Thèse de médecine

CHU-BREST

Université de Brest
DES de biologie médicale

# Description du microbiote pulmonaire chez les patients atteints de mucoviscidoses

*Auteur :*
Sacha SCHUTZ

*Responsable :*
Geneviève HERY-ARNAUD

16 avril 2017


# Engagement de non-plagiat

Je, soussigné Sacha SCHUTZ, interne en biologie moléculaire au CHU 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 :

ssh pub key fingerprint : a4 :e3 :da :87 :78 :2d :e1 :6f :bb :56 :5c :d1 :72 :f5 :50 :63

# Licence



# Table des matières

# 1 Avant-propos

Depuis Pasteur, les micro-organismes ont toujours été perçus négativement, car associés aux maladies. La mise en évidence des agents pathogènes allant de la syphilis jusqu'aux grandes pestes n'a pas aidé ces êtres microscopiques à sortir de ce stéréotype. La médecine s'est donc naturellement orientée à les combattre plus qu'à les étudier. Aujourd'hui, personne ne peut nier que les traitements anti-infectieux ont permis l'amélioration de notre santé. Les avancées majeures en ce qui concerne l'hygiène, la vaccination et les antibiotiques ont conduit à diminuer la prévalence des maladies infectieuses voir les faire disparaître. Cependant, la destruction systématique et massive des micro-organismes qui vivent et évoluent en nous depuis des milliers d'années pourrait bien être la cause de l'émergence de nouvelles maladies ( Figure 1).

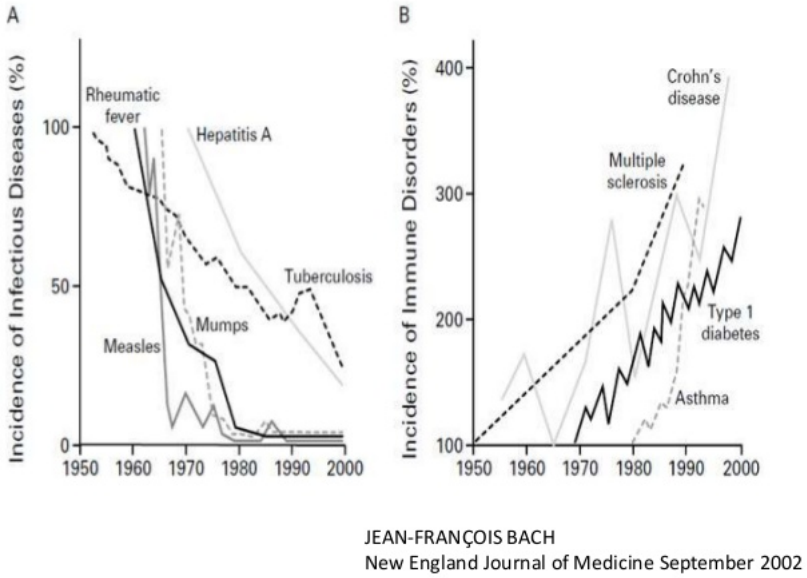


FIGURE 1 – Incidence des maladies infectieuses et auto-immunes en Europe au cours du temps. [@]

Les récentes méthodes 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 en permettant à la biomasse d'absorber le diazote atmosphérique. Les bactéries sont dans ce sens, la source primaire permettant aux organismes de construire leurs protéines et leurs ADN. Elles peuvent vivre dans les milieux les plus inhospitaliers. Les archées (anciennement archéobactérie) peuvent résister à des conditions d'acidités et de températures exceptionnelles. On les retrouve dans les fonds océaniques privés de lumière où elles sont la seule source

d'énergie pour la faune en utilisant la chimiosynthèse à l'instar de la photosynthèse. les archées nous ont par ailleurs éclairés sur l'origine des eucaryotes. [1] et nous ont permise un bond de géant en biologie moléculaire [2].

L'homme ne fait pas exception. La majorité des bactéries ont longtemps été indétectables par les méthodes de culture classiques. Mais à présent, les régions anatomiques autrefois considérées stériles foisonnent de bactéries. Elles sont retrouvées dans tous les territoires du corps exposé ou elles forment des communautés. La peau est colonisée par *Propionibacterium*, *Corynebacterium* et *Staphylococcus* ? . Le vagin contient des *Lactobacille* et la bouche principalement du *Streptococcus* ? . L'intestin est une flore bactérienne dominée par les anaérobies pouvant représenter jusqu'à 2 kg du poids corporel ? . En échange de son hospitalité, le microbiote contribue au bon fonctionnement de son hôte. Il aide à la digestion en dégradant par exemple les sucres du lait maternel chez le nouveau-né[@bifidus]. Il participe à la synthèse de vitamine essentielle (K, B12,B8)[@ref]. Il éduque notre système immunitaire et fait barrière à tout nouvel agent pathogène. Toute défaillance de notre microbiote ou *dysbiose*, peut être délétère pour notre santé. La liste des affections associées est longue. On retrouve par exemple la maladie de Crohn[@], la maladie coeliaque[@], le cancer de l'intestin[@], le syndrome du côlon irritable[@], l'obésité[@], le diabète de type 1[@], l'asthme[@], l'eczéma[@], la sclérose en plaque[@], la polyarthrite rhumatoïde[@], la maladie d'alzheimer[@] et même l'autisme[@].

La colite à *Clostridium Difficile* est un exemple de dysbiose avec une application clinique directe. Suite à une prise d'antibiotique, la flore intestinale est détruite laissant l'opportunité à *Clostridium Difficile* de s'installer. Un des traitements proposés est la transplantation fécale visant à régénérer le microbiote du patient.

Le microbiote amène donc à reconsidérer notre individualité. Nous ne sommes plus seulement un organisme multicellulaire composé d' un seul génome. Mais plutôt un écosystème ou les cellules eucaryotes et les micro-organismes vivent en symbiose. Cette relation n'étant pas figée dans le temps et pouvant varier entre le commensalisme, le parasitisme et le mutualisme. Les dernières études estiment que pour chaque individu il y a environ 30 billions de cellules humaines pour 39 billions de cellules microbiennes[@]. En associant les gènes bactériens, le génome d'un humain passe de 23 000 à 3,3 millions[@] de gènes avec toute la complexité des interactions que cela engendre[@]. Les scientifiques ont attribué le nom d'*holobionte* à 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

1. Le génome des archées est composé d'introns comme chez les eucaryotes
2. La Taq polymérase est une enzyme d'archée qui résiste à de hautes températures utilisées dans les PCR

confondent corrélation et causalité. Ces publications ont même conduit à la création du hashtag humoristique sur Twitter : *#GutMicrobiomeAndRandomSomething*. Les études sur le microbiote nécessitent d'être étayé par la métagénomique fonctionnelle afin de trouver les relations de cause à effet. Les corrélations doivent être réalisées sur des populations plus grandes avec un suivi dans le temps plus important. 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 de demain ou s' il s'agit d'un effet de mode. Mais au regard de l'évolution biologique, il y a fort à parier dessus. Car, ne l'oublions pas, ce sont bien des anciennes bactéries, qui permettent à l'ensemble de nos cellules de respirer et que nous appelons maintenant des mitochondries.

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

# 2 Définition

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

**Le microbiome** s'emploie selon deux définitions. En français, le microbiome est l'environnement qui héberge le microbiote. Dans sa définition angle-Saxonne, le microbiome fait référence à l'ensemble des génomes microbiens contenus dans un environnement. De façon générale, le microbiome est associé aux génomes bactériens. Les termes de Virome et de Mycobiome sont utilisés pour les génomes viraux et mycosiques.

**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.

**La métagénomique** est une méthode d'étude du contenu en ADN 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'individu. 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'ARN 16s est associé à tort au terme métagénomique, mais son usage reste courant. On lui préférera le terme de **metagénétique**

**Un read** est un terme bio-informatique désignant une séquence d'ADN issue d'un séquençage haut débit. Selon les technologies, les reads varient entre 150 et 300 paires de bases.

**Un OTU** (*Operational taxonomic Init*) est un terme utilisé en phylogénie, désignant un groupe d'individu proche faisant souvent référence à l'espèce dans la classification de Linée. En microbiologie, un OTU est défini par un groupe d'individu ayant une similarité dans leurs séquences d'ARN 16s supérieur à 97%.

**L'Abondance** absolu est le nombre de séquences d'ADN d'un OTU retrouvé dans un échantillon. L'abondance relative est le pourcentage en séquences d'ADN d'un OTU retrouvé dans un échantillon. Ce dernier permet de rendre les échantillons comparables entre eux.

**La table des OTU** correspond à un tableau à double entrée contenant l'abondance par OTU et par échantillon. Dans le tableau suivant, l'échantillon 1 contient 68% de l'OTU 1.

| | échantillon 1 | échantillon 2 | échantillon 3 |
|---|---|---|---|
| OTU 1 | 68% | 12% | 25% |
| OTU 2 | 40% | 24% | 25% |
| OTU 3 | 28% | 64% | 50% |

FIGURE 3 – La table des Hotus

**La diversité alpha** est une mesure de biodiversité au sein d'un échantillon. Elle correspond à l'étude d'une colonne dans la table des Hotus. Plusieurs indicateurs de diversité alpha existent.

**La diversité bêta** est une analyse descriptive de la biodiversité entre plusieurs échantillons. Elle correspond à l'étude de l'ensemble de la table des Hotus. 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 dessiner, en le projetant dans un espace à une ou deux dimensions.

**La richesse** est le nombre d'espèces présent dans un échantillon. Les deux échantillons suivant une richesse de 2.

échantillon 1 : 4 Streptoccus , 4 Escherichia
échantillon 2 : 432 Streptoccus, 12 Escherichia

**L'uniformité / équitabilité** indique si les espèces d'un échantillon sont réparties uniformément. L'uniformité du premier échantillon est plus grande que la seconde

échantillon 1 : 50 Streptococus , 50 Escherichia
échantillon 2 : 432 Streptoccus, 12 Escherichia

**L'indice Chao1** est une estimation de la richesse réelle (in vivo) par rapport à la richesse observée (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.

$$A = B \tag{1}$$

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

$$A = B \tag{2}$$

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

$$A = B \tag{3}$$

**Pipeline** Un pipeline est un ensemble d'étapes de calcul. Chaque étape prend en entrée des fichiers pour en produire des 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 fonde.( 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, qui permettent de générer un graphe des étapes( direct Acyclique Graph) et de trouver la meilleure façon d'optimiser les taches entre les différents cuisiniers (processeur).

**La courbe de raréfaction** est utilisé pour déterminer si la profondeur de séquençage est suffisante pour caractériser la diversité d'un échantillon. Pour générer cette courbe, des groupes de reads de taille croissante (1...n) sont tirés aléatoirement sans remise. Le groupe est reporté sur l'axe X et le nombre d'OTU correspondant est reporté sur l'axe Y. Une courbe s'aplatissant indique une bonne profondeur de séquençage? AMD?

# 3 Introduction

## 3.1 La mucoviscidose

### 3.1.1 Une maladie génétique

La mucoviscidose est une maladie génétique autosomique récessive grave qui frappe en France 1 naissance sur 5400 [@]. La Bretagne est la région la plus touchée avec une prévalence de 1/3000[@]. La loi de Hardy Weinberg estime qu'en Bretagne 1 patient sur 25 est porteur de la mutation à l'état hétérozygotes[@Heterozygote davantage]. Cette haute prévalence s'explique probablement par un effet fondateur associé à un avantage sélectif pour les individus porteurs de l'allèle muté. [3]

Le gène CFTR impacté se situe sur le chromosome 7 en position q31.2. Il est constitué de 27 exons pour 250 188 [@] paires de bases. Il code pour 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-)[@].

On dénombre à ce jour 2017 mutations mises en cause dans la mucovicidoses[@]. La perte d'une phénylalanine en position 508 par délétion du triplet c.1521-1523delCTT (anciennement $\Delta$F508) cause à elle seule 80% des mucoviscidoses. 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 retrouvés. L'atteinte de la fonction respiratoire est la plus bruyante. En effet au niveau de l'épithélium broncho-pulmonaire, l'absence d'un CFTR fonctionnel est à l'origine d'une déshydratation du mucus le rendant plus visqueux et empêche les cils bronchiques de jouer leurs rôles.[@]

La forte prévalence de la maladie nécessite de réaliser un dépistage précoce chez tous les nouveaux nés (test de Gutri) afin d'adapter au plus tôt la prise en charge. Seul le test à la sueur permet de poser le diagnostic. Le dépistage prénatal basé sur l'ADN circulant est actuellement à l'étude[@]. Le traitement repose avant tout sur une prise en charge respiratoire (kinésithérapie, dorasse, antibiothérapie). Les thérapies génétiques sont encore à l'étude[HAS] L'Ivacaftor est le seul traitement à ce jour qui agit directement sur le CFTR. Mais concerne uniquement certaine mutation rare comme la G551D.[has]La greffe pulmonaire est le dernier recours.

3. Plusieurs hypothèses ont été proposées, notamment lors des grandes épidémies de choléra en diminuant les pertes hydriques. D'autres suggèrent qu'il s'agit d'une pléiotropie antagoniste.[@]

### 3.1.2 Une maladie infectieuse

L'atteinte pulmonaire est caractérisée par des infections successives associées à une réaction inflammatoire qui dégrade progressivement la fonction respiratoire. Plusieurs pathogènes sont impliqués. Chez les jeunes enfants, *Haemophilus influenza* et *Staphilococcus Aureus* sont le plus souvent responsable. *Burkolderia Cepace* et *Stenotrophomonas Maltophilia* sont retrouvés parmi les sujets plus âgé. 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 stricte est un germe de l'environnement rarement retrouvé parmi les patients sains[@]. En revanche, dans la mucoviscidose, il est mis en évidence chez 60%[@] des patients jeunes, et plus de 90% des patients adultes[@].

La primocolonisation à *Pseudomnas Aeruginosa* est difficilement détectable, mais semble avoir lieu tôt dans l'enfance[@]. Il y a ensuite une phase de latence, variable entre les individus, marquée par des épisodes d'exacerbations et de rémissions. À ce moment, l'éradication[@] par des antibiotiques reste possible. Puis survient le passage à la chronicité. *Pseudomnas Aeruginosa* s'adapte à son milieu et s'installe à long terme. Il perd certains caractères de virulence, mais devient résistant aux antibiotiques[@]. 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 bactérie est responsable d'activation de certains gènes par un processus appelé *quorum sensing*[@]. Un processus dans lequel chaque bactérie communique avec ses voisines par des signaux. Le génome de *Pseudomnas Aeruginosa* devient aussi hyperpermutable afin de présenter une plus grande diversité génétique au regard de la sélection naturelle [4].

À ce stade, le traitement antibiotique n'est plus curatif et l'évolution tend inexorablement vers un déclin de la fonction respiratoire. L'approche clinique est donc préventive. Elle vise à éliminer *Pseudomnas Aeruginosa* dès qu'il est détecté en culture. Une surveillance rapprochée des patients avec un prélèvement mensuel ou bimensuel est préconisée selon l'HAS[@]. 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 [@]. La PCR ciblée associant les protéines bactériennes *OPRL1* , *GYRB1* et *ECFX1* s'est montrée plus sensible et plus spécifique que la culture[@].

En pratique, la colonisation chronique est définie lorsque 3 expectorations sont rendues positives en culture, successivement au cours d'un suivi mensuel ou bimensuel[@]. Une autre classification, celle de Lee a montré une forte liaison clinico-biologique. Elle est composée de 4 groupes :

**groupe chronique** $>$ 50% des cultures sont positives sur 12 mois

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

**groupe intermédiaire** $\leq 50\%$ des cultures sont positives sur 12 mois

**groupe Free** Toute culture négative sur 12 mois, avec des antécédents

**groupe Nevers** Toute culture négative sur 12 mois sans antécédents

On ne sait pas aujourd'hui pourquoi *Pseudomnas Aeruginosa* s'installe préférentiellement chez les patients atteints de mucoviscidose. Plusieurs hypothèses ont été proposées :

— La dysfonction ciliaire empêche les pseudo d'être viré par le haut
— L'hypersalinité du film muqueux désactive les peptides antimicrobiens
— Le CFTR est un récepteur de pyo qui les internalise et les viré
— L'inflammation de l'epithelimum augmente les métabolites qui permettent de se développer.
— Alanine et l'acta sont une source de carbone pour le pyo.
— Le microbiote influence la colonisation
— Le pyo stimule le Systeme immunitaire pour virer tous les autres concurrents

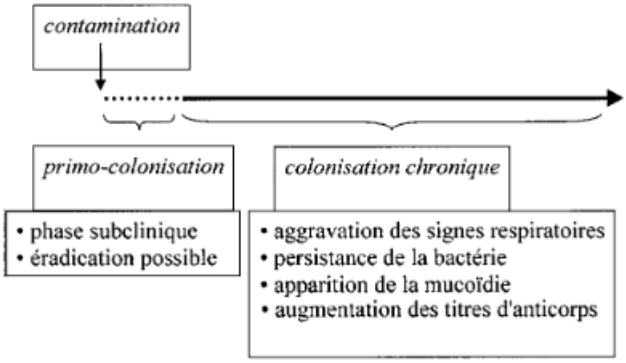


FIGURE 4 – Infection pulmonaire à pyo dans la mucoviscidose. ref []

## 3.2 Le microbiote pulmonaire

Bien qu'il soit en contacte avec le milieu extérieur, l'arbre respiratoire (comprenant la trachée, les bronches et les alvéoles) a longtemps été considéré comme stérile avec les méthodes de culture classique. Il a fallu attendre l'avènement du séquençage haut débit pour mettre en évidence le microbiote pulmonaire[ref Host-microorganism 1-3]. Le microbiote pulmonaire est beaucoup moins abondant que le microbiote digestif. Il est constitué d'une flore dynamique provenant de l'air ambiant, des voies supérieures, mais aussi du tube digestif par des micro-aspirations.[@] Le microbiote pulmonaire est dominé par le phylum des *Firmicutes* (*Streotpococcus*) et des *Bacteroidetes* (*Prevotella*). Les genres retrouvés majoritairement sont *Streptococcus*, *Prevotella*, *Fusobacteria*, *Veillonella*, *Haemophilus*, *Neisseria* et *Porphyromonas*. L'arbre respiratoire étant

en continuité direct avec les voies aériennes supérieur, certains genres bactériens sont communs, comme *Streptococcus*, *Staphilococcus*, *Haemophilus* et *Moraxella*. Tandis que d'autres genres comme *Corynébactérium* et *Dolosigranulum* ne sont retrouvés qu'au niveau du nez et de l'oropharynx.
Le microbiote pulmonaire varie dans l'espace et le temps.
Dans l'espace, du fait de sa structure, certaines régions de l'arbre bronchique peuvent présenter des microbiotes dissemblables. Un prélèvement au niveau d'un foyer infectieux sera nécessairement distinct d'un foyer sain. Le poumon montre également des différences physico-chimiques selon la localisation pouvant sélectionner certaines espèces. Les cavernes tuberculeuses par exemple se trouvent essentiellement dans le lobe supérieur en raison d'une concentration en oxygène plus élevé favorisant ce bacille aérobie stricte.
Dans le temps.....résilience .....
Le microbiote est variable entre les individus ...
Le microbiote est corrélé à la pathologie respiratoire. Plusieurs études suggèrent une différence entre patients sains et patients asthmatiques, BPCO ou atteint de mucoviscidose ..... [@].

## 3.3 Exploration du microbiote pulmonaire

### 3.3.1 Méthode de prélèvement

Le microbiote respiratoire est exploré en séquençant l'ensemble des ADN présent dans un échantillon. Toutes les méthodes de recueils sont possibles, mais les prélèvements protégés (combicath, LBA) sont recommandés afin d'éviter une contamination par les voies supérieures. Dans le cas contraire (ECBC) on peut évaluer la qualité du prélèvement en comptant le nombre de cellules épithélium ( normalement bas ) et de polynucléaire ( normalement haut ) dans le poumon. La meilleure méthode de prélèvement étant le prélèvement in situ réalisable lors des greffes pulmonaires.

### 3.3.2 Séquençage haut débit

Grâce à son haut débit, Le séquençage de nouvelle génération permet de séquencer l'ensemble des ADN présent dans un échantillon respiratoire et ainsi déterminer sa composition en bactérie. 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 capillaire sur la machine. À l'inverse, un séquenceur de nouvelle génération lit des fragments plus courts de l'ordre de 150pb. Mais cette lecture peut être parallélisé jusqu'à 20 billions de fois en un seul run sur un illumina Novaseq.
2 stratégies de séquençages sont utilisées en écologie microbienne :

**La stratégie shogun** consiste à séquencer l'ensemble des ADN présents dans l'échantillon sans discernement, que ce soit humain ou bactérien. 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. Il s'agit d'amplifier un gène uniquement bactérien et suffisamment variable pour discriminer une espèce. Pour les bactéries, il s'agit de l'ARN 16S.
L'ARN 16S est un ARN non codant participant à la structure de la petite sous unité des ribosomes bactériens. Il est composé de 1500 nucléotides et forme plusieurs boucles dans sa structure secondaire (Figure 6). L'alignement des séquences d'ARN 16S entre plusieurs espèces met en évidence des régions constantes et 9 régions variables (Figure 5). Les régions constante permette de designer des amorces s'hybridant à toutes les bactéries. Quand au séquences contenues dans les régions variables, elles apportent la spécificté taxonomique permettant d'identifier l'espèce bactérienne. Les études de [@] ont montrée que certaine région variable apportent plus de spécificité.

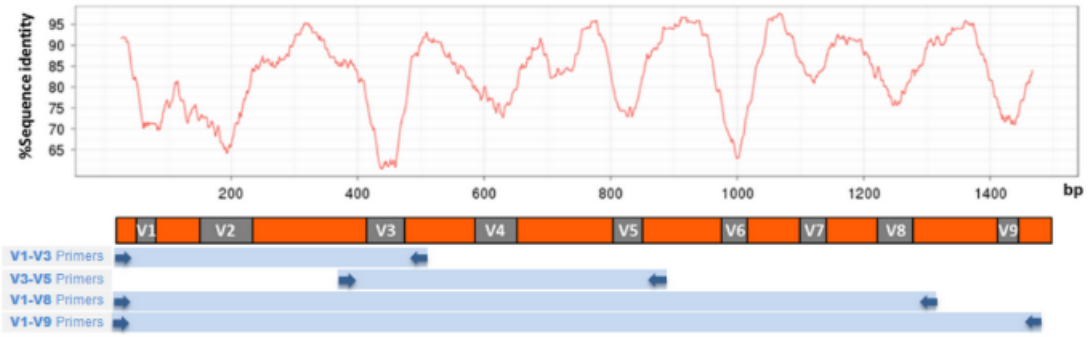


FIGURE 5 – région constante et variable de l'ARN 16S

La première stratégie est plus informative, car elle séquence l'ensemble des génomes bactériens et permet de prédire la fonction d'un microbiote. En effet, les transferts génétiques horizontaux amènent à dissocier l'espèce de sa fonction. 2 bactéries d'une même espèce peuvent avoir des fonctions différentes. L'inférence fonctionnelle réalisée à partir de la stratégie 16S est déconseillé[@]. 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.

## 3.4 Objectif de l'étude Mucobiome

L'objectif de notre étude est de savoir si le microbiote respiratoire influence la primo colonisation à *Pseudomnas Aeruginosa*. Pour cela, nous avons suivi pendant 3 ans, une

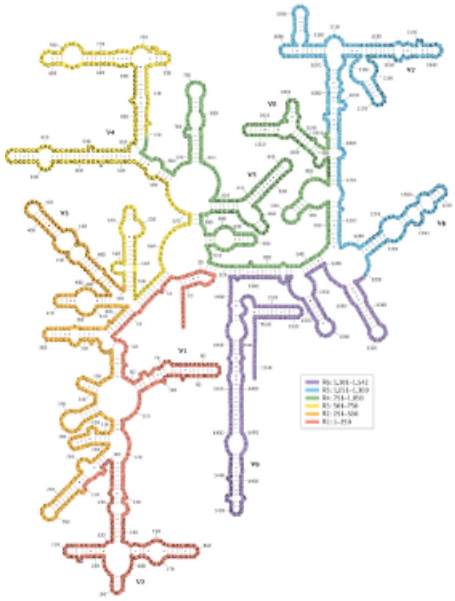

FIGURE 6 – Structure secondaire de l'ARN 16S

cohorte de 47 patients atteints de mucoviscidoses sans infection chronique. L'exploration de leurs microbiotes par la stratégie 16S a été réalisée sur leurs ECBC dans le cadre de leurs suivis réguliers. À partir des données générées du séquençage, nous avons effectué l'étude descriptive et analytique de leurs microbiotes en créant un pipeline bio-informatique dédié.

# 4 Matériel et Méthodes

## 4.1 Recueil des données

47 patients atteints de mucoviscidose ont été suivis sur 3 ans (2008-2011) dans une étude prospective multicentrique (Nantes,Brest,Roscoff) appelée mucobiome. La 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 ECBC des patients ont été recueillies lors des séances de kinésithérapie respiratoire tous les 3 mois, suivant 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 génotype CFTR et un test à la sueur positif. Les transplantés ont été exclus de l'étude. Une culture positive à *Pseudomnas Aeruginosa* était un critère de non-inclusion. Si pendant l'étude, une culture revenait positive à

ce dernier, le patient était sorti de l'étude pour être réinclus 1 an après en l'absence de colonisation chronique. 15 patients ont été ainsi réinclus. Chaque patient a été répertorié dans la catégorie Free ou Nevers (Lee et ale 2003). D'autres données ont été également recueillies ( tableau). Au total , 188 échantillons ont été récoltés, soit en moyenne 4 échantillons par patients. Pour chaque échantillon, une culture a été réalisée en suivant les procédures standards [ref]. Une qPCR ciblant le *Pseudomnas Aeruginosa* a parallèlement été réalisée en combinant les marqueurs gyrB/ecfX désignés au laboratoire[ref].

| | Never | Free |
|---|---|---|
| *N samples* | 114 | 74 |
| *Male (%)* | 77 (67.54) | 41 (55.41) |
| *Age mean (year)* | 22.45 | 19.65 |
| *Weight mean(kg)* | 39.3 | 33.66 |
| *Height mean (cm)* | 146.7 | 140.8 |
| *IMC mean (kg/cm²)* | 17.22 | 16.32 |
| *dF508/dF508 homo (%)* | 59 (51.75) | 35 (47.3) |
| *dF508/other hetero (%)* | 55 (48.25) | 39 (52.7) |
| *Taking Dornase (%)* | 78 (68.42) | 59 (79.73) |
| *Taking antibiotics (%)* | 48 (42.11) | 33 (44.59) |
| *Positif PA culture (%)* | 10 (8.77) | 12 (16.22) |
| *Positif PA PCR (OPRL1+) (%)* | 20 (17.54) | 26 (35.14) |

FIGURE 7 – Données associés par échantillons ( à refaire par patients)

## 4.2 Extraction de l'ADN

Les échantillons ont été liquéfiés avec du Dithiotrétiol. Les protéines ont été dégradées avec une Protéine kinase. 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 via QUIAamp DNA Minikit ( Quagen). Les extraits d'ADN ont été envoyés pour séquençage par un prestataire GATC.

## 4.3 Séquençage

La librairie [5] a été généré 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 généré sur Illumina MiSeq. Cette technologie permet de lire un couple de séquence de 300pb qui ensemble permet de lire une séquence plus longue de 535 pb. La figure 8 montre le chevauchement du couple de reads permettant de lire l'amplicon V3-V5.
Environ 25 millions de reads sont produits par run MiSeq. En multiplexant à l'aide de 94 index, les 188 échantillons ont été séquencé sur 2 runs. Au final 188 x 2 fichiers fastq ont été générés à l'issue du séquençage.

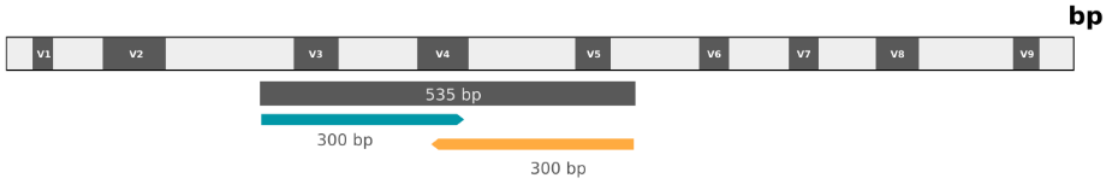


FIGURE 8 – le couple de séquence de 300pb permet de recouvrir l'ensemble de la région V3-V5

## 4.4 Analyse bio-informatique

L'analyse des 188 pairs de fichiers fastq a été réalisé grâce à un pipeline bio-informatique, appelé *mucobiome*, conçu et tester dans le cadre de cette étude. Par rapport aux autres logiciels comme **QIIME** ou **MOTHUR**, le pipeline mucobiome est spécialisé dans l'analyse des données 16S. Il est également plus rapide en raison d'un très haut niveau de parallélisation permis grâce à **Snakemake**. 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 fichiers fastq provenant du séquençage et produit un fichier BIOM contenant la table des OTUs. La figure 9 et 11 sont des graphe résumant les étapes du pipeline.

### 4.4.1 Prétraitement

### 4.4.2 Merging : Fusions des reads

*2 fastq en entrée 1 fastq en sortie.*

5. l'ensemble des fragments d'ADN à séquencer

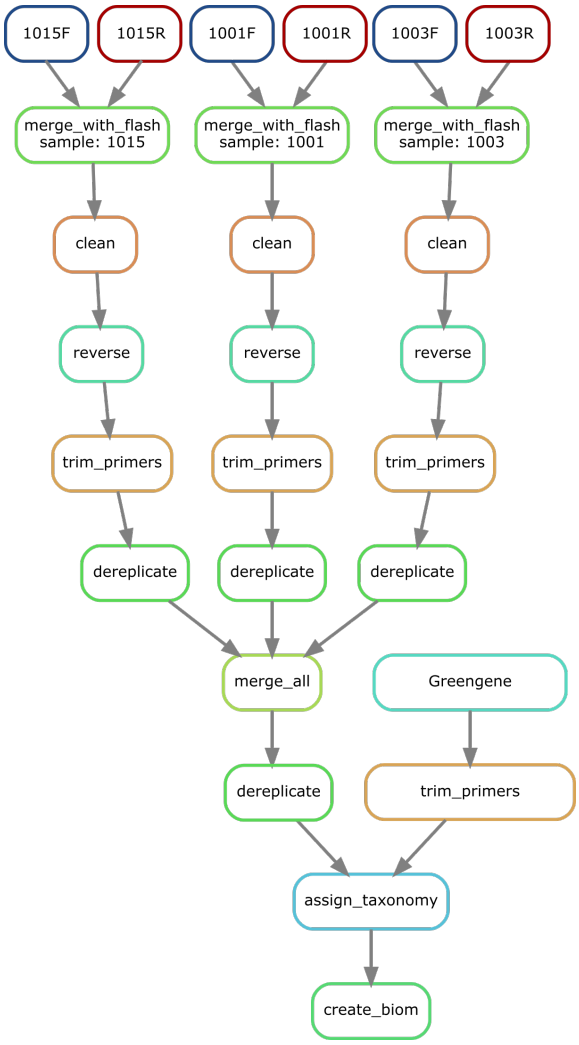


FIGURE 9 – Graphe du pipeline simplifié sur 3 échantillons 1015, 1001 et 1003. **merging** : Les reads pairs de 300 pb sont fusionnés pour produire un fichier fastq contenant des reads de 535pb. **cleanning** : les reads de mauvaise qualité sont supprimés. **reversing** : les reads sont transformés en leurs séquences complémentaires pour pouvoir être alignés. **Trimming** : seule la séquence entre les primes V3-V5 est conservée. **dereplicating** : les séquences dupliquées sont retirées. **merging** : L'ensemble des séquences est regroupé dans un seul fichier. **Taxonomie assignent** : Les séquences sont alignées sur la base de données greengene. **create_biom** : la table des OTU est créée

Les données brutes provenant du séquenceur sont des fichiers **Fastq**. Ils contiennent les séquences nucléotidiques et les scores de qualités par nucléotide lus par le séquenceur. Deux fichiers pair-end avec des reads de 300pb sont générés pour chaque échantillon. L'un correspond à la séquence lue dans le sens forward, l'autre dans le sens reverse. La première étape du pipeline consiste à fusionner ces deux fichiers afin de produire une plus longue séquence de 535pb correspondant à la région V3-V5 de l'ARN 16S. Le programme **Flash**[@] a été utilisé avec les paramètres par défaut. À partir de deux fichiers fastq, ce dernier recherche le meilleur alignement entre deux reads et produit

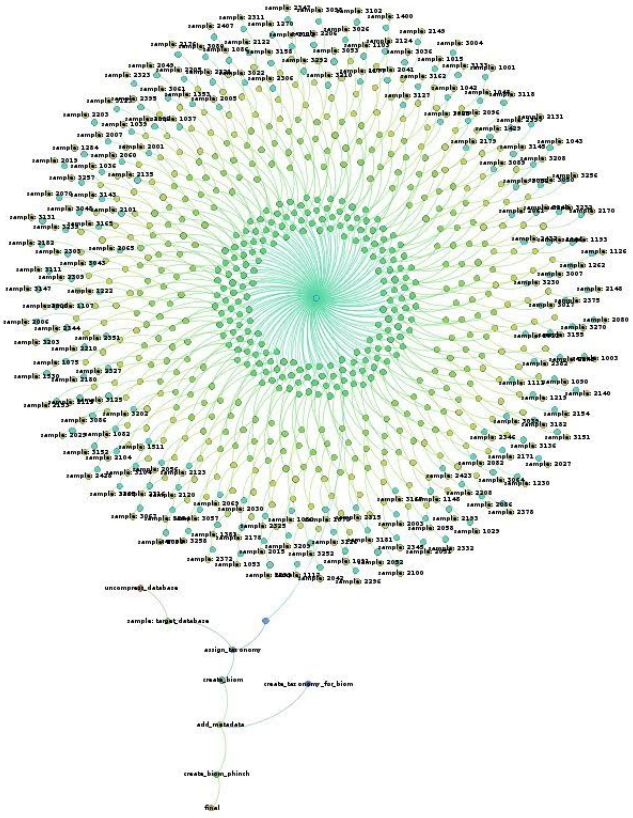


FIGURE 10 – le graphe du pipeline sur l'ensemble des échantillons

1 fichier fasta contenant les reads fusionnés. Une analyse qualitative des reads a été réalisée avec FastQt[?] avant et après fusions.

### 4.4.3 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. ( expliquer la qualité). Le filtrage des reads de mauvaise qualité est réalisé avec le programme **sickle**. 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 20pb. Une analyse qualitative des reads a été réalisée avec FastQt[?] après le filtrage.

### 4.4.4 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*. Pour permettre l'alignement, les séquences ont été remplacées par leurs séquences complémentaires grâce au programme **seqtk**. Par la même occasion les scores de qualités devenus inutiles sont supprimés. Les séquences sont sauvegardées dans un fichier fasta.

### 4.4.5 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 **cutadapts** est utilisé avec une tolérance de 0.1 par défaut.

### 4.4.6 dereplicating : Suppression des doublons

*1 fasta en entrée 1 fasta en sortie.*

Cette étape consiste à supprimer tous les reads dupliqués. En procédant ainsi, on s'assure de ne pas répéter l'assignement taxonomique plusieurs fois sur un même read. C'est une étape d'optimisation permettant d'économiser en temps de calcul. La déréplication a été réalisé avec **vsearch** et sa fonction *–derep_fulllength*

```
>sample1
ACGTTTTT
>sample1
ACGTTTTT
>sample1
GTAGAGT
>sample1
ACGTTTTT
```

*avant dereplication*

```
>sample1;size=3
ACGTTTTT
>sample1;size=1
GTAGAGT
```

*après dereplication*

FIGURE 11 – Exemple de déréplication d'un fichier fasta

### 4.4.7 Assignement taxonomique

*2 fasta en entrée, 1 fichier biom en sortie.*

L'assignement taxonomique consiste à labelliser 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 98% 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* version mai 2013 a été utilisé. Il s'agit d'un fichier fasta contenant 1 262 986 séquences et 203 452 OTUs.
L'autre stratégie d'assignement *de novo* n'a pas été utilisée. Cette dernière, de complexité $N^2$, consiste à comparer les reads entre eux pour former des groupes. Elle s'emploie de préférence pour détecter les bactéries absentes des bases de données.
L'assignement taxonomique a été réalisé avec vsearch et sa fonction *–usearch_global*.

#### 4.4.8 Analyse descriptive

L'analyse de la table des OTUs a été réalisée avec R et le package phyloseq.

# 5 Résultat

## 5.1 Pipeline Mucobiome

Après démultiplexage, 188 x 2 fichiers fastq ont été générés soit 2 fichiers paire par échantillons. La taille des reads pour chaque fichier est de 301 paires de bases. Au total 115 002 297 reads ont été produits sur 2 runs MiSeq. 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. La figure 14 illustre ces grandes variations entre échantillons.
Les analyses de qualité avec Fastqt montrent dans l'ensemble une baisse de qualité en fin de séquence. Les 50 derniers nucléotides ont des scores de qualités médiocres entre 10 et 20 points selon le Phred score.
La figure 12 représente le profil de qualité typique retrouvé. La figure 13 montre le profil de qualité des reads fusionnés avant et après filtrage des qualités. Après traitement des reads, c'est-à-dire merging et filtering, en moyenne 49,24 % des reads sont conservés avec des bornes allant de 37,30% à 61,13%.
99,88% de l'ensemble des reads analysés à reçu une assignation taxonomique.
Au total le pipeline mucobiome s'est exécuté en 1 h 29 sur 40 cœurs et 20 gigaoctets de mémoire contre 32h dans les tests précédant sans optimisation.

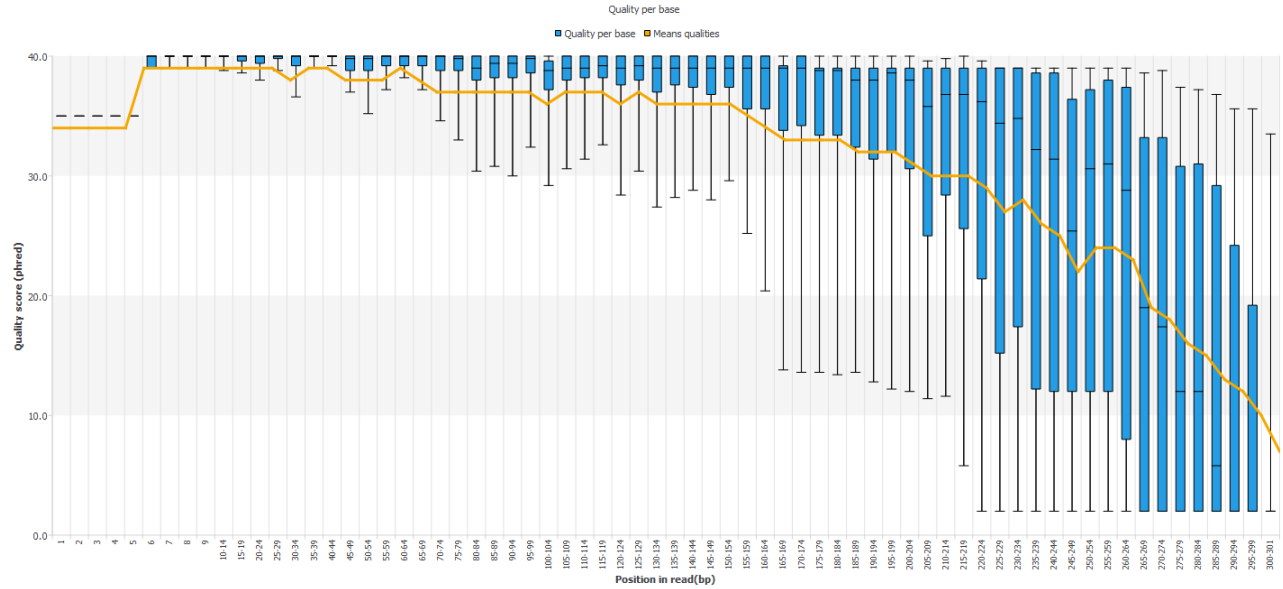


FIGURE 12 – Qualité par nucléotide des reads forward de l'échantillon 1003. **Axe X** : la position sur le read. **Axe Y** : La distribution des qualités

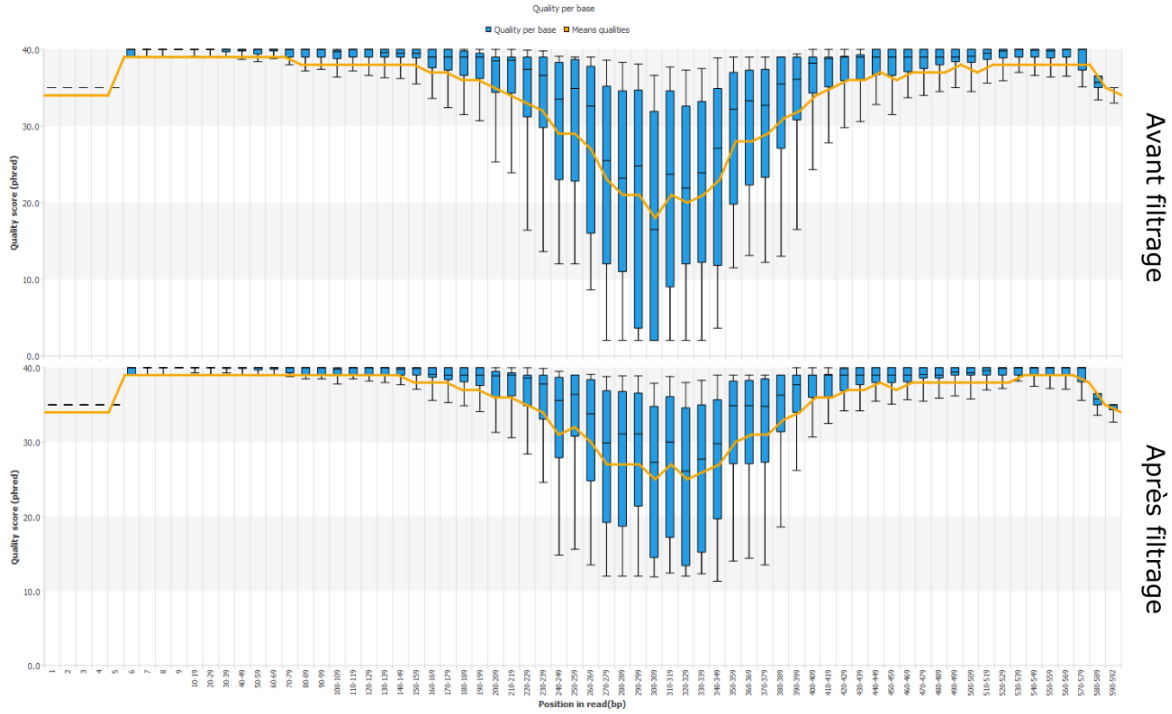


FIGURE 13 – Qualité par nucléotide sur l'ensemble des reads forward de l'échantillon 1003

## 5.2 Diversité du microbiote respiratoire

### 5.2.1 Courbe de raréfaction

Les courbes de raréfaction par échantillons (Figure 15) s'aplatissent précocement, témoignant d'un très bon niveau échantillonnage.

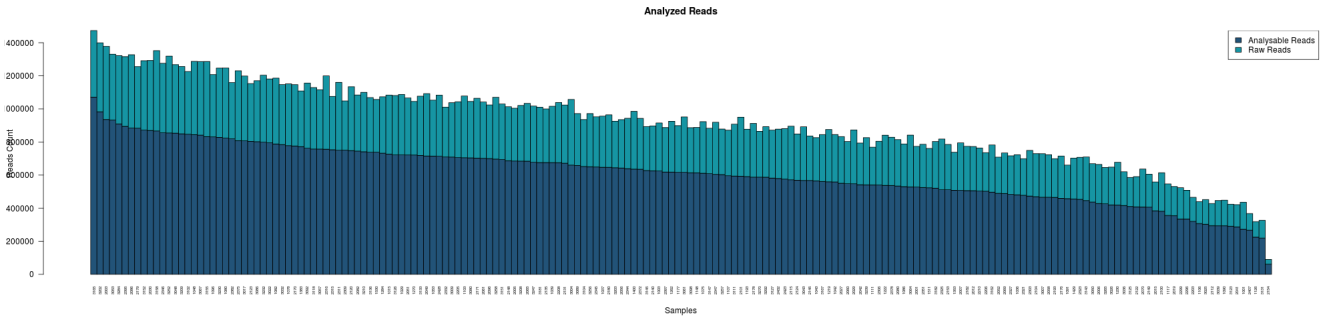


FIGURE 14 – Nombre de reads analysables avant et après filtrage

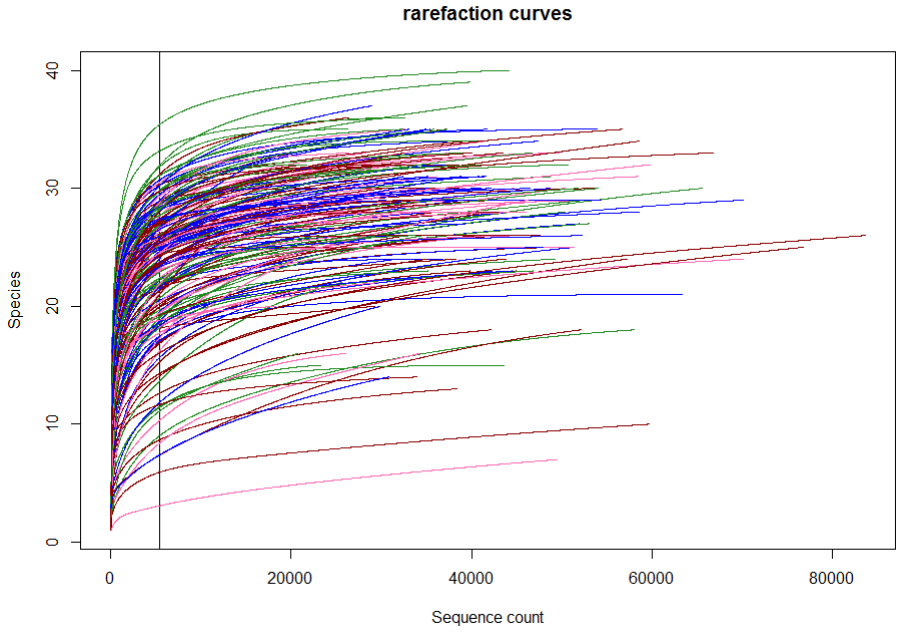


FIGURE 15 – Courbes de raréfactions des 188 échantillons

### 5.2.2 Diversité des échantillons

54 genres (Figure 17) et 7 phylums (Figure 16) bactériens sont retrouvés dans l'ensemble des échantillons. Les trois phylums majoritaires sont *Protéobacteria*(*Haemophilus*), *Firmicutes*(*Streptococcus*) et *Bacteroidete* (*Prevotella*).
Le tableau 20 montre la prévalence des genres bactériens dans les échantillons. Certains genres bactériens sont très prévalents, c'est à dire présent dans l'ensemble des échantillons. *Streptococcus*, *Neisseria*, *Prevotella*, *Granulicatella*, *Gemella*, *Veillonella* et *Fusobacterium* sont présents dans plus de 185 échantillons. D'autres sont dominantes, c'est-à-dire qu'ils représentent plus de 90% du microbiote pulmonaire dans certains échantillons. Il s'agit de *Streptococcus*, *Neisseria*, *Haemophilus* et *Staphyloccoccus* dans la majorité des cas. *Sténotrophomonas* et *Achromobacter* sont retrouvées dominantes

dans 64 et 8 échantillons.
Le core microbiota est défini comme l'ensemble des taxons retrouvé dans plus de 50% des échantillons et ayant une abondance > 0,1%. Il est constitué de 15 genres (Figure 18).
La figure 19 montre la variabilité des abondances entre échantillons. *Neisseria* et *Streptoccocus* ont des abondances très variables. *Staphiloccocus* et *Haemophilus* ont dans la plupart des échantillons des abondances relativement faibles. Leur variance est tirée vers le haut par quelque échantillon ou ils sont dominants.

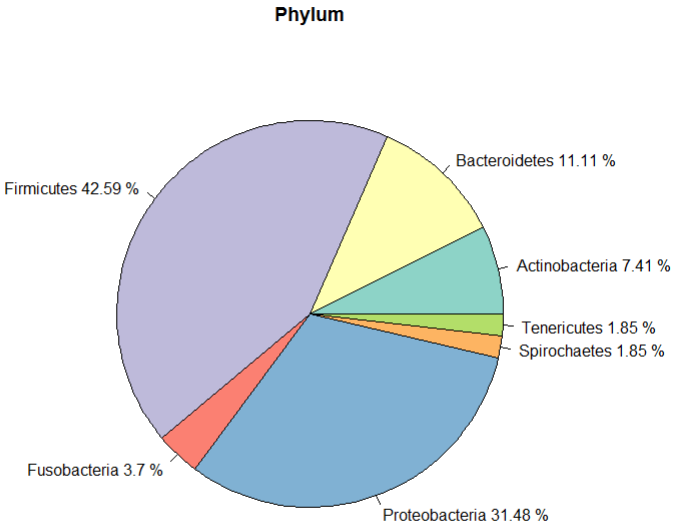


FIGURE 16 – truc

## 5.3 Évolution dans le temps

## 5.4 Comparaison entre catégories Free et Nevers

## 5.5 Sensibilité et Specificté du pyo

# 6 Discussion

# 7 Conclusion

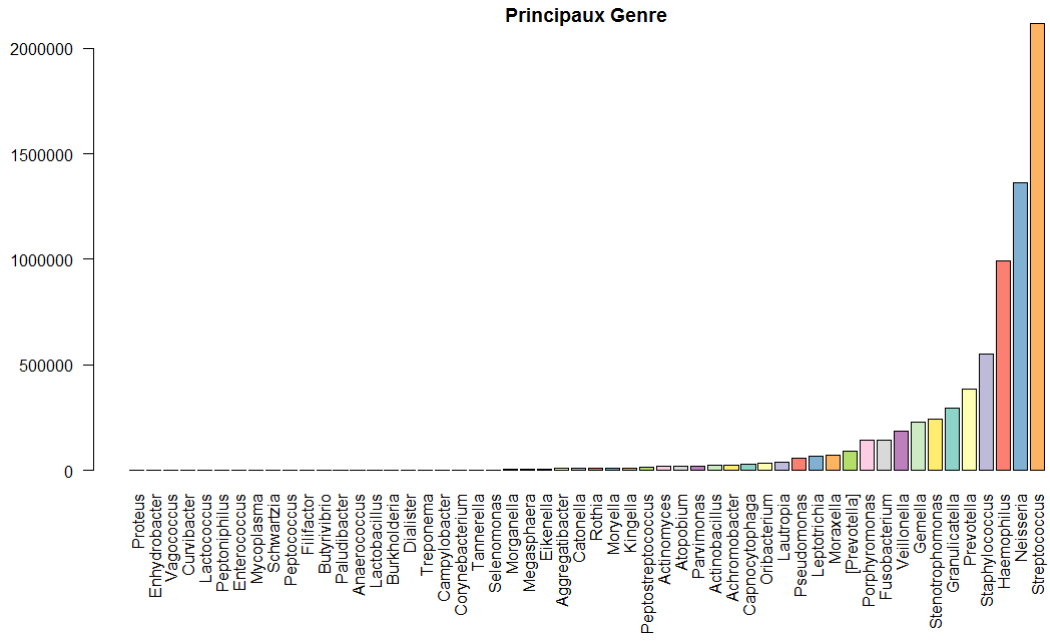


FIGURE 17 – truc

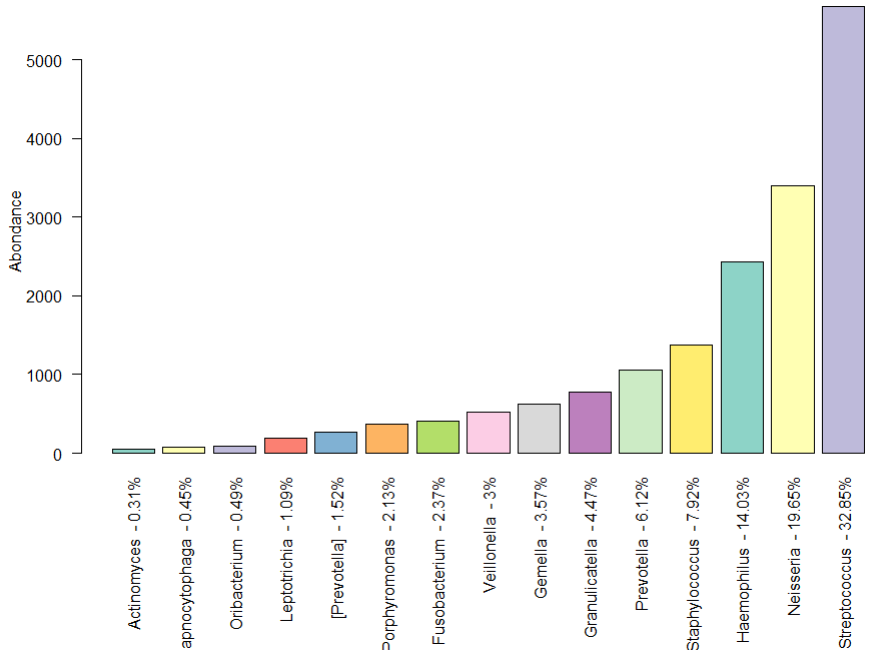


FIGURE 18 – truc

# Références

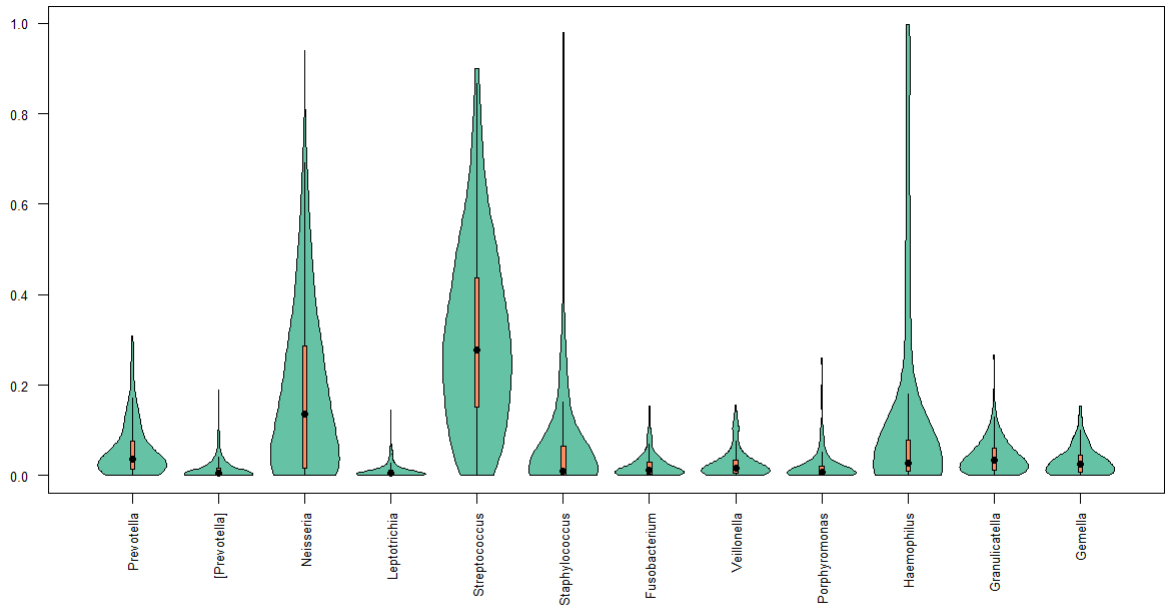


FIGURE 19 – truc

| | reads | reads.total.p | rel.abundance.mean | rel.abundance.sd | rel.abundance.min | rel.abundance.max | samples |
|---|---|---|---|---|---|---|---|
| *Streptococcus* | 2119507 | 29.32 | 30.22 | 20.04 | 0.01 | 89.87 | 188 |
| *Neisseria* | 1363473 | 18.86 | 18.08 | 18.49 | 0 | 93.96 | 187 |
| *Haemophilus* | 994209 | 13.75 | 12.91 | 23.97 | 0 | 99.78 | 184 |
| *Staphylococcus* | 549082 | 7.6 | 7.28 | 15.79 | 0 | 97.9 | 178 |
| *Prevotella* | 385850 | 5.34 | 5.63 | 5.98 | 0 | 30.85 | 188 |
| *Granulicatella* | 294632 | 4.08 | 4.12 | 3.83 | 0 | 26.73 | 185 |
| *Gemella* | 226403 | 3.13 | 3.29 | 3.26 | 0 | 15.34 | 186 |
| *Stenotrophomonas* | 241616 | 3.34 | 3.12 | 15.21 | 0 | 94.71 | 65 |
| *Veillonella* | 183601 | 2.54 | 2.76 | 3.18 | 0 | 15.62 | 188 |
| *Fusobacterium* | 144709 | 2 | 2.18 | 2.72 | 0 | 15.43 | 185 |
| *Porphyromonas* | 141016 | 1.95 | 1.96 | 3.46 | 0 | 26.08 | 167 |
| *[Prevotella]* | 92004 | 1.27 | 1.4 | 2.34 | 0 | 19.03 | 170 |
| *Leptotrichia* | 68581 | 0.95 | 1.01 | 1.69 | 0 | 14.44 | 172 |
| *Pseudomonas* | 56914 | 0.79 | 0.81 | 3.83 | 0 | 25.94 | 53 |
| *Moraxella* | 72796 | 1.01 | 0.79 | 6.33 | 0 | 75.51 | 51 |
| *Lautropia* | 39496 | 0.55 | 0.53 | 1.46 | 0 | 13.94 | 142 |
| *Achromobacter* | 25274 | 0.35 | 0.52 | 7.06 | 0 | 96.78 | 8 |
| *Oribacterium* | 31670 | 0.44 | 0.45 | 0.67 | 0 | 5.17 | 178 |
| *Capnocytophaga* | 26941 | 0.37 | 0.42 | 0.82 | 0 | 7.25 | 175 |
| *Actinobacillus* | 24228 | 0.34 | 0.37 | 2.35 | 0 | 31.19 | 95 |
| *Atopobium* | 19381 | 0.27 | 0.29 | 0.54 | 0 | 3.61 | 166 |
| *Actinomyces* | 19094 | 0.26 | 0.29 | 0.51 | 0 | 4.18 | 174 |
| *Parvimonas* | 19479 | 0.27 | 0.29 | 0.76 | 0 | 6.07 | 151 |
| *Peptostreptococcus* | 14961 | 0.21 | 0.23 | 0.49 | 0 | 2.93 | 129 |
| *Kingella* | 11943 | 0.17 | 0.17 | 0.47 | 0 | 4.55 | 149 |
| *Moryella* | 11912 | 0.16 | 0.16 | 0.39 | 0 | 3.62 | 143 |
| *Rothia* | 8346 | 0.12 | 0.12 | 0.25 | 0 | 1.83 | 138 |
| *Catonella* | 8130 | 0.11 | 0.12 | 0.17 | 0 | 1.16 | 159 |
| *Aggregatibacter* | 7597 | 0.11 | 0.11 | 0.32 | 0 | 2.51 | 86 |
| *Eikenella* | 6883 | 0.1 | 0.1 | 0.35 | 0 | 2.55 | 143 |
| *Megasphaera* | 5772 | 0.08 | 0.08 | 0.26 | 0 | 2.6 | 105 |
| *Morganella* | 5192 | 0.07 | 0.06 | 0.83 | 0 | 11.44 | 7 |
| *Selenomonas* | 2164 | 0.03 | 0.04 | 0.1 | 0 | 0.81 | 147 |
| *Tannerella* | 1240 | 0.02 | 0.02 | 0.04 | 0 | 0.31 | 120 |
| *Corynebacterium* | 854 | 0.01 | 0.01 | 0.03 | 0 | 0.18 | 113 |
| *Treponema* | 495 | 0.01 | 0.01 | 0.04 | 0 | 0.39 | 53 |
| *Burkholderia* | 415 | 0.01 | 0.01 | 0.11 | 0 | 1.36 | 2 |
| *Campylobacter* | 495 | 0.01 | 0.01 | 0.02 | 0 | 0.11 | 99 |
| *Lactobacillus* | 391 | 0.01 | 0.01 | 0.03 | 0 | 0.39 | 34 |
| *Dialister* | 461 | 0.01 | 0.01 | 0.02 | 0 | 0.16 | 53 |
| *Vagococcus* | 17 | 0 | 0 | 0 | 0 | 0.01 | 13 |
| *Mycoplasma* | 68 | 0 | 0 | 0 | 0 | 0.02 | 19 |
| *Anaerococcus* | 174 | 0 | 0 | 0.04 | 0 | 0.51 | 2 |
| *Enhydrobacter* | 16 | 0 | 0 | 0 | 0 | 0.02 | 9 |
| *Lactococcus* | 36 | 0 | 0 | 0 | 0 | 0.04 | 13 |
| *Paludibacter* | 158 | 0 | 0 | 0.02 | 0 | 0.18 | 18 |
| *Peptoniphilus* | 45 | 0 | 0 | 0 | 0 | 0.02 | 12 |
| *Schwartzia* | 85 | 0 | 0 | 0.01 | 0 | 0.12 | 11 |
| *Peptococcus* | 88 | 0 | 0 | 0 | 0 | 0.04 | 23 |
| *Butyrivibrio* | 146 | 0 | 0 | 0.02 | 0 | 0.19 | 10 |
| *Enterococcus* | 54 | 0 | 0 | 0.01 | 0 | 0.08 | 27 |
| *Proteus* | 12 | 0 | 0 | 0 | 0 | 0.01 | 9 |
| *Filifactor* | 93 | 0 | 0 | 0.01 | 0 | 0.11 | 5 |
| *Curvibacter* | 19 | 0 | 0 | 0 | 0 | 0.01 | 18 |

FIGURE 20 – truc

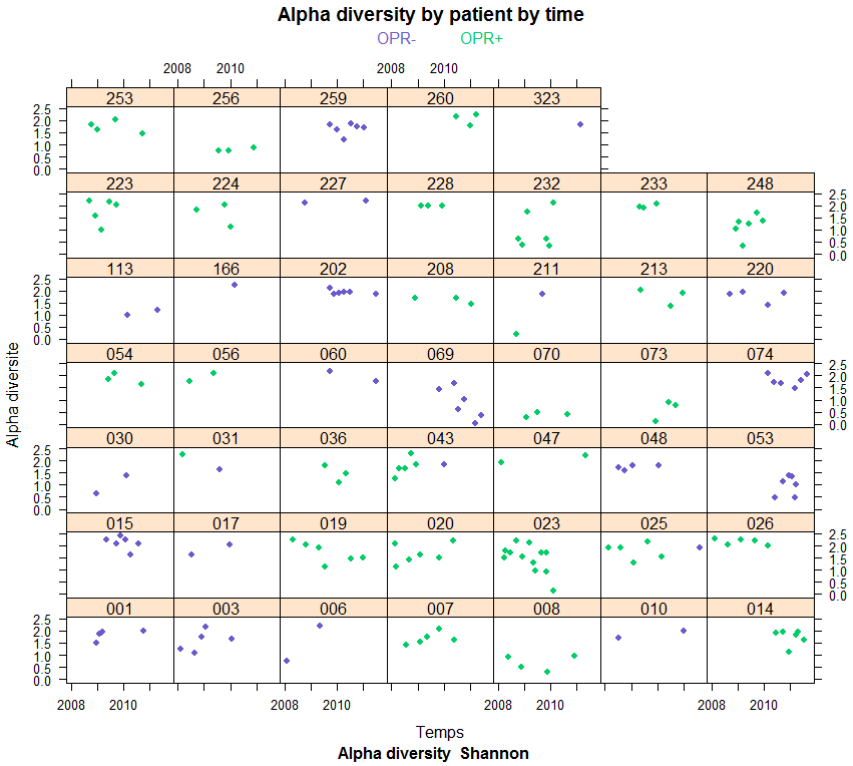


FIGURE 21 – truc

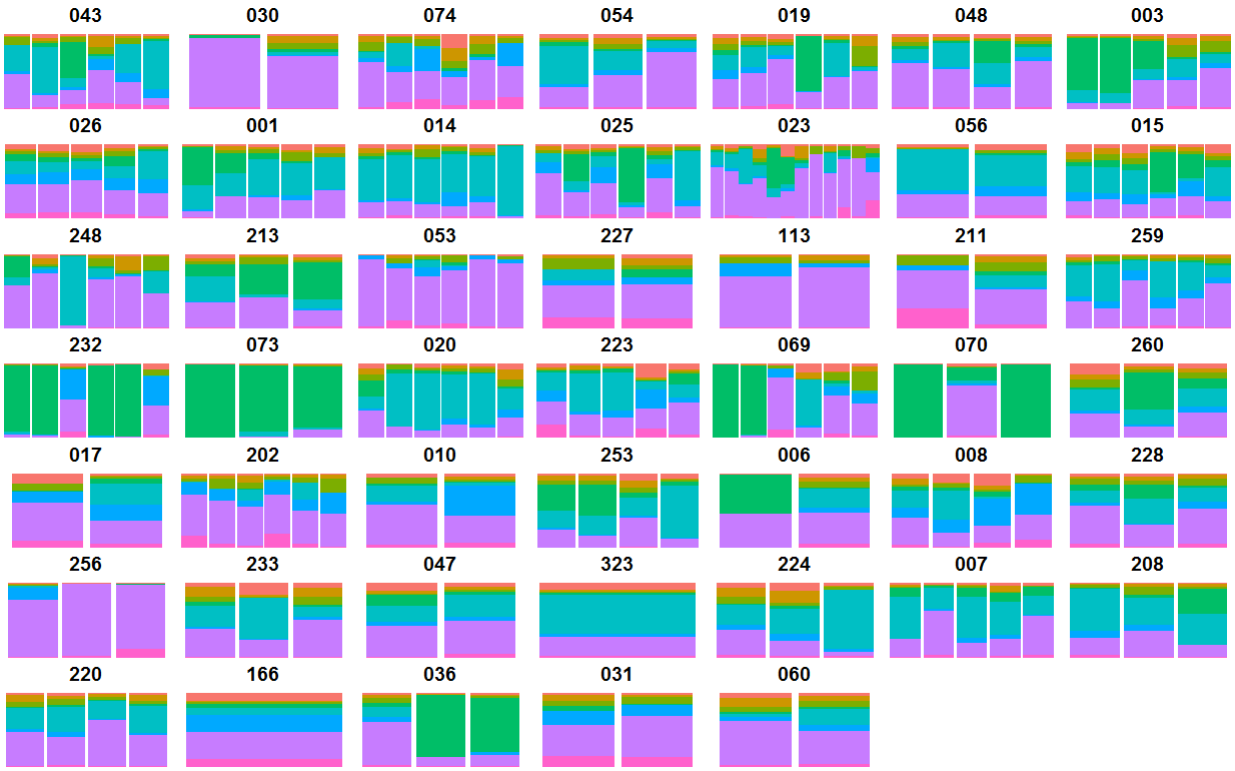

FIGURE 22 – truc