effet déjà été montré que les régions hypervariables prises isolément ne sont pas assez discriminantes[51]. Deux espèces peuvent par exemple varier d'un seul nucléotide dans une région hypervariable. D'autre part, l'assignation taxonomique dépend d'une similarité que nous avons fixé arbitrairement à 97 %. D'autres algorithmes permettent de s'affranchir de ce seuil. Par exemple la clusterisation par calcul d'entropie maxium (Figure 30) *Maximum entropy clusterisation*[1].

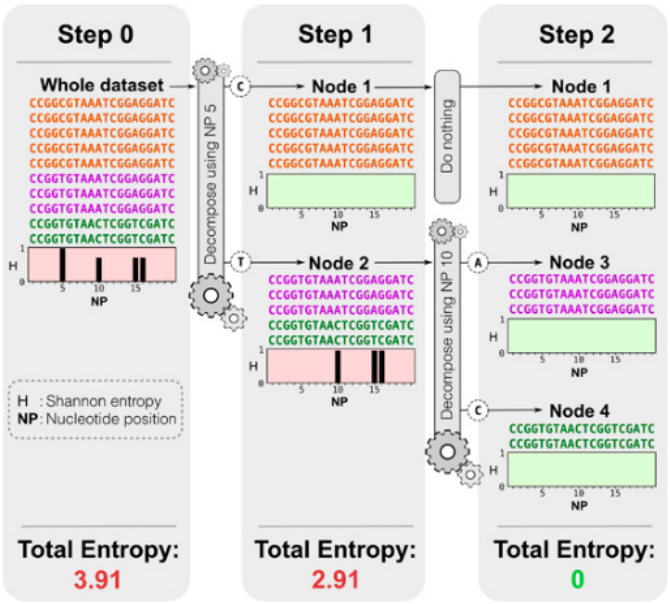

FIGURE 30 – Principe de la décomposition par entropie minimum[1]. Étape 1 : les reads sont alignés. L'entropie H est calculée pour chaque position nucléotidique. Étape 2 : La position ayant la plus forte entropie est utilisée pour discriminer deux groupes. Étape 3 : Le processus est répété jusqu'à atteindre une entropie totale de 0

Méthodes d'analyse

Parmi les analyses de bêta diversité, seule l'analyse en coordonnées principales a été présentée en utilisant la distance de Bray-Curtis. Nous avons également réalisé d'autres ordinations (données non montrées) qui n'ont pas été concluantes. Par exemple, en changeant le type de distance ou en utilisant un positionnement multidimensionnel (MDS) au lieu de la PCoA.

Par ailleurs, la proximité phylogénétique des bactéries n'a pas été prise en compte. Un microbiote contenant trois bactéries proches n'est pas la même chose qu'un microbiote contenant trois bactéries très éloignées. La distance UniFrac[28] pourrait alors remplacer la distance de Bray-Curtis en prenant en considération la phylogénie des bactéries.

Analyse des données

Principaux phyla bactériens

Par rapport aux données chez les sujets sains[30], nos résultats chez les patients mucoviscidosiques montrent une augmentation des *Proteobacteria* (31.48 %) aux dépens des *Bacteroidetes* (11.11 %). Et dans une moindre mesure, une augmentation des *Actinobacteria* (7.41 %). Ces résultats sont en accord avec ceux présentés dans la méta-analyse de *Nature*[30] retrouvant ces proportions dans la mucoviscidose, l'asthme et la BPCO. Cet excès s'explique principalement par la dominance d'*Haemophilus* et de *Neisseria*. Par

ailleurs, une large proportion d'anaérobies est retrouvée, avec notamment *Prevotella* et *Veillonella*. La corrélation forte entre ces deux genres (coefficient de Pearson = 0.60) a été décrite dans le microbiote intestinal[14]. La signification clinique de la présence des anaérobies est encore incertaine[47].

Dominance des pathogènes

Les autres pathogènes associés à la mucoviscidose sont tous retrouvés (*Haemophilus*, *Staphylococcus*, *Burkholderia*, *Stenotrophomonas*) et ont tendance à dominer les autres. Pourtant, la majorité des échantillons proviennent de patients sans détérioration clinique. Ceci remet en cause l'idée selon laquelle l'infection bactérienne soit à l'origine des exacerbations. En effet, plusieurs études[45;20] ont montré qu'il peut y avoir exacerbations sans infection et inversement. Une des hypothèses suggère que l'anomalie du *CFTR* est directement responsable d'un terrain pro-inflammatoire.

Présence de *Pseudomonas*

Ces résultats confirment les résultats déjà décrits au laboratoire[25] avec les marqueurs oprL et la nécessité de coupler une PCR plus spécifique avec les marqueurs gyrB/ecfX. Ces résultats doivent être nuancés, car la résolution taxonomique de notre méthode s'arrête au genre *Pseudomonas* et non à l'espèce *P. aeruginosa*.

Patient Free et Never

Les patients selon leur appartenance aux catégories Never et Free[27], qui permettent de distinguer les patients selon leur anamnèse vis-à-vis de P. aeruginosa (totalement exempts ou colonisation datant d'au moins 1 an), n'ont pas montré de différence statistiquement significative, ni dans leur diversité alpha ni dans la composition de leur microbiote. Les analyses de bêta diversité montre en revanche une homogénisation des microbiotes chez les patients Free, probablement causé par l'antibiothérapie d'éradication.

Hétérogénéité des microbiotes

Nos résultats de diversité alpha et bêta montrent la grande variabilité du microbiote respiratoire. Ils sont différents entre les patients et changent rapidement au cours du temps. Aucune des données recueillies (sexe, âge, IMC , type de mutation, score cytologique, prise d'antibiotique) n'a permis de distinguer des catégories de microbiote. Il semble donc que ce soit l'individu lui-même la principale source de variabilité. La proximité, sur le graphique d'ordination, de certains échantillons d'un même patient va dans ce sens.

Limites de l'étude

Plusieurs points sont à prendre en considération pour critiquer la justesse des résultats. Premièrement, les ADN séquencés proviennent d'expectorations bronchiques

ayant traversé les voies aériennes supérieures. Une contamination est toujours possible. Le score cytologique de Bartlett, n'a pas permis de mettre en évidence des différences dans la structure du microbiote qui soient imputables aux contaminations de l'oropharynx. Mais la présence de *Corynebacterium* dans certains échantillons est évocatrice. De plus, la stratégie 16S ne fait pas la distinction entre des bactéries mortes et vivantes ce qui biaise encore l'estimation des abondances. Dans l'idéal il faudrait réaliser des prélèvements invasifs assez difficiles à mettre en pratique. Une autre idée serait de réaliser parallèlement à l'ECBC l'analyse d'un prélèvement de l'oropharynx pour calculer un différentiel.
D'un point de vue du recueil, l'analyse longitudinale ne contient pas assez d'échantillons (en moyenne 3 échantillons par patient sur 3 ans). Et comme nous avons pu le voir, l'hétérogénéité du microbiote rend difficilement comparables les patients entre eux. Il serait intéressant d'augmenter la résolution temporelle, en augmentant le nombre d'échantillons par patient. Ce qui d'autant plus faisable car tous les échantillons sont conservés au laboratoire à -80 degrés. Enfin les nouvelles technologies de séquençage évoqué plus haut seraient idéales pour atteindre le niveau taxonomique de l'espèce.

Conclusion

Le séquençage de l'ADNr 16S a été réalisé sur 188 expectorations bronchiques de 47 patients atteints de mucoviscidose avec un pipeline bio-informatique dédié. Malgré la faible qualité du séquençage MiSeq (Illumina) dans la version 3 du kit, plus de 99 % des reads ont reçu une assignation taxonomique jusqu'au genre bactérien. L'augmentation des *Proteobacteria* évoquée dans d'autres études sur le microbiote respiratoire dans la mucoviscidose est retrouvée. Les analyses de diversité ont montré une grande hétérogénéité des microbiotes entre les patients et une variabilité dans le temps. Ni les données clinico-biologiques recueillies ni les catégories de Lee (statut vis à vis de *P. aeruginosa*) n'ont montré une différence significative en termes de diversité alpha. Toutefois en terme de bêta diversité, le microbiote des patients Free est moins divers que celui des patient Never, probablement à cause de l'antibiothérapie d'éradiction de *P. aeruginosa*. Une étude longitudinale avec une résolution temporelle plus importante serait intéressante pour évaluer la cinétique d'évolution du microbiote. Les séquenceurs de troisième génération permettraient l'assignation taxonomique jusqu'à l'espèce. Enfin d'autres approches seraient intéressantes pour évaluer la fonction du microbiote et aussi l'impact du virome respiratoire, qui ne l'oublions pas est tout aussi important. La métagénomique fonctionnelle ou encore la PCR digitale apporteraient ces éléments de réponse au prix de calculs informatiques parfois très élevés.

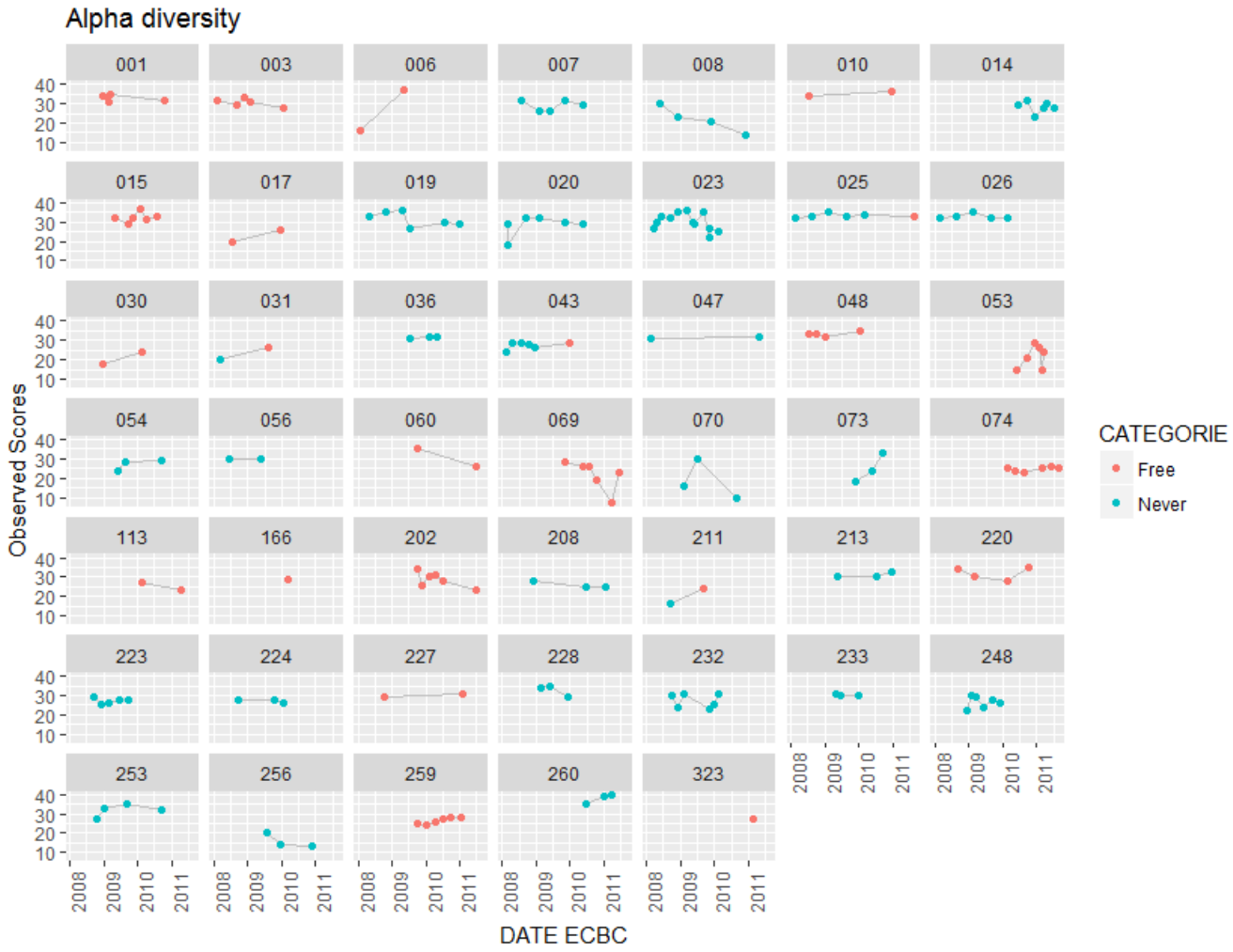

FIGURE 31 – Évolution du nombre d'espèces par patient en fonction du temps

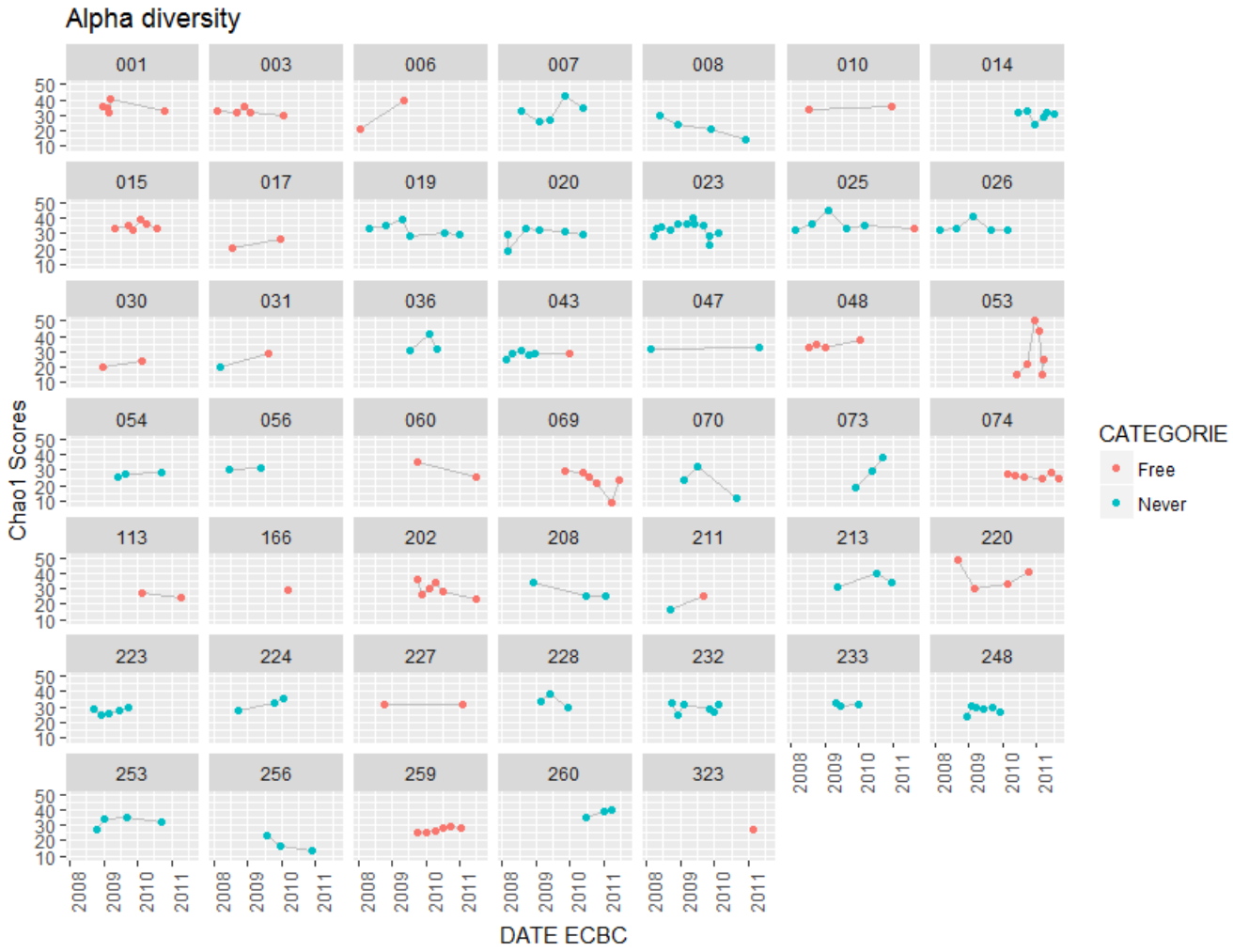

FIGURE 32 – Évolution de l'indice de Chao1 par patient en fonction du temps

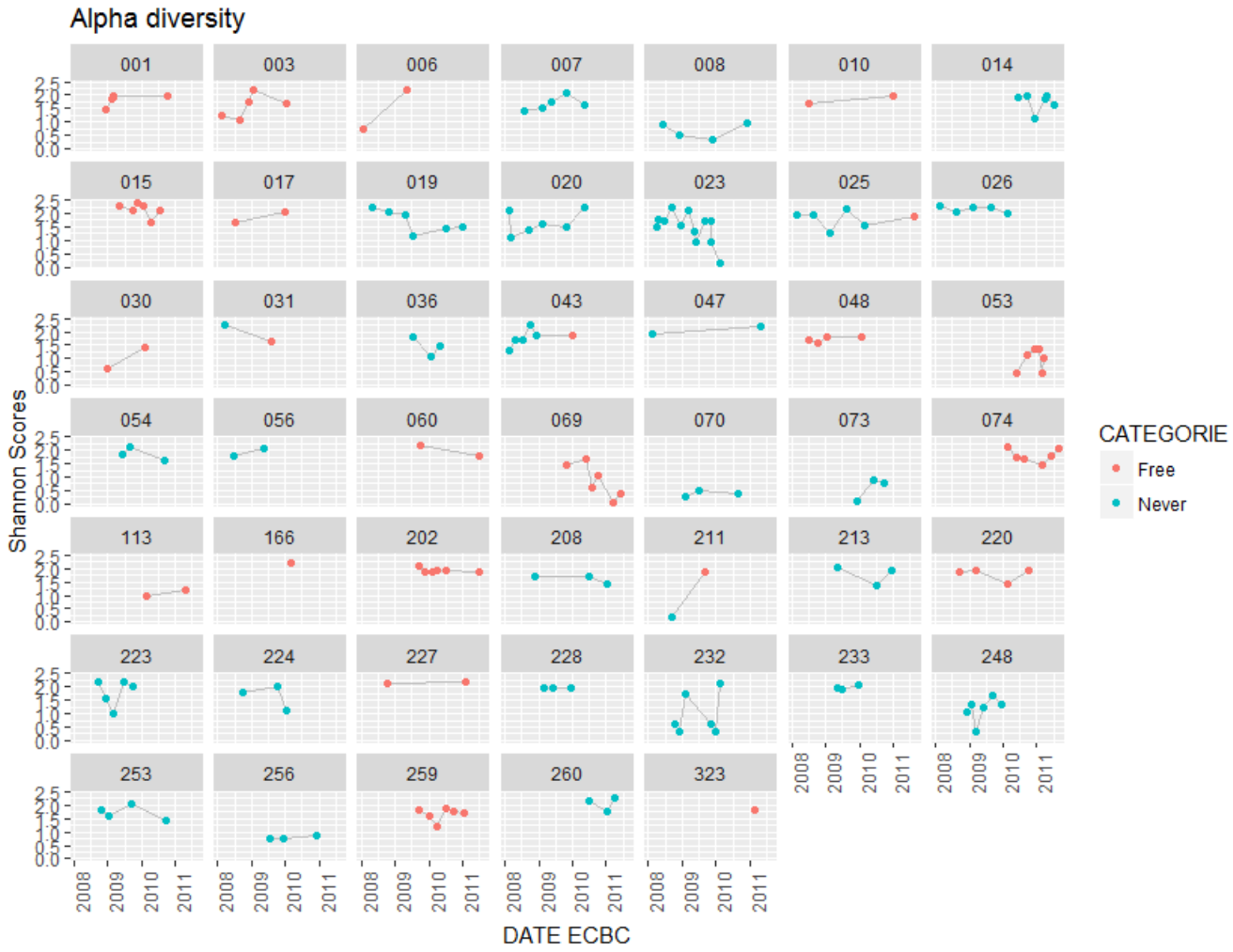

FIGURE 33 – Évolution de l'indice de Shanon par patient en fonction du temps

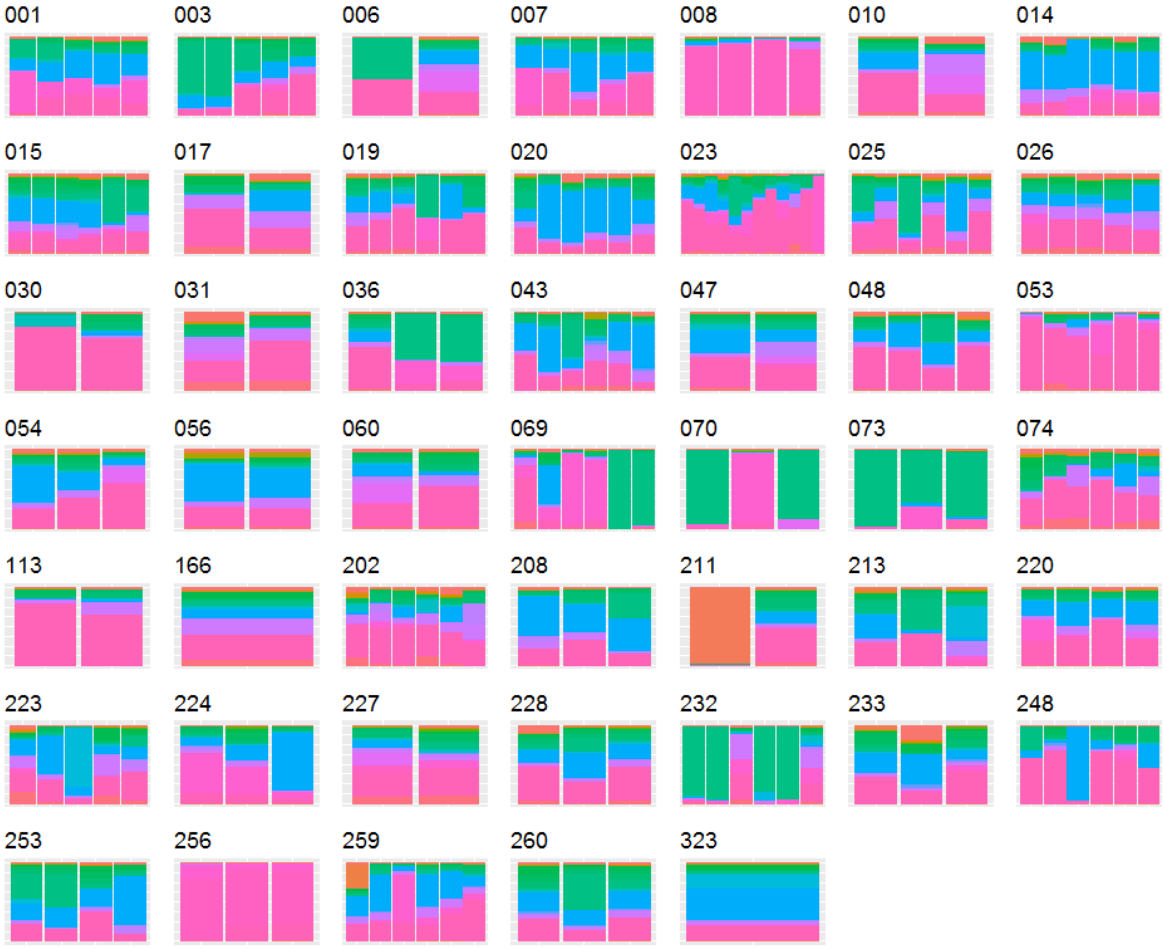

FIGURE 34 – Évolution des abondances par genre. Les échantillon sont ordonnés dans le temps

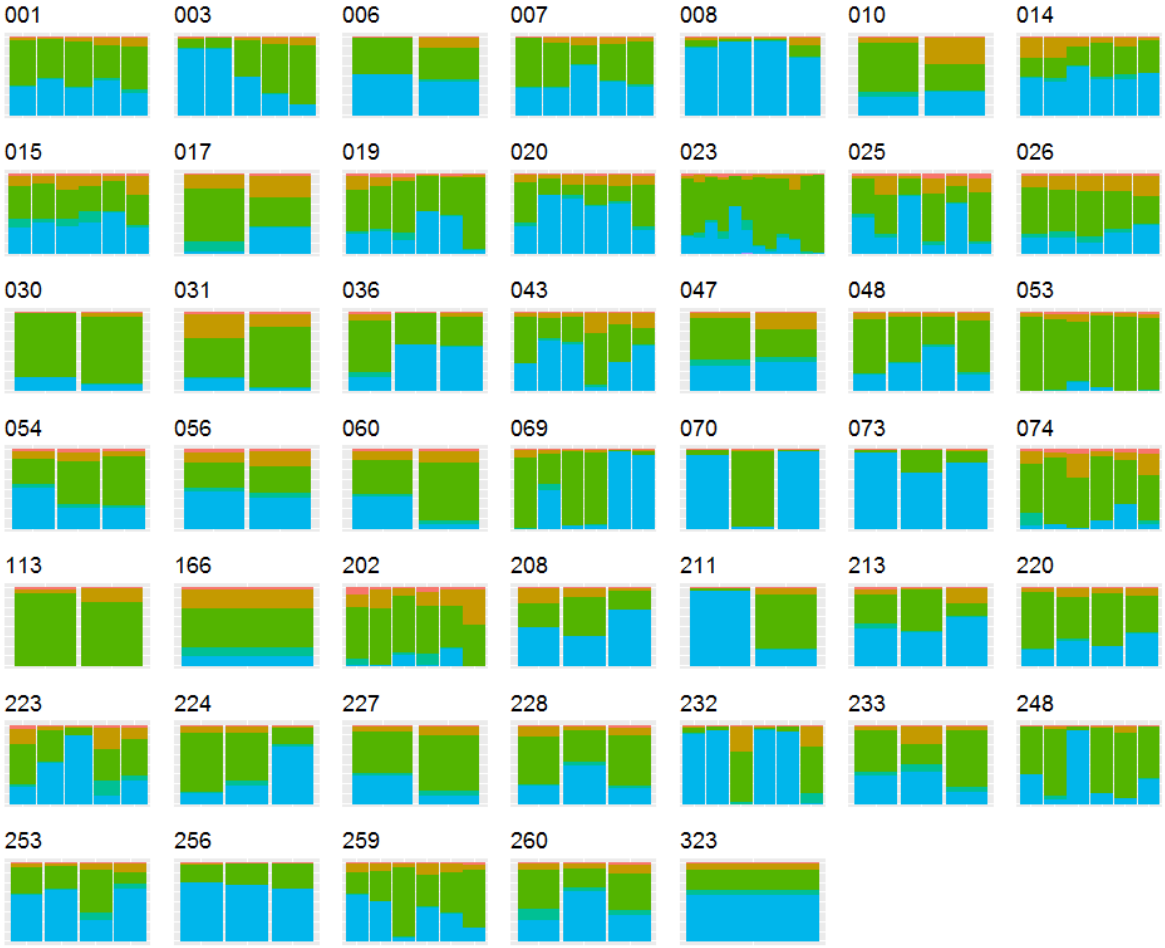

FIGURE 35 – Évolution des abondances par phylum. Les échantillon sont ordonnés dans le temps

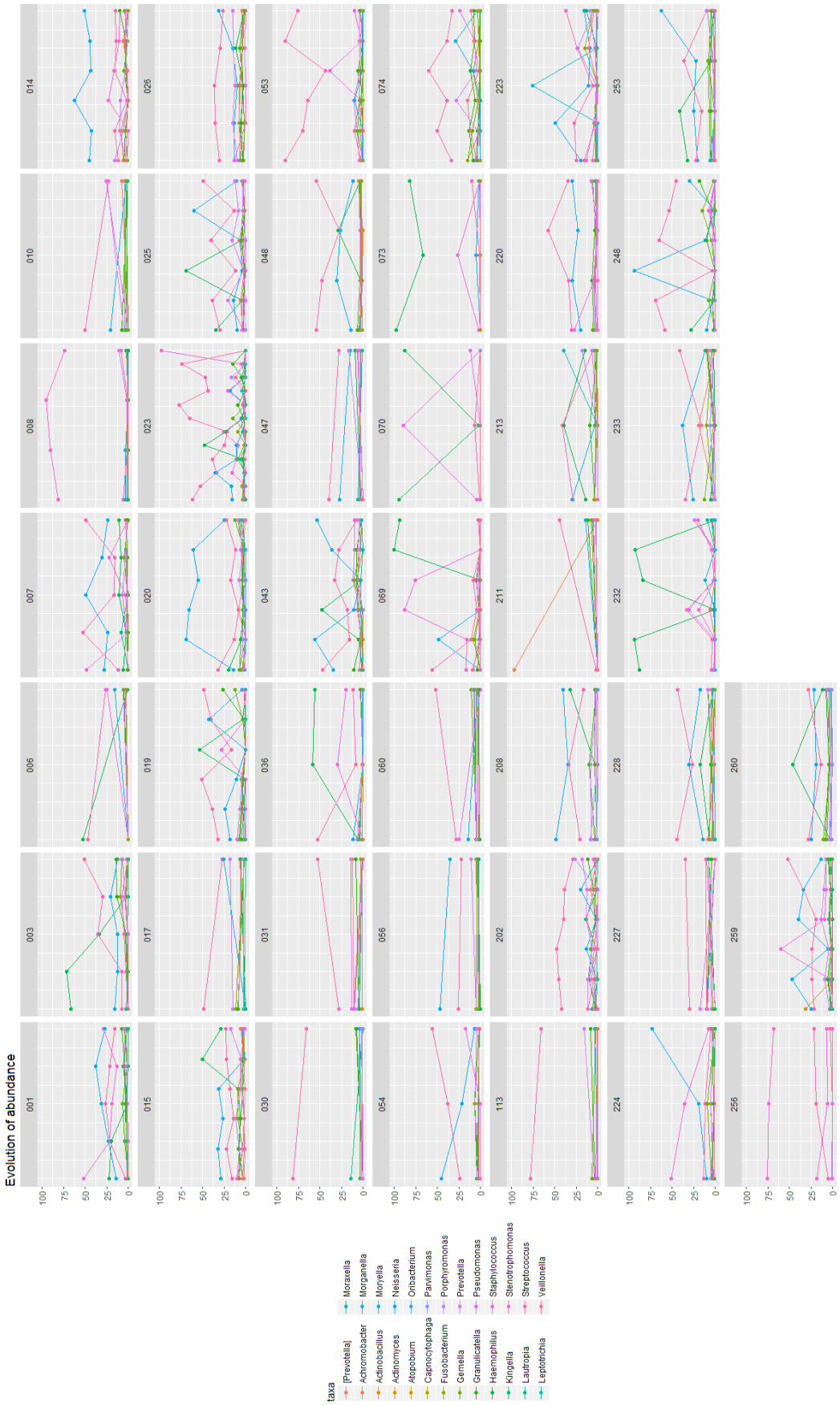

FIGURE 36 – Évolution des abondances par genre

Références

[1] Edwin Aldana-Bobadilla and Angel Kuri-Morales. A clustering method based on the maximum entropy principle. *Entropy*, 17(1) :151–180, 2015.

[2] Jean-François Bach. The Effect of Infections on Susceptibility to Autoimmune and Allergic Diseases. *New England Journal of Medicine*, 347(12) :911–920, sep 2002.

[3] James M Beck, Vincent B Young, and Gary B Huffnagle. The Microbiome of the Lung.

[4] Alfonso Benítez-Páez, Kevin J. Portune, and Yolanda Sanz. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION^{TM} portable nanopore sequencer. *GigaScience*, 5(1) :4, dec 2016.

[5] Lars Bode. Human milk oligosaccharides : every baby needs a sugar mama. *Glycobiology*, 22(9) :1147–62, sep 2012.

[6] Megan L Boulette, Patricia J Baynham, Peter A Jorth, Irena Kukavica-Ibrulj, Aissa Longoria, Karla Barrera, Roger C Levesque, and Marvin Whiteley. Characterization of alanine catabolism in Pseudomonas aeruginosa and its importance for proliferation in vivo. *Journal of bacteriology*, 191(20) :6329–34, oct 2009.

[7] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5) :335–336, may 2010.

[8] Bryan Coburn, Pauline W Wang, Julio Diaz Caballero, Shawn T Clark, Vijaya Brahma, Sylva Donaldson, Yu Zhang, Anu Surendra, Yunchen Gong, D Elizabeth Tullis, Yvonne C W Yau, Valerie J Waters, David M Hwang, and David S Guttman. Lung microbiota across age and disease stage in cystic fibrosis. *Scientific reports*, 5 :10241, may 2015.

[9] Jane C Davies. Pseudomonas aeruginosa in cystic ®brosis : pathogenesis and persistence THE MOLECULAR BASIS OF CYSTIC FIBROSIS.

[10] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. Greengenes, a Chimera-Checked

16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, 72(7) :5069–5072, jul 2006.

[11] Robert P Dickson, John R Erb-Downward, and Gary B Huffnagle. The Role of the Bacterial Microbiome in Lung Disease.

[12] Robert P Dickson, John R Erb-Downward, and Gary B Huffnagle. The Role of the Bacterial Microbiome in Lung Disease.

[13] Robert P Dickson and Gary B Huffnagle. The Lung Microbiome : New Principles for Respiratory Bacteriology in Health and Disease.

[14] Aldona Dlugosz, Björn Winckler, Elin Lundin, Katherina Zakikhany, Gunnar Sandström, Weimin Ye, Lars Engstrand, and Greger Lindberg. No difference in small bowel microbiota between patients with irritable bowel syndrome and healthy controls. *Scientific Reports*, 5(1) :8508, jul 2015.

[15] Jonathan Eisen. What does the term microbiome mean ? And where did it come from ? A bit of a surprise. *http ://www.microbe.net/2015/04/08/what-does-the-term-microbiome-mean-and-where-did-it-come-from-a-bit-of-a-surprise*.

[16] Katherine B Frayman, David S Armstrong, Rosemary Carzino, Thomas W Ferkol, Keith Grimwood, Gregory A Storch, Shu Mei Teo, Kristine M Wylie, and Sarath C Ranganathan. The lower airway microbiota in early cystic fibrosis lung disease : a longitudinal analysis. *Thorax*, pages thoraxjnl–2016–209279, 2017.

[17] Genet.sickkids.on.ca. Cystic Fibrosis Mutation Database.

[18] C. Guissart, C. Dubucs, C. Raynal, A. Girardet, F. Tran Mau Them, V. Debant, C. Rouzier, A. Boureau-Wirth, E. Haquet, J. Puechberty, E. Bieth, D. Dupin Deguine, P. Khau Van Kien, M.P. Brechard, V. Pritchard, M. Koenig, M. Claustres, and M.C. Vincent. Non-invasive prenatal diagnosis (NIPD) of cystic fibrosis : an optimized protocol using MEMO fluorescent PCR to detect the p.Phe508del mutation. *Journal of Cystic Fibrosis*, 16(2) :198–206, mar 2017.

[19] H. Li. Seqtk : a fast and lightweight tool for processing FASTA or FASTQ sequences, 2013. https ://github.com/lh3/seqtk.

[20] Harry Heijerman. Infection and inflammation in cystic fibrosis : A short review. 2005.

[21] Wing Ho Man, Wouter AA de Steenhuijsen Piters, and Debby Bogaert. The microbiota of the respiratory tract : gatekeeper to respiratory health. *Nature Publishing Group*, 2017.

[22] Fass JN. Joshi NA. Sickle : A sliding-window, adaptive, quality-based trimming tool for FastQ files https ://github.com/najoshi/sickle. 2011.

[23] J. Koster and S. Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19) :2520–2522, oct 2012.

[24] Muriel Le Bourgeois and Stéphanie Vrielynck. Infection bronchopulmonaire dans la mucoviscidose.

[25] Florence Le Gall, Rozenn Le Berre, Sylvain Rosec, Jeanne Hardy, Stéphanie Gouriou, Sylvie Boisramé-Gastrin, Sophie Vallet, Gilles Rault, Christopher Payan, and Geneviève Héry-Arnaud. Proposal of a quantitative PCR-based protocol for an optimal Pseudomonas aeruginosa detection in patients with cystic fibrosis.

[26] Jean Guy LeBlanc, Christian Milani, Graciela Savoy de Giori, Fernando Sesma, Douwe van Sinderen, and Marco Ventura. Bacteria as vitamin suppliers to their host : a gut microbiota perspective. *Current Opinion in Biotechnology*, 24(2) :160–168, apr 2013.

[27] Tim W.R. Lee, Keith G. Brownlee, Steven P. Conway, Miles Denton, and James M. Littlewood. Evaluation of a new definition for chronic Pseudomonas aeruginosa infection in cystic fibrosis patients. *Journal of Cystic Fibrosis*, 2(1) :29–34, mar 2003.

[28] Catherine Lozupone and Rob Knight. UniFrac : a New Phylogenetic Method for Comparing Microbial Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Applied and environmental microbiology*, 71(12) :8228–8235, 2005.

[29] Tanja Mago ? ? and Steven L. Salzberg. FLASH : Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21) :2957–2963, 2011.

[30] Benjamin J Marsland and Eva S Gollwitzer. Host–microorganism interactions in lung diseases. *Nature Publishing Group*, 14, 2014.

[31] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1) :pp. 10–12, 2011.

[32] Paul J. McMurdie, Susan Holmes, R Kindt, P Legendre, and RB O'Hara. phyloseq : An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, 8(4) :e61217, apr 2013.

[33] Tristan Montier, Pascal Delépine, Rémi Marianowski, Karine Le Ny, Morgane Le Bris, Danielle Gillet, Gaël Potard, Philippe Mondine, Irène Frachon, Jean-Jacques Yaouanc, Jean-Claude Clément, Harvé des Abbayes, and Claude Férec. CFTR Transgene Expression in Primary ΔF508 Epithelial Cell Cultures From Human Nasal Polyps Following Gene Transfer With Cationic Phosphonolipids. *Molecular Biotechnology*, 26(3) :193–206, mar 2004.

[34] Linh D N Nguyen, Eric Viscogliosi, and Laurence Delhaes. The lung mycobiome : an emerging field of the human respiratory microbiome. *Frontiers in microbiology*, 6 :89, 2015.

[35] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI : current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1) :D733–D745, jan 2016.

[36] Maria Elisa Perez-Muñoz, Marie-Claire Arrieta, Amanda E. Ramer-Tait, Jens Walter, F Ballester, MP Francino, C Mølgaard, KF Michaelsen, TR Licht, U Sauer, KD McCoy, AJ Macpherson, E Schober, C Ionescu-Tirgoviste, G Devoti, CE Beaufort, K Buschard, CC Patterson, C Colding, V Tremaroli, Y Yin, S Bergman, X Xu, L Madsen, K Kristiansen, J Dahlgren, and W Jun. A critical assessment of the "sterile womb" and "in utero colonization" hypotheses : implications for research on the pioneer infant microbiome. *Microbiome*, 5(1) :48, 2017.

[37] P Plésiat. Quels critères microbiologiques pour définir une colonisation ou une infection à Pseudomonas aeruginosa ?

[38] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, Daniel R. Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie,

Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H. Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, Maria Antolin, François Artiguenave, Hervé Blottiere, Natalia Borruel, Thomas Bruls, Francesc Casellas, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Miguel Forte, Carsten Friss, Maarten van de Guchte, Eric Guedon, Florence Haimet, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Michiel Kleerebezem, Jan Knol, Michel Kristensen, Severine Layec, Karine Le Roux, Marion Leclerc, Emmanuelle Maguin, Raquel Melo Minardi, Raish Oozeer, Maria Rescigno, Nicolas Sanchez, Sebastian Tims, Toni Torrejon, Encarna Varela, Willem de Vos, Yohanan Winogradsky, Erwin Zoetendal, Peer Bork, S. Dusko Ehrlich, and Jun Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285) :59–65, mar 2010.

[39] Paul M. Quinton. The neglected ion : HCO3|[minus]|. *Nature Medicine*, 7(3) :292–293, mar 2001.

[40] Registredelamuco.org. Registre Francais de la mucovicidoses 2015.

[41] R. Ruimy and A. Andremont. Quorum-sensing chez Pseudomonas aeruginosa : Mécanisme moléculaire, impact clinique, et inhibition, 2004.

[42] Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, Jason W Sahl, Blaz Stres, Gerhard G Thallinger, David J Van Horn, and Carolyn F Weber. Introducing mothur : open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23) :7537–41, dec 2009.

[43] Ron Sender, Shai Fuchs, Ron Milo, T Lee, H Ahn, and S Baek. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*, 14(8) :e1002533, aug 2016.

[44] Labsquare team. Fastqt : another quality control tool for high throughput sequence data. 2017.

[45] Rabindra Tirouvanziam, Sophie de Bentzmann, Cédric Hubeau, Jocelyne Hinnrasky, Jacky Jacquot, Bruno Péault, and Edith Puchelle. Inflammation and Infec-

tion in Naive Human Cystic Fibrosis Airway Grafts. *American Journal of Respiratory Cell and Molecular Biology*, 23(2) :121–127, aug 2000.

[46] Torbjørn Rognes. A peer-reviewed version of this preprint was published in PeerJ on 3 March 2015. *PeerJ*, (March), 2015.

[47] Michael M. Tunney, Tyler R. Field, Thomas F. Moriarty, Sheila Patrick, Gerd Doering, Marianne S. Muhlebach, Matthew C. Wolfgang, Richard Boucher, Deirdre F. Gilpin, Andrew McDowell, and J. Stuart Elborn. Detection of Anaerobic Bacteria in High Numbers in Sputum from Patients with Cystic Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 177(9) :995–1001, may 2008.

[48] Christopher J van der Gast, Alan W Walker, Franziska A Stressmann, Geraint B Rogers, Paul Scott, Thomas W Daniels, Mary P Carroll, Julian Parkhill, and Kenneth D Bruce. Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. *The ISME journal*, 5(5) :780–91, may 2011.

[49] Josef Wagner, Paul Coupland, Hilary P. Browne, Trevor D. Lawley, Suzanna C. Francis, and Julian Parkhill. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiology*, 16(1) :274, dec 2016.

[50] Fiona J. Whelan, Alya A. Heirali, Laura Rossi, Harvey R. Rabin, Michael D. Parkins, and Michael G. Surette. Longitudinal sampling of the lung microbiota in individuals with cystic fibrosis. *Plos One*, 12(3) :e0172811, 2017.

[51] Bo Yang, Yong Wang, and Pei-Yuan Qian. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC bioinformatics*, 17 :135, mar 2016.

[52] Ed Yong. *Moi, microbiote, maître du monde : Les microbes, 30 billions d'amis !* 2017.

SCHUTZ (Sacha) - Description du Microbiote réspiratoire chez des patients atteind de mucovicidoses - 41 pages, 36 figures, 4 tableaux

RESUME : Les nouvelles technologies de séquençage haut débit ont révélé l'existence du microbiote pulmonaire, y compris chez le sujet sain. COMPLETER UN PEU LE RATIONNEL, c'est trop short. Dire un mot sur infection pulmonaire microbiote et muco L'objectif principal de l'étude prospective multicentrique MUCOBIOME était de définir l'impact du microbiote respiratoire dans la primocolonisation à Pseudomonas aeruginosa. Une cohorte de 96 patients atteints de mucoviscidose et exempts de P. aeruginosa depuis au moins un an a été incluse dans trois centres (Brest, Roscoff et Nantes). Sur les trois ans de suivi, 707 expectorations ont été recueillies itérativement. L'exploration du microbiote respiratoire a été réalisée par le séquençage de l'ADNr 16S sur Illumina MiSeq à l'aide du kit V3. L'objectif précis de ce travail de thèse a été de mettre en place un pipeline bio-informatique spécialement dédié à l'analyse de ces données. COMPLETER SUR LE PIPELINE MIS EN PLACE et le nb de données analysées (188 prélèvements de x patients d'âge moyen = etc ...)

MOTS CLEFS :
Microbiote respiratoire
Mucoviscidose
Pseudomonas aeruginosa
ARNr 16S
Bioinformatique
Séquençage haut débit

JURY :
Président : M. FEREC
Membres : G. HERY-ARNAUD
E. GENIN
C. LE MARECHAL
P. LANOTTE

DATE DE SOUTENANCE :
23 Avril 2017

ADRESSE DE L'AUTEUR :
33 Rue d'aiguillon, 29200 Brest