Guide pratique à destination des biologistes, bioinformaticiens et statisticiens qui souhaitent s'initier aux analyses métabarcoding

Hélène Falentin, Lucas Auer, Mahendra Mariadassou, Géraldine Pascal, Olivier Rué, Eric Dugat-Bony, Céline Delbes, Aurélie Nicolas, Etienne Rifa, Samuel Mondy, et al.

Guide pratique à destination des biologistes, bioinformaticiens et statisticiens qui souhaitent s'initier aux analyses métabarcoding

Hélène Falentin[1], Lucas Auer[2,3], Mahendra Mariadassou[4], Géraldine Pascal[5], Olivier Rué[4], Eric Dugat-Bony[6], Céline Delbès[7], Aurélie Nicolas[1], Etienne Rifa[7], Samuel Mondy[8], Malo Le Boulch[5], Laurent Cauquil[5], Guillermina Hernandez-Raquet[2], Sébastien Terrat[8], Anne-Laure Abraham[4]

Résumé. Les méthodes d'analyse métabarcoding (également appelées métagénomique ciblée ou amplicon) sont de plus en plus utilisées pour étudier la diversité des espèces présentes dans un écosystème (micro organismes, plantes, animaux). Le principe consiste à extraire l'ADN d'un échantillon environnemental puis à amplifier par PCR un fragment cible à l'aide d'un couple d'amorces prédéfini. Ces produits PCR, après ajouts de barcodes (oligonucléotides uniques pour chaque échantillon) et adaptateurs de séquençage, sont ensuite séquencés. Après le séquençage, les séquences sont triées par échantillon grâce aux barcodes puis assignées à des taxons par comparaison avec des séquences de référence. Beaucoup de méthodes et outils d'analyse ont été développés pour obtenir une vision la plus précise possible des écosystèmes étudiés. Les techniques de préparation puis d'analyse des échantillons dépendent de l'écosystème, des questions auxquelles on souhaite répondre et de la technologie de séquençage utilisée. Nous proposons des conseils issus de nos expériences, discussions et lectures bibliographiques afin de guider les lecteurs depuis la planification expérimentale jusqu'à l'analyse des données, en détaillant les points de vigilance à chaque étape.

Mots-clés : métagénomique ciblée, metabarcoding, séquençage

Abstract : Metabarcoding analysis methods (also known as gene-based metagenomics or amplicon) are more and more used to study the diversity of species in ecosystems (microorganisms, plants, animals). DNA from an environmental sample is extracted, then a targeted fragment is amplified by PCR with a predefined couple of primers. After the addition of barcodes (unique nucleotides for each sample) and sequencing adapters, the PCR

1 *INRA, UMR 1253 STLO, 35042 Rennes Cedex, France ; Agrocampus Ouest, UMR1253 STLO, 35042 Rennes Cedex, France.*
2 *LISBP, Université de Toulouse, CNRS, INRA, INSA, Toulouse, France.*
3 *Université de Lorraine, INRA, IAM, F-54000 Nancy, France. (Adresse actuelle)*
4 *INRA, MaIAGE, Université Paris-Saclay, 78350 Jouy-en-Josas, France, anne-laure.abraham@inra.fr*
5 *GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan, France*
6 *INRA, UMR 782 GMPA, 78850 Thiverval-Grignon, France ; AgroParisTech, UMR 782 GMPA, 78850 Thiverval-Grignon, France.*
7 *Université Clermont Auvergne, INRA, VetAgro Sup, UMRF, 15000 Aurillac, France*
8 *Agroécologie, AgroSup Dijon, INRA, Univ. Bourgogne, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France*

products are sequenced. After sequencing, reads are separated by sample by using the barcodes, and are assigned to taxa after a comparison with reference sequences. A lot of methods and analysis tools have been developed to obtain the most accurate view of the studied ecosystems. Preparation and sample analysis techniques depend on the ecosystem, the research issue and the sequencing technology used. We propose some advice taken from our experience, discussions and bibliographic reading so as to guide readers from the experimental planning to data analysis detailing some watchfulness at each step.

Keywords: gene-based metagenomics, metabarcoding, sequencing

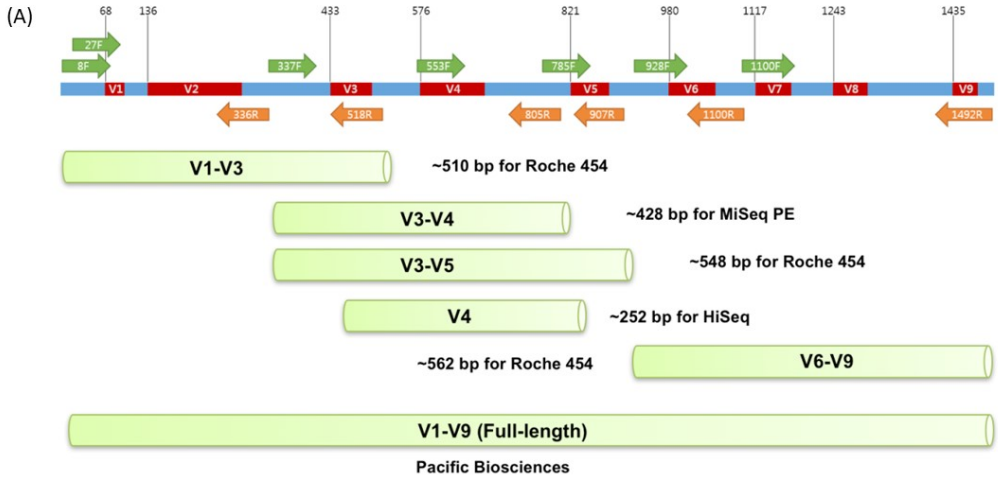
Introduction aux méthodes d'analyse de métagénomique ciblée

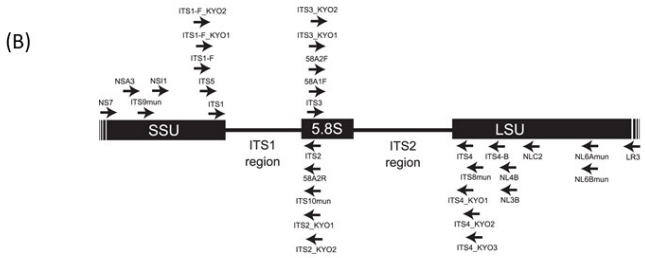
Choix du gène marqueur

Les gènes codant pour l'ARNr 16S et 18S sont les plus utilisés pour les analyses de métabarcoding ciblant des micro-organismes procaryotes et eucaryotes respectivement, et ce sont ces gènes qui seront plus particulièrement présentés dans ce document. Cependant, d'autres gènes marqueurs peuvent être utilisés (par exemple, l'ARNr 23S, gyrB, rpoB pour les procaryotes, ITS pour les eucaryotes, *rbc*L pour les plantes, COI pour les métazoaires...). Le choix du marqueur se fait en fonction de l'écosystème étudié et des bases de données disponibles. Il faut d'abord vérifier qu'il est possible de définir des amorces spécifiques pour le gène choisi, et que la taille du fragment (amplicon) à amplifier est compatible avec la longueur des séquences (reads) qu'il est possible d'obtenir avec la technologie de séquençage choisie.

Les gènes de l'ARNr 16S et 18S sont des séquences non codantes qui, une fois transcrites et repliées en boucle, vont participer à la constitution de la petite sous-unité des ribosomes des procaryotes (16S) et des eucaryotes (18S). Elles sont distribuées de manière ubiquitaire dans le monde vivant. Les séquences 16S et 18S sont constituées d'une succession de régions formant des épingles (très conservées, représentées en bleu sur la figure 1A) et des boucles (régions variables, représentées en rouge). Ces régions variables sont considérées comme des marqueurs reflétant les relations phylogénétiques entre les organismes (Woese *et al.* 1990). Les régions conservées permettent de définir des amorces pour amplifier ces gènes chez la plupart des micro-organismes (cf choix des amorces). Lors du dépôt d'une nouvelle espèce dans les bases de données, fournir les séquences 16S pour les procaryotes et 18S pour les eucaryotes est obligatoire. C'est pourquoi, toutes les espèces cultivables isolément et séquencées ont des séquences 16S ou 18S attribuées dans les bases de données. Les gènes codant pour l'ARN 16S/18S sont également peu soumis aux transferts horizontaux. Ces caractéristiques font des gènes 16S/18S des marqueurs de choix pour l'assignation taxonomique. Les gènes du 16S et 18S font respectivement environ 1500 et 1900 paires de bases (pb). Ils présentent 9 régions variables ; de manière générale, une ou deux de ces régions variables sont amplifiées pour les analyses de diversité (cf choix des amorces). Les séquences ITS1 et ITS2 sont très utilisées pour étudier les champignons. Ils sont de taille très variable, plusieurs amorces existent (figure 1 B et C). Une des difficultés est que les ITS les plus longs (> 560 bases) ne peuvent pas être séquencés en entier avec la technologie Illumina.

Le choix de ces gènes pour des analyses taxonomiques a aussi des limites qui seront discutées dans la suite de ce document. Un marqueur idéal est un marqueur universel, spécifique des espèces étudiées, de taille adéquate par rapport à la technologie de séquençage utilisée, à copie unique et suffisamment étudié pour être renseigné dans des bases de données. Un compromis doit être fait en fonction de l'écosystème et des questions biologiques d'intérêt.





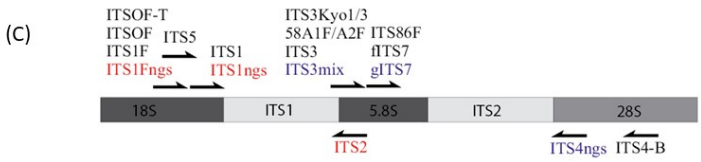


Figure 1A : Schéma représentatif de l'ADNr 16S des procaryotes (d'après ezbiocloud.net): les 9 régions variables apparaissent en rouge et les régions conservées en bleu. Les flèches indiquent la position des amorces.

Figures 1B et 1C : Représentation de l'opéron ribosomique eucaryote d'après Toju et al. 2012 et Tedersoo et al. 2015 respectivement (SSU : petite sous-unité ribosomique 18S, LSU : grande sous-unité ribosomique 28S, ITS : séquence interne du transcrit)

Extraction d'ADN

La méthode d'extraction est à optimiser pour chaque type d'échantillon, en fonction de la question scientifique et des espèces attendues dans l'échantillon. Une bonne revue résume une comparaison de méthodes d'extraction pour la métagénomique (Knudsen *et al.* 2016). De plus, il est connu que les kits d'extractions d'ADN sont plus ou moins performants selon la matrice à extraire (sol, aliments gras…) et les espèces en présence. L'étape de lyse doit être optimisée en fonction des espèces attendues. Il peut être judicieux d'ajouter des enzymes spécifiques dans le tampon de lyse (ex. lysostaphine pour extraire le genre *Staphylococcus*). Il est conseillé de faire des tests sur des communautés synthétiques de compositions connues afin d'identifier le kit le mieux adapté.

Par ailleurs, les kits eux-mêmes peuvent parfois être source de contamination. Celle-ci n'est en général visible qu'en cas de très faibles concentrations d'ADN à extraire et elle peut être différente d'un kit à l'autre (Salter *et al.* 2014). Si le problème est négligeable et même invisible dans la plupart des cas (quand l'ADN à extraire est suffisamment concentré), il faut en tenir compte pour des échantillons difficiles à extraire (Glassing *et al.* 2016). Dans ce cas, il est conseillé d'introduire au sein des échantillons à extraire un témoin négatif (remplacer l'échantillon par de l'eau) pour identifier les potentielles espèces contaminantes dans les kits.

Méthodes de séquençage

Le séquençage Sanger a permis de séquencer de longs fragments d'ADN préalablement clonés dans *E. coli*, mais cette méthode a un faible débit d'analyse et n'est pas adaptée à des études de communautés bactériennes complexes. Les NGS (Next Generation Sequencing) regroupent l'ensemble de technologie de séquençage haut débit : elles permettent de séquencer plusieurs milliers de séquences différentes à la fois. Historiquement, la technologie de pyroséquençage ROCHE 454 a été largement utilisée. Elle pouvait produire jusqu'à 200 000 lectures de 700 pb chacune. Avec l'évolution des technologies de séquençage, elle est maintenant remplacée par les technologies Illumina (MiSeq ou HiSeq). Le MiSeq produit actuellement 30M de lectures de 2x300 pb, en paired-end (la molécule amplifiée est séquencée par les deux extrémités). Alors que la technologie 454 présentait un taux d'insertions/délétions élevé du fait d'erreurs de lecture dans les homopolymères, la technologie Illumina présente un taux d'erreur inférieur à 2 %, majoritairement des substitutions. Les dernières améliorations de la chimie Illumina ont permis d'obtenir des fragments proches en taille de ceux du 454 avec un coût par base amplifiée beaucoup plus faible. Cette technologie est actuellement la plus utilisée dans les études d'écologie microbienne. D'autres technologies de séquençage haut débit existent dont le PGM Ion Torrent permettant l'obtention de séquences allant jusqu'à 400 pb mais avec un taux d'erreur élevé au niveau des homopolymères. Enfin, de nouvelles technologies de séquençage ont fait leur apparition ces dernières années comme celle d'Oxford Nanopore ou encore de Pacific BioSciences qui permettent l'obtention d'amplicons de plusieurs kb (il existe par exemple un kit 16S développé par Oxford Nanopore permettant l'obtention de la séquence de l'ADNr 16S dans son intégralité). A l'heure actuelle, leur application est limitée à cause des taux d'erreurs encore trop élevés (D'Amore *et al.* 2016) et au faible volume de données en sortie de séquençage. L'élaboration des banques de séquences et la réalisation du séquençage sont le plus souvent confiés à des plateformes de séquençage haut

débit ou à des prestataires de services. Les méthodes d'élaboration des banques et le séquençage lui-même ne seront pas détaillés dans ce document. Quelques explications utiles sont disponibles sur les sites des fournisseurs (New England Biolabs, Illumina) pour le séquençage Illumina ou dans des articles traitant du métabarcoding Nanopore (Krehenwinkel *et al.* 2019) ou PacBio (Tendersoo *et al.* 2018).

Qu'est-ce qu'une OTU ?

Le nombre de copies des gènes 16S/18S/ITS varie selon les organismes (de 1 à 21 chez les bactéries (16S), de 1 à 4 chez les archées (16S), de 14 à 1442 chez les champignons (ITS), et ces copies peuvent avoir des séquences différentes. D'autre part, plusieurs génomes du même genre peuvent avoir des copies identiques (Vetrovsky *et al.* 2013 ; Angly *et al.* 2014, Smets *et al.* 2016, Lofgren *et al.* 2019). De plus, différents biais techniques (erreurs de PCR, de séquençage etc.) génèrent des erreurs dans les lectures. Enfin, les banques de données ne sont pas exhaustives. Pour toutes ces raisons, il n'est pas possible de faire une assignation taxonomique directement sur les séquences obtenues, ni de compter directement les séquences pour estimer l'abondance des communautés.

Il est donc nécessaire d'avoir une méthode pour regrouper les lectures correspondant à un même groupe taxonomique. C'est le principe de la création d'unités taxonomiques opérationnelles, ou OTU, d'après leur acronyme anglais « Operational Taxonomic Unit » (Moyer *et al.* 1994). Les lectures sont alignées et comparées entre elles, et les relations entre deux séquences sont exprimées en pourcentage de similarité ou de divergence entre les bases nucléiques qui les composent. Les séquences qui présentent un pourcentage de divergence inférieur à un seuil donné sont alors regroupées au sein du même OTU. Sur la base des premières études portant sur l'ARNr 16S, le seuil de 97 % d'identité est généralement utilisé comme permettant de différencier deux espèces bactériennes (Stackebrandt and Goebel 1994). La comparaison des régions homologues du génome à celui des gènes de l'ARNr 16S de 7000 génomes a permis de proposer un seuil plus précis, de 98,65 %, pour définir deux espèces (Kim *et al.* 2014). Cependant, on sait aujourd'hui que ce seuil est variable entre branches de l'arbre phylogénétique, et influencé par les traits de vie des bactéries. Ainsi, certaines espèces (parfois certains genres) bien différenciées peuvent présenter des séquences d'ARNr 16S avec plus de 99 % d'identité. Il est par ailleurs à noter que ces seuils ont été définis sur la base de l'ARNr 16S complet, alors qu'en général c'est une longueur de 300 pb qui est séquencée, et que celle-ci peut comporter des proportions variables et de régions constantes. Il existe quelques méthodes pour regrouper les séquences sans seuil fixe (Mahé *et al.* 2014). Elles permettent d'obtenir des OTUs plus proches des espèces biologiques mais séparent moins bien le bruit, qui doit donc être filtré par ailleurs.

Qu'est-ce qu'un ASV ?

De nouvelles méthodes dédiées au débruitage (denoising) des séquences ont émergé récemment pour obtenir des ASV (Amplicon Sequence Variants) (Callahan *et al.* 2017) ou oligotypes (Eren *et al.* 2014). L'objectif est de corriger le bruit introduit dans les fragments séquencés plutôt que d'agglomérer les séquences proches. On parle d'une approche divisive plutôt qu'agglomérative. Le principe est le suivant : les lectures appartiennent initialement

toutes au même groupe. La diversité des lectures au sein de ce groupe est ensuite comparée à la diversité attendue uniquement en présence d'erreur de séquençages. Si elle est trop grande, le groupe initial est divisé en sous-groupes plus homogènes. Le processus est répété dans chacun des sous-groupes jusqu'à obtenir des groupes homogènes, c'est-à-dire correspondant à une « vraie » lecture et aux variations autour de celle-ci engendrées par les erreurs de séquençage. Ces approches permettent d'obtenir des OTUs très fines, ne différant que d'un nucléotide dans le cas le plus extrême, mais sont très sensibles à des biais systématiques lors du séquençage.

Biais techniques et biologiques des méthodes d'analyse de métagénomique ciblée

L'objectif principal du traitement bioinformatique est de traiter les lectures issues du séquençage afin d'obtenir un tableau de comptage qui indique, pour chaque échantillon, l'abondance des OTUs qu'il contient.

De l'extraction des échantillons au séquençage, chaque étape est susceptible d'introduire des erreurs dans les séquences (erreurs lors des PCR, erreurs lors du séquençage), ou des biais de quantification (biais d'extraction, d'amplification par PCR ou de profondeur de séquençage) (Figure 1). Il peut également y avoir formation de séquences chimériques, assemblages artéfactuels de plusieurs séquences. Enfin, il peut exister des sources de contamination, qui vont aboutir à la détection de lectures qui ne devraient pas être présentes dans l'échantillon (biais de la méthode d'extraction d'ADN, contamination technique lors de l'amplification par PCR ou lors du séquençage entre échantillons du même run ou entre runs successifs). Il est important de connaître ces biais biologiques, techniques et humains pour les corriger dans la mesure du possible. Certains biais sont détaillés ci-dessous.

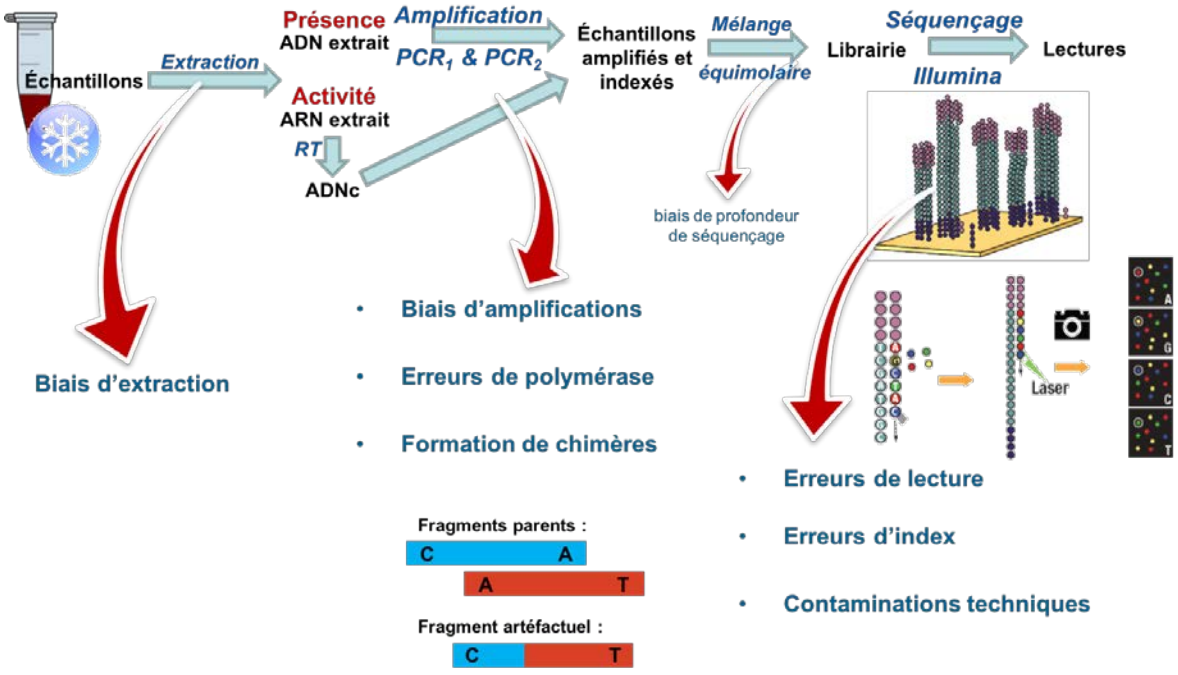


Figure 2. Possibles sources d'erreurs de l'extraction des échantillons au séquençage.

Diversité et nombre de copies 16S

Le nombre de copies des gènes 16S/18S/ITS varie selon les organismes, et donc le nombre de copies trouvées ne peut pas servir à quantifier directement un genre/une espèce dans un échantillon, mais peut être utilisé pour comparer des échantillons entre eux. Ce nombre de copies peut être différent au sein d'un même taxon, et des taxons proches peuvent avoir des séquences identiques. Ce phénomène complexifie l'assignation taxonomique et la détermination de l'abondance des espèces (Větrovský and Baldrian 2013; Angly *et al.* 2014 ; Smets *et al.* 2016).

Les erreurs séquençage et PCR

Des erreurs de séquençage de type substitution et insertion/délétion peuvent être introduites à deux niveaux, pendant la préparation des librairies ou au moment du séquençage. Les erreurs introduites au moment de la préparation des librairies sont principalement des erreurs de PCR. Leur fréquence est dépendante des protocoles propres à chaque technologie de séquençage (PCR en émulsion, …) et aux polymérases utilisées.

Les erreurs de séquençage *sensus stricto* sont des erreurs de lecture par le séquenceur. Le profil de ces erreurs est donc dépendant de la technologie de séquençage. La technologie 454 est réputée pour introduire facilement des insertions/délétions dans les homopolymères (D'Amore *et al.* 2016) alors que les erreurs de lecture sont les plus fréquentes dans la technologie Illumina (Nelson *et al.* 2014).

Les chimères

Les chimères sont des séquences "physiques" artéfactuelles, produites au cours de la PCR. Lorsqu'une séquence en cours d'élongation se détache de sa matrice, elle est susceptible de se ré-hybrider à un brin matrice différent lors des cycles suivants, et d'ainsi former une molécule chimérique. Cette chimère, composée de la séquence de deux (ou plus) brins matrices, n'a aucune existence dans l'échantillon de départ, mais devient présente dans la librairie et sera donc séquencée. Les chimères sont des sous-produits de PCR et leur abondance est donc nécessairement plus faible que les séquences réelles. Mais elles peuvent conduire à la création d'OTUs chimériques, en abondance faible, mais qui peuvent représenter une part très importante du nombre d'OTUs. (Bonder *et al.* 2012 ; Shin *et al.* 2015 ; Bachy *et al.* 2013).

Les biais d'amplification PCR

Le rendement de l'amplification par PCR dépend de la séquence : l'hybridation des primers est influencée par d'éventuels mésappariements avec la matrice, la polymérase peut patiner davantage sur certaines régions du fait de la séquence ou de structures secondaires. Sans qu'on sache encore le prédire, on peut donc avoir des différences d'efficacité d'amplification selon les séquences matrices, ce qui conduit à des abondances biaisées avant et après PCR. Ce type de biais ne peut être corrigé à l'heure actuelle (Aird *et al.* 2011).

Les contaminations

Méthode d'extraction ADN

Les biais liés à l'extraction d'ADN ont été détaillés dans le paragraphe 1b.

Pendant le séquençage au sein d'un run et entre runs

L'un des intérêts des NGS est le séquençage multiplex, c'est-à-dire, le séquençage simultané de plusieurs échantillons (ou librairies), chacun étant identifié par un barcode unique. Avec la technologie Illumina MiSeq, il a été plusieurs fois signalé la présence inattendue de séquences, lors du séquençage d'échantillons témoins de composition connue (Nelson *et al.* 2014). Ces séquences inattendues ont été identifiées comme provenant d'autres librairies séquencées dans le même run, et seraient donc mal attribuées. Leur abondance reste assez faible (<0,1 %) mais très dépendante de la composition du run de séquençage. Si celui-ci présente une faible diversité et des populations très dominantes, celles-ci peuvent atteindre des abondances de contamination élevée dans tous les autres échantillons du run.

Choix des amorces et de la région amplifiée

Il est important de prendre le temps de bien choisir les amorces d'amplification PCR de l'amplicon à séquencer. Les amorces doivent réunir les qualités suivantes :

- les amorces 'forward' et 'reverse' doivent avoir la même température de fusion (+/- 2°C près) et ne pas s'hybrider entre elles (possibilité de vérifier sur le site Primer3).
- elles doivent s'hybrider au mieux sur la séquence cible 16S ou 18S. Pour obtenir une amplification optimale sur l'ensemble des espèces de l'échantillon, les amorces sont souvent dégénérées. Dans ce cas, les amorces contiennent d'autres lettres que ATGC. L'amorce synthétisée correspond alors à une population d'amorces ne contenant pas toutes les mêmes séquences mais représentant toutes les possibilités selon les bases utilisées.
- selon votre prestataire de séquençage, il peut être nécessaire d'ajouter en 5' de l'amorce des séquences adaptateurs. Ces adaptateurs ne s'hybrident pas sur la séquence cible, leur composition en bases n'affecte donc pas la température de fusion de l'amorce.
- les amorces permettant l'amplification PCR vont être choisies dans les régions conservées et de manière à encadrer les régions variables.
- la région variable cible doit être judicieusement choisie pour être discriminante entre les espèces attendues dans l'écosystème, de façon à maximiser les chances d'assignation taxonomique précise au moment de l'analyse bioinformatique. Quand on dispose des informations préalables sur les espèces dominantes de l'écosystème étudié, il est possible de récupérer les séquences 16S ou 18S sur la base de données RDP ou LTP, faire un alignement multiple des séquences (avec Mega5 par exemple) et choisir la zone variable qui différencie le mieux les espèces. Il faut également vérifier que les amorces

dégénérées permettront de bien amplifier les espèces étudiées. La bibliographie portant sur des écosystèmes proches peut également être utile pour définir les amorces à utiliser. A des fins de publication, il est indispensable de justifier le choix de la région variable ciblée.

- il faut également faire attention aux amorces dégénérées car elles peuvent être sources d'amplification non spécifique.

Enfin, une très bonne revue peut vous guider dans le choix des amorces (Klindworth *et al.* 2013). N'hésitez pas également à demander conseil auprès de votre prestataire de séquençage. Pour aller plus loin les références suivantes sont très informatives : Liu *et al.* 2008 ; Kumar *et al.* 2011; Cruaud *et al.* 2014.

Les différentes étapes d'une analyse, avec quelques exemples d'outils

Preprocessing

Le preprocessing a pour but d'éliminer les lectures contenant des bases inconnues ("N"), les lectures trop courtes ou trop longues (les lectures trop longues peuvent être des chimères et les lectures trop courtes ne seront pas assez informatives pour l'assignation taxonomique), les lectures ne contenant pas les amorces, ou avec des erreurs dans les amorces. Il permet également, grâce au repérage des barcodes, d'associer les lectures à l'échantillon d'origine quand plusieurs échantillons ont été séquencés dans une même piste d'Illumina (démultiplexage). Il est aussi possible de contiguer les lectures qui viennent d'un séquençage paired-end. Enfin, la déréplication permet de garder un exemplaire d'une séquence qui serait vue plusieurs fois. L'information d'abondance est conservée (souvent dans la description des séquences au format FASTA). Cette étape est purement technique et permet d'accélérer les calculs suivants.

Déchimerisation

Il existe deux principes pour éliminer les chimères : la déchimérisation sur référence ou la déchimérisation *de novo*.

La déchimerisation sur référence part du principe que les séquences 'parents' de la chimère sont présentes dans les bases de données. La chimère est détectée par alignement des lectures sur une base de données de 16S de référence (Silva, Greengenes, LTP). Une lecture mosaïque qui génère des alignements partiels sur 2 ou 3 ou 4 16S appartenant à des taxons différents sera éliminée.

La déchimérisation *de novo* part du principe que les séquences 'parents' de la chimère sont présentes dans l'échantillon séquencé en plus grande quantité que la chimère elle-même. La chimère est détectée par alignement de chaque lecture sur les autres lectures de l'échantillon. Une lecture mosaïque qui génère des alignements partiels (sur une partie de la longueur) sur 2 ou 3 ou 4 autres lectures différentes sera éliminée.

Une des limites de la déchimérisation sur référence, est que les banques de références contiennent des chimères qui ne pourront donc pas être détectées (Mysara *et al.* 2015). D'autre part, cette méthode n'est pas adaptée pour des écosystèmes peu connus dont la plupart des séquences ne sont pas dans les banques.

Clustering

Le principe du clustering est de regrouper les lectures qui ont un fort pourcentage d'identité en clusters. Dans les étapes ultérieures, seule une lecture (souvent la plus abondante) sera assignée taxonomiquement et toutes les lectures du même cluster serviront au calcul d'abondance de ce taxon. Ainsi, le clustering permet de réduire le nombre de séquences à traiter, et donc le temps de calcul. De plus, cela permet de simplifier les analyses ultérieures en regroupant les séquences similaires (souvent de la même espèce, cf. biais du nombre de copies, et erreurs de PCR et séquençage). Les analyses seront donc faites sur les OTUs et non sur les séquences individuelles.

Il existe plusieurs méthodes de clustering. Le plus connu est celui basé sur un seuil d'identité de 97 %. Une autre méthode développée récemment, SWARM, agrège les lectures, non pas sur un seuil global d'identité, mais en agrégeant progressivement les séquences qui ont jusqu'à X différences (valeur conseillée de X : entre 1 et 3) (Mahé *et al.* 2014). Il existe également d'autres méthodes qui ne sont pas basées sur un pourcentage d'identité : les MED Nodes (Eren *et al.* 2014) et les Amplicon Sequence Variants (ASV) (Glassman and Martiny 2018). En général, les OTUs produits par toutes les méthodes de clustering arrivent à bien identifier les grands patrons écologiques et les changements de composition taxonomique de haut niveau. En revanche, le paramétrage du clustering a une forte influence sur le nombre de taxons différents (indice de richesse) identifiés à l'issue de l'analyse.

Denoising

L'approche inverse du clustering est le denoising qui est proposé par dada2 (Callahan *et al.* 2017) ou Deblur (https://github.com/biocore/deblur). L'objectif étant de corriger les erreurs de séquençage en se basant sur des modèles d'erreurs construits en utilisant des modèles paramétriques sur les données de séquençage.

Filtres

À l'issue de la construction des OTUs, il est de coutume d'appliquer des filtres basés sur l'abondance des OTUs. En effet, les OTUs en faible abondance sont majoritairement des chimères ou des lectures avec des erreurs de séquençage, qui n'ont pas été détectées avant. Le groupe de Boukulich recommande de supprimer les OTUs dont l'abondance (relative) dans le jeu de données est inférieure à 0.005 % (Bokulich *et al.* 2013). Si on dispose de réplicats biologiques pour chaque condition d'intérêt et qu'on souhaite conserver toutes les bactéries "typiques" d'au moins une condition, on peut appliquer le filtre précédent *par condition* et garder tous les OTUs sélectionnés dans *au moins* une condition. On peut également appliquer un filtre de prévalence et ne sélectionner que les OTUs présents dans au moins la moitié (ou X% avec X à choisir) des réplicats d'au moins une condition. On s'assure ainsi de ne conserver que les OTUs dont la présence dans une condition donnée est répétable.

Enfin, si on dispose de réplicats techniques, on peut déterminer empiriquement le seuil à partir duquel les OTUs identifiées sont liés à des biais techniques, et supprimer les OTUs dont l'abondance est inférieure à ce seuil dans tous les échantillons.

Les filtres sont très importants, ils permettent de réduire drastiquement le nombre de faux OTUs, rendent l'interprétation biologique des résultats plus facile et accélèrent les traitements (statistiques ou bioinformatiques) ultérieurs.

Assignation taxonomique

L'assignation taxonomique se fait par similarité de séquences par rapport à des bases de données de séquences de référence (Silva, Greengenes, LTP) ou des bases de données spécifiques à un écosystème (HOMD pour le microbiote oral de l'homme par exemple). Les méthodes d'assignation se font par alignement de séquences (BLAST (Altschul *et al.* 1990)). Le paramétrage concerne les seuils minimaux de pourcentage d'identité et de séquence couverte que l'on autorise pour l'assignation taxonomique.

Table d'OTUs et format BIOM

La table d'OTUs est généralement mise à disposition par les pipelines d'analyse au format biom (http://biom-format.org/). Il s'agit du standard général mis en place par la communauté pour manipuler et stocker des données de comptages sous forme compacte. Ce format de fichier peut contenir à la fois la matrice qui contient les comptages de chaque OTU dans chaque échantillon, les métadonnées associées aux échantillons et des métadonnées associées aux OTUs (généralement la taxonomie).

La souplesse du format BIOM fait néanmoins que la taxonomie peut prendre différentes formes (une chaîne de caractères, un vecteur de noms taxonomiques) et qu'il faut parfois utiliser des fonctions d'import différentes pour lire correctement les informations taxonomiques. De façon générale, le format BIOM est un format destiné à des machines, il a vocation à être lu plus facilement par une machine plutôt que par un être humain.

Les bases de données

Le choix des marqueurs à utiliser pour l'identification taxonomique dépend de la qualité des banques de données utilisée. Les séquences présentes dans les banques servent de référence pour l'assignation taxonomique, et la richesse et la précision des banques est donc déterminante pour une bonne identification. Les marqueurs moléculaires peu courants ont donc pour inconvénient supplémentaire de ne souvent disposer que de banques de petite taille. Les marqueurs très courants, comme l'ARNr 16S et 18S disposent parfois de plusieurs banques de données différentes, qui ont toutes leurs avantages et inconvénients.

Silva est une des banques de référence d'ARNr 16S et 18S les plus riches qui, de plus, est mise à jour très fréquemment (Quast *et al.* 2013). Dans ses dernières versions, une étape de non-redondance a été appliquée pour contrer une richesse en trop forte augmentation qui la rendait coûteuse en ressources. Un critère de 99 % d'identité

est maintenant appliqué à l'aide de l'outil UCLUST afin d'éliminer les séquences très similaires. Le nombre d'entrées est fortement réduit mais la représentativité est ainsi conservée. La dernière version, la 123, contient 597 607 séquences, contre près de cinq millions au total et 1,7 millions de séquences uniques. Un ensemble d'outils (ARB) est distribué par Silva.

La base de données LTP (All-Species Living Tree Project) est une sous-partie 'nettoyée' de Silva et contient un nombre beaucoup plus faible de séquences (11 900 dans sa dernière version) (Munoz *et al.* 2011). Les séquences répertoriées correspondent uniquement à des souches types d'espèces bien classifiées d'Archées et de Bactéries. Son avantage est un poids très faible, ce qui a un impact important sur les temps de calcul et une 'propreté' inégalée avec uniquement des séquences vérifiées définies jusqu'à l'espèce voire la souche. Cependant, pour des études environnementales ou concernant des communautés faiblement caractérisées et peu connues, elle est peu informative.

RDP (Ribosomal Database Project) est à la fois une banque de données et un ensemble d'outils, dont une méthode d'assignation par k-mer (Cole *et al.* 2014). Cette banque est pour l'instant non nettoyée et regroupe plus de trois millions de séquences ; ceci la rend lourde à manipuler. D'autre part, RDP met à disposition une infrastructure d'assignation en ligne, et sa méthode est une des plus reconnues, elle est ainsi l'une des plus utilisées.

Greengenes est une base réputée pour sa 'propreté', avec des séquences vérifiées manuellement et ne contient aucune chimère, un problème courant dans d'autres bases de données (DeSantis *et al.* 2006). Avec un peu plus d'un million de séquences, elle contient moins de séquences que Silva ou RDP mais sa représentativité pour tous les phyla est très bonne. En plus, son format est compatible avec l'outil ARB de Silva. Cependant, les mises à jour sont assez rares, la dernière remontant à 2013. Elle reste assez utilisée.

EZBioCloud (Yoon *et al.* 2017) est une base de données regroupant des séquences 16S curées ou extraites de génomes complets provenant du NCBI et du JGI. Les fichiers de séquence et de taxonomie sont disponibles sur demande après enregistrement et gratuits pour les académiques.

UNITE est une base de référence, assez complète et diversifiée pour les champignons (Nilsson *et al.* 2018) qui contient un peu plus de 800 000 séquences d'ITS. Les fichiers préformatés pour Qiime, Mothur et Usearch sont mis à disposition sur le site web (https://unite.ut.ee/index.php).

Bases spécialisées

Enfin, de nombreux travaux utilisent des bases de données spécifiques à leur objet d'étude, soit construites *de novo*, soit enrichies à partir d'une base existante. C'est le cas de DictDB, une banque dédiée à l'étude des flores digestives d'insectes (Mikaelyan *et al.* 2015b). Basée sur Silva, DictDB incorpore les données de clones obtenus à partir d'études de microbiotes intestinaux d'insectes, pour lesquels la résolution et la précision en sont ainsi augmentées. L'avantage de ce type d'approche est de pouvoir inclure des séquences déjà trouvées dans d'autres échantillons (même si elles sont inconnues), tout en pouvant se servir d'une base de données plus légère,

contenant que des séquences pertinentes pour l'environnement étudié. DairyDB (Meola *et al.* 2018) est une base de données 16S spécialisée pour les produits laitiers qui a été curée manuellement et permet des assignations taxonomiques plus précises que des banques généralistes.

Une étude comparative des bases de données 16S a été réalisée par Balvočiūtė and Huson (2017).

ISHAM (Irinyi and Meyer 2015) est une base de données ITS spécialisée dans les champignons pathogènes de l'homme et des animaux contenant environ 4000 séquences représentant 640 espèces fongiques pathogènes.

Principaux pipelines bioinformatiques existants

MOTHUR

À l'origine conçu pour traiter des données 454, Mothur (Schloss *et al.* 2009) incorpore des outils spécifiques du 454 comme des débruiteurs pour corriger les erreurs d'homopolymères. Chaque outil étant optionnel, il est complètement adaptable à des données Illumina. Les outils qu'il contient permettent de passer de données brutes à des tables d'abondances avec assignation taxonomique ; quelques banques de données dans le format compatible Mothur sont disponibles au téléchargement. Mothur propose une large gamme de choix pour la plupart des étapes clés citées précédemment. Il propose aussi l'estimation des courbes de raréfaction, des indices de diversité, des diagrammes de Venn ou autres statistiques, notamment de test d'hypothèses. Les possibilités sont très vastes, mais une des faiblesses de Mothur réside dans sa méthode de clustering. Celle-ci repose sur la construction d'une matrice de distances entre séquences qui nécessite d'être complètement chargée en mémoire pour que l'algorithme de clustering, très lent, puisse tourner. Cette étape est la seule pour laquelle il n'est proposé qu'un seul outil. Il est possible de paramétrer le seuil global ainsi que la définition de cluster (voisin le plus proche, le plus éloigné ou moyen). Il est possible de réduire le nombre de séquences à clusteriser à l'aide d'un outil de pré-clustering qui permet de réduire le bruit de séquençage. Après un alignement multiple, si une séquence très rare, présente moins de différences avec une abondance, elle est considérée comme artéfactuelle. Son abondance est alors ajoutée à la séquence abondante et sa séquence éliminée du jeu de données.

Dans le cas de jeux de données de plusieurs millions de séquences, il devient techniquement impossible d'utiliser Mothur selon la procédure standard conseillée par les auteurs (Schloss *et al.* 2009) du fait de cette étape limitante de clustering. Une astuce possible pour dépasser ces limitations consiste à filtrer en amont du clustering toutes les lectures singletons (monocopie) pour réduire la complexité du jeu de données. Ce filtre permet de réduire les besoins en mémoire et de gagner en temps de calcul, tout en ayant un impact quasiment indétectable sur les résultats finaux (Auer *et al.* 2017). Mothur est un outil disponible en ligne de commande mais il existe une interface utilisateur graphique (GUI) disponible pour Mac, utilisable uniquement en local avec de petits jeux de données. Mothur est disponible sur l'instance Galaxy de Migale, de Genouest et d'ABIMS - Roscoff.

QIIME

QIIME (Quantitative Insights Into Microbial Ecology) est un pipeline polyvalent d'analyse de données de type métabarcoding. Il inclut de nombreux outils pour chaque étape d'analyse : le démultiplexage, les filtres qualités, la détection de chimères, le clustering, l'assignation taxonomique, la phylogénie, ainsi que des analyses d'alpha et beta diversité et des outils de visualisation des résultats (Caporaso *et al.* 2010). Pour cela, il s'appuie sur un ensemble de scripts codés en python et faisant appel à de nombreux logiciels libres (par exemple : Blast, Uclust, Usearch, ChimeraSlayer, RDP Classifier, Swarm, R). L'inconvénient est qu'il peut être difficile à utiliser pour des personnes qui débutent.

Les possibilités de Mothur et QIIME sont équivalentes car ils reposent sur l'utilisation des mêmes outils, mais les formats qu'ils utilisent ne sont pas les mêmes. La différence majeure entre les deux réside dans le choix d'outils proposés pour certaines étapes : par exemple, QIIME ne dispose pas d'un outil de pré-clustering, mais inclut plusieurs outils de clustering dont UPARSE, DADA2 et Swarm, qui peuvent permettre de traiter des jeux de données de grande taille.

QIIME est disponible en ligne de commande (par exemple sur les plateformes Migale et Genotoul), ou via des instances Galaxy (par exemple Migale). La version QIIME2, disponible depuis 2018 (https://docs.qiime2.org/2018.8/) inclut une interface graphique utilisateur. Il existe un guide de QIIME 1 pour ceux qui souhaitent lancer une première analyse, avec un choix fixe d'outils (http://nbviewer.jupyter.org/github/biocore/qiime/blob/1.9.1/examples/ipynb/illumina_overview_tutorial.ipynb). Attention, ce guide ne propose pas de traiter les chimères. Il existe également des guides pour QIIME2 (https://docs.qiime2.org/2018.8/tutorials/). La communauté d'utilisateurs de QIIME est très active et de nombreux tutoriels sont disponibles (http://qiime.org/tutorials/index.html). De plus, il existe un forum très utile, sur lequel des utilisateurs ainsi que les développeurs de QIIME répondent aux questions liées à cet outil (https://groups.google.com/forum/#!forum/qiime-forum).

UPARSE

Le workflow UPARSE inclut les étapes suivantes : assemblage de reads pairés, filtre qualité, filtre sur la taille, déréplication, suppression des singletons (optionnel) et clustering des OTUs avec l'algorithme UPARSE (ci-dessous) (Edgar 2013). C'est historiquement le premier outil à fournir des résultats proches des résultats attendus sur des échantillons de composition connue. UPARSE propose une approche associant la mise à l'écart des lectures singleton et un algorithme de clustering et détection de chimères réalisés simultanément, qui permet d'obtenir un nombre d'OTU très proche de l'attendu, là où les autres méthodes surestimaient systématiquement de 5 à 10 fois la diversité. Le filtrage des lectures singleton est optionnel mais proposé par défaut. Il permet une forte réduction du nombre d'OTUs détectés et participe à l'exceptionnelle rapidité d'UPARSE. En effet, en plus de sa précision validée sur des mélanges contrôlés de souches, UPARSE est capable de traiter en quelques heures des jeux de données de grande taille, dont le traitement pouvait prendre plusieurs semaines avec d'autres outils. Cependant, il ne résout pas les problèmes liés au clustering hiérarchique à seuil global (Edgar 2013).

FROGS

FROGS (Find Rapidly OTU with Galaxy Solution), développé par l'INRA de Toulouse et Sigenae, est un logiciel facile à utiliser pouvant être aussi bien utilisé sous Galaxy qu'en ligne de commande (Escudié *et al.* 2018). FROGS permet l'analyse de grands ensembles de séquences d'amplicons d'ADN avec précision et rapidité. Il produit une table d'abondance des OTUs avec leurs affiliations taxonomiques. FROGS est le seul outil capable de gérer les paires de reads non chevauchantes, comme cela peut être le cas pour les marqueurs ITS, RPOB, LSU ou tout autre marqueur de taille supérieure à 600 nucléotides.

FROGS a été conçu pour prendre en charge des séquences multiplexées ou démultiplexées. L'étape de prétraitement se charge d'apparier les lectures entre elles, de nettoyer et dérépliquer les séquences. L'étape de clustering des séquences utilise Swarm qui fonctionne avec un seuil de clustering local et non pas global (souvent 97 % par défaut) comme le font les autres logiciels de clusterisation. L'étape d'élimination des chimères utilise VSEARCH combiné à une validation croisée des chimères originales. L'étape de filtration permet de supprimer la majorité restante des séquences artéfactuelles. La dernière étape permet d'affilier chaque OTUs à un taxon (jusqu'à l'espèce) avec la possibilité d'une annotation "multi-affiliation" qui permet de générer une liste de tous les taxons qui présentent la même séquence. De nombreuses analyses statistiques et d'illustrations graphiques sont générées tout au long de l'analyse, permettant à l'utilisateur de suivre graphiquement l'effet de chaque étape. Les différentes étapes de FROGS peuvent être utilisées comme outils indépendants ou comme pipeline (http://frogs.toulouse.inra.fr/, SOP présenté figure 2). Les scripts sont disponibles (https://github.com/geraldinepascal/FROGS).

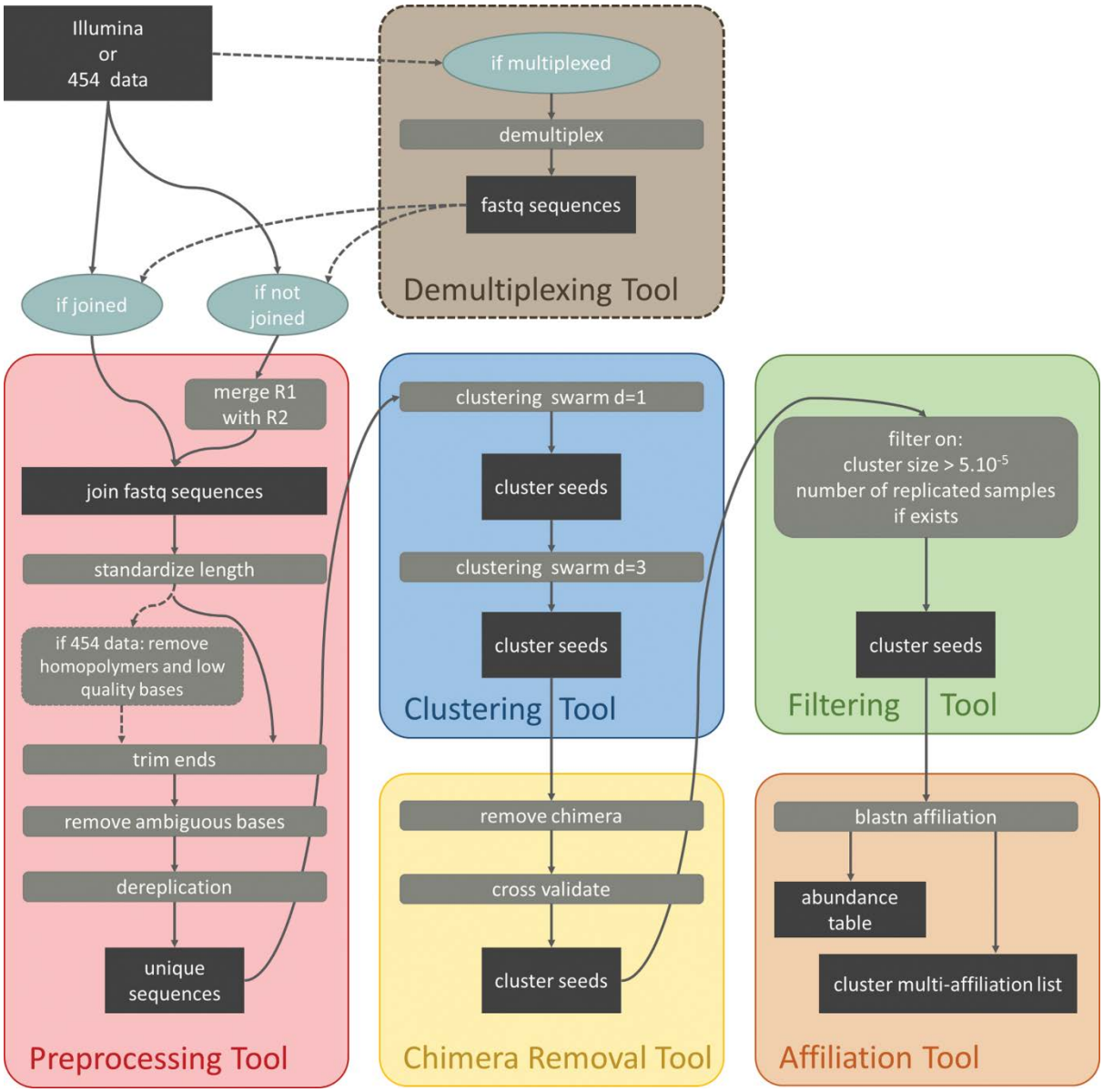


Figure 3. SOP de FROGS.

FROGS propose également huit outils dédiés à l'analyse statistique des tables d'abondances obtenues. Un premier pour la reconstruction d'arbre phylogénétique avec Mafft ou Pynast suivis de FastTree et phangorn, et sept autres outils pour l'analyse statistique à proprement parler :

- FROGS Phyloseq Import Data : import des fichiers dans un objet Phyloseq.
- FROGS Phyloseq Composition Visualization : pour la visualisation brute de la composition.
- FROGS Phyloseq Alpha Diversity : visualisation et analyse de la variance d'alpha diversité (6 indices).
- FROGS Phyloseq Beta Diversity distance matrix : calcul de matrices de distances (plus de 40 méthodes possibles).
- FROGS Phyloseq Structure : visualisation des structures de communautés via des heatmaps et graphiques d'ordination (3 méthodes possibles).
- FROGS Phyloseq Clustering : génération d'un clustering hiérarchique des échantillons.

- FROGS Phyloseq Manova : analyse multivariée.

Chaque outil produit un fichier HTML regroupant les différentes visualisations. Il permet également de récupérer le code R pour aller plus loin dans les analyses.

FROGS est disponible, entre autres, sur les instances Galaxy de Genotoul Bioinfo, Migale, GenoOuest, Ifremer, Roscoff, Plateforme de bioinformatique de Lyon et de Bordeaux, de l'INRA de Montpellier. FROGS est installable sur toutes les plateformes Galaxy avec peu de contraintes grâce au packaging bioconda (https://anaconda.org/bioconda/frogs) et les wrappers Galaxy (https://testtoolshed.g2.bx.psu.edu/view/oinizan/frogs_2_0_0).

MG RAST

MG-RAST (*Metagenomic Rapid Annotations using Subsystems Technology,* https://www.mg-rast.org/) est un serveur web open-source qui propose une analyse phylogénétique et fonctionnelle des données métagénomiques. Un de ses principaux avantages est son interface graphique. Par contre, le choix d'outils et de paramètres est limité. Il est nécessaire de s'inscrire pour avoir un compte (Keegan *et al.* 2016).

Pipeline de l'EBI

L'European Bioinformatic Institut propose un pipeline d'analyse de séquences métagénomiques (shotgun ou amplicon 16S) (Mitchell *et al.* 2016).

Vous devez créer un compte European Nucleotide Archive (ENA) pour soumettre vos reads bruts et les métadata associées (https://www.ebi.ac.uk/metagenomics/submission). Un numéro d'accession vous sera fourni et pourra être utilisé dans vos publications. L'ENA n'accepte que les données qui feront l'objet d'une publication. Puis l'équipe de l'EBI traite vos séquences avec le pipeline EBI. Le pipeline accepte les séquences 454, Illumina et Ion Torrent. Après un contrôle qualité des fichiers, si les données sont pairées, elles sont contiguées avec SeqPrep. Ensuite, l'outil rRNAselector trie les séquences 16S et les envoie dans un pipeline QIIME pour l'assignation taxonomique en utilisant la base de données Greengenes (actuellement version 13.8) et le 'closed-reference OTU picking protocol' par défaut (https://www.ebi.ac.uk/metagenomics/pipelines/3.0). L'ensemble des données peuvent être gardées confidentielles pendant deux ans maximum puis elles sont rendues publiques et accessibles via l'ENA.

Dada2

Dada2 est un package R qui est utilisable depuis qiime2 ou directement sous R. Les auteurs recommandent une succession d'étapes (https://benjjneb.github.io/dada2/tutorial.html) pour obtenir la table d'abondance. Tout d'abord, il est conseillé de visualiser la qualité des bases des reads pour évaluer leurs caractéristiques (longueur, qualité). Ensuite une étape de filtres permet de tronquer et de filtrer les reads suivant leur taille et leur qualité. En effet, l'étape d'apprentissage des taux d'erreur est plus sensible si les données d'entrée sont de bonne qualité. Dada2 utilise un modèle paramétrique pour évaluer les erreurs de séquençage. Une fois l'apprentissage effectué, les reads sont corrigés à partir de ce modèle. Le modèle par défaut est calibré pour la technologie Illumina mais est paramétrable en cas d'utilisation d'une autre technologie. Ensuite, si nécessaire, les paires sont contiguées, en choisissant la longueur et le nombre d'erreurs dans l'overlap. Les ASVs (Amplicon Sequence Variant) sont ainsi formés. De façon classique, une étape de suppression de chimères est effectuée, puis les ASVs sont assignés avec RDP. Dada2 fournit sur son site internet les fichiers de base de données formatés pour RDP, Greengenes, Silva et UNITE. Dada2 a l'avantage de proposer des ASV qui ont une résolution plus fine que les OTUs, et peut être adapté à certaines études pour lesquelles la discrimination de souches très proches entre elles sont nécessaires (Callahan *et al.* 2017). Dada2 est un package R qui peut s'installer facilement, il est disponible sur la plateforme Migale (Jouy en Josas).

Conclusion

Les analyses de métagénomique ciblée permettent de détecter les micro-organismes présents dans des échantillons complexes avec un niveau de précision jamais atteint par les autres techniques, et ceci sans passer par des étapes de culture ou d'enrichissement. La détection de microorganismes sous-dominants est possible dans la mesure où la technique peut générer plusieurs dizaines voire centaines de milliers de séquences par échantillon. Elle peut donc être utilisée pour créer des inventaires très exhaustifs de micro-organismes à partir d'échantillons d'origines variées (échantillons environnementaux, associés à un hôte, aliments, etc…). De plus, l'énorme capacité de multiplexage offerte par l'utilisation des nouvelles techniques de séquençage (>300 échantillons peuvent être séquencés à la fois sur un seul run de MiSeq), fait de cette technique l'une des plus adaptée pour des études en écologie microbienne.

Cette technique permet d'avoir une vision approximative de l'abondance relative des différentes entités biologiques présentes dans les échantillons. Etant donné les biais évoqués précédemment, elle ne peut en aucun cas permettre de quantifier de manière absolue les différents micro-organismes présents dans les échantillons. Si vous souhaitez avoir une vision quantitative de l'écosystème, vous pouvez utiliser des approches de Digital PCR ou qPCR qui sont plus coûteuses. Une autre méthode pour améliorer la quantification est d'ajouter un "spike" (souche pure qui n'est pas présente dans l'écosystème étudié) en concentration identique dans tous les échantillons, avant l'extraction d'ADN. La quantification de cette espèce dans le tableau d'OTU final permet d'ajuster les quantifications des OTUs entre échantillons.

Afin de valider vos résultats, il est conseillé de prévoir des réplicats biologiques, et si possible des échantillons contrôles (milieu de culture, souche pure connue, échantillon précédemment séquencé par exemple), qui vous

permettront de mieux évaluer les biais techniques (comme proposé par The Microbiome Quality Control Project Consortium *et al.* 2017). Il est également conseillé de préparer des témoins négatifs d'extraction d'ADN et de PCR.

L'affiliation taxonomique des OTUs 16S est généralement précise jusqu'au niveau du genre et parfois jusqu'à l'espèce. Les résultats d'affiliation obtenus seront bien entendu liés à la complétude et à la précision de la base de données utilisée. Il est donc recommandé de bien choisir sa base de données et même de combiner plusieurs sources si cela s'avère nécessaire. Dans certains cas, des séquences appartenant à deux espèces différentes, voire à deux genres différents, peuvent avoir une séquence identique sur la portion du biomarqueur utilisé (par exemple la séquence du gène codant pour l'ARNr 16S est strictement identique pour les membres des genres Escherichia et Shigella). Conscient de cette limitation, il est conseillé de bien vérifier les affiliations proposées automatiquement par les outils d'affiliations et de tenir compte des cas particuliers pour l'interprétation des résultats. Il est également possible de combiner plusieurs marqueurs afin d'affiner ou de confirmer les résultats d'assignation taxonomiques ou d'utiliser des marqueurs discriminants pour l'écosystème d'intérêt tels que le gyrB pour les bactéries (Poirier *et al.* 2018).

Lorsque la métagénomique ciblée est appliquée sur de l'ADN environnemental extrait directement d'un échantillon, il n'est pas possible de distinguer l'ADN provenant de cellules vivantes de celui provenant de cellules mortes (ADN pouvant être parfois très ancien et donc refléter la présence passée de certains micro-organismes). Il est néanmoins possible de traiter les échantillons préalablement à l'extraction d'ADN avec du propidium monoazide (PMA) pour bloquer l'amplification de l'ADN libre, c'est-à-dire non protégé à l'intérieur d'une cellule. Une autre alternative peut consister à travailler à partir d'ARN au lieu d'ADN, ceci nécessitant de rajouter une étape de transcription reverse dans le protocole.

La métagénomique ciblée permet donc d'avoir accès à la composition de la communauté microbienne mais ne peut en aucun cas renseigner sur le rôle fonctionnel des micro-organismes détectés. Bien que certains outils, tels que PiCRUST (phylogenetic investigation of communities by reconstruction of unobserved states ; Languille *et al.* 2013) utilisent les données génomiques disponibles pour tenter de proposer une analyse fonctionnelle des écosystèmes à partir de données metabarcoding, ce type d'analyse est loin de faire l'unanimité dans la communauté scientifique. Il convient donc de rester extrêmement prudent quant à l'extrapolation faite à partir de tels résultats.

Remerciements

Nous souhaitons remercier Emilie Chancerel et Franck Salin pour leurs précieux conseils pour améliorer ce document. Nous remercions également tous les membres du pôle d'animation « Métagénomique, identification d'espèces, phylogénie » du PEPI IBIS pour les discussions lors des réunions du pôle, et en particulier Stéphane Chaillou et Jean-Pierre Gauthier.

Références bibliographiques

Aird D, Ross MG, Chen W-S, et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* 12:R18. doi: 10.1186/gb-2011-12-2-r18

Altschul SF, Gish W, Miller W, et al. (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215(3):403–410. doi: 10.1016/S0022-2836(05)80360-2.

Angly FE, Dennis PG, Skarshewski A, et al. (2014) CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* 2:11. doi: 10.1186/2049-2618-2-11

Auer L, Mariadassou M, O'Donohue M, et al. (2017) Analysis of large 16S rRNA Illumina data sets: Impact of singleton read filtering on microbial community description. *Molecular Ecology Resources* 17(6):e122–e132. doi: 10.1111/1755-0998.12700

Bachy C, Dolan JR, López-García P, et al. (2013) Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. *ISME Journal* 7:244–255. doi: 10.1038/ismej.2012.106

Balvočiūtė M, Huson DH (2017) SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC Genomics* 18:S2 doi: 10.1186/s12864-017-3501-4

Bokulich NA, Subramanian S, Faith JJ, et al. (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods* 10:57–59. doi: 10.1038/nmeth.2276

Bonder MJ, Abeln S, Zaura E, Brandt BW (2012) Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* 28(22):2891–2897. doi: 10.1093/bioinformatics/bts552

Callahan BJ, McMurdie PJ, Holmes SP (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal 11:2639–2643. doi: 10.1038/ismej.2017.119

Caporaso JG, Kuczynski J, Stombaugh J, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7:335–336. doi: 10.1038/nmeth.f.303

Cole JR, Wang Q, Fish JA, et al (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* 42:D633–D642. doi: 10.1093/nar/gkt1244

Cruaud P, Vigneron A, Lucchetti-Miganeh C, et al. (2014) Influence of DNA extraction method, 16S rRNA targeted hypervariable regions, and sample origin on microbial diversity detected by 454 pyrosequencing in marine chemosynthetic ecosystems. *Applied and Environmental Microbiology* 80(15):4626–4639. doi: 10.1128/AEM.00592-14

D'Amore R, Ijaz UZ, Schirmer M, et al. (2016) A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* 17:55. doi: 10.1186/s12864-015-2194-9

DeSantis TZ, Hugenholtz P, Larsen N, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* 72(7):5069–5072. doi: 10.1128/AEM.03006-05

Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* 10:996–998. doi: 10.1038/nmeth.2604

Eren AM, Morrison HG, Lescault PJ, et al. (2014) Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME Journal* 9:968–979. doi: 10.1038/ismej.2014.195

Escudié F, Auer L, Bernard M, et al. (2018) FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics* 34(8):1287–1294. doi: 10.1093/bioinformatics/btx791

Glassman SI, Martiny JBH (2018) Broadscale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units. *mSphere* 3(4):e00148-18. doi: 10.1128/mSphere.00148-18

Glassing A, Dowd SE, Galandiuk S, et al. (2016) Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathogens* 8: 24. doi: 10.1186/s13099-016-0103-7

Irinyi L, Serena C, Garcia-Hermoso D, et al. (2015) International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database--the quality controlled standard tool for routine identification of human and animal pathogenic fungi. Medical Mycology 53(4):313–337. doi: 10.1093/mmy/myv008

Keegan KP, Glass EM, Meyer F (2016) *MG-RAST, a metagenomics service for analysis of microbial community structure and function, in :* Microbial Environmental Genomics (MEG). Edition Springer, New York, NY, pp 207–233.

Kim M, Oh H-S, Park S-C, Chun J (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* 64:346–351. doi: 10.1099/ijs.0.059774-0

Klindworth A, Pruesse E, Schweer T, et al. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research* 41(1):e1. doi: 10.1093/nar/gks808

Knudsen BE, Bergmark L, Munk P, et al. (2016) Impact of sample type and DNA isolation procedure on genomic inference of microbiome composition. *mSystems* 1(15):e00095-16. doi: 10.1128/mSystems.00095-16

Krehenwinkel H, Pomerantz A, Henderson JB, et al. (2019) Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. GigaScience 8:1-16. doi: 10.1093/gigascience/giz006

Kumar PS, Brooker MR, Dowd SE, Camerlengo T (2011) Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PLoS ONE* 6(6):e20956. doi: 10.1371/journal.pone.0020956

Langille MGI, Zaneveld J, Caporaso JG, et al. (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* 31:814–821. doi: 10.1038/nbt.2676

Liu Z, DeSantis TZ, Andersen GL, Knight R (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research* 36(18):e120–e120. doi: 10.1093/nar/gkn491

Lofgren LA, Uehling JK, Branco S, et al. (2019) Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles. Molecular Ecology 28:721–730. doi: 10.1111/mec.14995

Mahé F, Rognes T, Quince C, et al. (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2:e593. doi: 10.7717/peerj.593

Meola M, Rifa E, Shani et al. (2019) DAIRYdb: A manually curated gold standard reference database for improved taxonomy annotation of 16S rRNA gene sequences from dairy products. BMC Genomics 20:560 doi: 10.1186/s12864-019-5914-8

Mikaelyan A, Köhler T, Lampert N, et al. (2015) Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database (DictDb). *Systematic and Applied Microbiology* 38(7):472–482. doi: 10.1016/j.syapm.2015.07.004

Mitchell A, Bucchini F, Cochrane G, et al. (2016) EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research* 44(D1):D595–D603. doi: 10.1093/nar/gkv1195

Moyer CL, Dobbs FC, Karl DM (1994) Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16S rRNA genes from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaiit. *Applied and Environmental Microbiology* 60(3):871-879.

Munoz R, Yarza P, Ludwig W, et al. (2011) Release LTPs104 of the All-Species Living Tree. *Systematic and Applied Microbiology* 34(3):169–170. doi: 10.1016/j.syapm.2011.03.001

Mysara M, Saeys Y, Leys N, et al. (2015) CATCh, an ensemble classifier for chimera detection in 16S rRNA sequencing studies. *Applied and Environmental Microbiology* 81(5):1573–1584. doi: 10.1128/AEM.02896-14

Nelson MC, Morrison HG, Benjamino J, et al. (2014) Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS ONE* 9(4):e94249. doi: 10.1371/journal.pone.0094249

Nilsson RH, Taylor AFS, Adams RI, et al. (2018) Taxonomic annotation of public fungal ITS sequences from the built environment – a report from an April 10–11, 2017 workshop (Aberdeen, UK). MycoKeys 28:65–82. doi: 10.3897/mycokeys.28.20887

Poirier S, Rué O, Peguilhan R, et al. (2018) Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using gyrB amplicon sequencing: A comparative analysis with 16S rDNA V3-V4 amplicon sequencing. PLOS ONE 13:e0204629. doi: 10.1371/journal.pone.0204629

Quast C, Pruesse E, Yilmaz P, et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Reseach* 41(D1):D590–D596. doi: 10.1093/nar/gks1219

Salter SJ, Cox MJ, Turek EM, et al. (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* 12: doi: 10.1186/s12915-014-0087-z

Schloss PD, Westcott SL, Ryabin T, et al. (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75:7537–7541. doi: 10.1128/AEM.01541-09

Shin S-K, Kim J, Ha S, et al. (2015) Metagenomic Insights into the Bioaerosols in the Indoor and Outdoor Environments of Childcare Facilities. *PLOS ONE* 10:e0126960. doi: 10.1371/journal.pone.0126960

Smets W, Leff JW, Bradford MA, et al. (2016) A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biology Biochemisty* 96:145-151.

Stackebrandt E, Goebel BM (1994) Taxonomic Note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology* 44(4):846–849. doi: 10.1099/00207713-44-4-846

Tedersoo L, Anslan S, Bahram M, et al. (2015) Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. MycoKeys 10:1–43. doi: 10.3897/mycokeys.10.4852

Tedersoo L, Tooming-Klunderud A, Anslan S (2018) PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytologist 217:1370–1385. doi: 10.1111/nph.14776

The Microbiome Quality Control Project Consortium, Sinha R, Abu-Ali G, et al (2017) Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nature Biotechnology 35:1077–1086. doi: 10.1038/nbt.3981

Toju H, Tanabe AS, Yamamoto S, Sato H (2012) High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples. PLOS ONE 7:e40863. doi: 10.1371/journal.pone.0040863

Větrovský T, Baldrian P (2013) The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* 8:e57923. doi: 10.1371/journal.pone.0057923

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *PNAS* 87(12):4576–4579. doi: 10.1073/pnas.87.12.4576

Yoon S-H, Ha S-M, Kwon S, et al. (2017) Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. International Journal of Systematic and Evolutionary Microbiology 67(5):1613–1617. doi: 10.1099/ijsem.0.001755