Solutions d'amélioration des études de métagénomique ciblée

Léa Siegwald

Solutions d'amélioration des études de métagénomique ciblée

Thèse de Doctorat

soutenue le 23 mars 2017

en vue de l'obtention du grade de Docteur

dans la discipline « Génétique, génomique, bioinformatique »

par

Léa SIEGWALD

UNIVERSITÉ DE LILLE - ÉCOLE DOCTORALE BIOLOGIE ET SANTÉ

Composition du jury

Rapporteurs :	**Georges DAUBE**	PR, Université de Liège
	Guy PERRIÈRE	DR CNRS, Université Lyon 1
Examinateurs :	**Benoît FOLIGNÉ**	PR, Université de Lille
	Evguenia KOPYLOVA	CR, Clarity Genomics, Gand
Directeurs de thèse :	**Yves LEMOINE**	PR, Université de Lille
	Hélène TOUZET	DR CNRS, CRIStAL, Lille
Encadrants :	**Christophe AUDEBERT**	CR, Gènes Diffusion, Douai (responsable CIFRE)
	Ségolène CABOCHE	IR, Université de Lille

Plateforme PEGASE-biosciences — Institut Pasteur de Lille
1, rue du Professeur Calmette —59000 LILLE

Remerciements

Je me suis engagée dans ce projet de thèse en le considérant comme une opportunité unique d'évoluer dans mon domaine d'expertise et d'y apporter ma propre contribution. Trois années plus tard, c'est emplie d'humilité que j'achève la rédaction de ce manuscrit, consciente d'avoir tiré de cette expérience bien plus d'enseignements que ce que je pouvais en espérer. Ségolène m'indiquait un jour que l'impact d'une thèse, outre les résultats et innovations qu'elle présente, pouvait aussi s'évaluer par les différences entre le doctorant qui la débute et celui qui la soutient. Pour ma part, je dois cette évolution toute autant scientifique que personnelle à chacune des personnes suivantes, envers qui je suis profondément reconnaissante.

Tout d'abord, je tiens à remercier sincèrement Yves Lemoine et Hélène Touzet, mes directeurs de thèse, pour la confiance qu'ils ont placée en moi et leur suivi tout au long de ce projet. Un grand merci également aux membres de mon Comité de Suivi de Thèse, Georges Daube et Pierre Peterlongo, pour leur regard expert et avisé sur mon travail. Votre bienveillance à tous les quatre m'a été bénéfique à de nombreuses reprises tout au long de ces trois ans.

J'ai eu la chance d'être encadrée au quotidien par Christophe et Ségolène, dont la supervision et la disponibilité ont été sans égales ; je n'aurais pu rêver de meilleurs tuteurs de thèse. Merci à Ségo de m'avoir apporté un sens de la rigueur et un art d'aller à l'essentiel qui m'étaient fondamentalement lacunaires. Merci à Christophe d'avoir cherché à me faire sortir de mon fonctionnement d'ingénieure, en me poussant toujours à réfléchir plus loin que le bout de mon nez. Votre duo complémentaire a été la clef de voûte de ce projet. Je suis honorée d'avoir pu bénéficier de vos nombreux enseignements, et vous dois beaucoup.

Merci à l'ensemble des membres de la plate-forme PEGASE-biosciences, Pasteuriens comme GDScanners, présents comme passés. Vous m'avez permis d'accomplir cette thèse dans un environnement de travail toujours stimulant et propice à la bonne humeur. J'ai une forte pensée envers chacun d'entre vous.

Merci aussi à tous les bioinformaticiens de BONSAI de m'avoir accueillie une journée par semaine comme si j'étais membre intégrante de leur équipe, que ce soit par les nombreux échanges scientifiques tout comme les discussions des mardi midi.

J'adresse également mes remerciements aux biologistes que j'ai pu côtoyer ces trois années, et grâce à qui mes travaux prennent tout leur sens. Merci notamment à David pour ses conseils avisés et réponses à mes nombreuses questions, à Émilie pour la première application qu'elle a apportée à Harpon, et à Magali pour sa gentillesse et son expertise en parasitologie.

Par ailleurs, je pense fort aux personnes qui ont contribué, il y a plus ou moins longtemps, à m'orienter vers le chemin que je parcours aujourd'hui. Je tiens à mentionner Nicolas pour ces quelques minutes de discussion après un cours, qui m'ont fait envisager pour la première fois la possibilité d'un doctorat. Un clin d'œil à Christine pour m'avoir ouvert les portes de Lille et de la métagénomique. Je dois également beaucoup à Marie-Christine, pour son écoute et ses conseils durant les moments les plus tumultueux.

Merci énormément à mes amis pour leurs encouragements, en particulier dans la dernière ligne droite. Vos nombreux messages m'ont été un carburant précieux, j'ai hâte de vous retrouver et de pouvoir vous rendre la pareille.

J'adresse tout mon amour à ma famille pour leur indéfectible soutien et pour leur sollicitude malgré la distance. Je dédie tout particulièrement cette thèse à mes parents, qui m'ont toujours poussée à aller plus loin et à tenir bon. Cette ténacité, c'est à vous que je la dois. *Ich han's gepackt !*

Merci enfin à K, de tenir la barre avec moi par beau temps comme par tempête, et de m'être une ancre lorsque je vais à la dérive. Je ne voudrais d'aucun autre paysage que l'horizon qui se dessine devant nous.

Résumé

Solutions d'amélioration des études de métagénomique ciblée

La métagénomique ciblée, étude de la composition et de la diversité des communautés microbiennes présentes dans différents échantillons biologiques sur la base d'un marqueur génomique, a connu un véritable essor lors de cette dernière décennie grâce à l'arrivée du séquençage haut-débit. Faisant appel à des outils de biologie moléculaire et de bioinformatique, elle a été à l'origine de substantiels progrès dans les domaines de l'étude de l'évolution et de la diversité microbienne. Cependant, de nouvelles problématiques sont apparues avec le séquençage haut-débit : la génération exponentielle de données soulève des problèmes d'analyse bioinformatique, qui doit être adaptée aux plans d'expérience et aux questions biologiques associées.

Cette thèse propose des solutions d'amélioration des études de métagénomique ciblée par le développement d'outils et de méthodes innovantes, apportant une meilleure compréhension des biais d'analyse inhérents à de telles études, et une meilleure conception des plans d'expérience.

Tout d'abord, une expertise du pipeline d'analyse utilisé en production sur la plate-forme PEGASE-biosciences a été menée. Cette étude a révélé la nécessité de mettre en place un protocole d'évaluation formelle de pipelines d'analyses de données de métagénomique ciblée. Cette méthode a été développée sur la base de données simulées et réelles, et de métriques d'évaluation adaptées. Cette méthode a été utilisée sur plusieurs pipelines d'analyse couramment utilisés par la communauté, tout comme sur de nouvelles approches d'analyse jamais utilisées dans un tel contexte. Cette évaluation a permis de mieux comprendre les biais du plan d'expérience qui peuvent affecter les résultats et les conclusions biologiques associées. Un de ces biais majeurs est le choix des amorces d'amplification de la cible ; un logiciel de design d'amorces adaptées au plan d'expérience a été spécifiquement développé pour minimiser ce biais. Enfin, des recommandations de montage de plan d'expérience et d'analyse ont été émises afin d'améliorer la robustesse des études de métagénomique ciblée.

Mots-clefs : métagénomique, métagénétique, microbiote, séquençage haut-débit, amorces, bioinformatique.

Abstract

Solutions to improve targeted metagenomics studies

Targeted metagenomics is the study of the composition of microbial communities in diverse biological samples, based on the sequencing of a genomic locus. This application has boomed over the last decade thanks to the democratisation of high-throughput sequencing, and has allowed substantial progress in the study of microbial evolution and diversity. However, new problems have emerged with high-throughput sequencing: the exponential generation of data must be properly analyzed with bioinformatics tools fitted to the experimental designs and associated biological questions.

This dissertation provides solutions to improve targeted metagenomics studies, by the development of new tools and methods allowing a better understanding of analytical biases, and a better design of experiments.

Firstly, an expert assessment of the analytical pipeline used on the PEGASE-biosciences plateform has been performed. This assessment revealed the need of a formal evaluation protocol of analytical pipelines used for targeted metagenomics analyses. This method has been developed with simulated and real datasets and adequate evaluation metrics. It has been used on several analytical pipelines commonly used by the scientific community, as well as on new analytical methods which have never been used in such a context before. This evaluation allowed to better understand experimental design biases, which can affect the results and biological conclusions. One of those major biases is the design of amplification primers to target the genomic locus of interest. A primer design software, adaptable to different experimental designs, has been specifically developed to minimize this bias. Finally, analytical guidelines and experimental design recommendations have been formulated to improve targeted metagenomics studies.

Keywords: metagenomics, metagenetics, microbiota, high-throughput sequencing, primers, bioinformatics.

Table des matières

Remerciements.....3

Résumé.....5

Abstract.....6

Contexte de la thèse.....9

Introduction.....13
Le besoin d'observer ce qui est invisible à l'œil nu.....14
L'utilisation de l'ADN pour identifier, comparer et classer de nouveaux microbes.....16
Les débuts de la métagénomique.....17
La généralisation du concept par le séquençage haut-débit.....18
Une application qui sort des laboratoires de recherche.....21
Les limites de la métagénétique.....22

Chapitre 1 - Des échantillons à la description des microbiotes.....27
1.1 - Déroulement global d'une étude métagénétique.....27
1.2 - Technologies de séquençage haut-débit de seconde génération.....28
1.3 - De l'échantillonnage aux données séquencées.....36
1.4 - Analyse primaire des données séquencées.....51
1.5 - Normalisation des tables de comptages.....71
1.6 - Analyse secondaire.....75

Chapitre 2 - Expertise du pipeline d'analyse métagénétique développé sur la plate-forme PEGASE-biosciences.....89
2.1 - Évaluation du pipeline PEGASE v1 sur une communauté microbienne artificielle générée *in vitro*.....89
2.2 - Évaluation du pipeline PEGASE v1 sur une communauté microbienne artificielle générée *in silico*.....97
2.3 - Améliorations du pipeline PEGASE.....101

Chapitre 3 - Évaluation formelle de pipelines d'analyse de données métagénétiques.....105
3.1 - Mise en place d'un protocole d'évaluation de pipelines et contexte d'utilisation.....106
3.2 - Évaluation des pipelines selon différentes variables du plan d'expérience 122
3.3 - Bilan de l'étude et perspectives.....152

Chapitre 4 - Impact de la variation de pipeline d'analyse dans les conclusions d'une étude métagénétique du microbiote intestinal humain.....157
4.1 - Contexte biologique.....157
4.2 - Description des données de séquençage et analyses associées.....159
4.3 - Évaluation de l'impact du changement de pipeline sur l'interprétation biologique des résultats.....168
4.4 - Conclusion.....184

Chapitre 5 - Harpon : Design d*e novo d*'amorces dégénérées à façon selon un microbiote d'intérêt.....187
5.1 - Problématique du design d'amorces dans un contexte métagénétique. 187
5.2 - Méthodes intégrées à Harpon.....194
5.3 - Validation de Harpon sur différents contextes métagénétiques.....204
5.4 - Conclusion et perspectives.....219

Chapitre 6 - Recommandations d'analyse de données métagénétiques issues d'un séquençage de librairies bidirectionnelles Ion Torrent PGM.....223
6.1 - Première évaluation des lectures.....225
6.2 - Pré-traitement des lectures par QIIME.....228
6.3 - Pré-analyse *assignment-first*.....230
6.4 - Élimination des lectures contaminantes (optionnel).....232
6.5 - Élimination des échantillons aberrants (optionnel).....233
6.6 - Exécution de l'analyse QIIME.....234
6.7 - Normalisation des données.....236
6.8 - Conclusion.....236

Conclusions & perspectives.....239

Glossaire.....244

Bibliographie.....246

Annexe 1 : Principe du séquençage Ion Torrent PGM.....261

Annexe 2 : Lignes de commande exécutées pour chaque pipeline. 269
A2.1 - BMP.....269
A2.2 - mothur.....271
A2.3 - QIIME SortMeRNA + Sumaclust.....274

Annexe 3 : Identification des espèces non référencées dans les banques de séquences d'ADNr.....279

Annexe 4 : Espèces de champignons d'intérêt.....281

Contexte de la thèse

Ce projet de thèse CIFRE a été financé par la bourse n°2013/0920 de l'Association Nationale de la Recherche et de la Technologie, ainsi que par Gènes Diffusion SAS. Il s'inscrit dans le cadre d'une collaboration entre Gènes Diffusion, l'équipe de Transcriptomique & Génomique Appliquées (TAG) du Centre d'Infection et d'Immunité de Lille (CIIL) de l'Institut Pasteur de Lille, l'équipe BONSAI affiliée au Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL, UMR CNRS 9189, Université de Lille) et l'Institut National de Recherche en Informatique et en Automatique (INRIA).

Depuis 50 ans, Gènes Diffusion (http://www.genesdiffusion.com) développe son activité en génétique et reproduction animale. Travaillant sur les espèces bovine, porcine, équine et lapine, Gènes Diffusion est actuellement l'un des seuls groupes génétiques, au niveau mondial, spécialisé sur plusieurs marchés. En 2009, Gènes Diffusion a lancé la sélection génomique sur le marché de l'élevage bovin et porcin en France. Ce programme de sélection permet de prédire le potentiel d'un animal sur plusieurs caractères d'intérêt agro-économique ou de santé (par exemple chez les vaches : production laitière, résistance aux maladies des pieds, comportement vis-à-vis du veau, ...) à partir de son empreinte génomique (génotypage haut-débit de plusieurs dizaines de milliers de marqueurs de type SNP).

Ce développement a été accompagné de l'établissement de la plate-forme génomique GDScan en 2003, sur le site de l'Institut Pasteur de Lille. Cette équipe est constituée de quatre personnes à temps plein : un chargé de recherche génomique, un bioinformaticien, une technicienne supérieure spécialisée en biologie moléculaire et un généticien spécialisé en génétique quantitative et biostatistiques. Elle est associée à la plate-forme TAG (*Transcriptomics and Applied Genomics*) de l'Institut Pasteur de Lille (équipe constituée de cinq personnes à temps plein : un chargé de recherche, deux bio-informaticiennes, une technicienne supérieure ainsi qu'une ingénieure spécialisée en biologie moléculaire). Cette plate-forme propose des services et collaborations à la communauté scientifique en développant des solutions et

applications de génomique, par puces à ADN et séquençage haut-débit.

Les deux équipes GDScan et TAG proposent de nouveaux axes de recherches en lien avec l'exploitation des données biologiques à haut-débit sous l'appellation PEGASE-biosciences (Plate-forme d'Expertises Génomiques Appliquées aux Sciences Expérimentales, http://www.pegase-biosciences.com). La plate-forme est équipée notamment d'un séquenceur haut-débit de paillasse (technologie PGM, Ion Torrent, ThermoFisher Scientific). La métagénomique est devenue un axe primordial de développement au sein de PEGASE-biosciences. En effet, plus de 50 % des expertises menées sur la plate-forme concernent cette application du séquençage haut-débit, à l'origine de nombreuses collaborations menées par l'équipe TAG. En outre, la métagénomique est un nouveau sujet d'investigation de Gènes Diffusion, souhaitant évaluer l'intérêt du microbiote intestinal bovin comme nouveau critère à inclure dans ses modèles prédictifs (modèle de sélection génomique permettant de prédire le potentiel génétique d'un reproducteur). Ainsi, la plate-forme PEGASE-biosciences a la nécessité de développer un outil d'analyse métagénomique robuste venant renforcer son corpus d'expertises.

Cette thèse s'est également inscrite dans le cadre d'une collaboration avec l'équipe de recherche BONSAI, affiliée au Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL) et à l'INRIA Lille. Cette équipe de recherche en bioinformatique, composée d'une vingtaine de personnes, mène des recherches méthodologiques et des développements d'algorithmes et de modèles adaptés au traitement des données issues de séquençage de deuxième et troisième génération. Elle possède une expertise dans l'annotation des génomes, la phylogénie, et le traitement de données métagénomiques *shotgun*. Cette expertise algorithmique a ainsi pu être appliquée aux données de métagénomique ciblée, au cœur du sujet de thèse.

Ce projet, débuté le 3 mars 2014, a été mené sous la co-direction d'Yves Lemoine, professeur des universités (Université de Lille) et ancien chef d'équipe de la plate-forme TAG de l'Institut Pasteur de Lille, ainsi que d'Hélène Touzet, directrice de recherche CNRS au Centre de Recherche en Informatique, Signal

et Automatique de Lille (CRIStAL). L'activité au laboratoire a été suivie par Ségolène Caboche (Ingénieur de recherche de l'Université de Lille), et la tutelle entreprise a été accomplie par Christophe Audebert (Chargé de recherches Gènes Diffusion). Le temps de travail a été partagé entre 80 % sur la plate-forme PEGASE-biosciences, et 20 % dans l'équipe BONSAI, afin d'assurer le transfert d'expertise entre les recherches fondamentales sur le traitement de données métagénomiques et leur application sur la plate-forme.

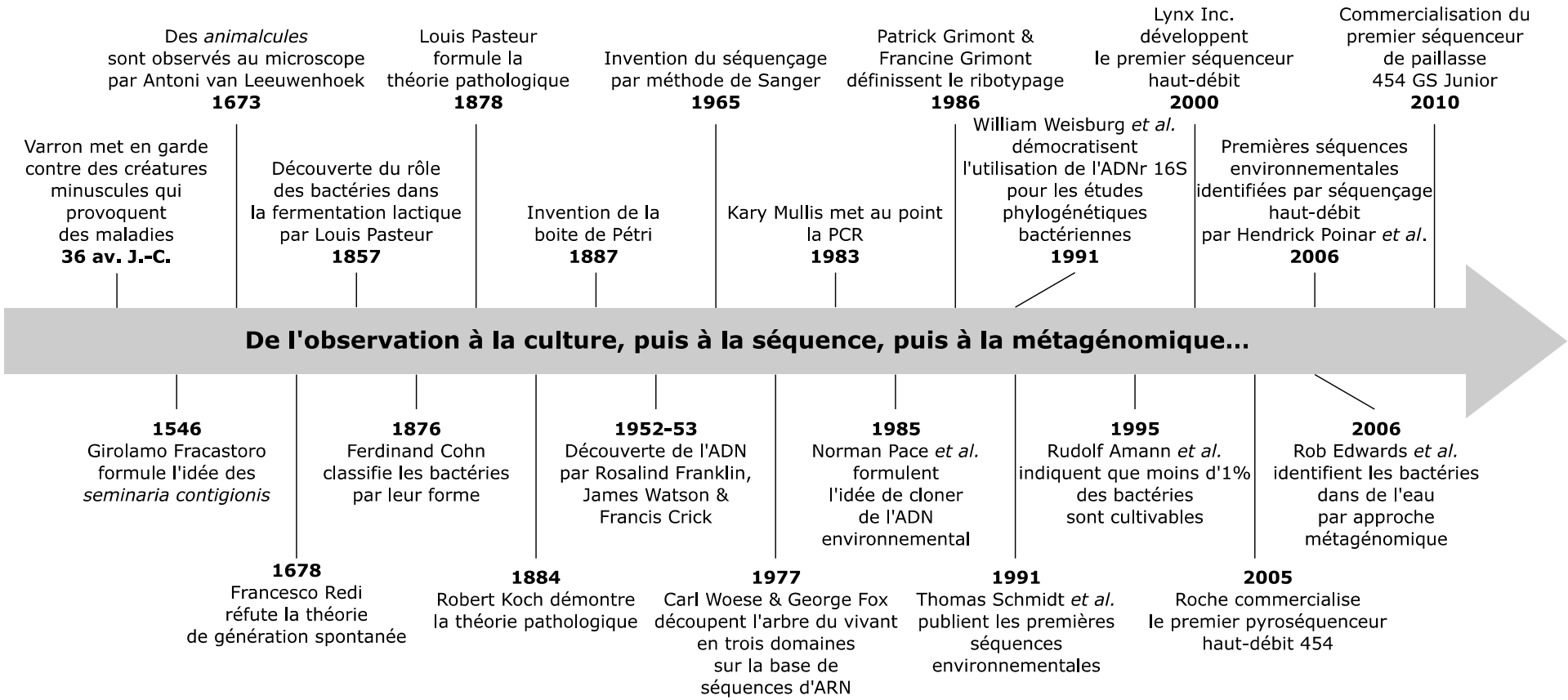


Figure 1 : Frise chronologique des dates importantes de la microbiologie.

Introduction

« J'étais très surpris par ce spectacle si merveilleux, n'ayant jamais vu aucune créature vivante comparable en petitesse à celles-ci. Je ne pouvais imaginer que la nature ait créé des animaux dans des proportions si excessivement minuscules. »

Robert Hooke, suite aux premières observations de micro-organismes au microscope par Antoni van Leeuwenhoek en 1678. [Hooke, 1678]

Depuis l'observation d'un monde vivant invisible à l'œil nu, le développement de la microbiologie a permis ces 350 dernières années de faire évoluer le stéréotype du microbe d'un statut d'ennemi parfois mortel à celui d'un allié indispensable à notre santé. L'étude des micro-organismes et de leurs relations à leur hôte humain a grandement accru le corpus de connaissances en santé humaine. Par exemple, certaines bactéries symbiotiques de notre intestin ont un rôle majeur dans le maintien de la barrière immunitaire innée propre à cet organe [Thaiss *et al.* 2016].

À partir des découvertes de Louis Pasteur sur le rôle des micro-organismes en tant qu'ingrédients actifs de recettes séculaires (principalement par l'action de la fermentation), le champ d'investigation de la microbiologie s'est élargi et a permis leur exploitation à des fins de production industrielle. Ils tiennent ainsi une place prépondérante dans l'industrie alimentaire bien sûr, que ce soit dans l'utilisation des bactéries lactiques dans la production de produits laitiers ou l'utilisation de levures dans les processus de fermentation alcoolique, mais aussi dans l'industrie pharmaceutique, par exemple pour la production d'antibiotiques, de probiotiques, ou d'hormones comme l'insuline. L'exploitation des micro-organismes, considérés comme un système-usine à part entière, émerge également dans le secteur des énergies renouvelables (production de bioéthanol), et prend une place prépondérante dans des stratégies de bioremédiation (utilisation de micro-organismes pour neutraliser ou supprimer des polluants).

Leur étude est ainsi essentielle non seulement pour appréhender leur impact, positif et négatif, sur notre santé et notre environnement, mais également pour développer de nouvelles solutions biotechnologiques basées sur leur exploitation.

Le besoin d'observer ce qui est invisible à l'œil nu

L'idée d'entités invisibles responsables de maladies épidémiques avait déjà été formulée en 36 av. J.-C. par Varron dans son traité d'agriculture, mentionnant qu'il était préférable de ne pas développer une ferme proche de marécages, *« parce qu'il s'y développe certaines créatures minuscules que les yeux ne peuvent voir, qui flottent dans l'air et entrent dans le corps par la bouche et le nez, et causent de graves maladies. »* [Cato & Varro, 36 av. J.-C.]. L'existence de ces créatures a également été supposée par Girolamo Fracastoro en 1546, qui les avait alors baptisées *« seminaria contigionis* ». La confirmation de leur existence n'a toutefois pu être possible qu'au XVII^{e} siècle grâce à l'invention du microscope, permettant de franchir la barrière d'un monde invisible. La première description formelle de l'existence de micro-organismes a été rapportée à la Royal Society de Londres par Antoni van Leeuwenhoek dès 1673. Ce drapier avait fabriqué un microscope lui-même, permettant d'agrandir jusque 300 fois, dans le but de vérifier la pureté des étoffes qu'il vendait. En y observant des échantillons d'eau, il a décrit ce qu'il appelait des « *animalcules* », et a détaillé les premières représentations précises de bactéries, champignons et protozoaires, inaugurant ainsi le domaine de la microbiologie.

Ces observations n'étaient à l'époque que d'un intérêt descriptif, les micro-organismes étant considérés comme issus de génération spontanée. Cette théorie a été remise en question pour la première fois par Francesco Redi en 1668, démontrant que des larves n'émergeaient pas spontanément de viande avariée si cette dernière était couverte pour empêcher la ponte des mouches.

L'expérimentation pour cultiver et isoler afin de mieux pouvoir étudier

Louis Pasteur a mené de nombreuses expériences vers la fin du XIX^{e} siècle pour discréditer définitivement la théorie de génération spontanée. Il a en outre associé pour la première fois en 1857 les bactéries à la fermentation lactique, justifiant qu'une contamination bactérienne pouvait faire tourner le vin. Pasteur a également formulé la théorie pathogénique en 1878, indiquant que les micro-organismes sont à l'origine des maladies infectieuses (confirmant ainsi l'hypothèse de Girolamo Fracastoro), en mettant en parallèle l'idée que si les bactéries peuvent altérer le vin, elles peuvent peut-être affecter l'humain de la même manière. Robert Koch a réussi à démontrer cette théorie en 1884, montrant que des souris inoculées par des souches cultivées de *Bacillus antracis* sont systématiquement atteintes d'anthrax.

Cette découverte a engendré ce qu'on appelle l'âge d'or de la microbiologie, durant lequel de nombreux pathogènes ont été identifiés et liés à différentes maladies infectieuses. Julius Petri a permis, sur la base des travaux de Koch, l'isolement et la culture de colonies bactériennes par l'invention de la boîte éponyme en 1887. Le développement des méthodes culturales a été précieux pour l'étude de ces micro-organismes de façon isolée, afin de pouvoir décrire leurs caractéristiques phénotypiques.

En parallèle, l'amélioration de la résolution des microscopes a permis à Ferdinand Cohn de classifier les premières cellules bactériennes par forme (sphérique, allongée, en bâtonnet, ou en spirale) en 1876, posant les fondations de la classification des micro-organismes par caractéristiques phénotypiques, qui a été la méthode de classification de référence jusqu'à la moitié du XX^{e} siècle.

L'utilisation de l'ADN pour identifier, comparer et classer de nouveaux microbes

La découverte de l'ADN comme support de l'information génétique (par Franklin & Gosling en 1952 et Watson & Crick et 1953 [Franklin & Gosling 1952, Watson & Crick 1953]) et son séquençage permis par Frederick Sanger vingt ans plus tard [Sanger 1977] ont révolutionné l'étude du vivant en la basculant d'observations phénotypiques vers des observations génotypiques. La théorie d'horloge moléculaire (définie par Zuckerkandl et Pauling en 1965 [Zuckerkandl & Pauling 1965]) a été pour la première fois appliquée à l'étude des relations évolutives entre les organismes par Woese & Fox, après 10 ans de lectures de séquences de la petite sous-unité d'ARN ribosomique sur gels d'électrophorèse. En effet, cet ARN avait été identifié par Woese comme étant indispensable à la machinerie cellulaire de tout être vivant, tout en étant assez long pour pouvoir observer des différences de séquences entre différents organismes. Sur la base de ces séquences, Woese & Fox ont séparé l'arbre du vivant en trois domaines majeurs en 1977, proposant pour la première fois une étude de la phylogénie des organismes sur la base d'une séquence d'ARN ubiquitaire [Woese & Fox 1977].

Cette approche a été étendue à une cible génomique par Grimont & Grimont, qui ont défini en 1986 le ribotypage, comparaison de profils de restriction de l'ADN ribosomique 16S (abrégé ADNr 16S), comme méthode de classification taxonomique des bactéries [Grimont & Grimont 1986]. L'automatisation du séquençage Sanger (1986) croisée au développement des techniques de réaction en chaîne par polymérase (PCR, envisagée par Kjell Kleppe en 1971 et mise au point par Kary Mullis en 1983) ont permis à Weisburg *et al.* (1991) de standardiser cette approche en utilisant la séquence d'ADNr 16S comme marqueur phylogénétique bactérien, démarche qui s'est imposée comme standard dans l'étude phylogénétique bactérienne [Weisburg *et al.* 1991]. Cette approche a engendré une augmentation drastique de séquençages d'ADNr 16S chez la plupart des bactéries cultivables.

Toutefois, de nombreuses divergences étaient observées entre les micro-organismes révélés par culture, et une observation directe par microscope d'une communauté microbienne complexe issue d'un échantillon environnemental [Staley & Konopka, 1985]. Cette révélation a fait émerger l'idée que certains micro-organismes ne pouvaient se développer que dans certaines conditions, associant leur environnement à leur fonctionnement. Begon *et al.* ont ainsi défini en 1986 une communauté microbienne comme un ensemble de micro-organismes coexistant à un instant donné dans un même environnement [Begon *et al.* 1986]. Il a été estimé que les méthodes culturales classiques rendent compte de moins d'1 % de la diversité bactérienne de la plupart des échantillons environnementaux [Amann *et al.* 1995], ce qui laisse une immense quantité de « matière noire » [Filée *et al.* 2005] inaccessible à l'étude.

Les débuts de la métagénomique

Pace *et al.* ont proposé en 1985 l'idée de cloner de l'ADN directement extrait d'échantillons environnementaux afin de pouvoir accéder aux organismes non-cultivables [Pace *et al.* 1985]. Leur technique impliquait l'extraction de l'ADN total présent dans un échantillon, le clonage de cet ADN dans des organismes cultivables, et le séquençage des inserts pour rechercher de nouveaux gènes. Cette idée a été mise en application pour la première fois en 1991 sur une communauté bactérienne de picoplancton marin [Schmidt *et al.* 1991].

Le terme « métagénomique » a été utilisé pour la première fois par Jo Handelsman *et al.* en 1998, le définissant comme la *« collecte de tous les génomes des membres d'une communauté microbienne à partir d'un certain environnement »*, et appliquant cette approche de clonage à des microbiotes du sol [Handelsman *et al.* 1998]. Le microbiote, défini par Joshua Lederberg en 2001, est « *la communauté écologique des micro-organismes commensaux, symbiotiques et pathogènes qui se partagent littéralement notre corps »* [Lederberg 2001]. Cette définition a ensuite été étendue aux communautés présentes dans un environnement donné (le microbiome), à un temps donné.

La généralisation du concept par le séquençage haut-débit

La société Lynx Therapeutics a proposé en 2000 la première technologie de séquençage haut-débit, le MPSS *(massively parallel signature sequencing)* basé sur du séquençage par ligation, bien qu'elle n'ait jamais été commercialisée. L'arrivée du séquençage haut-débit par synthèse vers 2005 a été permise par la découverte du pyroséquençage et la commercialisation de pyroséquenceurs haut-débit par la société 454. Cette technologie a provoqué une nouvelle révolution dans l'étude des micro-organismes, permettant de séquencer de nombreux fragments d'ADN physiquement isolés de façon parallèle, et donc une augmentation drastique du volume de séquences générées. Ce principe s'est ainsi substitué à la fastidieuse nécessité de multi-cloner un échantillon, et permet également de s'affranchir d'une obligation de culture.

Cette adaptation technologique a provoqué une augmentation considérable des études métagénomiques, en permettant le séquençage direct d'une population hétérogène de différents génomes issus d'un même échantillon. Poinar *et al.* ont ainsi publié, en 2006, les premières séquences environnementales issues d'un séquençage métagénomique haut-débit d'échantillons de mâchoires de mammouth préservé dans de la glace [Poinar *et al.* 2006]. La même année, Edwards *et al.* ont séquencé par pyroséquençage 454 le microbiote d'échantillons d'eau prélevés au fond de mines de fer [Edwards *et al.* 2006]. Dans cette dernière étude, les auteurs ont identifié les séquences d'ADNr 16S présentes dans ces données en les comparant à une banque de séquences d'ADNr 16S de référence, afin de pouvoir les identifier.

La métagénomique telle qu'on la connaît aujourd'hui rassemble en réalité deux méthodes bien distinctes, nommées métagénomique WGS (*whole genome shotgun*), et la métagénomique ciblée. La métagénomique WGS consiste au séquençage de l'ensemble de l'ADN contenu dans un échantillon donné, générant ainsi une importante quantité de données mélangeant des

fragments génomiques de tous les organismes en présence. Une panacée serait de pouvoir reconstruire le génome individuel de chaque organisme à partir d'un tel mélange, mais la complexité de telles données rend inapplicables les algorithmes d'assemblages de génomes actuels, et difficile le développement de nouveaux algorithmes d'assemblage. Ces derniers doivent en effet être capables de gérer un mélange de fragments de génomes bactériens, eucaryotes et/ou viraux, dans des proportions variables et comprenant de nombreuses régions répétées, similaires, et/ou de faible complexité [Ghurye *et al.* 2016]. Sans être assemblés, les génomes séquencés peuvent tout de même être identifiés, en y recherchant des signatures propres à différents taxons afin de pouvoir les étiqueter.

Actuellement, l'objectif principal de la métagénomique WGS n'est toutefois pas d'identifier les organismes en présence, mais plutôt d'identifier les gènes qu'ils expriment potentiellement. La métagénomique WGS est ainsi principalement utilisée pour reconstruire des voies métaboliques présentes dans un microbiote, en recherchant et/ou prédisant des séquences de gènes et leurs fonctions associées parmi les données séquencées. Cette approche ne permet toutefois pas de sélectionner les gènes effectivement exprimés, ne permettant pas de ce fait l'étude différentielle d'expression de gènes sur un même microbiote soumis à des conditions environnementales différentes. La métatranscriptomique pourrait répondre à cette problématique, en séquençant non pas l'ensemble des génomes en présence, mais l'ensemble des transcrits exprimés dans un microbiote. Néanmoins, cette approche est restreinte par de nombreuses limitations techniques : par exemple, l'isolement d'ARN messager à partir d'une matrice biologique et son court temps de demi-vie rendent sa manipulation très complexe [Bashiardes *et al.* 2016].

Une autre approche moins complexe permettant d'obtenir une identification des organismes en présence dans un microbiome est de se concentrer non pas sur leurs génomes complets, mais sur un *locus* d'intérêt bien précis. Cette approche est appelée métagénomique ciblée, métagénomique amplicon ou métagénétique. Nous favorisons l'emploi de ce dernier terme

[Esposito *et al.* 2014], que nous utiliserons dans la suite de ce manuscrit. La métagénétique consiste en l'amplification pré-séquençage d'un *locus* génomique considéré comme étant ubiquitaire chez l'ensemble des organismes d'intérêt du microbiote étudié (afin de pouvoir être amplifié par un même couple d'amorces pour tout le microbiote), et comportant des régions variables d'un organisme à un autre (comme l'ADNr 16S bactérien par exemple). L'interprétation bioinformatique de ces régions variables après séquençage sur la base de leurs différences de séquences entre organismes vise à identifier la nature et la diversité des organismes en présence.

Cette approche d'analyse moins complexe (sans pour autant être triviale) est rapidement devenue une méthode de référence de profilage d'un microbiote, comme le montre la volumétrie bibliographique associée représentée dans la Figure 2. Sa démocratisation a été accélérée par l'arrivée sur le marché de séquenceurs haut-débit de paillasse (en 2010 par le pyroséquenceur 454 GS Junior de Roche), permettant de mener des études métagénétiques à plus petite échelle et à prix réduit, ce qui les rend accessibles aux hôpitaux et industries.

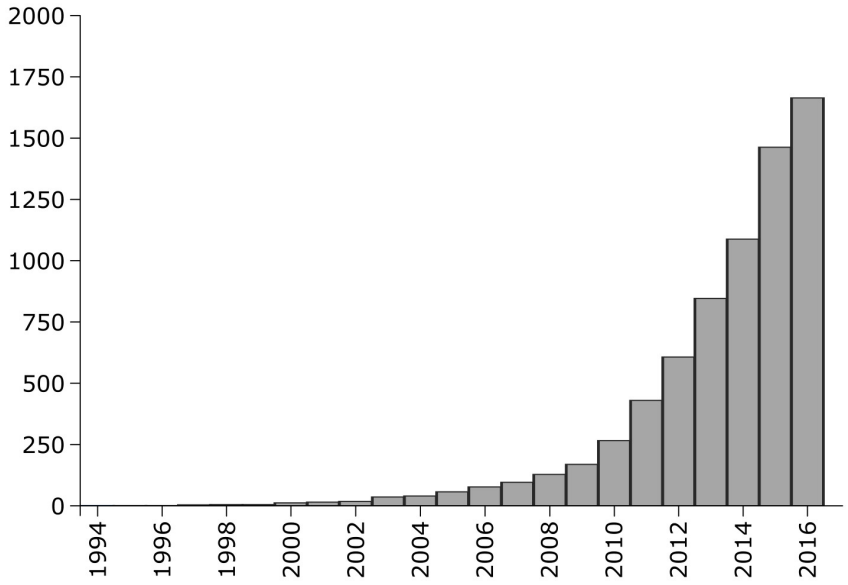


Figure 2 :Nombre de publications mentionnant la métagénétique entre 2004 et 2016 (requête PubMed : ((16S or 18S or ITS or targeted or amplicon) (metagenomics or metagenetics or microbiome or microbiota))).

Une application qui sort des laboratoires de recherche

La démocratisation récente de la métagénétique a engendré ces dernières années un transfert technologique de ces études vers des applications industrielles, notamment alimentaires. En effet, la métagénétique propose une méthode rapide de profilage de différents micro-organismes dans un produit, permettant la détection de profils aberrants pouvant être associés à des risques d'altération ou de dangerosité du produit. Par exemple, Benson *et al.* ont décrit en 2014 le changement de flore bactérienne d'un produit de viande réfrigéré sur plusieurs jours, y observant des vagues successives de différentes espèces bactériennes co-dominantes par approche métagénomique [Benson *et al.* 2014]. Une telle approche par métagénomique peut aujourd'hui être envisagée pour étudier et contrôler différentes méthodes de conservation des viandes [Bozec *et al.* 2016].

En 2015, IBM Research s'est associé à Mars Inc. afin de créer le Consortium pour le séquençage de la chaîne d'approvisionnement agroalimentaire (*Consortium for sequencing the food supply chain*), dont un des projets majeurs consiste en l'application de la métagénomique permettant la caractérisation des microbiotes bactérien, fongique et viral de différents environnements de la chaîne de production de Mars [IBM & Mars 2015]. En 2015 toujours, le projet européen MicroWine a été initié à hauteur de presque 4 millions d'euros, entre 13 acteurs mondiaux académiques et industriels, afin d'étudier par approches métagénomiques les nombreux microbiotes associés à l'industrie viticole (par études de microbiotes issus des sols, de la vigne, des différents stades de fermentation du vin, … [Microwine 2015])

Autre exemple d'application industrielle, un profilage métagénétique d'échantillons de lait a permis d'évaluer l'impact d'un antibiotique comme traitement contre les mammites chez les vaches [Ganda *et al.* 2016]. L'utilisation de la métagénétique peut ainsi permettre d'optimiser une stratégie de traitement d'animaux d'élevage. Plus généralement, le profilage du

microbiote intestinal de ces animaux peut être directement associé à la santé de ces derniers, comme chez le poulet, le porc ou la vache [Deusch *et al.* 2015]. L'association de ce profil à des variables environnementales pourrait ainsi permettre une optimisation des conditions d'élevage.

Les limites de la métagénétique

La métagénétique par séquençage haut-débit peut être considérée comme étant une discipline très jeune, ayant émergé il y a un peu plus de dix ans seulement. Elle possède ainsi de nombreux verrous parfois difficiles à appréhender. Une de ces premières limites est constituée par le fait que cette approche apparaît comme étant purement descriptive (bien qu'on puisse l'extrapoler vers de l'inférence fonctionnelle en supposant la nature des gènes potentiellement exprimés à partir de la description des taxons en présence). Ces descriptions ne sont en effet pas suffisantes pour comprendre le fonctionnement et les variations des microbiotes étudiés. Par exemple, un profil de flore peut être un élément facilitant l'interprétation de données cliniques, mais ne peut pas à lui seul expliquer l'étiologie d'une maladie, ne sachant si ce profil est une cause ou une conséquence de la condition évaluée.

En outre, la vulgarisation de la métagénétique est freinée par l'absence d'une méthode standard de référence avec laquelle la valider. Dans le cas d'une étude transcriptomique, il est en effet possible d'utiliser des méthodes de qRT-PCR comme validation de la variation d'une cible ARN entre deux conditions. Une approche similaire en métagénétique n'est toutefois pas standardisée : une validation de présence/absence ou variation de proportions d'un groupe taxonomique par qPCR est bien plus complexe et onéreuse à mettre en place (biais de choix d'amorces, nécessité d'un ADN normalisateur exogène, …)

Un autre biais majeur de l'approche métagénétique est la sélection d'un *locus* cible suffisamment discriminant entre les organismes d'intérêt, fortement limitée par la courte taille des lectures de séquençage (< 500 nt). Au vu de cette quantité d'information restreinte, la métagénétique ne permet ainsi pas d'obtenir une image détaillée du microbiote d'intérêt, l'information contenue

dans les lectures n'étant pas souvent suffisante pour discriminer les espèces entre elles. En outre, le design d'amorces permettant l'amplification du *locus* d'intérêt tout en étant les plus universelles possibles est une problématique fondamentale : si ces amorces ne s'hybrident pas sur le génome d'un organisme dans le milieu étudié, ce dernier ne pourra pas être détecté.

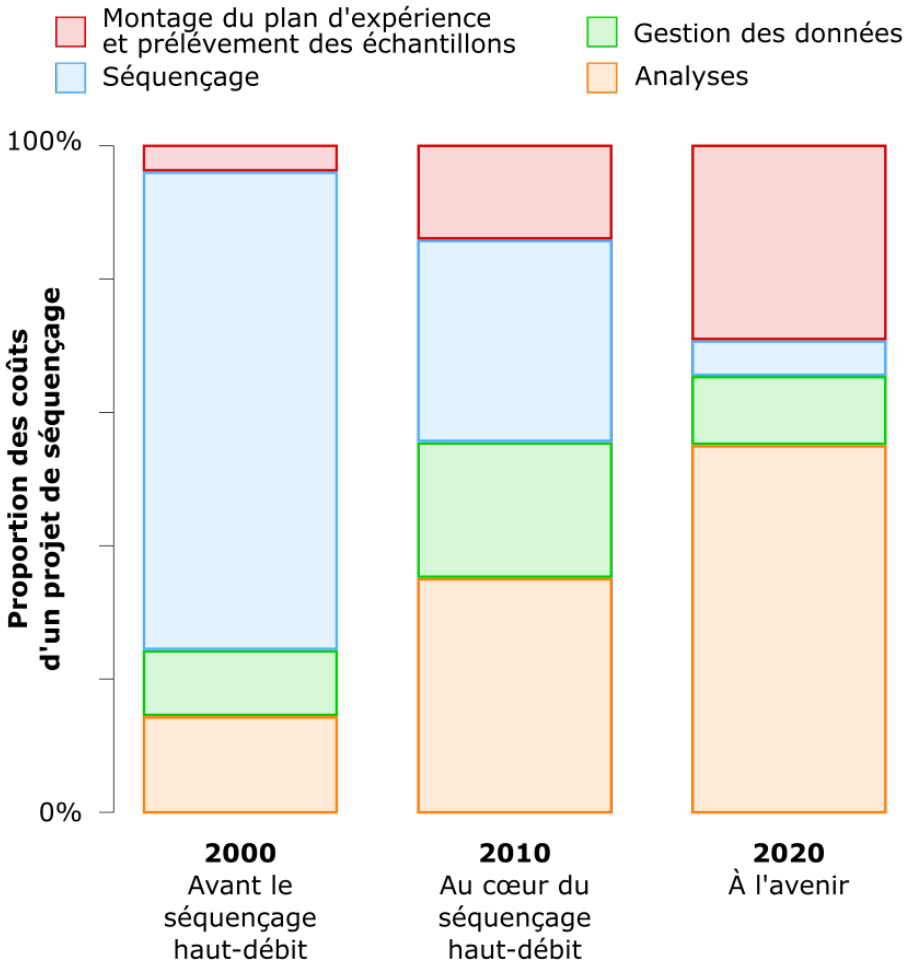


Figure 3 : Proportion des coûts d'un projet de séquençage (adapté de Sboner et al. 2011).

Le développement de nombreuses applications associées à la métagénétique est lié à l'accessibilité des technologies de séquençage, rendue possible par la diminution des coûts et associée à l'augmentation des débits. Cette démocratisation a toutefois engendré un déplacement du goulot d'étranglement de la génération de ces données vers leur analyse. Ces données

doivent en effet être convenablement interprétées afin de révéler les informations biologiques d'intérêt qu'elles contiennent. La majeure partie du coût global d'un projet de séquençage glisse ainsi du séquençage en lui-même vers les analyses des données générées, comme montré dans la Figure 3.

Malgré l'existence de nombreux logiciels d'analyse, il n'existe actuellement aucun consensus sur l'approche analytique à mener dans un contexte métagénétique. Certains de ces logiciels sont de source ouverte, et proposent un enchaînement d'étapes d'analyse sous la forme d'un pipeline nécessitant de bonnes connaissances informatiques pour être correctement paramétré et exécuté. D'autres logiciels commerciaux misent sur la convivialité de leur outil pour attirer un public moins expert, fournissant une solution clef-en-main sans possibilité de l'adapter à certaines spécificités des plans d'expérience. Il est difficile pour un bioanalyste de s'orienter vers une solution adaptée à ses besoins et qu'il pourra considérer comme étant fiable. Par exemple, la plate-forme PEGASE-biosciences est équipé d'un séquenceur de paillasse Ion Torrent PGM, pour lequel aucune solution analytique n'était validée. Il est pourtant essentiel de pouvoir comprendre comment les pipelines d'analyse réagissent face aux erreurs de séquençage propre à cette technologie, afin de pouvoir choisir une solution d'analyse robuste et d'interpréter les résultats en toute connaissance de cause.

Le projet de thèse

Cette thèse s'inscrit dans la nécessité de mieux appréhender les méthodes d'analyse bioinformatiques de jeux de données métagénétiques, afin de pouvoir proposer des solutions et recommandations d'analyse automatisables, reproductibles, et adaptées aux plans d'expérience. Son objectif principal est de mieux appréhender les biais d'analyse intervenant dans de telles études, et de proposer des solutions permettant de limiter l'impact de ces biais sur les résultats.

Ce manuscrit se compose de six chapitres dédiés aux différents projets et résultats obtenus durant la thèse. Le premier chapitre est un état de l'art décrivant les méthodes actuelles d'une étude métagénétique. Il décrit les méthodes d'échantillonnage ainsi que les techniques de séquençage, et présente les différentes méthodes d'analyse existantes pour interpréter les jeux de données. Dans cette partie, une attention particulière est portée sur les biais inhérents à chaque étape. Le second chapitre présente les résultats de l'expertise de la solution d'analyse métagénétique déployée sur la plate-forme PEGASE-biosciences afin d'évaluer ses performances et limites. Cette expertise a fait émerger la nécessite de mettre en place un protocole formel d'évaluation de pipelines d'analyse, qui a permis d'étudier l'impact de différents contextes expérimentaux sur les performances de plusieurs pipelines, et est présenté dans le Chapitre 3. Cette approche a ensuite été transférée à une étude de microbiote intestinal humain menée précédemment sur la plate-forme. Les données de cette étude ont été ré-interprétées par différents pipelines, afin d'évaluer la variation des conclusions biologiques selon le pipeline utilisé. Le Chapitre 4 présente les résultats de cette évaluation appliquée à un jeu de données réelles. Les différentes évaluations des méthodes d'analyses ont monté que l'un des biais majeurs d'une étude métagénétique est la sélection d'un couple d'amorces adaptées au microbiote d'intérêt, captant le maximum d'organismes en présence et ciblant une région génomique la plus variable possible. Nous avons développé une solution logicielle répondant à cette problématique, décrite dans le Chapitre 5. Enfin, l'expertise acquise durant la thèse a permis d'établir des recommandations pour l'analyse de données métagénétiques qui ont été mises en place sur la plate-forme PEGASE-biosciences et qui sont présentées dans le dernier chapitre.

Une page Internet dédiée au projet de thèse et partageant des données complémentaires à ce manuscrit est accessible à l'adresse suivante :

http://www.pegase-biosciences.com/2013-0920

Chapitre 1 - Des échantillons à la description des microbiotes

1.1 - Déroulement global d'une étude métagénétique

Une étude métagénétique se déroule en de nombreuses étapes successives, aussi bien techniques qu'analytiques, résumées dans la Figure 1.1. Son objectif est de donner des éléments de réponse à une question biologique associée à un microbiote d'intérêt. On peut donner pour exemples les questions suivantes :

- Comment se répartissent différentes espèces de micromycètes sur des séquoias de différents sites californiens ? [Harrisson *et al.* 2016]
- Le microbiote bactérien du fromage de Herve est-il différent lorsque ce fromage est produit à partir de lait cru ou de lait pasteurisé ? [Delcenserie *et al.* 2014]
- Quelle est la composition des microbiotes bactériens prélevés dans différents organes de cadavres humains à différents stades de décomposition ? [Hyde *et al.* 2013]

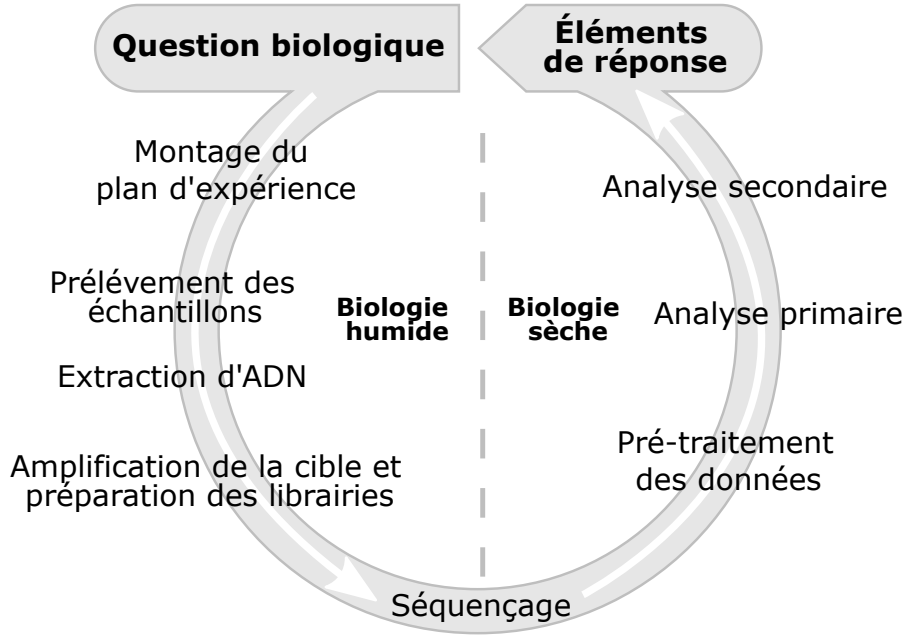


Figure 1.1 : Étapes principales d'une étude métagénétique.

De la question biologique posée aux éléments de réponses, les différentes étapes d'une étude métagénétique peuvent être partagées en deux catégories : avant, et après séquençage. On considère souvent que l'avant-séquençage implique des techniques de biologie humide (*wet lab*), manipulant une matrice biologique, tandis que l'après-séquençage est réservé à des techniques de biologie sèche (*dry lab*), manipulant des données informatiques. En réalité, il existe une interdépendance entre ces deux champs d'expertise, la bioinformatique pouvant aider à limiter des biais techniques, tandis que les choix techniques opérés au laboratoire peuvent directement influer sur l'analyse bioinformatique.

Le noyau d'une étude métagénétique est le run de séquençage haut-débit, permettant de passer d'une information moléculaire (fragments d'ADN) à une information textuelle (séquences d'ADN). Afin de mieux comprendre cette transition, il est nécessaire de décrire les technologies de séquençage haut-débit actuelles, en particulier celles utilisées par la plate-forme PEGASE-biosciences.

1.2 – Technologies de séquençage haut-débit de seconde génération

Les séquenceurs d'ADN sont communément répartis en quatre générations. Walter Gilbert et Frederick Sanger ont inventé dans les années 70 les premières techniques de séquençage d'ADN rapide, sans pour autant permettre d'atteindre des débits conséquents. Bien qu'automatisable, cette approche de séquençage est considérée comme à bas/moyen débit. Le MPSS [Brenner *et al.* 2000], publié en 2000 et produit par la société Lynx Therapeutics sans jamais avoir été commercialisé, est considéré comme la première génération de séquençage haut-débit dit NGS (*next generation sequencing)*. Mais le séquençage haut-débit n'a réellement diffusé que depuis la deuxième génération de séquenceurs haut-débit (454 GS-FLX et GS-FLX+ de Roche, Illumina HiSeq & MiSeq, Ion Torrent PGM) qui sont actuellement les séquenceurs haut-débit les plus démocratisés. Le principe du séquençage de deuxième génération repose sur une amplification clonale par PCR d'une

librairie de fragments d'ADN à séquencer. La résultante de cette amplification clonale est la matrice de séquençage. Les séquenceurs de troisième génération (Pacific Biosciences SMRT, Oxford Nanopore MinION, ...) permettent le séquençage direct d'un ADN matrice sans étape d'amplification par PCR. Ces technologies montrent un accroissement substantiel de la taille des lectures produites, accompagné d'un accroissement des erreurs de séquençage. Actuellement utilisés dans un contexte de recherche uniquement, ces séquenceurs et leurs données ne sont pas abordés dans ce projet de thèse.

1.2.1 - Préparation des librairies de séquençage

Dans une étude métagénétique, les fragments d'ADN à séquencer sont générés par PCR sur le mélange d'ADN génomique initial, en utilisant des d'amorces ciblant le *locus* d'intérêt. Ces amplicons doivent également contenir à leurs extrémités des séquences artificielles nécessaires au séquençage, comme représenté dans la Figure 1.2. Pour toutes les technologies de séquençage, ces séquences artificielles sont composées :

- des séquences d'adaptateur, permettant d'ancrer par complémentarité les fragments au support de séquençage et d'initier ce dernier ;
- des séquences d'index, codes-barres artificiels permettant de marquer les fragments selon leur échantillon d'origine, et ainsi de mélanger plusieurs échantillons en un seul run de séquençage (dit multiplex).

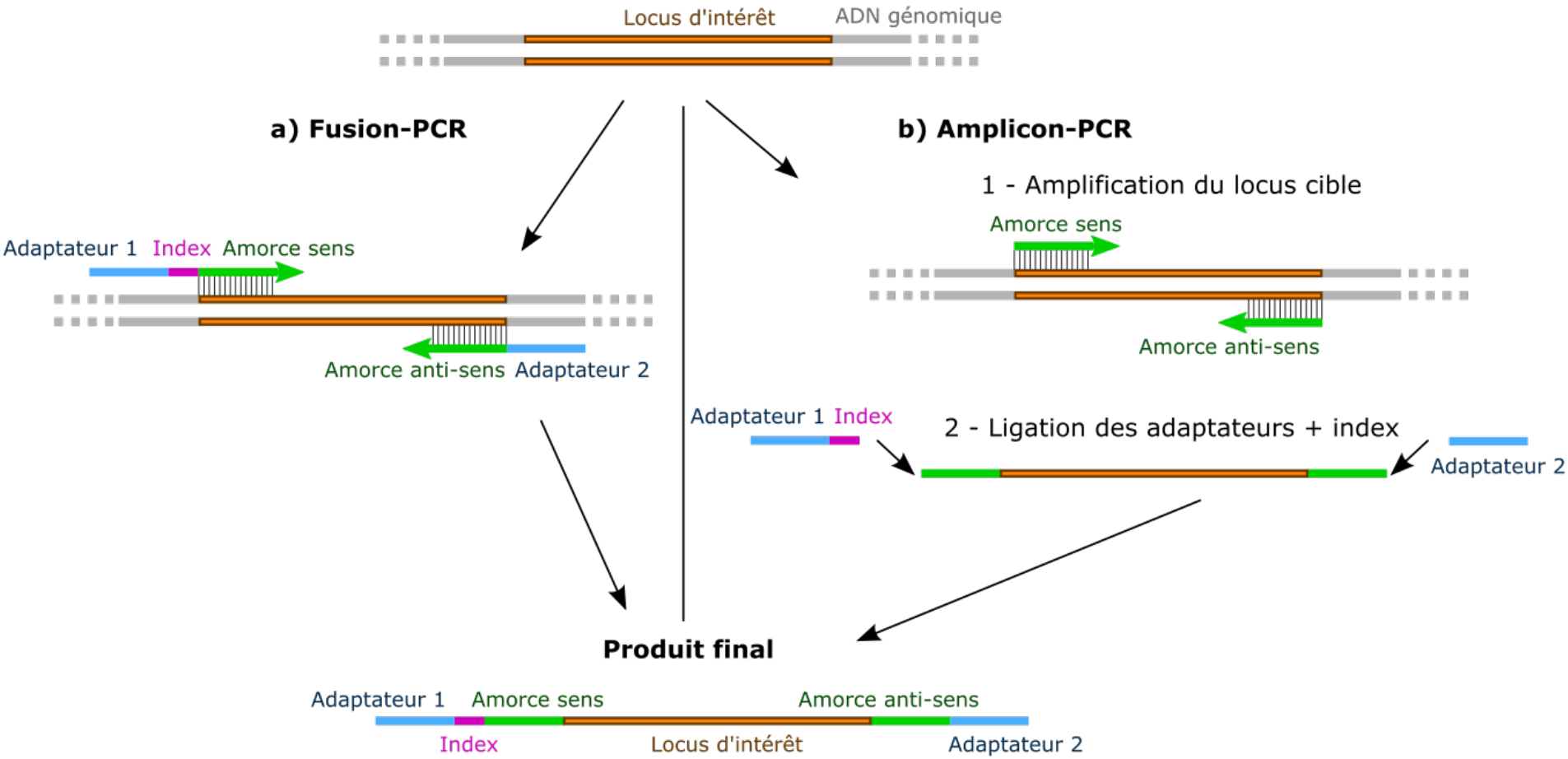


Figure 1.2 : Préparation des librairies de séquençage par a) Amplicon-PCR ou b) Fusion-PCR.

L'ajout de ces séquences artificielles aux fragments d'ADN à séquencer peut être effectué par Fusion-PCR (Figure 1.2.a) ou par Amplicon-PCR (Figure 1.2.b). La Fusion-PCR amplifie le locus d'intérêt en utilisant des amorces auxquelles sont liés les adaptateurs de séquençage et index associés. Ainsi, la librairie de séquençage est générée en une seule réaction d'amplification. À l'inverse, la méthode Amplicon-PCR est réalisée en deux étapes : le *locus* cible est tout d'abord amplifié par PCR, et les séquences d'adaptateurs et index sont ajoutées aux amplicons par ligation. La solution de Fusion-PCR étant plus facile à pratiquer qu'une ligation (qui est moins efficace), cette première est couramment utilisée sur la plate-forme PEGASE-biosciences. Ces librairies de séquençage sont adaptées à être traitées par des séquenceurs haut-débit de deuxième génération, afin de générer des lectures (*reads*) par un run de séquençage. Cette génération de séquenceurs est la première à avoir permis l'utilisation du séquençage haut-débit sur des machines tenant sur une paillasse, d'où leur nom de *benchtop sequencers*. Il existe actuellement trois modèles distincts de séquenceurs de paillasse : le 454 GS Junior (dont la production a été arrêtée en 2016), l'Ion Torrent PGM, et l'Illumina MiSeq, dont les caractéristiques principales sont détaillées dans le Tableau 1.3.

	454 GS Junior	**Ion Torrent PGM**	**Illumina MiSeq**
Date de sortie	Fin 2009	2010	2011
Méthode d'amplification	PCR par émulsion (hors séquenceur)	PCR par émulsion (hors séquenceur)	PCR en ponts (dans le séquenceur)
Mode de détection de la polymérisation	Pyroséquençage (détection de lumière)	Détection de variation de pH	Terminateur fluorescent réversible
Taille maximale des lectures	400 nt	450 nt	2 x 300 nt
Débit	35 Mb	Chip 314™ ~ 40 Mb Chip 316™ ~ 200 Mb Chip 318™ ~ 1 Gb	1,5 Gb
Temps d'un run (hors préparation des librairies)	8 h	3 h	27 h

Tableau 1.3 : Caractéristiques des séquenceurs haut-débit de paillasse.

1.2.2 - Séquençage 454

Avant l'avènement des séquenceurs de paillasse, le séquenceur haut-débit majoritairement utilisé pour des études de métagénétique était le 454, premier séquenceur haut-débit décrit en 2005 [Margulies *et al.* 2005] et mis sur le marché en 2006 par 454 Life Sciences (société rachetée par Roche Diagnostics en 2007). Le 454 repose sur la technique du pyroséquençage : la polymérisation nucléotidique (élongation du brin complémentaire au brin matrice) libère un pyrophosphate. Ce dernier est transformé en ATP par une ATP sulphurylase. Cet ATP est alors utilisé, couplé à une luciférine, par une luciférase. S'en suit une production d'oxyluciférine et d'un signal lumineux. Lors du pyroséquençage, les désoxyribonucléotides sont ajoutés de façon séquentielle, permettant de détecter lequel d'entre eux est intégré lors d'une émission de lumière. L'intérêt principal de cette technologie était sa capacité à générer les lectures les plus longues en comparaison avec ses concurrents (atteignant les 400 nt avec le séquenceur 454 GS-FLX et 700 nt avec le séquenceur 454 GS-FLX+), ce qui est particulièrement intéressant dans un contexte de métagénétique (plus la cible séquencée est grande, plus elle permet de potentiellement discriminer les taxons présents dans l'échantillon).

Fin 2009, la société Roche Diagnostics a mis sur le marché le premier séquenceur dit de paillasse, le GS Junior, plus petit et avec moins de débit que le GS FLX, permettant ainsi à de plus petites structures d'accéder à une technologie de séquençage haut-débit à coût moins élevé. Toutefois, Roche Diagnostics a arrêté la production de séquenceurs 454 en 2016, rendant cette technologie désormais obsolète.

1.2.3 - Séquençage Ion Torrent PGM

Life Technologies (racheté en 2013 par Thermo Fisher Scientific) a mis sur le marché son séquenceur Ion Torrent PGM (*personal genome machine)* en 2010, différant du pyroséquençage par sa méthode de détection des nucléotides incorporés. Cette dernière se base sur l'utilisation d'une puce semi-conductrice, permettant de réduire les coûts de l'appareil et des consommables par rapport

au 454 de Roche Diagnostics.

Après préparation de la librairie de séquençage, chaque fragment d'ADN est fixé par l'adaptateur trP1 sur des billes de silice (une seule molécule par bille) en suspension dans une émulsion eau/huile, chaque goutte d'huile contenant idéalement une bille et formant ainsi un microréacteur de PCR, permettant d'effectuer une réaction d'amplification isolée. Cette amplification (PCR en émulsion) permet de recouvrir chaque bille de nombreuses copies du même fragment d'ADN. Les billes sont ensuite déposées sur une puce semi-conductrice composée de puits permettant ainsi à chaque bille d'être présente dans un seul puits. L'Ion Torrent PGM peut être utilisé avec 3 puces différentes, chacune correspondant à un certain débit de lectures (Chip 314™ ~ 40 Mb, Chip 316™ ~ 200 Mb, Chip 318™ ~ 1 Gb). Après ajout de réactifs, le séquençage par polymérysation peut débuter à partir de l'adaptateur A, par cycles successifs d'ajout d'un type de désoxynucléotide, capture de signal, et lavage. Dans un cycle, chaque nucléotide incorporé à un brin néosynthétisé provoque le relargage d'un proton, engendrant une variation de pH dans le puits. Cette variation de pH est détectée par le séquenceur comme signal d'incorporation d'un ou plusieurs nucléotides du cycle en cours dans le puits concerné. La succession de cycles permet ainsi, pour chaque puits, d'obtenir une séquence de mesure de pH (appelée ionogramme), qui est ensuite convertie en séquence par l'Ion Torrent PGM (étape dite de *basecalling*). Chacune de ces étapes est illustrée dans l'Annexe 1.

La technologie proposée par Ion Torrent est connue pour générer un taux d'erreur relativement plus élevé que ses concurrents, principalement des insertions/délétions dans des homopolymères. En effet, l'incorporation d'une succession de nucléotides identiques provoque une forte variation de pH saturant le signal, et rendant plus difficile la détermination du nombre exact de nucléotides identiques incorporés lors du même cycle.

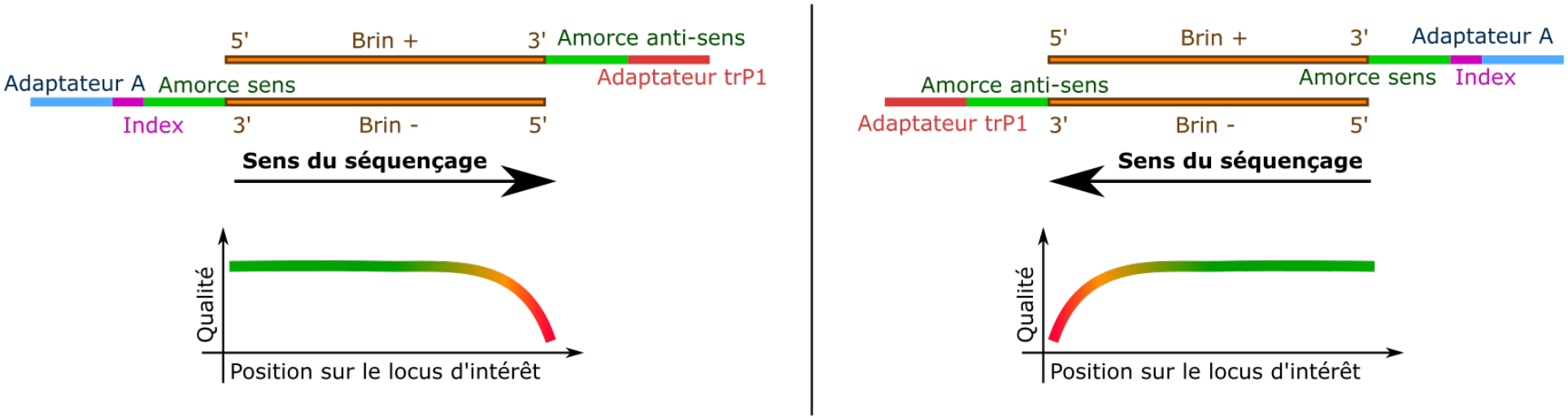


Figure 1.4 : Principe d'une préparation d'une librairie bidirectionnelle Ion Torrent PGM

La plate-forme PEGASE-biosciences est la première plate-forme en Europe à avoir été équipée d'un séquenceur de paillasse Ion Torrent PGM, en 2011, et l'a utilisé sur des applications métagénétiques dès 2012. Elle y a notamment appliqué la technique de préparation de librairie dite bidirectionnelle. En effet, une caractéristique des lectures Ion Torrent PGM est d'être de moins bonne qualité en 3'. Ainsi, lors d'un séquençage métagénétique classique, la région en 3' du *locus* d'intérêt séquencé contiendra systématiquement plus d'erreurs. Une préparation de librairie bidirectionnelle peut être effectuée afin de pallier ce problème (Figure 1.4), en utilisant deux couples d'amorces pour la Fusion-PCR : un couple où l'adaptateur d'initiation de séquençage est associé à l'amorce sens, et un couple où cet adaptateur est associé à l'amorce anti-sens. Ainsi, le *locus* d'intérêt sera séquencé dans les deux sens, générant des lectures qui seront de bonne qualité en 5' du *locus* ciblé, et d'autres qui seront de bonne qualité en 3' du *locus* ciblé (correspondant à la région 5', donc de bonne qualité, des lectures séquencées).

1.2.4 - Séquençage Illumina MiSeq

L'Illumina MiSeq est arrivé sur le marché en 2011, bien que la technologie Illumina ait été introduite en 2006. Cette technologie est également basée sur du séquençage par synthèse, mais son principe fait intervenir des nucléotides modifiés chimiquement et porteurs d'un fluorochrome différent (quatre fluorochromes pour les quatre bases nucléiques) pour proposer un séquençage médié par des terminateurs réversibles. En outre, les fragments d'ADN ne sont pas fixés sur une bille, mais directement sur le support de séquençage (une lame de verre nommée *flow cell*), sur laquelle l'amplification des fragments se fait en ponts (bridge-PCR), formant des clusters de fragments identiques sur la lame. L'intérêt du séquençage Illumina est sa plus grande précision, cette technologie générant moins d'erreurs de séquençage que ses concurrents. Le séquençage Illumina a longtemps été restreint à de courtes tailles de lectures, de l'ordre d'une centaine de nucléotides. Les développements des chimies de séquençage permettent actuellement à l'Illumina MiSeq de générer des lectures de 300 nucléotides, rendant cette technologie intéressante

dans un contexte métagénétique. Cependant, cet allongement des tailles de lectures a également entraîné une chute de qualité en 3', lié au bruit ajouté dans le signal qui se cumule à chaque cycle du séquençage [ecSeq 2017].

Afin de permettre un séquençage de fragments plus longs, Illumina a introduit la technologie de séquençage *paired-end*, consistant à séquencer les deux extrémités d'un même fragment comme présenté dans la Figure 1.5. L'intérêt est qu'à chaque fragment va correspondre deux lectures, une de son extrémité 5', et une de son extrémité 3'. Si le fragment est suffisamment court, les deux lectures seront chevauchantes, et pourront ainsi être assemblées en une lecture plus longue. Ainsi, sur un séquençage *paired-end* MiSeq générant des lectures de 300 nt, on peut espérer obtenir des lectures de 500 à 550 nt (avec une région chevauchante de 100 et 50 nt respectivement).

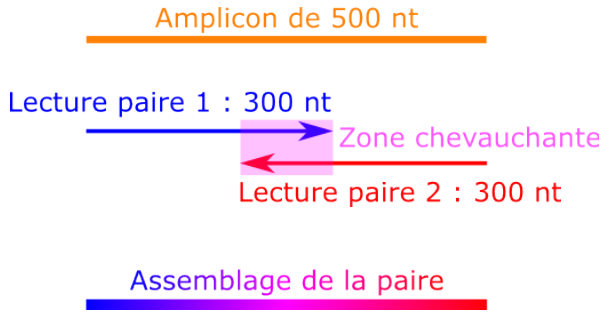


Figure 1.5 : Séquençage Illumina paired-end. Le séquençage de la cible est effectué dans les deux sens, générant deux lectures paires par amplicon qui se chevauchent, permettant de les assembler en une lecture plus longue sur la base de cette zone chevauchante.

1.3 - De l'échantillonnage aux données séquencées

Que ce soit dans l'établissement du plan d'expérience, l'utilisation de différents protocoles techniques ou solutions analytiques, chaque choix du biologiste a un impact sur l'image qu'il obtiendra de la composition de ses échantillons après analyse. Certains de ces choix sont inhérents à la nature même des échantillons : par exemple, des protocoles adaptés aux matrices biologiques seront à privilégier pour l'extraction d'ADN bactérien issu de crachats humains, ou lorsque la matrice biologique est présente en très faible quantité dans un prélèvement d'air. D'autres décisions sont contraintes par des limitations financières : on peut comprendre qu'une étude cas/témoin sur un modèle animal soit limitée en nombre d'individus par les contraintes éthiques et

les frais d'animalerie associés, ou que la quantité de lectures par échantillon soit réduite en faveur du nombre d'échantillons séquencés sur un même run. Certains choix peuvent enfin être guidés par l'habitude d'utilisation, l'autorité de méthodes de référence dans la littérature, ou la disponibilité des équipements : le biologiste n'a pas toujours le choix de la plate-forme de séquençage dont il dispose, et utilise souvent les réactifs dont il a l'habitude et qui sont compatibles avec le matériel dont il est tributaire. De même, ses choix d'amorces et de pipelines d'analyse se font souvent sur la base d'études bibliographiques, sans toujours être validés dans son propre contexte expérimental.

Le biologiste doit limiter au maximum les biais liés à ces contraintes et choix, en mettant en place des méthodes de contrôle (par l'emploi de témoins négatifs par exemple), en choisissant les protocoles techniques qui lui sont accessibles tout en étant les plus adaptés à ses échantillons, et en optimisant son plan d'expérience. De nombreux leviers analytiques sont également à sa disposition pour l'accompagner dans l'appréhension de ces biais.

Cette partie a pour objectif de présenter les différentes étapes, de l'échantillonnage à l'obtention des séquences à analyser, pouvant causer des biais dans les jeux de données ainsi que les méthodes analytiques disponibles pour diminuer leur impact sur l'analyse.

1.3.1 - Mise en place du plan d'expérience

La mise en place du plan d'expérience est une étape fondamentale de toute étude métagénétique. Définir une question biologique claire est indispensable pour monter un plan d'expérience robuste qui permettra au biologiste d'y répondre. Le nombre d'échantillons à séquencer, les variables à évaluer, la profondeur de séquençage, le nombre de réplicats sont autant de facteurs à prendre en compte ; ces derniers pourront guider le choix des technologies, les protocoles techniques et les méthodes d'analyse les plus adaptés. Par exemple, une étude cas/témoin doit être représentée par suffisamment d'échantillons pour être statistiquement valide. De même, il est préférable de renforcer le plan d'expérience par l'ajout de réplicats biologiques à

un surséquençage qui n'apportera pas plus d'information utile à l'analyse. À cette étape, l'intervention d'un bioinformaticien et biostatisticien est importante : leur expertise sur les données permettra de valider un plan d'expérience robuste, et de générer des données dont l'analyse permettra de répondre à la question biologique initialement posée.

Cette réflexion doit également être accompagnée d'une évaluation des connaissances actuelles sur les microbiotes d'intérêt. Toute information sur le milieu étudié et les organismes qui y sont associés (conditions environnementales, contaminants potentiels, caractéristiques cellulaires, ...) peut guider le choix de protocoles techniques adaptés. De même, le recueil de métadonnées sur chaque échantillon permet de les replacer dans un contexte biologique sur lequel s'appuyer pour le montage de plan d'expérience, et qui peut étayer les conclusions de l'analyse. Le bioanalyste doit cependant prendre garde à utiliser ces informations comme données complémentaires, et en aucun cas comme *a priori* qui pourrait finir par biaiser son interprétation en influant sur les méthodes mises en place.

Des initiatives telles que le Genomics Standard Consortium proposent des recommandations de métadonnées à recueillir, telles que les MIMS (Minimum Information about a Metagenome Sequence [Field *et al.* 2008]) ; ces standards sont toutefois peu référencés, et donc rarement suivis. Toutes ces informations sont également importantes dans l'établissement d'une stratégie d'échantillonnage. Chaque prélèvement doit être représentatif du milieu étudié, ce qui soulève de nombreuses questions. Peut-on transposer les conclusions tirées de l'étude d'un microbiote fécal à une flore intestinale alors que les deux milieux ont des conditions environnementales bien différentes ? Un sol n'étant pas homogène, quelle stratégie d'échantillonnage mettre en place pour être le plus représentatif de sa composition microbienne globale ? Comprendre la variabilité des communautés microbiennes dans le milieu étudié est indispensable à l'établissement d'une stratégie d'échantillonnage efficace. L'allocation de ressources à une étude pilote préliminaire (comparant différentes méthodes par exemple) sur un nombre d'échantillons restreint peut

permettre une meilleure appréhension du microbiote d'intérêt, et un ajustement du plan d'expérience en conséquence.

Une fois le plan d'expérience mis en place, les étapes de biologie humide peuvent être exécutées, du prélèvement des échantillons au séquençage. Chacune de ces étapes est sujette à des biais potentiels, résumés dans la Figure 1.6 et détaillés dans la suite de ce chapitre. Certains de ces biais peuvent être limités par des leviers analytiques faisant intervenir des méthodes bioinformatiques.

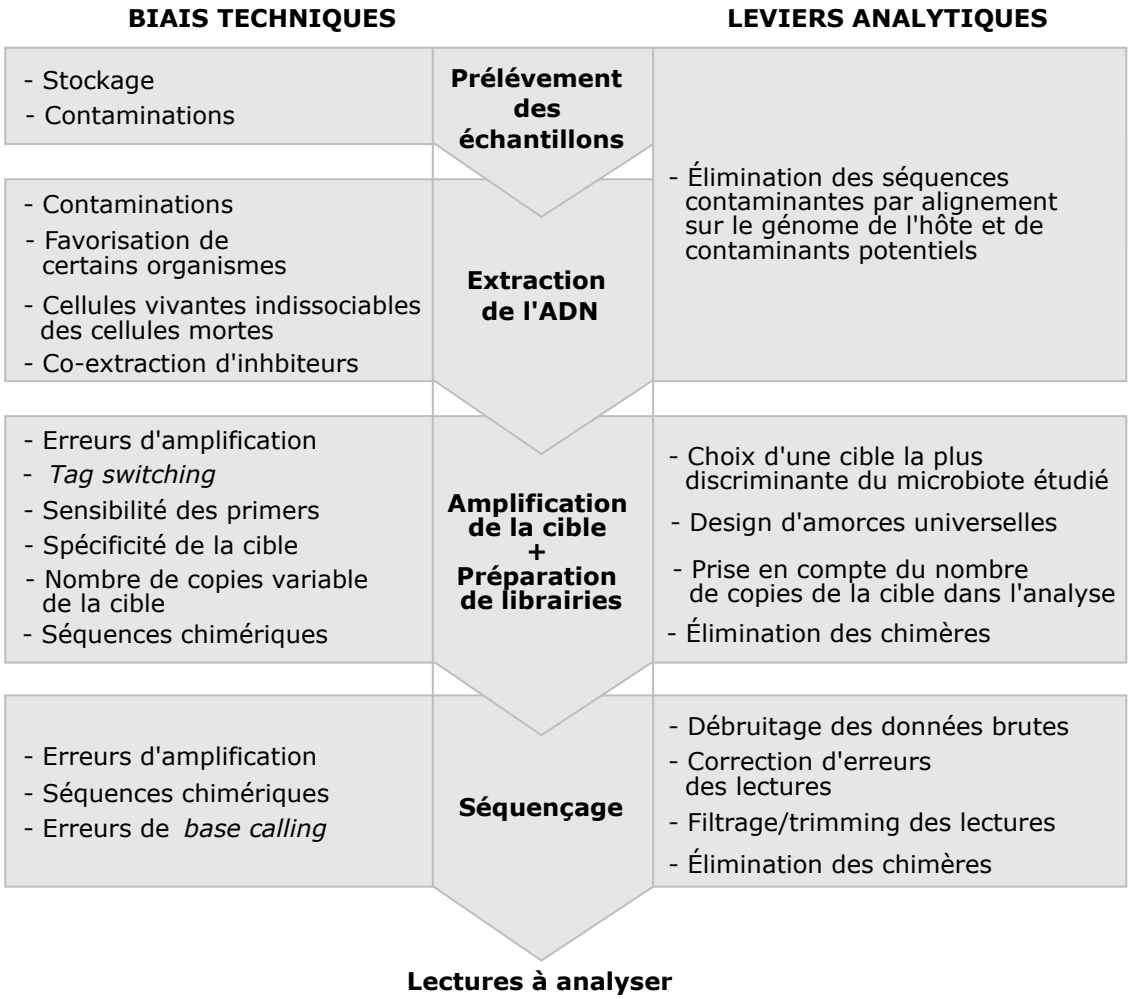


Figure 1.6 : Biais techniques dans les différentes étapes intervenant de l'échantillon aux lectures à analyser, et voies d'intervention possibles du bioanalyste pour les minimiser.

1.3.2 - Prélèvement des échantillons

La première étape technique consiste à prélever et traiter les échantillons à séquencer. Idéalement, le résultat d'une analyse métagénétique doit refléter la composition globale du microbiote d'intérêt. Cette matrice étant vivante, elle est sujette à évolution après prélèvement. Dans le cas d'un prélèvement humain par exemple, une composition microbienne peut évoluer très rapidement en passant de son milieu chaud anaérobie d'origine à un milieu à température ambiante. L'idéal est ainsi d'extraire l'ADN sitôt l'échantillon prélevé. Toutefois, la nécessité de transport ou de stockage lors d'études à grandes échelles impose souvent une nécessité de préservation (par congélation à -80°C et ajout de préservateur par exemple) pour figer la composition du microbiote. Ces méthodes de préservation peuvent tout de même altérer la composition des échantillons, certaines études décrivant une composition différente entre un échantillon congelé (avec ou sans préservateur) et un échantillon frais [Choo *et al.* 2015]. En outre, effectuer des prélèvements dans des conditions non stériles peut augmenter le risque d'introduction d'ADN exogène contaminant dans les échantillons d'intérêt.

1.3.3 Extraction de l'ADN

Une fois tous les échantillons à disposition, une étape d'extraction et de purification va permettre d'accéder à l'ADN des organismes en présence, idéalement sans contamination de l'ADN hôte dans le cas d'un prélèvement organique. L'ADN doit être extrait en une quantité suffisante pour la préparation de la librairie de séquençage, et dans des proportions respectant celles des organismes dans le milieu. La méthode d'extraction doit être adaptée à la nature des cellules de ces derniers. Leur hétérogénéité implique des propriétés variables (telles que leur taille ou la structure de leur paroi), qui les rendent par exemple plus ou moins sensibles à différentes méthodes de lyse. Une approche enzymatique douce n'extraira pas l'ADN des cellules plus résistantes (archées et Gram+ par exemple), mais une approche mécanique brutale risque de fragmenter l'ADN qui ne pourra pas être amplifié

correctement. Malgré l'existence de kits commerciaux adaptés à différents microbiotes, aucun protocole d'extraction ne peut assurer une homogénéité de traitement entre les organismes, et ainsi un respect de leurs proportions initiales dans les proportions de l'ADN obtenu. L'évaluation de cette étape se base sur des mesures de rendement et de pureté, mais il est impossible d'évaluer la correspondance entre l'ADN extrait et les organismes présents dans l'échantillon. En outre, de l'ADN exogène issu de l'hôte tout comme de certains réactifs peut être à l'origine de contaminations. D'où l'importance de mettre en place des contrôles avec des témoins négatifs, et en contrôlant la présence de contaminant par qPCR si possible. L'étape d'extraction ne permet pas de distinguer entre cellules vivantes et cellules mortes dans le milieu, ce qui est à garder à l'esprit dans le cas d'une étude d'inférence fonctionnelle complémentaire. Enfin, la co-extraction de molécules inhibitrices peut impacter l'efficacité même de la PCR, et donc porter un préjudice à l'élaboration d'une librairie de séquençage.

1.3.4 Amplification de la cible

L'étape d'amplification avant séquençage va permettre de cibler le *locus* génomique d'intérêt et d'en synthétiser une quantité suffisante pour le séquençage. Cette cible est un marqueur taxonomique choisi pour son universalité (présent dans tous les génomes extraits) et sa spécificité (à séquence variable selon les taxons). Différents *loci* se sont imposés comme des marqueurs de référence pour différents règnes, souvent présents dans l'opéron ribosomique (ADNr 16S pour les bactéries, ITS pour les champignons, ADNr 18S pour les eucaryotes, ...) Par exemple, l'ADNr 16S est un marqueur métagénétique historique sur lequel les biologistes se sont appuyés depuis les années 80 pour identifier et classifier les différentes espèces bactériennes. Ce gène est en effet le plus conservé des trois gènes de l'opéron ribosomique au sein d'une même espèce, et contient des régions hypervariables qui permettent de ségréguer les espèces bactériennes en se basant sur sa séquence (Figure 1.7). Il a ainsi été proposé comme marqueur évolutif de référence pour le règne bactérien. Il peut être amplifié dans de nombreuses bactéries différentes d'un

même échantillon en une seule réaction, grâce à ses régions hautement conservées entre taxons, permettant la sélection d'amorces d'amplification universelles. En outre, son utilisation historique a permis l'enrichissement des banques de séquences sur cette cible.

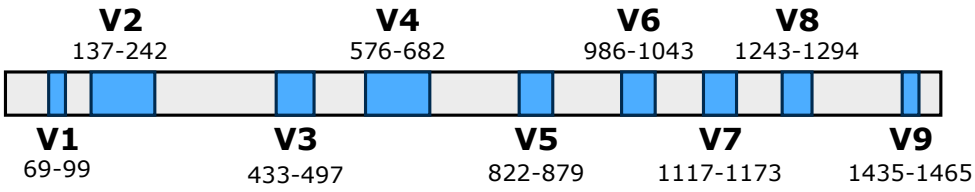


Figure 1.7 : Régions conservées (gris) et hypervariables (bleu) de l'ADNr 16S bactérien. Ces dernières sont numérotées de V1 à V9, et leurs positions sont tirées de Baker et al. (2003), utilisant la nomenclature E. coli définie par Brosius et al. (1978).

Pourtant, utiliser l'ADN ribosomique dans une étude métagénétique peut être à l'origine de biais dans les compositions bactériennes déduites après analyse. En effet, des événements de transferts horizontaux de gènes entre différents genres bactériens peuvent remettre en question sa spécificité [Acinas *et al.* 2004]. Toutefois, la structure des gènes ribosomiques en opéron semble limiter ce phénomène dans le cas de l'opéron ribosomique, comme formulé par Jain *et al.* dans l'hypothèse de complexité [Jain *et al.* 1999].

Une autre limite de cet opéron est sa présence en un nombre de copies variables selon les génomes, pouvant aller de 1 à 15 copies selon l'espèce [Klappenbach *et al.* 2000]. L'hétérogénéité des séquences de ces différentes copies rend plus complexe l'identification de ces organismes sur la base de ce seul marqueur : certains génomes présentent une variation de séquences intragénomiques entre ces différentes copies pouvant atteindre jusque 11 % de la séquence de l'opéron [Acinas *et al.* 2004]. De plus, la variation du nombre de copies de l'opéron ribosomique introduit un biais d'estimation de l'abondance relative des différents organismes dans l'échantillon : une espèce à grand nombre de copies sera favorablement amplifiée et représentée dans les lectures issues du séquençage. Ainsi, la variation du nombre de copies du *locus* cible entre génomes provoque une estimation biaisée de la diversité microbienne.

Ce biais peut être appréhendé *a posteriori* en corrigeant l'abondance des bactéries à l'issue de l'analyse par leur nombre de copies d'ADNr 16S s'il est connu, ou par celui d'une espèce proche [Kembel *et al.* 2012]. Cette information n'est pas toujours disponible, si l'on étudie des organismes non séquencés par exemple, et ne permet pas de corriger un biais d'amplification éventuel induit par une variation du nombre de copies de la cible selon les organismes composant le microbiote d'intérêt. Les banques de séquences d'ADNr 16S sur lesquelles se basent les pipelines d'analyse ne contiennent pas l'information du nombre de copies par génome, qui est de ce fait rarement pris en considération dans l'analyse.

D'autres gènes à copie unique ont été proposés comme marqueurs métagénétiques pour estimer une meilleure abondance relative des différentes espèces bactériennes. En effet, certains gènes de ménage tels que ceux de la famille *recA* ou le gène *rpoB* sont universellement présents chez les bactéries et ont un degré de variation de copies moindre entre espèces, comparé à l'ADNr 16S. Ces gènes semblent en outre discriminer la taxonomie à un niveau plus élevé, en présentant plus de variabilité [Thompson *et al.* 2004, Vos *et al.* 2012]. Cependant, il est plus difficile de designer des amorces universelles sur des gènes codant pour des protéines. En effet, des régions protéiques conservées n'impliquent pas que les régions génétiques correspondantes le soient également ; les amorces désignées sur ces cibles doivent ainsi contenir plus de dégénérescences pour s'aligner sur un maximum d'organismes, sans garantir leur universalité. Ces gènes sont également plus sensibles à des événements de transferts horizontaux. Enfin, la représentation de ces gènes dans les banques de séquences est minime en comparaison aux séquences d'ADNr 16S, gène qui reste une référence par son grand nombre de séquences annotées dans les banques.

Plus une cible est étudiée, plus elle est référencée, plus on favorisera son étude ; ce cercle vicieux rend difficile le développement de nouvelles méthodes basées sur d'autres cibles. Ce problème se présente par exemple pour les études de métagénomique fongique : les séquences ribosomiques interstitielles ITS

sont un marqueur évolutif plus discriminant des Fungi que l'ADNr 18S [Schoch *et al.* 2012]. Or, il existe peu de banques spécifiques à ce marqueur ; les séquences d'ITS qui s'y trouvent sont souvent incomplètes et/ou mal annotées. Le manque de connaissances sur ce marqueur freine ainsi le développement des applications de métagénétique fongique, qui sont pourtant précieuses dans l'étude de certaines pathologies humaines par exemple.

Le séquençage haut-débit de seconde génération ne permet de couvrir qu'un fragment de la région génomique d'intérêt, au maximum de 450 nucléotides (Ion Torrent PGM) à 550 nucléotides (2x300 *paired-end* chevauchant Illumina MiSeq). En reprenant l'exemple de l'ADNr 16S bactérien, les limitations de longueurs des lectures ne permettent d'obtenir la séquence que de deux à trois régions hypervariables au maximum. Une lecture de quelques centaines de nucléotides va nécessairement contenir moins de sites discriminants entre taxons que toute la séquence de l'opéron ; cette diminution d'information rend complexe la tâche d'identifier une région ayant le même pouvoir de discrimination pour toutes les familles bactériennes.

Différentes études se contredisent sur le choix d'une telle région, indépendamment du type de microbiote (V2-V3 [Liu *et al.* 2008] ou V4-V6 [Yang *et al.* 2016]). Ces propositions se basent sur l'idée reçue que les positions de ces régions hypervariables [Baker *et al.* 2003] sont fixes quel que soit le génome bactérien. Or, certaines familles bactériennes (les *Acetobacteraceae* par exemple [Chakravorty *et al.* 2015]) présentent ces régions hypervariables à des positions différentes ; elles seront moins facilement discriminées en se basant sur ces recommandations. Ces dernières ont d'ailleurs été validées sur l'étude de la variabilité des séquences dans les banques de séquences ribosomiques existantes. Or, ces banques contiennent une majorité de séquences d'organismes cultivables, historiquement plus étudiés, que d'organismes non-cultivables issus de milieux peu décrits dans la littérature. Le choix d'une cible sur la base de toutes les séquences présentes dans une banque spécialisée est ainsi biaisé par la surreprésentation de certains organismes et types de microbiotes dans ces banques.

Il semble utopique de valider une région comme étant suffisamment discriminante entre tous les taxons quel que soit le microbiote. En effet, les régions hypervariables ont été soumises à différentes pressions évolutives selon les branches taxonomiques [Kumar *et al.* 2011], ce qui les rend plus ou moins informatives selon les organismes étudiés, et donc selon la composition microbienne du milieu d'intérêt. En sachant que différentes régions hypervariables peuvent donner une image différente de la composition d'un microbiote et de sa diversité [Birtel *et al.* 2015], une certaine connaissance *a priori* de sa composition semble ainsi nécessaire pour sélectionner la région la plus discriminante entre les taxons concernés.

Plusieurs études proposent des recommandations de certaines régions hypervariables plus adaptées à certains microbiotes. Par exemple, certains auteurs privilégient V1 et V2 car elles permettent de discriminer un maximum de genres dans des échantillons issus d'eaux usées [Guo *et al.* 2013], tandis que d'autres recommandent V3 et V4 dans le même contexte [Cai *et al.* 2013]. Autre exemple, Nossa *et al.* recommandent également l'emploi de V3 et V4 pour l'intestin antérieur humain en se basant sur un jeu de données simulé représentatif d'un tel microbiote, et en validant que ces régions permettent une identification la plus précise des organismes présents [Nossa *et al.* 2010].

L'utilisation d'échantillons réels pour valider une région d'intérêt ne permet pas d'évaluer sa sensibilité, tandis que l'utilisation de jeux de données simulées est biaisée par un *a priori* de composition. Une solution idéale serait d'amplifier plusieurs cibles afin de couvrir un maximum de régions informatives [Soergel *et al.* 2012]. Néanmoins, les résultats d'une telle approche sont difficiles à interpréter : Prend-on en compte uniquement les taxons retrouvés par toutes les régions ? Que conclure d'une diversité variable entre les régions étudiées ? Une approche de séquençage *paired-end* Illumina sur deux régions différentes permettrait de résoudre ces problèmes en conservant l'information que chaque paire de lectures appartient au même génome. Néanmoins, les pipelines d'analyse actuels ne sont pas capables d'interpréter ces jeux de données appariés, ce qui limite grandement l'intérêt d'une telle approche.

Le design des amorces amplifiant la région cible d'intérêt est essentiel pour capter un maximum d'organismes en présence, sans amplifier de l'ADN contaminant. Le choix des amorces se fait souvent sur base bibliographique, en utilisant un couple déjà validé par des études précédentes sur le même type de microbiote. Pour l'ADNr 16S, de nombreuses amorces ont été désignées il y a plus de vingt ans, sur la base d'une connaissance qui n'est plus actuelle. Plusieurs études ont évalué la sensibilité des amorces les plus utilisées [Mao *et al.* 2012, Klindworth *et al.* 2013], qui est toutefois directement dépendante de la composition du microbiote. Un biologiste souhaitant valider son couple d'amorces ou effectuer un nouveau design va utiliser un alignement de quelques séquences des organismes qu'on s'attend à retrouver dans le milieu étudié. Il se basera alors sur l'idée préconçue que des amorces suffisamment universelles sur ces séquences le seront pour tous les organismes de son microbiote d'intérêt.

L'universalité des amorces n'est pas le seul critère de validation : leur compatibilité et spécificité sont tout aussi importantes, et le biologiste devra souvent s'accommoder de ces propriétés pour réaliser un compromis aboutissant à un design optimal. L'ajout de bases dégénérées permet d'augmenter la sensibilité d'une amorce, aux dépens de sa spécificité (augmentant le risque d'amplification de séquences contaminantes – par exemple de séquences chloroplastiques ou mitochondriales dans le cas d'amorces ciblant l'ADNr 16S bactérien).

La réaction d'amplification en elle-même est source de nombreux biais techniques. En effet, il est délicat d'assumer que les séquences amplifiées sont représentatives des abondances initiales des organismes. Les paramètres de PCR, habituellement optimisés pour un rendement optimal, doivent également prendre en considération la préservation du ratio des différents génomes de la matrice. Un nombre de cycles PCR trop élevé ou une concentration de l'ADN matrice trop importante augmente les risques de déséquilibrer ce ratio, et la formation de séquences chimériques (lorsqu'un amplicon en cours de synthèse glisse vers une autre séquence matrice). Ces séquences hybrides seront au mieux écartées par les pipelines d'analyse, au pire mal identifiées. Les erreurs

d'amplification (substitutions causées par de mauvais appariements et délétions dues à des glissements de la polymérase) peuvent avoir un impact considérable sur les résultats, si par exemple elles touchent un nucléotide discriminant entre deux taxons. Utiliser une polymérase à activité correctrice peut limiter la quantité d'erreurs dans les amplicons, mais semble favoriser l'apparition de chimères [Ahn *et al.* 2012].

Les amplicons obtenus sont complétés par l'ajout de séquences artificielles nécessaires au séquençage, comme les séquences d'adaptateurs et d'index (voir Chapitre 1, Section 1.2.1). L'index peut également être à l'origine de biais dans les résultats : que ce soit en Illumina [Esling *et al.* 2015] ou 454 [Carlsen *et al.* 2012], et Ion Torrent par extension, une inversion des code-barres (phénomène de *barcode switching*) peut atteindre jusqu'à 1 % des lectures. Celles-ci sont attribuées à un échantillon différent de leur échantillon d'origine, introduisant ainsi un biais de composition dans les résultats.

1.3.5 Le séquençage

L'étape de séquençage en elle-même est également à l'origine de biais, causés par les étapes d'amplification durant le séquençage ou par de mauvais *basecalls* (étape de conversion du signal brut en séquence). Ainsi, différentes technologies de séquençage peuvent aboutir à une image différente du même microbiote séquencé [Luo *et al.* 2012, Clooney *et al.* 2016]. Historiquement, la technologie de choix pour du séquençage métagénétique était le pyroséquenceur 454 de Roche, car il permettait d'avoir les lectures les plus longues. Il était toutefois à l'origine d'un taux d'erreurs non négligeables, principalement des insertions/délétions dans les homopolymères. Une augmentation d'erreurs de *basecalling* était également observée vers la fin des lectures, faisant chuter leur qualité [Balzer *et al.* 2010]. L'Ion Torrent PGM, également basé sur du séquençage par PCR en émulsion sur billes, est souvent considéré comme ayant un profil d'erreurs similaire au 454. Cette technologie semble toutefois générer un taux d'insertions/délétions plus élevé et sous-estimer les espèces à très fort ou très faible taux de GC. Actuellement, le séquençage Illumina est la plate-

forme de prédilection pour du séquençage métagénétique, par son faible taux d'erreurs et ses développements permettant d'atteindre une taille de lectures comparable au 454. Cependant, la fin de ses lectures présente aussi une chute de qualité [ecSeq 2017]. Une revue de Laehnemann *et al.* répertorie l'ensemble des types d'erreurs de séquençage induites par chaque technologie, ainsi que les différents logiciels développés pour y palier [Laehnemann *et al.* 2016].

Afin de limiter ces erreurs dans les jeux de données, quelques filtres triviaux peuvent être appliqués sur les lectures en sortie de séquenceur en se basant sur leurs propriétés intrinsèques, notamment de qualité. Les lectures contenant des bases indéterminées (N, présentes dans les données 454 et Illumina) peuvent être filtrées. Les lectures issues de plate-formes à risque d'erreurs dans les homopolymères (454, Ion Torrent), peuvent être filtrées au-delà d'une certaine longueur d'homopolymère dans la séquence. Pour toutes les technologies, les séquences peuvent être coupées lorsque leur qualité chute sous un certain seuil [Kunin *et al.* 2010, Schloss *et al.* 2011]. Les séquences contenant des nucléotides erronés dans leurs amorces peuvent également être éliminées. Enfin, les lectures peuvent aussi être sélectionnées en fonction de la taille d'amplicon attendue. En effet, Wommak *et al.* ont démontré dans le contexte de métagénomique WGS que de petites séquences (100-200 nt) ne portent pas assez d'information taxonomique par rapport à des lectures longues (> 400 nt), et qu'il est préférable d'avoir peu de lectures entre 150 et 400 nt qu'une forte couverture de lectures à 100 nt [Wommack *et al.* 2008]. Enfin, les séquences contaminantes (hôte, chloroplastes, ...) peuvent être éliminées en alignant les lectures sur des séquences de références identifiées comme étant des contaminants potentiels.

De nombreux algorithmes ont été développés dans le but d'éliminer les erreurs ponctuelles induites par les différentes technologies de séquençage, que ce soit dans les données brutes de mesure de signal, ou dans les lectures issues du *basecalling* effectuée par le séquenceur. Cette étape, nommée le *denoising* ou débruitage, a été initialement développée pour le pyroséquençage, technologie générant le plus d'erreurs. AmpliconNoise [Quince *et al.* 2011] (dont

l'étape PyroNoise a été implémentée dans l'étape *shhh.flows* du pipeline mothur) ainsi que FlowClus [Gaspar & Thomas 2015] se basent sur le *clustering* de *flowgrams* (fichiers de signal brut issus du séquençage) pour identifier les signaux dus à des erreurs. Ces méthodes sont limitées au format *sff* des *flowgrams,* propre au pyroséquençage 454. Pour les autres technologies, différents algorithmes tels que Denoiser [Reeder *et al.* 2010], Acacia [Bragg *et al.* 2012] ou encore DADA [Rosen *et al.* 2012] corrigent les lectures après *basecalling* en évaluant leur abondance relative dans les jeux de données. Huse *et al.* proposent une étape de pré-*clustering* à un fort taux de similarité pour réduire le bruit présent dans les séquences [Huse *et al.* 2010]. À noter que des méthodes telles que Fiona [Schulz *et al.* 2014] ou Pollux [Marinier *et al.* 2015] ne peuvent être appliquées dans un contexte de métagénétique, car ces algorithmes se basent sur des caractéristiques de séquences aléatoires issues de génomes complets.

Une dernière étape possible du pré-traitement des séquences consiste à éliminer celles qui sont chimériques (pouvant atteindre jusqu'à 20 % de séquences pour des résultats de séquençages métagénétiques 454 [Haas *et al.* 2011]). Les chimères, séquences hybrides de deux matrices biologiques différentes, sont à l'origine d'une sur-représentation du nombre de taxons estimés, donc de la diversité de l'échantillon. Elles peuvent être éliminées après apprentissage sur un jeu de données de référence (ChimeraSlayer [Haas *et al.* 2011]) ou *de novo* (UCHIME [Edgar *et al.* 2011], Perseus [Quince *et al.* 2011], Decipher [Wright *et al.* 2012]). Ces derniers algorithmes partent de l'idée selon laquelle les séquences chimériques sont sous-représentées par rapport à leurs séquences parentes qui ont subi au moins un cycle de PCR supplémentaire. Ces algorithmes ont été validés sur des séquences ribosomiques, mais doivent être validés sur d'autres cibles [Kim *et al.* 2013].

Toutes ces méthodes analytiques sont autant de leviers à actionner pour essayer de minimiser au maximum les erreurs dans les séquences avant analyse. Cependant, essayer de trop corriger les données peut aussi avoir un impact négatif : en effet, les analyses métagénétiques se basent sur les différences entre

séquences pour pouvoir les identifier. Ainsi, un filtrage trop drastique ou mal adapté à la technologie utilisée peut supprimer des variations de séquences ayant une réalité taxonomique, représentant jusqu'à des taxons entiers. Optimiser les différents protocoles (par exemple en utilisant des enzymes de haute confiance, en minimisant le nombre de cycles de PCR) et valider ses choix sur des contrôles internes permet de diminuer le risque d'erreurs tout en préservant cette information biologique précieuse [Gaspar *et al.* 2013]. Dans ce sens, plusieurs initiatives cherchent à définir des standards ou tout du moins des recommandations techniques [Knight *et al.* 2012] adaptées à des études cliniques [Sinha *et al.* 2015] ou environnementales [Ju & Zhang 2015]. L'utilisation d'échantillons artificiels (*mock communities*) permet une comparaison directe entre les résultats de l'analyse et la composition connue des organismes initialement présents dans l'échantillon, donnant ainsi une évaluation de la distorsion de l'image obtenue due aux biais techniques et analytiques [Brooks *et al.* 2015, Singer *et al.* 2016]. Cependant, ces communautés basées sur un nombre limité de bactéries connues et cultivables ne reflètent pas la réalité d'un microbiote complexe.

Malgré l'existence de plusieurs méthodes permettant de limiter les biais intervenant entre le prélèvement d'échantillon et l'obtention de séquences à analyser, le biologiste n'obtiendra jamais une image exacte de la composition de ses microbiotes d'intérêt ; il ne pourra pas tirer de conclusions sur les proportions d'organismes initiaux à partir des proportions de séquences identifiées *per se*. Cependant, les études métagénétiques ont en général pour but de comparer différents échantillons (cas/témoin dans l'étude d'une maladie, évolution temporelle d'une condition, …) Si tous les échantillons sont traités en utilisant les mêmes protocoles, séquencés sur le même run, et analysés avec les mêmes méthodes, ils partageront tout ou partie des biais dans les mêmes proportions, et pourront ainsi être comparés pour révéler des variations entre les conditions étudiées.

1.4 - Analyse primaire des données séquencées

L'objectif principal de l'analyse bioinformatique est de rattacher chaque lecture à son taxon d'origine, afin d'estimer la composition et la diversité de l'échantillon dont elle provient. L'analyse d'un jeu de données métagénétiques se découpe en plusieurs étapes, chacune étant une problématique bioinformatique bien distincte, à laquelle répondent de nombreuses approches algorithmiques différentes. Cette multiplicité rend difficile la maîtrise d'une analyse complète de métagénétique, le bioanalyste devant être capable d'évaluer toutes ces solutions afin de choisir celles qui seront les plus adaptées à ses jeux de données et à la question biologique posée.

Le développement de pipelines d'analyse a émergé dans le but d'alléger cette démarche : ces pipelines sont des surcouches logicielles reliant ces différentes étapes successives, en intégrant les algorithmes et banques de séquences jugés comme étant les plus adaptés à un contexte d'étude. Ces pipelines, souvent accompagnés de recommandations d'utilisation, offrent à l'utilisateur la facilité de lancer une seule commande exécutant toutes les étapes successives de l'analyse. Cependant, cette accessibilité et automatisation sont souvent privilégiées aux dépens d'un regard critique de l'utilisateur sur les méthodes intégrées. En effet, certains de ces pipelines sont actuellement des références dans la littérature par leur autorité historique qui les a rendus populaires. Ils sont de ce fait souvent un choix par défaut pour les bioanalystes, qui ne maîtrisent pas toujours les méthodes intégrées et n'ont pas conscience de leur applicabilité à leur propre contexte d'étude ni les biais que ces méthodes peuvent introduire dans les résultats. En outre, de nombreuses méthodes émergentes prometteuses sont peu utilisées car elles n'ont pas été développées initialement pour traiter des données métagénétiques, et n'ont jamais été validées dans ce contexte. Cette partie propose un tour d'horizon des différentes approches d'analyse métagénétique ainsi que les différentes solutions bioinformatiques existantes à chaque étape, et pouvant être intégrées dans un pipeline d'analyse.

Actuellement, les pipelines de métagénétique peuvent être séparés en deux catégories distinctes présentées en Figure 1.8 :

- Les pipelines *clustering-first* regroupent d'abord les lectures en clusters appelés OTUs (*Operational Taxonomic Units)* en se basant sur leur similarité, puis identifient leur taxonomie en comparant une séquence représentant chaque cluster à une banque de séquences annotées.
- Les pipelines *assignment-first* ont la démarche opposée : ils annotent tout d'abord toutes les lectures par leur taxonomie en les comparant à une banque de séquences de référence, puis ils les regroupent en clusters sur la base d'annotations partagées.

La vue d'ensemble proposée par la Figure 1.8 permet d'identifier certains avantages et inconvénients propres à chaque catégorie. Par exemple, les pipelines *clustering-first* sont capables de séparer les taxons qui ne peuvent pas être annotés (taxons vert et violet dans la Figure 1.8). Cette ségrégation permet ainsi d'identifier la diversité des organismes en présence, même s'ils ne sont pas identifiés. Ceci est impossible pour les pipelines *assignment-first* : sans annotation, ces lectures ne peuvent être départagées. À l'inverse, un inconvénient majeur des pipelines *clustering-first* est le seuil de similarité fixe définissant les OTUs, risquant de regrouper des lectures qui appartiennent à des taxons différents.

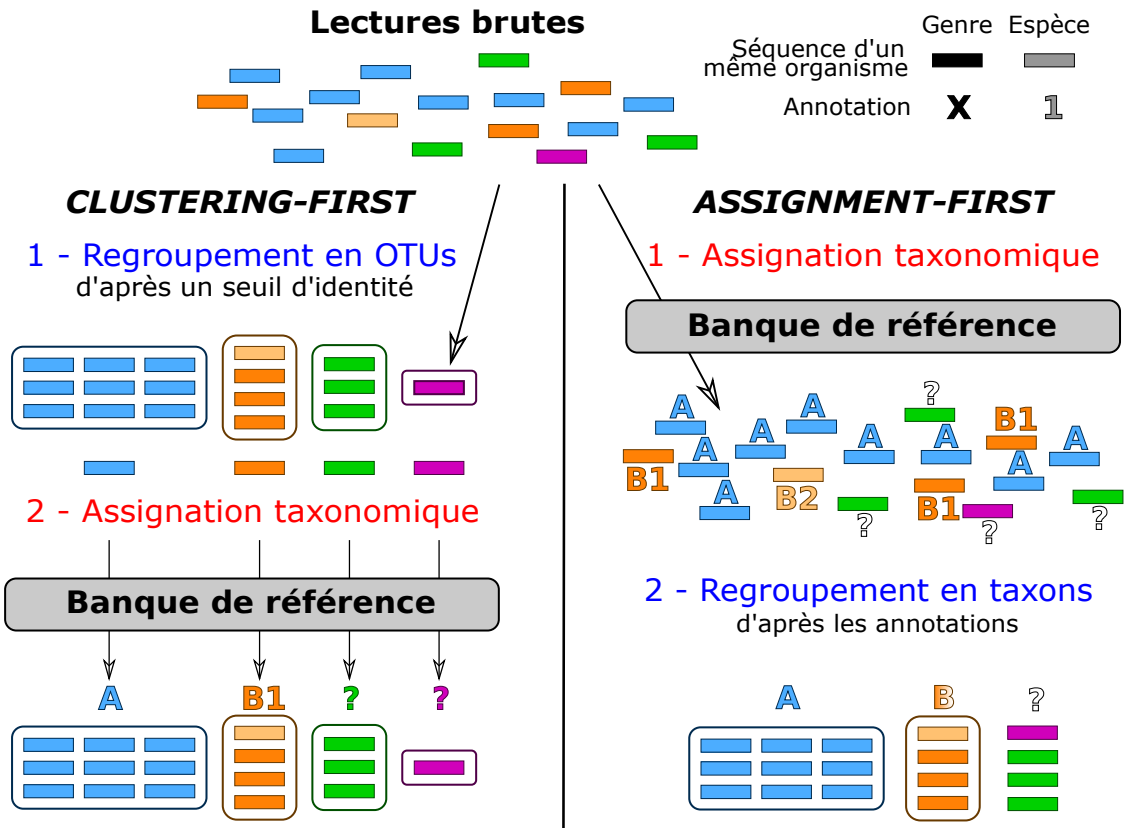


Figure 1.8 : Représentation simplifiée des catégories d'analyse clustering-first et assignment-first (adapté de Siegwald et al. 2017).

Par exemple dans la Figure 1.8, les séquences orange appartiennent au même genre B, mais concernent deux espèces différentes : B1 en orange foncé, B2 en orange clair. Imaginons que le *locus* d'intérêt ait une séquence très proche pour ces deux espèces : la similarité des lectures issues de B1 et B2 sera supérieure au seuil de définition des OTUs pour les pipelines *clustering-first*. Ces pipelines regrouperont les lectures des deux espèces dans le même cluster orange. Si le choix de la lecture sur laquelle baser l'annotation est celui de la lecture la plus abondante, ici B1, l'OTU sera annoté B1 et passera ainsi sous silence l'existence de B2 dans l'échantillon étudié. Cet écueil est évité par les méthodes *assignment-first* qui vont distinguer les deux espèces par leur annotation. Si leur similarité est trop proche pour distinguer B1 et B2, les méthodes *assignment-first* les regrouperont à un niveau taxonomique moins précis, dans cet exemple sous leur genre commun B.

1.4.1 Approches clustering-first

Les premiers outils d'analyse de données de métagénomique ciblée entrent dans la catégorie *clustering-first*, à laquelle appartiennent de nombreux pipelines majoritairement utilisés (tels que mothur [Schloss *et al.* 2009] et QIIME [Caporaso *et al.* 2010]). Ces pipelines adoptent différentes approches pour regrouper les séquences en OTUs [Navas-Molina *et al.* 2013], schématisées dans la Figure 1.9.

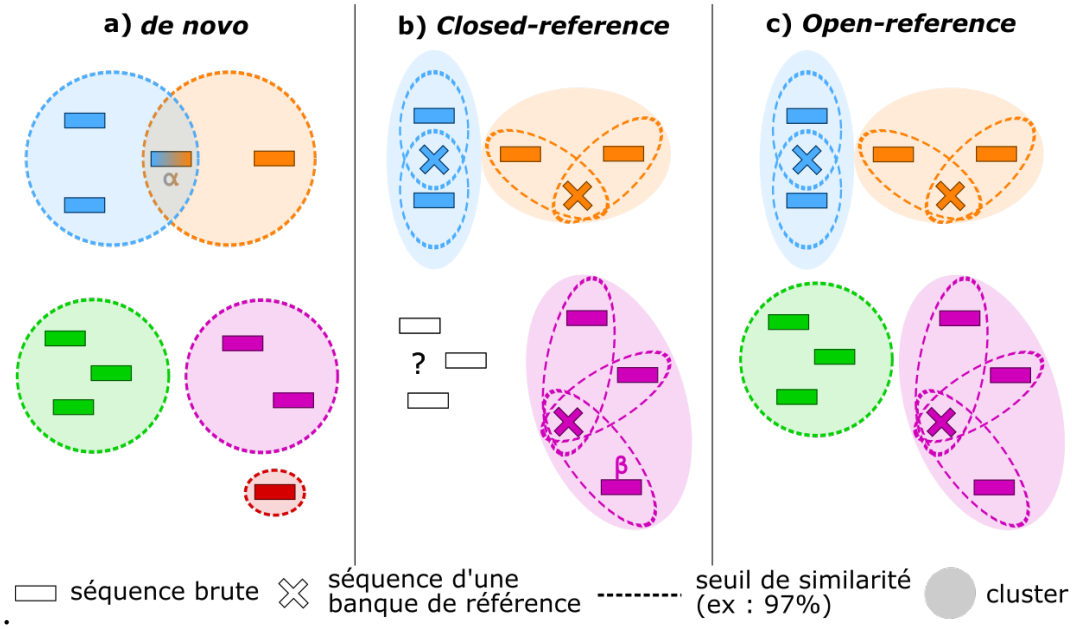


Figure 1.9 : Approches principales d'analyses clustering-first.

Le *clustering de novo* (Figure 1.9.a) est une approche qui se base uniquement sur les propriétés intrinsèques des lectures, en utilisant un *clustering* hiérarchique ou un *clustering* par centroïdes. DOTUR [Schloss & Handelsman 2005] est le premier logiciel ayant déployé une telle approche en transposant le *clustering* hiérarchique UPGMA (*Unweighted pair group method with arithmetic mean*) utilisé historiquement dans des analyses phylogénétiques (Figure 1.10.a).

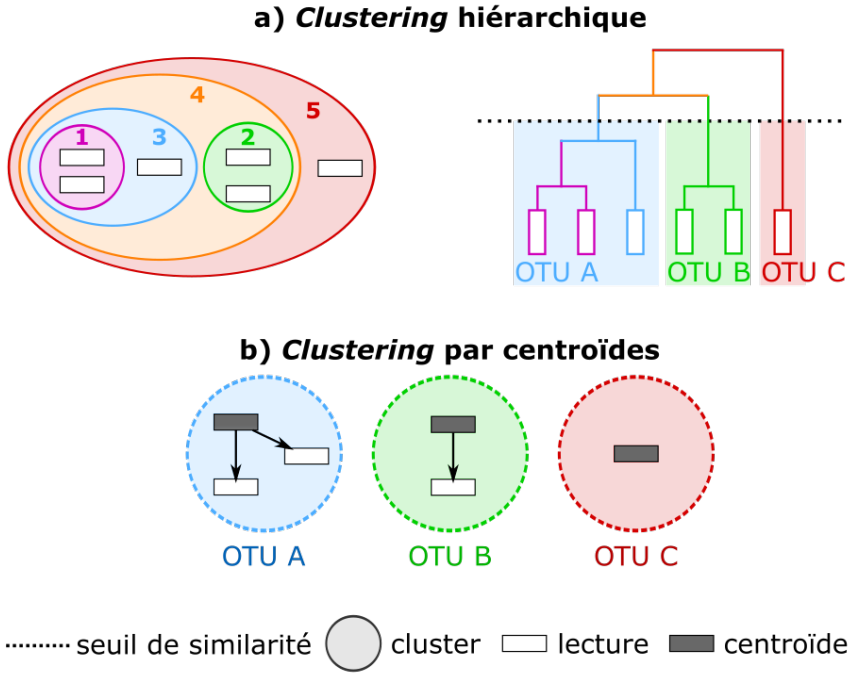


Figure 1.10 : Méthodes de clustering de novo des lectures en OTUs.

Les lectures sont tout d'abord alignées entre elles afin de générer une matrice de distances. Dans un contexte métagénétique, les alignements deux-à-deux (PSA) de type Needleman-Wunsch sont plus précis que les alignements multiples (MSA) [Sun *et al.* 2012]. En effet, les alignements multiples sont guidés par la présence de positions préservées comme marqueurs de conservation évolutive entre des séquences homologues, ce qui n'est pas le cas dans le contexte métagénétique où l'on évalue des régions les plus variables possibles entre des organismes potentiellement éloignés phylogénétiquement parlant. Le *clustering* hiérarchique va regrouper les lectures de manière itérative sur la base de cette matrice de distances. Au départ, chaque lecture est un cluster isolé. Cette matrice de distance va permettre de définir les clusters de lectures de manière ascendante. Le *clustering* hiérarchique va regrouper les clusters les plus proches en un seul cluster, de manière itérative jusqu'à ce qu'il n'y ait plus qu'un seul cluster contenant toutes les lectures (dans la Figure 1.10.a, les lectures sont regroupées du cluster n°1 au n°5). Ce *clustering* itératif peut être représenté sous la forme d'un dendrogramme. Les OTUs sont ensuite

définis en sélectionnant un seuil de pourcentage de similarité (par exemple 97 %), correspondant à un niveau du dendrogramme où chaque branche forme un OTU contenant des séquences similaires à au moins ce seuil (dans la Figure 1.10.a, le seuil en pointillés permet de découper l'arbre en trois OTUs).

De nouvelles méthodes de *clustering de novo* hiérarchique ont été dérivées d'approches phylogénétiques plus précises, comme le *clustering* par maximum de vraisemblance qui utilise un modèle d'évolution pour former les clusters.

Le regroupement des lectures en clusters pour aboutir à la construction du dendrogramme peut être effectué sur la définition de différentes distances entre clusters, dont les plus courantes sont représentées sur la Figure 1.11 :

- en simple lien (*single linkage*, Figure 1.11.a*)*, les clusters fusionnés à chaque étape sont ceux dont les lectures les plus proches de chaque cluster ont la plus faible distance entre elles. Cette méthode est sensible au bruit et aura plutôt tendance à sous-évaluer le nombre de clusters en les sur-agrégeant.
- en lien complet (*complete linkage*, Figure 1.11.b), les clusters fusionnés à chaque étape sont ceux dont les lectures les plus éloignées ont la plus faible distance entre elles. Cette méthode est sensible aux séquences éloignées et aura plutôt tendance à surévaluer le nombre de clusters.
- en lien moyen (*average linkage,* Figure 1.11.c), les clusters fusionnés à chaque étape sont ceux dont la moyenne de la distance entre toutes les paires de lectures est la plus faible. Cette méthode est intermédiaire entre les deux précédentes et est la plus robuste, ce qui en fait la méthode la plus utilisée.

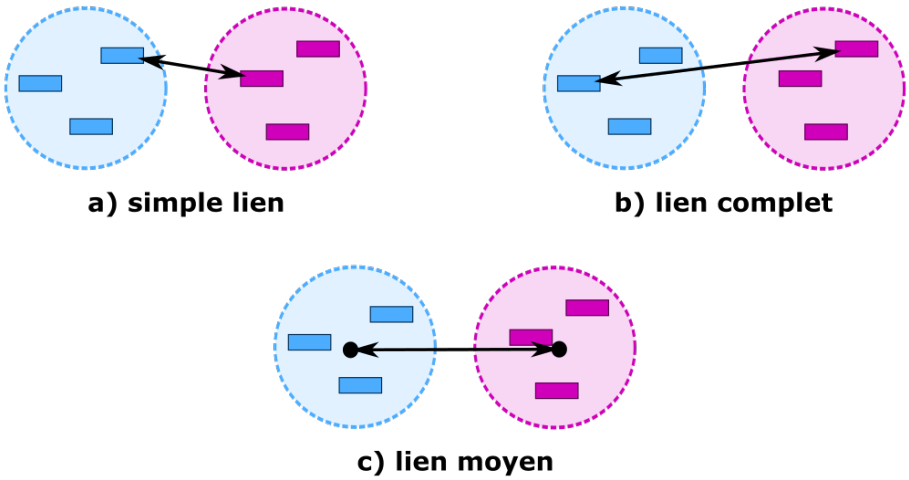


Figure 1.11 : Représentation des distances principales pouvant être évaluées entre deux clusters lors du clustering hiérarchique.

Le *clustering de novo* par approche hiérarchique a une complexité au moins quadratique, car il nécessite de comparer toutes les lectures entre elles avant de les agréger : ce goulot d'étranglement le rend difficilement applicables à de larges jeux de données, qui génèrent d'immenses matrices de distance. Les développeurs de ces méthodes recommandent ainsi fortement l'utilisation de logiciels de débruitage des lectures pour éliminer les erreurs potentielles qui augmentent la complexité des jeux de données (avec en contrepartie l'introduction de biais potentiels mentionnés dans le chapitre précédent). Le *clustering* hiérarchique est en outre souvent précédé d'une étape de dé-duplication des lectures ou de pré-*clustering* optimisé [Huse *et al.* 2010], pour réduire au maximum la quantité de lectures à comparer entre elles. Afin de réduire davantage la matrice de distances à générer, le logiciel ESPRIT [Sun *et al.* 2009] ne se base pas sur des distances entre lectures complètes, mais sur des distances entre mots (*k-mers*). ESPRIT permet ainsi d'éliminer les *k-mers* redondants entre séquences, mais a toujours une complexité quadratique.

Une approche heuristique du *clustering de novo* permet d'accélérer la construction des clusters : des programmes tels que UCLUST [Edgar 2010] et son équivalent libre VSEARCH [Rognes *et al.* 2016], CD-HIT [Fu *et al.* 2012] ou encore Sumaclust [Mercier *et al.* 2013] ont implémenté une méthode

heuristique de *clustering* glouton par centroïdes (Figure 1.10.b). Les lectures sont d'abord triées par longueur et/ou par abondance décroissante, partant de l'hypothèse que les séquences les plus longues et les plus abondantes contiennent un signal biologique fort. La première lecture de la liste est considérée comme étant le centroïde du premier cluster. La lecture suivante est comparée à ce centroïde : si leur identité est supérieure au seuil choisi pour définir un cluster (97 % dans notre exemple), alors la lecture est ajoutée au cluster existant. Sinon, elle devient le centroïde d'un nouveau cluster. Toutes les lectures sont ainsi comparées aux centroïdes définis au fur et à mesure. Cette méthode permet d'éviter la comparaison de toutes les séquences entre elles comme dans le cas du *clustering* hiérarchique. Néanmoins, elle est fortement dépendante de l'ordre des séquences : une lecture peut être assignée à un centroïde alors qu'elle présente plus de similarité avec un centroïde qui sera créé ultérieurement car plus loin dans la liste des séquences [Koeppel et Wu 2013, He *et al.* 2015]. Par exemple dans la Figure 1.9..a, la séquence α a 97 % d'identité avec une séquence bleue et une séquence orange. Cette séquence appartiendra ainsi soit au cluster bleu, soit au cluster orange selon l'ordre dans lequel les séquences sont traitées. Le *clustering* glouton par centroïdes (initialement créé pour accélérer la comparaison de séquences à des banques de référence) ne semble ainsi pas optimal pour révéler le partitionnement d'un jeu de données, car son choix de centroïdes est empirique et n'est pas ré-évalué au fur et à mesure du *clustering*.

Les auteurs d'ESPRIT ont développé une approche intermédiaire entre *clustering* hiérarchique et *clustering* par centroïdes, qui est ESPRIT-Tree [Cai & Sun 2011]. Cet algorithme utilise une première approche gloutonne et plusieurs optimisations heuristiques des différentes étapes du *clustering* hiérarchique, promettant d'atteindre une précision équivalente mais avec une complexité quasi linéaire. ESPRIT-Tree n'est toutefois pas encore intégré dans la plupart des pipelines populaires reposant sur un *clustering de novo*.

L'approche *closed-reference* (Figure 1.9.b) est similaire à l'approche *de novo* par centroïdes. Elle est toutefois supervisée, en utilisant non plus des lectures du jeu de données comme centroïdes, mais les séquences d'une banque de référence. Chaque lecture est ainsi comparée à tous les centroïdes, donc toutes les séquences de la banque, et est assignée au cluster dont le centroïde y est le plus similaire (ex : UCLUST_ref [Edgar 2010], SortMeRNA [Kopylova *et al.* 2012]). Les centroïdes sont plus robustes dans cette approche, car ils correspondent à une référence biologique sans être dépendants du jeu de données à partitionner. En outre, cette indépendance permet une comparaison directe des OTUs entre différents jeux de données sur la base des centroïdes communs, ce qui est impossible dans le cas d'un *clustering de novo*. Cette approche permet enfin une annotation taxonomique directe des lectures, chaque OTU pouvant être annoté avec l'annotation de sa séquence centroïde. À noter que ceci ne la classe pas dans la catégorie *assignment-first* car son *clustering* se base sur les similarités entre séquences, et non sur leur annotation (même si certaines approches hybrides utilisent les informations taxonomiques pour guider le *clustering* [Schloss & Westcott. 2011]).

La définition d'un cluster n'est pas la même dans l'approche *closed-reference* que dans l'approche *de novo* : dans l'approche *closed-reference*, les séquences d'un même cluster peuvent être à une plus grande distance l'une de l'autre qu'elles le sont de la séquence de référence. Par exemple dans la Figure 1.9.b, la séquence β est à 97 % de similarité de la séquence de référence violette. Pourtant elle est à une similarité inférieure des autres séquences violettes du même cluster, comme représenté dans la Figure 1.9.a où cette séquence est isolée (en rouge). Un autre inconvénient majeur de l'approche cl*osed-reference e*st d'être dépendante d'une banque de référence, ce qui induit un a *priori d*e connaissance. Dans le cas d'un environnement peu décrit dans les banques, des séquences non référencées (en blanc dans la Figure 1.9.b) ne pourront pas être assignées à un OTU. L'approche op*en-reference (*Figure 1.9.c) répond à ce problème en mélangeant les deux approches précédentes : on effectue tout d'abord une analyse cl*osed-reference*, puis une analyse d*e novo s*ur les

séquences qui ne s'alignent pas avec la banque de référence. Il faut néanmoins garder à l'esprit que les OTUs générés ne reposent pas sur les mêmes définitions, ce qui ne permet pas de les comparer entre eux (par exemple par une analyse phylogénétique complémentaire entre ces OTUs).

Toutes ces approches *clustering-first* nécessitent de fixer un seuil de similarité (abusivement appelé seuil d'identité) pour définir un OTU. Le seuil de 97 % s'est imposé comme un standard universel pour représenter différentes espèces bactériennes, sans correspondre à une réalité taxonomique. Ce seuil a été établi sur la base d'études d'hybridation ADN-ADN, sur lesquelles reposaient la définition de nouvelles espèces bactériennes, et la similarité des séquences d'ADNr correspondantes [Stackebrandt & Goebel 1994]. Ainsi, deux génomes sont définis comme espèces bactériennes différentes s'ils ont une valeur d'hybridation ADN-ADN inférieure à 70 %, ce qui équivaut à moins de 97 % d'identité entre leurs séquences d'ADNr 16S. Les auteurs de cette comparaison précisent que dans le cas où deux séquences ont plus de 97 % d'identité, seule une évaluation du taux d'hybridation entre génomes permet d'estimer leur degré d'homologie. La valeur de 97 % d'identité de séquence a été généralisée pour définir que deux séquences d'ADNr 16S appartiennent à la même espèce, négligeant ce dernier point de confirmation pourtant nécessaire. De même, ces conclusions ont été tirées de séquences d'ADNr 16S complètes, et non de fragments de gènes tels qu'étudiés en métagénétique. Différentes régions hypervariables de l'ADNr 16S présentant différents taux de variabilité [Baker *et al.* 2003], ce seuil fixe de 97 % ne peut être universel pour toutes les régions étudiées. Ce dernier est ainsi régulièrement remis en question, les OTUs générés sur ce critère ne pouvant pas être mis en corrélation avec un niveau taxonomique donné. En effet, les lignées bactériennes évoluant à différents rythmes, aucun seuil fixe ne peut représenter la séparation entre toutes les espèces bactériennes [Clarridge 2004, Koeppel & Wu 2013, Rossi-Tamisier *et al.* 2015]. Enfin, ce seuil ne peut être appliqué à l'étude d'autres cibles génomiques. Par exemple, dans des études métagénétiques fongiques, les régions ITS utilisées comme marqueurs taxonomiques varient entre 76 % et

99 % d'identité entre espèces [Nilsson *et al.* 2008]. De nouveaux algorithmes de *clustering de novo* cherchent à contrer cet écueil, en évitant l'utilisation d'un seuil fixe d'identité globale. Par exemple, SWARM [Mahé *et al.* 2014] se base sur un nombre maximal de différences locales entre amplicons, ce qui lui permet une formation d'OTUs plus fine. Ses résultats sont en outre indépendants de l'ordre des séquences d'entrée. Les OTUs générés ne peuvent toutefois pas être mis en corrélation avec un rang taxonomique précis.

Une fois les OTUs formés, une séquence représentative est sélectionnée pour chacun d'eux afin de les annoter en comparant cette séquence à une banque de référence, cette annotation étant étendue à toutes les lectures appartenant à l'OTU. Cette comparaison peut être effectuée par un alignement de séquences (de type BLAST), ou encore par une comparaison de *k-mers* (par exemple avec RDP Classifier [Wang *et al.* 2007]) dont le fonctionnement sera développé dans la partie suivante concernant les pipelines de la catégorie *assignment-first*. La séquence représentative pour chaque OTU peut être la plus longue (contenant le plus de sites informatifs), la plus représentée (contenant potentiellement moins d'erreurs), le centroïde de l'OTU s'il a été défini, ou un consensus entre toutes les lectures constitutives de l'OTU. Ce choix dépend fortement de la méthode d'assignation taxonomique envisagée, propre à chaque pipeline (par exemple, une séquence consensus n'est pas compatible avec une assignation taxonomique par BLAST, qui ne prend pas en compte les nucléotides dégénérés).

Le choix d'une banque de séquences de référence est crucial : cette banque doit être adaptée au *locus* cible d'intérêt, correctement annotée, aussi exhaustive que possible et doit suivre une taxonomie standardisée. Il existe trois banques principales de référence pour l'ADNr 16S bactérien [Santamaria *et al.* 2012], dont les caractéristiques sont résumées dans le Tableau 1.12.

	SILVA SSU Parc	**SILVA SSU Ref**	**Greengenes**	**RDP**
Version actuelle	128 (septembre 2016)		13.5 (mai 2013)	11.5 (septembre 2016)
Organismes	Bactéries, archées, eucaryotes		Bactéries, archées	Bactéries, archées
Origine des séquences	European Nucleotide Archive		Genbank	European Nucleotide Archive
Nombre de séquences	5 616 941	1 922 213	1 262 986	3 356 809
Taille minimale des séquences	300	1200 (bactéries/archées) 900 (eucaryotes)	1250	500
Sélection & validation des séquences	Alignement ≥ 50 % d'identité avec au moins une autre séquence de la banque	Alignement ≥ 70 % d'identité avec au moins une autre séquence de la banque	Score d'alignement positif avec au moins une autre séquence de la banque + élimination des séquences chimériques	Au moins 30 % de 7-mers partagés avec une autre séquence de la banque + score d'alignement positif sur un alignement de référence
Taxonomie	SILVA [Yilmaz *et al.* 2013]		Greengenes [McDonald *et al.* 2012]	RDP [Cole *et al.* 2014]
Licence	Utilisation gratuite académique / non-commerciale Licence payante non-académique / commerciale		Creative Commons BY-SA 3.0	Creative Commons BY-SA 3.0
Référence	[Quast *et al.* 2013]		[DeSantis *et al.* 2006]	[Cole *et al.* 2014]

Tableau 1.12 : Comparaison des trois principales banques de séquences d'ADN ribosomique.

SILVA [Quast *et al.* 2013] propose un ensemble de séquences alignées et annotées, tirées de l'ENA (European Nucleotide Archive), des gènes codant pour la grande et la petite sous-unité d'ADNr chez les bactéries, archées et eucaryotes. Deux versions des banques proposées par SILVA sont actuellement disponibles pour la petite sous-unité (SSU) de l'ADNr : SSU Parc contient plus de 5,5 millions de séquences, et SSU Ref contient une sélection curée d'un peu moins de 2 millions de séquences sélectionnées pour leur grande taille et haute qualité d'alignement. L'utilisation de cette dernière version permet une plus grande confiance dans les séquences de la banque (qui sont plus longues donc ayant plus de chances de couvrir le *locus* d'intérêt), aux dépens d'une certaine exhaustivité (SILVA contient de nombreuses séquences environnementales trop courtes pour être présentes dans SSU Ref, mais essentielles pour une estimation correcte de la diversité de tels milieux). SILVA met à jour sa banque en ajoutant incrémentalement les nouvelles séquences à l'alignement de la version existante, et à l'arbre taxonomique associé.

Greengenes [DeSantis *et al.* 2006] est l'équivalent américain de SILVA, proposant des séquences d'ADNr 16S uniquement issues de Genbank. Greengenes propose dans sa dernière version un peu plus d'un million de séquences de taille supérieure à 1 250 nucléotides. Chaque nouvelle version de la banque est accompagnée d'un nouvel alignement complet *de novo* de toutes ses séquences : cette approche permet de prendre en compte toutes les informations évolutives incluses dans les nouvelles séquences, mais est plus sensible à des séquences erronées ou de moins bonne qualité (d'où la non-inclusion de séquences courtes dans Greengenes). Greengenes a pour particularité de vérifier que chaque nouvelle séquence ajoutée à la banque n'est pas une séquence chimérique, puisqu'au moins 3 % des séquences publiques d'ADNr sont en réalité des séquences chimériques [Ashelford *et al.* 2005]. Il est important de noter que Greengenes n'a pas été mise à jour depuis 2013, et ne contient ainsi aucune séquence découverte au-delà de cette année.

Enfin, RDP (Ribosomal Database Project) [Cole *et al.* 2014] est une autre initiative américaine proposant dans sa version actuelle un peu plus de 3,3 millions de séquences d'ADNr 16S dont environ 90 000 d'archées. RDP a la particularité d'avoir un alignement et un arbre taxonomique de référence, basés sur un ensemble restreint de 10 000 séquences issues du séquençage de souches types. En outre, RDP utilise la structure secondaire des séquences d'ARN associées à ces séquences de référence pour guider l'alignement de nouvelles séquences ajoutées à la banque (en utilisant le logiciel Infernal [Nawrocki & Eddy 2013]). Toutefois, l'assignation taxonomique des séquences de cette banque est limitée au genre. À noter que RDP propose également depuis la version 11 une autre banque d'ADNr 28S fongique.

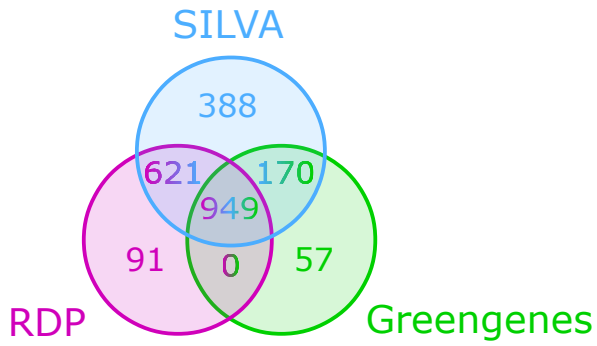


Figure 1.13 : Nombre de genres bactériens partagés entre SILVA, RDP et Greengenes (mars 2013, adapté de Ylmaz et al. 2014).

La Figure 1.13 compare les genres retrouvés dans ces trois banques [Yilmaz *et al.* 2014] : même si presque deux tiers des genres bactériens y sont communs, de nombreux taxons sont spécifiques à chaque banque. En effet, chaque banque utilise sa propre taxonomie et méthode pour classifier de nouvelles séquences, surtout pour les divisions dites « candidates ». Ces dernières sont des groupes taxonomiques sans souche type cultivable, souvent constituées de séquences environnementales. Leurs séquences d'ADNr 16S sont suffisamment divergentes des séquences de groupes taxonomiques existant pour être considérés comme de potentielles nouvelles branches taxonomiques. Ces dernières seront intégrées différemment à un arbre taxonomique bactérien

selon les banques considérées. Par exemple, certaines séquences de la division candidate NC10 du NCBI sont classées en tant que NC10 dans Greengenes (qui conserve la notion de division candidate pour les séquences environnementales), en tant que la division Nitrospirae pour SILVA (dont les curateurs estiment que ces séquences environnementales appartiennent à cette division), et en tant que la division Firmicutes dans RDP (ces séquences étant les plus proches de séquences souches type au niveau des Firmicutes). Ainsi, une même séquence pourra potentiellement être classée dans un groupe taxonomique différent selon la banque de référence utilisée pour l'identifier, surtout lorsque cette séquence est issue d'organismes peu décrits dans la littérature, dont la classification ne fait pas consensus.

La plupart des pipelines sont associés à une banque privilégiée : mothur recommande l'utilisation de SILVA (actuellement, version 119) pour l'alignement de séquences (jugeant que l'alignement de référence de SILVA est de bonne qualité) mais utilise les séquences des souches types de RDP pour la classification des séquences de référence des OTUs (ce jeu de données étant plus petit, la classification est plus rapide). À l'inverse, QIIME intègre la banque Greengenes (actuellement, version 13.8) et sa taxonomie associée. Ce choix est guidé par les développeurs des pipelines selon leurs méthodes algorithmiques et/ou des préférences personnelles. En effet, le développeur de mothur reproche à Greengenes son mauvais alignement dans les régions hypervariables ; mothur se basant sur l'intégration des lectures à une banque alignée pour guider le *clustering*, il est logique qu'il privilégie des séquences de la banque les mieux alignées possibles [Schloss 2009]. À l'inverse, une explication possible du choix de Greengenes pour QIIME est l'effort de ses développeurs pour intégrer des données et logiciels open-source (ce qui n'est pas le cas de SILVA). QIIME est en outre moins dépendant de la qualité de l'alignement de la banque, puisqu'il utilise l'approche par *k-mers* SortMeRNA pour le *clustering open-reference*, *closed-reference* ainsi que l'assignation taxonomique des OTUs.

Les approches *clustering-first* ne sont pas applicables dans un contexte de métagénomique WGS, où l'hétérogénéité des lectures (couvrant des génomes

complets) ne permet pas de les regrouper en OTUs. De nouvelles approches dites *assignment-first* ont ainsi dû être développées pour l'analyse taxonomique de tels jeux de données.

1.4.2 Approches assignment-first

Les approches *assignment-first* vont tout d'abord annoter individuellement chaque lecture, avant de les regrouper par taxon sur la base de ces annotations. Beaucoup de ces approches ne sont pas applicables à des données amplicon, comme par exemple l'assemblage de génomes complets adapté à des données métagénomiques (ex : GenoMeta [Davenport *et al.* 2012]), des méthodes utilisant un ensemble de différents marqueurs génétiques (ex : MetaPhyler [Liu *et al.* 2011]), ou se basant sur la prédiction de motifs protéiques (ex : CARMA [Krause *et al.* 2008], TreePhyler [Schreiber *et al.* 2010], Kaiju [Menzel *et al.* 2016]). D'autres approches *assignment-first* basées sur des comparaisons de motifs entre les lectures et des banques de séquences de référence peuvent être applicables à l'analyse de données métagénétiques, même si elles n'ont jamais été utilisées dans ce contexte. Enfin, il existe également des méthodes *assignment-first* qui s'appuient sur des approches phylogénétiques (ex : pplacer [Matsen *et al.* 2010]) ou de *machine learning* [Soueidan & Nikolski 2015] ; ces dernières approches n'ont pas été évaluées dans le contexte de cette thèse.

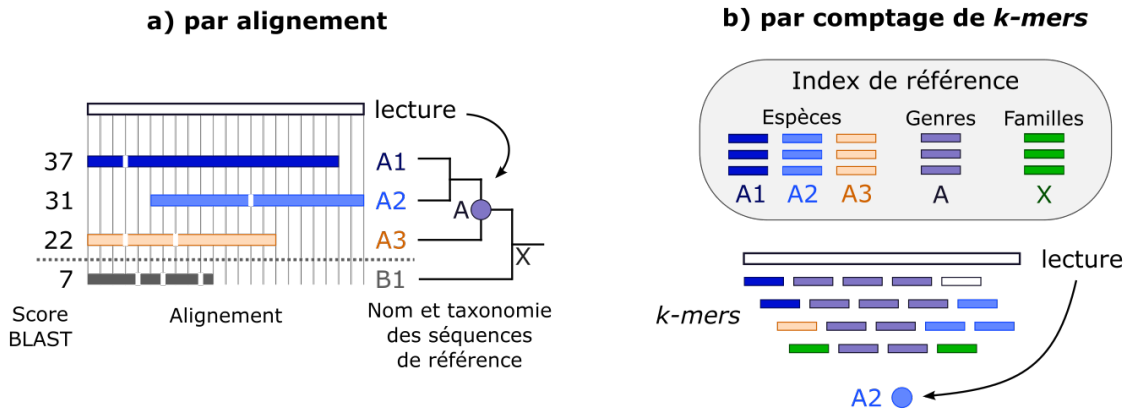


Figure 1.14 : Méthodes assignment-first utilisables en métagénétique.

La première approche *assignment-first* développée pour l'assignation taxonomique d'amplicons se base sur un alignement de chaque lecture sur une banque de séquences de référence (historiquement par BLAST [Altschul *et al.* 1990]). Ces alignements sont ensuite interprétés pour annoter chaque lecture par le taxon dont elle est le plus similaire (Figure 1.14.a). Une interprétation naïve des résultats d'alignement est de transférer sur chaque lecture l'annotation de son meilleur hit BLAST. Dans la Figure 1.14.a, la lecture serait ainsi assignée à l'espèce A1 avec laquelle elle présente le meilleur alignement en termes de score. Cette méthode ne prend pas en compte la possibilité d'une divergence entre une lecture (issue d'un génome inconnu par exemple) et les génomes présents dans la banque de référence, pouvant généraliser de fausses assignations taxonomiques trop précises. Par exemple, en regardant l'aspect des alignements de la Figure 1.14.a, la lecture pourrait tout aussi bien être assignée à l'espèce A2 – le score plus bas de l'alignement entre la lecture et A2 pouvant simplement être causé par le fait que la séquence A2 est tronquée dans la banque, et ne couvre pas le début de la lecture.

Une interprétation plus fine des alignements a été permise par l'algorithme LCA (*Lowest Common Ancestor*), introduit dans le logiciel MEGAN [Huson *et al.* 2007], et intégré par la suite dans de nombreux pipelines. Cet algorithme interprète, pour chaque lecture, une sélection de plusieurs hits BLAST validés comme étant significatifs sur la base de leur score. L'algorithme LCA assigne la lecture au taxon qui est le plus bas ancêtre commun parmi les hits significatifs. Dans la Figure 1.14.a, les hits significatifs sont les séquences des espèces A1, A2 et A3. La lecture est assignée à leur ancêtre commun, le genre A. Cette méthode permet une assignation plus sûre en prenant en compte l'incertitude due à l'évaluation d'alignements très proches. Néanmoins, elle se base également sur le seul score BLAST pour évaluer quels hits sont significatifs, ce qui peut être source d'erreurs. En effet, ce score est dépendant de la qualité et longueur des alignements, et n'est pas comparable d'une lecture à une autre. La méthode LCA a été affinée par d'autres logiciels : par exemple, MTR [Gori *et al.* 2011] utilise l'information taxonomique partagée entre les lectures à différents

niveaux taxonomiques, tandis que Sort-ITEMS [Monzoorul *et al.* 2009] affine l'assignation taxonomique en effectuant un BLAST réciproque entre le meilleur hit, la lecture et les hits significatifs. La principale limite de ces méthodes est leur dépendance d'une étape d'alignement, demandant beaucoup de temps de calcul pour de gros jeux de données. Par exemple, le pipeline MG-RAST [Meyer *et al.* 2008] basé sur une telle approche a été le premier pipeline d'analyse de données métagénomiques libre d'accès sur Internet dès 2008, proposant une interface d'analyse conviviale. Ce pipeline, automatisant les alignements et leur interprétation, souffre toutefois d'un délai important entre soumission des lectures et réception des résultats : en 2015, le temps d'attente médian était entre 7 et 10 jours pour un échantillon WGS, et 24h pour un échantillon amplicon [Wilke *et al.* 2016]. Ce délai conséquent dû au temps d'analyse est démultiplié par la popularité du pipeline, recevant actuellement 4 térapaires de base de séquences à analyser par mois, ce qui impose une file d'attente de plus en plus longue à ses utilisateurs.

De nouvelles méthodes *assignment-first* ont émergé ces dernières années pour pallier ce problème, en comparant les lectures à une banque de référence sans alignement. Elles se basent sur l'étude de la composition des lectures en *k-mers* pour retrouver des signatures spécifiques des génomes auxquels elles appartiennent (Figure 1.14.b). Par exemple, kraken [Wood *et al.* 2014] construit d'abord un index de tous les *k-mers* trouvés dans la banque de référence, et assigne à chacun d'eux l'ancêtre commun à tous les organismes contenant ce *k-mer*. Chaque lecture est alors elle-même découpée en *k-mers,* et comparée à cet index. Par exemple dans la Figure 1.14.b, la lecture partage le plus de *k-mers* avec le genre A, elle est ainsi au moins assignée à ce dernier. En évaluant les *k-mers* partagés entre la lecture et les espèces du genre A, il s'avère que la lecture a une majorité de *k-mers* similaires à l'espèce A2 : on peut ainsi préciser l'annotation en assignant la lecture à l'espèce A2. CLARK [Ounit *et al.* 2015] a amélioré cette approche en ne référençant que les *k-mers* discriminants entre taxons dans la banque à un niveau taxonomique donné, ce qui minimise le bruit induit par les *k-mers* communs entre taxons ainsi que la complexité de

recherche. Ces pipelines retournent en résultat une assignation taxonomique par lecture ; l'utilisateur doit ensuite regrouper les lectures en taxons pour générer une table d'OTUs, par un script fourni par le pipeline ou un logiciel tiers. Ces méthodes pouvant traiter plusieurs millions de lectures par minute sont bien plus rapides que les méthodes par alignement ; elles sont ainsi de plus en plus utilisées dans l'assignation taxonomique de données métagénomiques WGS. Elles sont même devenues la base du modèle économique de certains prestataires de service d'analyse : par exemple, One Codex (http://www.onecodex.com) ou encore Gaia (http://www.metagenomics.cloud) sont des pipelines commerciaux intégrant des algorithmes similaires sous une interface web ergonomique, rendant accessible l'analyse de grands jeux de données métagénomiques de manière rapide avec des résultats graphiques intuitifs. En contrepartie, ces pipelines ne sont pas open-source, peu paramétrables et dépendent souvent d'une banque de séquences propriétaire.

Le point limitant principal des pipelines *assignment-first* est de se baser sur une banque de séquences de référence, représentant un *a priori* de connaissance. L'assignation taxonomique des lectures étant limitée par les séquences de la banque et la qualité de leur annotation, une analyse *assignment-first* est impossible à appliquer sur des microbiotes peu décrits dans la littérature et peu représentés dans les banques de référence. En effet, les lectures de tels microbiotes seront annotées comme « non identifiées », n'ayant aucune séquence de référence proche dans les banques. Les pipelines *assignment-first* étant développés pour des données WGS, leur banque de référence doit couvrir un maximum de séquences annotées pour un maximum d'organismes, tous *loci* confondus. L'INSDC (International Nucleotide Sequence Database Collaboration) [Cochrane *et al.* 2016] est actuellement la plus large source de séquences nucléotidiques annotées, regroupant des informations de l'ENA (European Nucleotide Archive) [Toribio *et al.* 2016], Genbank du NCBI [Clark *et al.* 2016] et DDBJ (DNA Data Bank of Japan) [Mashima *et al.* 2016]. Ces banques sont certes exhaustives, mais beaucoup trop hétérogènes en termes de qualité et redondantes pour être utilisables par un pipeline *assignment-first*.

Ces banques ont avant tout une vocation d'archive, et non de référence. En effet, elles contiennent des entrées directement soumises par les utilisateurs, sans validation ni correction ; leurs séquences peuvent être erronées, tout comme leurs annotations. RefSeq [O'Leary *et al.* 2016], sous-ensemble de Genbank, est actuellement la banque la plus utilisée pour les pipelines *assignment-first* basés sur la comparaison de *k-mers*. Cette banque est non-redondante, contient des entrées validées, et dont l'annotation est standardisée : RefSeq s'appuie sur la taxonomie définie par le NCBI, curée manuellement à chaque ajout de séquence dans la banque. Certains pipelines *assignment-first* utilisent d'autres banques nucléotidiques. Par exemple, One Codex propose l'utilisation de sa banque propriétaire One Codex 28k, qui enrichit RefSeq d'une sélection d'environ 20 000 génomes issus de Genbank.

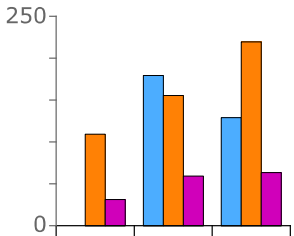
L'inconvénient d'utiliser de telles banques est leur taille, pouvant atteindre plusieurs dizaines de gigaoctets. Certains pipelines nécessitant de conserver toute la banque en mémoire ne peuvent ainsi être déployés sur des ordinateurs de bureau. Pour répondre à ce problème, les auteurs de kraken proposent une banque alternative, Minikraken, contenant 10 000 *k-mers* sélectionnés comme étant représentatifs de la banque RefSeq. Les utilisateurs de pipelines *assignment-first* peuvent également utiliser leur propre banque de référence, à condition qu'elle soit compatible avec la taxonomie du NCBI. Ce dernier point rend difficile l'utilisation de banques spécifiques dans l'application de ces pipelines à un contexte métagénétique, car les banques associées sont souvent régies par leur propre taxonomie (voir Tableau 1.12). Les pipelines *assignment-first* pourraient tout de même être utilisés sur des données de métagénétique, puisque les banques de référence utilisées par ces pipelines contiennent aussi des séquences d'ADNr 16S. Une telle utilisation n'a toutefois jamais été décrite dans la littérature.

1.5 - Normalisation des tables de comptages

À l'issue de l'analyse primaire, le bioanalyste obtient une table initiale d'observations (Figure 1.15.a), aussi appelée table de comptages ou table d'OTUs, contenant le nombre de lectures attribuées à chaque OTU pour chaque échantillon. Malgré une normalisation de la quantité d'ADN entre échantillons lors de la préparation des librairies, le nombre total de lectures issues du séquençage peut varier d'un échantillon à l'autre (à cause de biais d'amplification par exemple). Comparer directement deux échantillons avec un nombre total de lectures différentes risque de révéler des différences de proportions d'OTUs qui sont dues à cette variation de profondeur de séquençage, et non à une différence biologique réelle. Par exemple dans la Figure 1.15.a, l'OTU 2 semble moins abondant dans l'échantillon A que dans l'échantillon C. Cette différence est uniquement due au fait que l'échantillon A a un nombre total de lectures bien plus faible (8000) que l'échantillon C (20000). Normaliser les comptages entre échantillons est ainsi essentiel pour pouvoir les comparer. Cette partie présente différentes méthodes de normalisation couramment utilisées en métagénétique, et les potentiels biais associés.

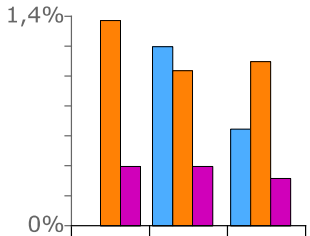
1.5.1 – Méthodes intuitives de normalisation

a) Table initiale d'observations



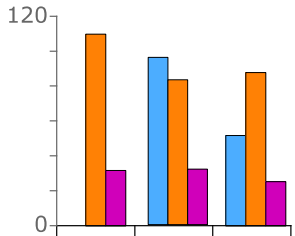
	A	B	C
OTU1	0	180	130
OTU2	110	156	220
OTU3	32	60	64
⋮	⋮	⋮	⋮
Total	8k	15k	20k

b) Normalisation TSS



	A	B	C
OTU1	0	1,2	0,65
OTU2	1,375	1,04	1,1
OTU3	0,4	0,4	0,32
⋮	⋮	⋮	⋮
Total	100%	100%	100%

c) Normalisation par raréfaction



	A	B	C
OTU1	0	96	52
OTU2	110	83,2	88
OTU3	32	32	25,6
⋮	⋮	⋮	⋮
Total	8k	8k	8k

Figure 1.15 : Méthodes intuitives de normalisation.

La méthode de normalisation la plus intuitive est d'utiliser non pas les abondances absolues, mais les fréquences des OTUs entre échantillons, en divisant chaque observation par le nombre total de lectures de l'échantillon (méthode TSS, *total-sum normalization,* Figure 1.15.b). Transformer ces valeurs en proportions relatives n'est toutefois pas statistiquement robuste : lors d'une analyse comparative, les échantillons normalisés auront tendance à se regrouper par profondeur de séquençage initiale bien plus que par similarité biologique, puisque les OTUs minoritaires seront tellement minimisés qu'ils n'auront plus de poids dans la comparaison entre échantillons [Friedman *et al.* 2012]. La méthode actuellement la plus populaire est de normaliser la table de comptages par raréfaction, c'est-à-dire en réduisant chaque échantillon à un même nombre de séquences (souvent choisi par le nombre de lectures total de l'échantillon le plus petit, Figure 1.15.c). Cette méthode a été fortement discréditée par McMurdie & Holmes (2013), qui affirment que « *raréfier des données de comptage biologique est statistiquement inadmissible* » [McMurdie & Holmes 2013]. En effet, raréfier les données équivaut à éliminer une grande quantité d'observations, ce qui diminue fortement la puissance statistique des données restantes, notamment pour les OTUs qui ont un nombre de lectures très faibles.

1.5.2 – Méthodes dérivées de la transcriptomique

Évaluer la composition différentielle entre échantillons en métagénétique est un problème comparable à celui d'évaluer une expression différentielle entre échantillons en transcriptomique. Dans les deux cas, les comparaisons se font sur la base d'une table de comptages de lectures : des méthodes de normalisation ont ainsi été directement transposées de l'analyse de données transcriptomiques vers l'analyse de données métagénétiques. Ces méthodes se basent en général sur le calcul d'un facteur d'échelle propre à chaque échantillon, par lequel sera multipliée chaque observation pour la corriger. Par exemple, comme représenté dans la Figure 1.16, DESeq [Anders & Huber 2010] (et son successeur DESeq2 [Love *et al.* 2014]) construisent un

échantillon représentatif moyen, contenant la moyenne géométrique du nombre de lectures par OTU sur tous les échantillons (Figure 1.16.a). Dans la table de comptages initiale, chaque observation est alors divisée par son équivalent dans l'échantillon moyen (Figure 1.16.b). Pour chaque échantillon, DESeq calcule ensuite la médiane de ces ratios, qui sera le facteur d'échelle par lequel diviser chaque observation (Figure 1.16.c). La normalisation par DESeq pose toutefois problème sur des observations nulles, incompatibles avec le calcul d'une moyenne géométrique. Les tables de comptages d'OTUs sont dites clairsemées : elles contiennent de nombreuses valeurs nulles, représentatives d'OTUs absents de certains échantillons, ou présents en trop faible quantité pour être détectés. Ceci force l'utilisateur de DESeq à modifier ses données, en remplaçant manuellement les valeurs nulles par des valeurs très faibles, afin de pouvoir appliquer cette normalisation (par exemple dans la Figure 1.16.a, le nombre de lectures de l'OTU1 dans l'échantillon A a été ajusté à 1 au lieu de 0).

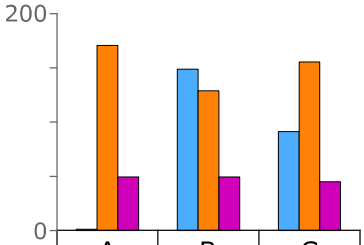
a) Table initiale d'observations

	A	B	C	Échantillon moyen
OTU1	1	180	130	28,60
OTU2	110	156	220	155,71
OTU3	32	60	64	49,72
⋮	⋮	⋮	⋮	⋮

b) Ratios des observations sur l'échantillon moyen

	A	B	C
OTU1	0,03	6,29	4,55
OTU2	0,71	1,00	1,41
OTU3	0,64	1,21	1,27
⋮	⋮	⋮	⋮
Médiane	0,64	1,21	1,41

c) Mise à l'échelle des observations par la médiane des ratios



	A	B	C
OTU1	1,55	149,15	92,01
OTU2	170,90	129,26	155,71
OTU3	49,72	49,72	45,30
⋮	⋮	⋮	⋮

Figure 1.16 : Normalisation par DESeq.

CSS (*cumulative sum scaling*) [Paulson *et al.* 2013] est une autre méthode adaptée de l'analyse de données transcriptomiques, qui est plus

adaptée à des données clairsemées. CSS part de l'hypothèse que la majorité des observations ne sont pas différentiellement représentées d'un échantillon à un autre. Ces observations devraient ainsi être présentes en proportions égales dans tous les échantillons : elles peuvent de ce fait être utilisées comme référence pour normaliser les données. CSS utilise ainsi pour facteur d'échelle un certain pourcentage des observations, estimé en fonction de la distribution des comptages des lectures dans les échantillons, et correspondant à la proportion de lectures pour laquelle cette distribution ne varie pas d'un échantillon à un autre (Figure 1.17.b). CSS divise ensuite chaque valeur de la table de comptages par ce quantile, correspondant à la somme cumulée du nombre de lectures jusqu'à ce pourcentage (Figure 1.17.c).

a) Table initiale d'observations

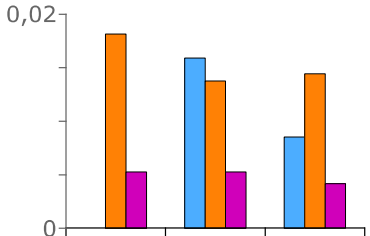
	A	B	C
OTU1	0	180	130
OTU2	110	156	220
OTU3	32	60	64
⋮	⋮	⋮	⋮
Total	8k	15k	20k

b) Calcul du facteur d'échelle

Pour chaque échantillon, le facteur d'échelle est la somme cumulée des observations jusqu'à ce que cette somme dépasse le centile prédit (par exemple arbitraire ici, 75%)

	A	B	C
75% du total	6000	11250	15000
Facteur d'échelle	6050	11300	15200

c) Mise à l'échelle des observations



	A	B	C
OTU1	0	0,015	0,009
OTU2	0,018	0,014	0,014
OTU3	0,005	0,005	0,004
⋮	⋮	⋮	⋮

Figure 1.17 : Normalisation par CSS

Comme vu sur l'exemple simplifié de données normalisées dans les Figures 1.15 à 1.17, la plupart des méthodes de normalisation donnent des résultats similaires si les variations de proportions entre OTUs sont importantes. La normalisation par raréfaction est une méthode plus adaptée à une forte variation de profondeur de séquençage entre échantillons ou dans

l'étude des OTUs rares, et ce malgré la perte d'information qu'elle provoque [Weiss *et al.* 2015]. DESeq et CSS permettent de conserver toute l'information biologique ; néanmoins, il se peut que ces méthodes ne réussissent pas à correctement normaliser les échantillons de faible profondeur de séquençage, ce qui peut parasiter les analyses statistiques secondaires. Dans tous les cas, il est recommandé de vérifier lors des analyses statistiques secondaires que la profondeur de séquençage initiale de chaque échantillon ne soit pas un facteur de confusion [Weiss *et al.* 2015].

La normalisation des données peut également être envisagée lors du montage du plan d'expérience, par une autre approche méthodologique directement inspirée des études transcriptomiques. L'ajout dans chaque échantillon d'un ADN exogène de quantité connue (*spike-in*) permet d'évaluer sa variation de proportions dans les tables d'observations finales [Stämmler *et al.* 2016], et donc de les normaliser en conséquences. Cette approche s'appuie toutefois sur la seule hypothèse que cet ADN exogène soit soumis aux mêmes biais techniques et analytiques que l'ADN endogène.

1.6 - Analyse secondaire

L'analyse secondaire s'appuie sur la table de comptages normalisée pour répondre aux questions biologiques posées lors du montage du plan d'expérience. Cette partie présente les différentes métriques et outils statistiques utilisables par le bioanalyste dans ce but, tout en soulignant les écueils propres à chaque méthode.

1.6.1 Alpha, beta et gamma-diversité

Comparer la composition en organismes de différents écosystèmes est une problématique fondamentale en écologie, pour mesurer les variations entre différentes populations d'organismes (par exemple pour comparer des populations d'animaux en différents points géographiques). Cette problématique est transposable à des études métagénétiques, dont le but est souvent de mesurer la variation de microbiotes entre des conditions différentes

(eau saine/polluée, patient malade/sain, ...) Ainsi, les indicateurs biologiques utilisés par les écologistes pour décrire et comparer la diversité spécifique de différents écosystèmes ont été transposés aux études métagénétiques. Robert Harding Whittaker a introduit en 1960 les termes d'alpha-diversité (α), beta-diversité (β) et gamma-diversité (γ) pour décrire la diversité d'espèces observées dans un environnement sur différentes échelles (Figure 1.18). L'alpha-diversité représente la diversité locale d'un échantillon, la beta-diversité représente la moyenne des différences de diversité entre échantillons, et la gamma-diversité représente l'alpha-diversité totale sur l'union des échantillons. Les trois mesures sont réunies par la formule suivante : $\beta = \frac{\gamma}{\alpha}$

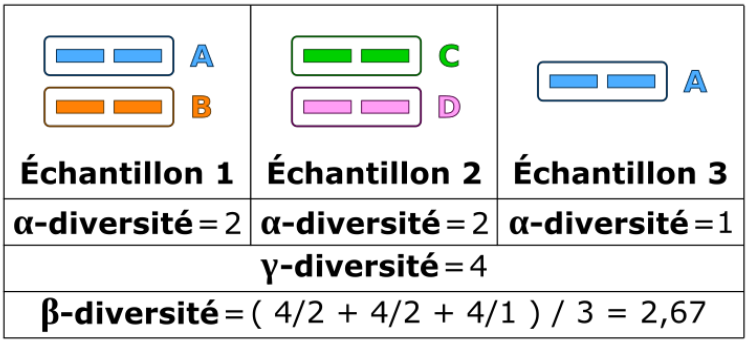


Figure 1.18 : Alpha, beta et gamma-diversité d'un exemple simplifié de 3 échantillons. Ici, l'alpha-diversité correspond au nombre d'OTUs par échantillon. La gamma-diversité totale est de 4 (il y a 4 OTUs distincts dans l'ensemble des échantillons). Puisque $\beta = \frac{\gamma}{\alpha}$ *, la beta-diversité est de 2 pour les échantillons 1 et 2, et de 4 pour l'échantillon 3. La beta-diversité totale est la moyenne des beta-diversités, égale à 2,67.*

1.6.2 Alpha-diversité

L'alpha-diversité peut être calculée pour chaque échantillon à partir de la table de comptages, et peut être déclinée en deux types de métriques : la richesse (comptage des différents OTUs), et la diversité proprement dite (comptage des différents OTUs prenant en compte leurs proportions). La richesse est un comptage des différents OTUs représentés, sans prendre en compte leur abondance relative. Ce nombre d'OTUs peut être sous-estimé par rapport à la richesse réelle du microbiote, par exemple si la profondeur de séquençage n'a pas été suffisante pour capter tous les organismes en présence.

Ce phénomène peut être observé par une courbe de raréfaction (Figure 1.19) : cette courbe est établie en sous-échantillonnant un échantillon à différents intervalles de profondeur (en abscisse), et en comptant le nombre d'OTUs représentés dans chaque sous-échantillon (en ordonnée).

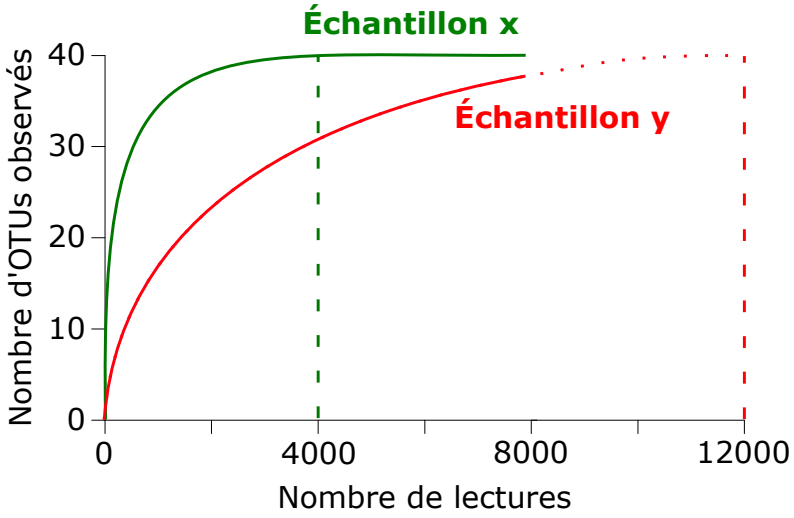


Figure 1.19 : Courbes de raréfaction de deux échantillons. L'échantillon x atteint une asymptote, et a de ce fait une profondeur de séquençage suffisante. L'échantillon y n'atteint pas d'asymptote même à profondeur maximale (8 000 lectures).

La pente de la courbe de raréfaction représente, pour une profondeur donnée, la probabilité d'obtenir de nouveaux OTUs en augmentant la profondeur de séquençage. Cette pente est logiquement très forte pour les plus petits sous-échantillons, ces derniers n'ayant pas une profondeur suffisante pour couvrir tous les organismes en présence. Idéalement, à profondeur de séquençage suffisante, la courbe de raréfaction atteint une asymptote qui représente la richesse réelle de l'échantillon (la pente de la courbe est nulle, puisqu'augmenter la profondeur de séquençage ne permet pas de découvrir de nouveaux OTUs). Par exemple, dans la Figure 1.19, les échantillons x et y sont séquencés à une profondeur de 8 000 lectures. La courbe de raréfaction de l'échantillon x atteint son asymptote dès 4 000 lectures : on peut estimer que les 40 OTUs observés sont proches du nombre d'OTUs réels, puisqu'augmenter la profondeur de séquençage (de 4 000 à 8 000 lectures) n'augmente pas le

nombre d'OTUs observés. À l'inverse, la courbe de raréfaction de l'échantillon y n'atteint pas d'asymptote au nombre maximal de lectures (35 OTUs observés pour 8 000 lectures). Le nombre d'OTUs observés dans l'échantillon est vraisemblablement sous-estimé par rapport au nombre d'OTUs réels. Par extrapolation (prolongation de la courbe en pointillés, en se basant sur l'hypothèse que l'échantillon y contient le même nombre d'OTUs que l'échantillon x), on estime qu'il aurait fallu un séquençage d'au moins 12000 lectures pour couvrir l'ensemble des 40 OTUs présents dans l'échantillon y.

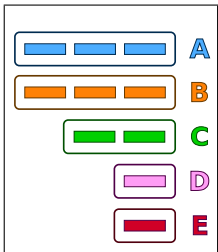
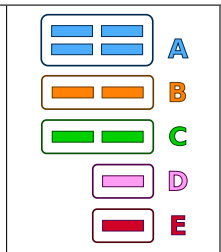
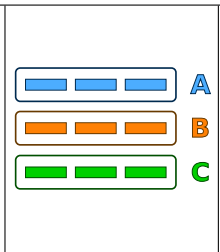
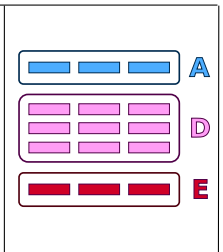
		 **Échantillon 1**	 **Échantillon 2**	 **Échantillon 3**	 **Échantillon 4**
RICHESSE	**Nombre d'OTUs**	5	5	3	3
	Chao1	5,5	5,33	3	3
	ACE	6,25	6,53	3	3
DIVERSITÉ	**Shannon**	2,17	2,12	1,58	1,37
	Simpson	0,24	0,26	0,33	0,44
	Simpson inverse	4,17	3,85	3,03	2,27

Figure 1.20 : Calcul de différents indices de richesse et de diversité pour quatre échantillons.

Différentes métriques de richesse comme Chao1 [Chao, 1984] ou ACE (*abundance-based coverage estimators*) [Chao, 2004] permettent de corriger le nombre d'OTUs observés en y ajoutant une estimation de la proportion d'OTUs rares non couverts. Cette estimation se base sur les fréquences des OTUs rares, partant du principe qu'une forte abondance d'OTUs rares signifie que de nombreux autres OTUs rares n'ont pas été séquencés. Par exemple dans la Figure 1.20, l'échantillon 1 contient 2 OTUs (D et E) contenant une seule lecture.

Ces observations révèlent qu'il existe sans doute d'autres OTUs rares dans l'échantillon qui n'ont pas été séquencés : les indices de richesse Chao1 et ACE vont ainsi estimer une richesse plus élevée que le nombre d'OTUs observés. Gotelli & Chao détaillent dans leur revue l'historique et la construction de telles métriques [Gotelli & Chao 2013]. Chao1 corrige le nombre d'OTUs observés de la façon suivante :

$$Chao1 = n_{total} + \frac{n_1^2}{2n_2} \quad \text{(A)}$$

$$Chao1 = n_{total} + \frac{n_1(n_1-1)}{2(n_2+1)} \quad \text{(B)}$$

où n_{total} est le nombre total d'OTUs de l'échantillon, n_1 est le nombre d'OTUs composés d'une seule lecture (singletons), et n_2 est le nombre d'OTUs composés de deux lectures (doubletons). Différentes formules dérivées de la formule (A) ont été proposées par Colwell & Chao, selon le nombre de singletons et de doubletons présents dans l'échantillon. Par exemple, la formule (B) est une version corrigée applicable dans le cas où n_2 est nul.

ACE est un autre indice de richesse corrigeant le nombre d'OTUs observés en considérant que les OTUs rares sont ceux ayant un nombre de lectures inférieur à 10 sur l'ensemble des échantillons (choix empirique) :

$$ACE = N_{>10} + \frac{N_{\leq 10}}{C} + \frac{f_1}{C} + \gamma$$

$$\text{où} \quad C = 1 - \frac{f_1}{\sum_{i=1}^{10} i f_i}$$

$$\text{et} \quad \gamma = max\left\{ \frac{N_{\leq 10}}{C} \frac{\sum_{i=1}^{10} i(i-1) f_i}{(\sum_{i=1}^{10} i f_i)(\sum_{i=1}^{10} i f_i - 1)} - 1{,}0 \right\}$$

où $N_{\leq 10}$ est le nombre d'OTUs dans l'échantillon considérés comme rares, c'est-

à-dire dont la somme des lectures sur l'ensemble de la table de comptages est inférieure ou égale à 10 (dans la Figure 1.20, ces OTUs sont B, C et E). $N_{>10}$ est le nombre d'OTUs restant dans l'échantillon, qui ne sont pas considérés comme rares (A et D dans la Figure 1.20). f_i est le nombre d'OTUs contenant i lectures dans l'échantillon. À noter qu'ACE n'est pas calculable lorsque tous les OTUs rares sont des singletons (C=0).

Deux échantillons ayant une richesse similaire peuvent tout de même présenter une forte variabilité de composition : par exemple dans la Figure 1.20, les échantillons 3 et 4 ont une richesse identique, et pourtant présentent des variations de proportions d'OTUs. Des mesures de diversité permettent de prendre en compte ces variations de proportions entre les deux échantillons. La métrique de diversité la plus populaire est l'entropie de Shannon [Shannon 1948], calculée de la façon suivante :

$$Shannon = -\sum_{i=1}^{N} p_i \log_2 p_i$$

où N est le nombre total d'OTUs, et p_i est la fréquence de l'OTU i dans l'échantillon. Cette entropie est tirée de la théorie de l'information, où on estime que plus une source émet d'informations différentes, plus son entropie est élevée. On peut ainsi estimer qu'une valeur élevée de l'indice de Shannon représente un microbiote à richesse élevée et dont les OTUs ont des proportions similaires (dans la Figure 1.20, l'échantillon 3 a une entropie de Shannon plus élevée que l'échantillon 4, car les proportions de ses OTUs sont plus proches entre elles. Par contre l'entropie de Shannon de l'échantillon 1 est plus élevée, car sa richesse est plus importante). Un autre indice de diversité couramment utilisé est l'indice de Simpson [Simpson 1949] qui calcule la probabilité que deux lectures tirées indépendamment d'un échantillon appartiennent au même OTU. L'indice de Simpson se calcule de la façon suivante :

$$Simpson = \sum_{i=1}^{N} p_i^2$$

où N est le nombre total d'OTUs, et p_i est la fréquence de l'OTU i dans

l'échantillon. Plus cette probabilité est élevée, plus la diversité est faible. Pour rendre l'interprétation plus intuitive, on utilise généralement l'indice de Simpson inverse. L'indice de Shannon accorde plus d'importance à des OTUs rares, tandis que l'indice de Simpson est plus impacté par les OTUs majoritaires [Krebs 2014]. Il existe enfin la diversité phylogénétique (PD, *phylogenetic diversity*) qui prend en compte la phylogénie des OTUs pour évaluer leur diversité sur un arbre phylogénétique, en partant du principe que deux OTUs phylogénétiquement proches auront moins de poids dans le calcul de diversité. Cette mesure de diversité est la plus proche de la composition biologique du microbiote, mais nécessite le calcul d'un arbre phylogénétique entre tous les OTUs, ce qui ajoute une étape chronophage à l'analyse. Par exemple, on peut généraliser l'indice de Shannon pour y inclure la distance phylogénétique entre OTUs [Allen *et al.* 2009] :

$$\textit{Shannon phylogénétique} = -\sum_{i=1}^{T} L_i a_i \log_2 a_i$$

où T est le nombre de branches dans l'arbre L_i est la longueur de la branche i dans l'arbre, et a_i est la somme des fréquences de tous les OTUs descendant de la branche i. Une fois ces indices d'alpha-diversité calculés pour tous les échantillons, un test statistique non-paramétrique (les observations par échantillon ne suivant pas une distribution normale [Wagner *et al.* 2011]) peut être utilisé pour estimer s'il existe une différence significative de richesse et/ou de diversité entre différents groupes d'échantillons (par exemple, Mann-Whitney pour la comparaison de deux conditions, ou Kruskal-Wallis pour la comparaison de plusieurs conditions).

L'utilisation d'indices de diversité apporte plus d'informations que les indices de richesse, car elle permet la prise en compte des proportions relatives des OTUs dans chaque échantillon. Elle est toutefois moins intuitive à interpréter : Pallmann *et al.* donnent l'exemple d'un échantillon contenant 100 OTUs aux proportions identiques. La disparition de 50 OTUs (soit la moitié) entraîne une diminution de l'indice de Simpson inverse de 0,99 à 0,98 ; cette

variation n'est ainsi pas représentative de l'ampleur du phénomène biologique associé [Pallmann *et al.* 2012].

1.6.3 Beta-diversité

La beta-diversité permet d'estimer la différence de diversité inter-échantillons. La beta-diversité globale sur l'ensemble des échantillons peut être calculée comme étant la moyenne de beta-diversité entre toutes les paires d'échantillons [Izsak & Price, 2001]. Il existe de nombreux indices de beta-diversité référencés [Lozupone & Knight 2008] ; ceux présentés ci-dessous sont les plus couramment rencontrés dans la littérature. Sur la base de la définition de Whittaker, la beta-diversité entre deux échantillons peut être calculée par l'indice de Jaccard [Jaccard 1902] de la façon suivante :

$$d_{Jaccard} = \frac{b+c}{a+b+c} = 1 - \frac{a}{(a+b+c)}$$

où a est le nombre d'OTUs partagés entre les deux échantillons, b est le nombre d'OTUs spécifiques au premier échantillon, et c est le nombre d'OTUs spécifiques au deuxième échantillon. Cet indice va de de 0 (les deux échantillons partagent les mêmes OTUs) à 1 (les deux échantillons n'ont aucun OTU en commun). L'indice de dissimilarité de Bray Curtis [Bray & Curtis, 1957] ajoute à l'indice de Jaccard l'information de proportions d'OTUs, de la façon suivante :

$$d_{Bray-Curtis} = \frac{\sum_{i=1}^{N} |p_{iA} - p_{iB}|}{\sum_{i=1}^{N} (p_{iA} + p_{iB})}$$

où p_{iA} et p_{iB} sont les abondances relatives de l'OTU i dans l'échantillon A et B respectivement.

Tout comme pour l'alpha-diversité, d'autres indices de beta-diversité comme la distance UniFrac [Lozupone & Knight 2005] prennent en compte la phylogénie des OTUs. Suite à la construction d'un arbre phylogénétique de tous les OTUs, la distance UniFrac entre deux échantillons se base sur les longueurs

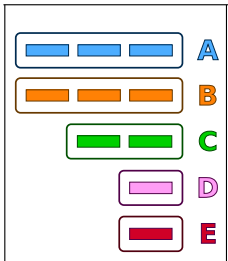
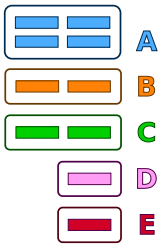
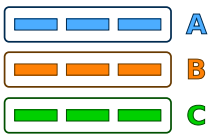
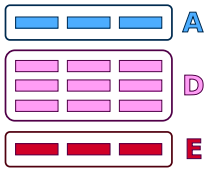
des branches de l'arbre partagées entre ces deux échantillons. La distance UniFrac non-pondérée (*unweighted UniFrac*) ne considère que la présence/absence des OTUs de la façon suivante :

$$d_{UniFrac} = \frac{unique}{totale}$$

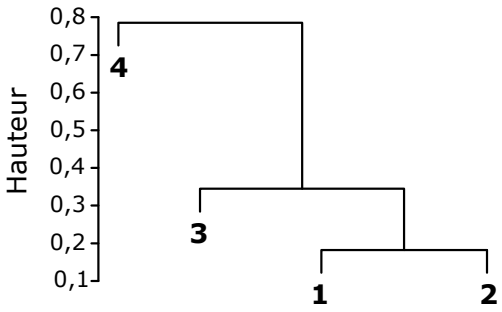
où *unique* est la somme des longueurs des branches propres à un seul échantillon, tandis que *totale* est la somme de toutes les longueurs des branches de ce même échantillon. Dans le cas où deux échantillons partagent tous leurs OTUs, *unique*=0, donc $d_{UniFrac}$=0. À l'autre extrême, lorsque deux échantillons ne partagent aucun OTU, *unique=totale=1*, donc $d_{UniFrac}$=1. La distance UniFrac non-pondérée accorde plus d'importance aux OTUs rares. La distance UniFrac pondérée (*weighted UniFrac*) ajoute à chaque branche un poids selon l'abondance relative des OTUs correspondant dans chaque échantillon ; elle accorde plus d'importance aux OTUs abondants. Une distance UniFrac unifiée a été proposée par Chen *et al.* en 2012, corrigeant les limitations des distances UniFrac pondérée et non-pondérée.

Un calcul de beta-diversité entre toutes les paires d'échantillons d'une étude métagénétique permet de générer une matrice de distances qui peut être interprétée par *clustering* hiérarchique ou par analyse en coordonnées principales (ACoP), afin d'identifier les échantillons partageant les mêmes profils de diversité. Par exemple, la Figure 1.21 reprend l'exemple de la Figure 1.20 : la beta-diversité calculée (a) entre chaque paire d'échantillons permet, par *clustering* hiérarchique (b) et/ou par ACoP (c), de constater que l'échantillon 4 a un profil très différent des autres échantillons, tandis que les échantillons 1 et 2 sont les plus proches. L'ACoP peut également permettre de représenter les variables associées à chaque échantillon pour évaluer leur impact.

a) Indices de dissimilarité de Bray-Curtis

	 Échantillon 1	 Échantillon 2	 Échantillon 3	 Échantillon 4
Échantillon 1	0			
Échantillon 2	0,100	0		
Échantillon 3	0,158	0,263	0	
Échantillon 4	0,600	0,600	0,750	0

b) Clustering hiérarchique



c) Analyse en coordonnées principales

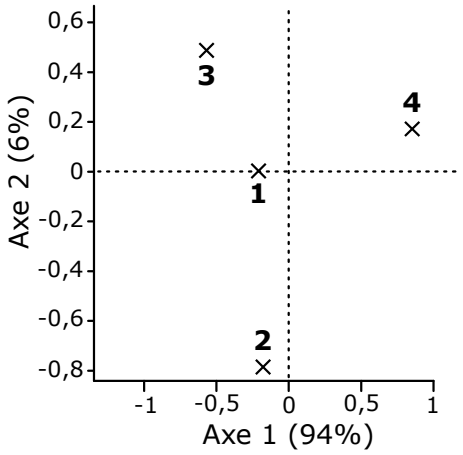


Figure 1.21 : a) Calcul des indices de dissimilarité de Bray-Curtis (beta-diversité) entre chaque paire d'échantillons. b) Clustering hiérarchique moyen entre échantillons à partir de ces indices. c) Analyse en coordonnées principales à partir de ces indices.

1.6.4 Tests statistiques de différences de composition

Les comparaisons de mesures d'alpha et beta-diversité entre échantillons permettent d'évaluer quels échantillons semblent différents les uns des autres ; elles ne permettent toutefois pas de déterminer quels sont les OTUs responsables d'une telle différence. Une analyse statistique plus fine est

nécessaire pour évaluer la nature des variations entre échantillons, et leur corrélation avec les variables biologiques du plan d'expérience. Ce dernier est souvent monté pour répondre à plusieurs questions biologiques, et pour étudier simultanément l'impact de plusieurs variables (dans le contexte d'études cliniques par exemple). Cette complexité des données nécessite des méthodes d'analyse statistique multivariée adaptées à leur nature clairsemée, rendant impossible l'application de tests statistiques paramétriques standards. En outre, le nombre d'observations par variable biologique étudiée est souvent réduit à cause de difficultés d'échantillonnage ou de plans d'expériences pas assez robustes. Enfin, la normalisation des données par raréfaction rend leur structure compositionnelle, ce qui empêche leur analyse statistique par des méthodes conventionnelles [Mandal *et al.* 2015 ; Tsilimigras & Fodor 2016]. DESeq et CSS contournent ce problème en appliquant une transformation logarithmique sur les valeurs normalisées, suivant la recommandation d'Aitchison [Aitchison, 1984].

De nombreuses méthodes statistiques adaptées à ces données ont été développées pour évaluer les différences de composition entre groupes d'échantillons, selon différentes variables, telles que Metastats [White *et al.* 2009], DESeq [Anders & Huber 2010], ou encore metagenomeSeq [Paulson *et al.* 2013]. Ces tests permettent d'estimer la significativité de la variation du nombre de lectures attribuées à un OTU dans deux échantillons ou deux groupes d'échantillons différents. La comparaison de ces méthodes sort du contexte de cette thèse ; il est néanmoins important de souligner l'importance de l'interprétation des résultats qu'elles rendent. En effet, il est essentiel de différencier les résultats liés à une réelle variation biologique entre deux groupes, des résultats statistiquement significatifs sans corrélation avec un phénomène biologique. L'interprétation des résultats se fait souvent sur la seule base de la p-valeur du test, couramment définie comme étant la probabilité d'obtenir les mêmes observations si l'hypothèse nulle du test est vraie (dans le contexte métagénétique, l'hypothèse nulle est en général l'absence de variabilité pour un OTU entre deux groupes d'échantillons). Le test est dit significatif si

cette p-valeur est inférieure à un certain seuil, communément fixé de façon arbitraire à 0,05. L'American Statistical Association a publié en 2016 une déclaration destinée à mettre en garde les scientifiques sur l'interprétation abusive de cette p-valeur comme seul critère de validation [Wasserstein & Lazar 2016]. Un test statistiquement significatif n'implique pas que la différence observée est biologiquement pertinente ; il est indispensable de prendre en compte les tailles d'effet et les intervalles de confiance associés pour évaluer son importance [Nuzzo 2014].

L'utilisation de ces nouvelles méthodes d'analyse statistique et leur interprétation est difficile d'accès pour un bioanalyste n'ayant pas de solides connaissances en statistiques adaptées à ce type bien particulier de données, ni dans les langages de programmation associés. Le logiciel STAMP [Parks *et al.* 2014] répond à ce problème en proposant une interface graphique intuitive et des recommandations d'analyses statistiques adaptées à la nature des données. Il permet la comparaison d'échantillons deux à deux, entre deux groupes, ou multi-groupes, et propose plusieurs représentations graphiques affichant les tailles d'effet des variables étudiées. L'utilisation de tels logiciels peut toutefois être un frein à la reproductibilité des analyses en rendant leur traçabilité moins automatisable [Santori 2016] Une telle accessibilité à de nombreuses méthodes d'analyse statistique peut également pousser le bioanalyste à la tentation du « p-hacking », définie par Nuzzo comme étant l'application de différentes méthodes jusqu'à l'obtention d'un résultat désiré [Nuzzo 2014].

1.7 - Bilan

Une étude métagénétique se déroule en de nombreuses étapes de biologie humide (prélèvement, préparation des échantillons et séquençage) et de biologie sèche (analyse bioinformatique primaire pour aller des lectures à une table de comptages, analyse statistique secondaire pour interpréter ces résultats). Comme indiqué en début de chapitre, ces différentes étapes doivent être prises en compte dès la mise en place du plan d'expérience : pour une même étude, l'utilisation de différentes solutions techniques et/ou d'analyse

peuvent aboutir à des résultats biologiques drastiquement différents [Hiergeist *et al.* 2016]. Le plan d'expérience doit être monté sur la base de questions biologiques claires qui guideront le choix des différentes méthodes d'analyse utilisées, et le bioanalyste doit avoir une connaissance suffisante des biais qui y sont associés, afin de pouvoir valider ou infirmer ses hypothèses en connaissance de cause et de façon pertinente.

Chapitre 2 - Expertise du pipeline d'analyse métagénétique développé sur la plate-forme PEGASE-biosciences

Lors de l'acquisition d'un séquenceur de paillasse Ion Torrent PGM par la plate-forme PEGASE-biosciences en 2011, les lectures générées par cette technologie étaient beaucoup plus courtes qu'actuellement (de l'ordre de la centaine de nucléotides). Aucun pipeline d'analyses métagénétiques publié n'avait été développé à l'époque pour prendre en compte ce profil de données bien particulier (lectures courtes et riches en erreurs). Par conséquent, l'équipe bioinformatique PEGASE-biosciences a développé un pipeline d'analyse *clustering-first* propre à la plate-forme et adapté à de telles données. Ce pipeline, nommé pipeline PEGASE v1, a été établi sur la base de recommandations formulées par Jünemann *et al.* pour l'analyse de données métagénétiques Ion Torrent PGM [Jünemann *et al.* 2012].

Le premier projet de cette thèse a consisté à expertiser ce pipeline PEGASE v1, dans un court délai, afin d'appréhender les méthodes d'analyse intégrées et les problématiques associées. Cette expertise a permis l'amélioration de ce pipeline vers une nouvelle version, le pipeline PEGASE v2, utilisable pour les analyses routinières des projets de métagénétique de la plate-forme [Loywick *et al.* 2014]. Cette première expertise d'une méthode d'analyse a engendré la nécessité de développer un protocole formel d'évaluation de pipelines d'analyses métagénétiques.

2.1 - Évaluation du pipeline PEGASE v1 sur une communauté microbienne artificielle générée *in vitro*

2.1.1 - Description de la communauté microbienne artificielle et du jeu de données test_bio associé

Un pipeline d'analyse métagénétique doit être évalué sur sa sensibilité, autrement dit sa capacité à identifier un maximum de lectures présentes dans le

jeu de données de départ, et sa spécificité, c'est-à-dire la justesse de cette identification. Un pipeline peu sensible risque de négliger des organismes en présence qui peuvent avoir un intérêt biologique dans les conclusions de l'étude menée, tandis qu'un pipeline peu spécifique risque de fausser ces conclusions par l'identification d'organismes qui ne sont en réalité pas présents dans l'échantillon.

L'évaluation de la qualité des résultats d'un pipeline ne peut se faire sans référence absolue, permettant d'estimer objectivement les différences entre la composition du microbiote déterminée en résultat du pipeline, et la composition réelle des jeux de données de référence avant analyse. De nombreuses études utilisent comme référence une communauté microbienne artificielle (*mock community,* [Bowers *et al.* 2015, Brooks *et al.* 2015]). Ces communautés sont composées de mélanges *in vitro* de cellules ou génomes d'organismes choisis dans des proportions prédéterminées, permettant de maîtriser la composition initiale du microbiote étudié.

La plate-forme avait à disposition les données de séquençage d'un tel échantillon artificiel simplifié, créé *in vitro* au laboratoire pour un projet antérieur. Cet échantillon contenait un mélange de génomes de 5 espèces bactériennes dans les proportions connues (Tableau 2.1), prenant en compte le nombre de copies d'ADNr 16S pour chaque organisme. L'échantillon contenait également de l'ADN d'organismes eucaryotes (plantes, animaux et champignons), mais cet ADN n'a pas été pris en compte dans l'évaluation du pipeline PEGASE, qui avait été développé pour les études métagénétiques bactériennes. Cet échantillon artificiel a été séquencé de façon bidirectionnelle par Ion Torrent PGM (voir Chapitre 1, Section 1.2.3, Figure 1.4), en ciblant l'amplicon 400(V4-V5) pour les bactéries, bordé par les amorces 519F et 907R (Lane, 1991; Stubner, 2002) qui encadrent les régions hypervariables V4 et V5 de l'ADNr 16S (~400 nt). Les autres organismes eucaryotes ont également été ciblés et séquencés (région 18S pour l'ADN de plantes et animaux, région ITS2 pour les champignons). Le jeu de données issu de ce séquençage est le jeu de

données *test_bio*, comportant des lectures de la cible 400(V4-V5) bactérienne d'intérêt (taille moyenne des lectures : 200 nt) ainsi que des lectures d'autres amplicons issus des organismes eucaryotes. Ce jeu de données a été utilisé pour évaluer le pipeline PEGASE v1.

Espèce	**Pourcentage de génomes bactériens**
Bacillus subtilis	80 %
Bordetella pertussis	5 %
Escherichia coli	5 %
Salmonella typhimurium	5 %
Yersinia pseudotuberculosis	5 %

Tableau 2.1 : Proportions des génomes de 5 différentes espèces bactériennes dans le jeu de données test_bio.

2.1.2 - Présentation du pipeline PEGASE v1 et de ses performances sur le jeu de données artificiel test_bio

La Figure 2.2 représente les différentes étapes du pipeline PEGASE v1, intégrées dans un workflow Galaxy [Afgan *et* al. 2016] permettant d'automatiser son exécution via une interface graphique. L'objectif de l'évaluation de ce pipeline était de valider rapidement les différentes étapes et leur paramétrage sur ce jeu de données *test_bio*. Le pipeline a dans un premier temps été évalué en utilisant une banque de séquences réduite (appelée GOLD), contenant uniquement les amplicons de chaque espèce bactérienne présente dans *test_bio*, séquencés au laboratoire par technique de Sanger. Cette solution permet d'accélérer les tests, de pouvoir relancer un grand nombre de fois l'exécutions de différentes étapes avec différents paramétrages, et de réduire la complexité de leur interprétation.

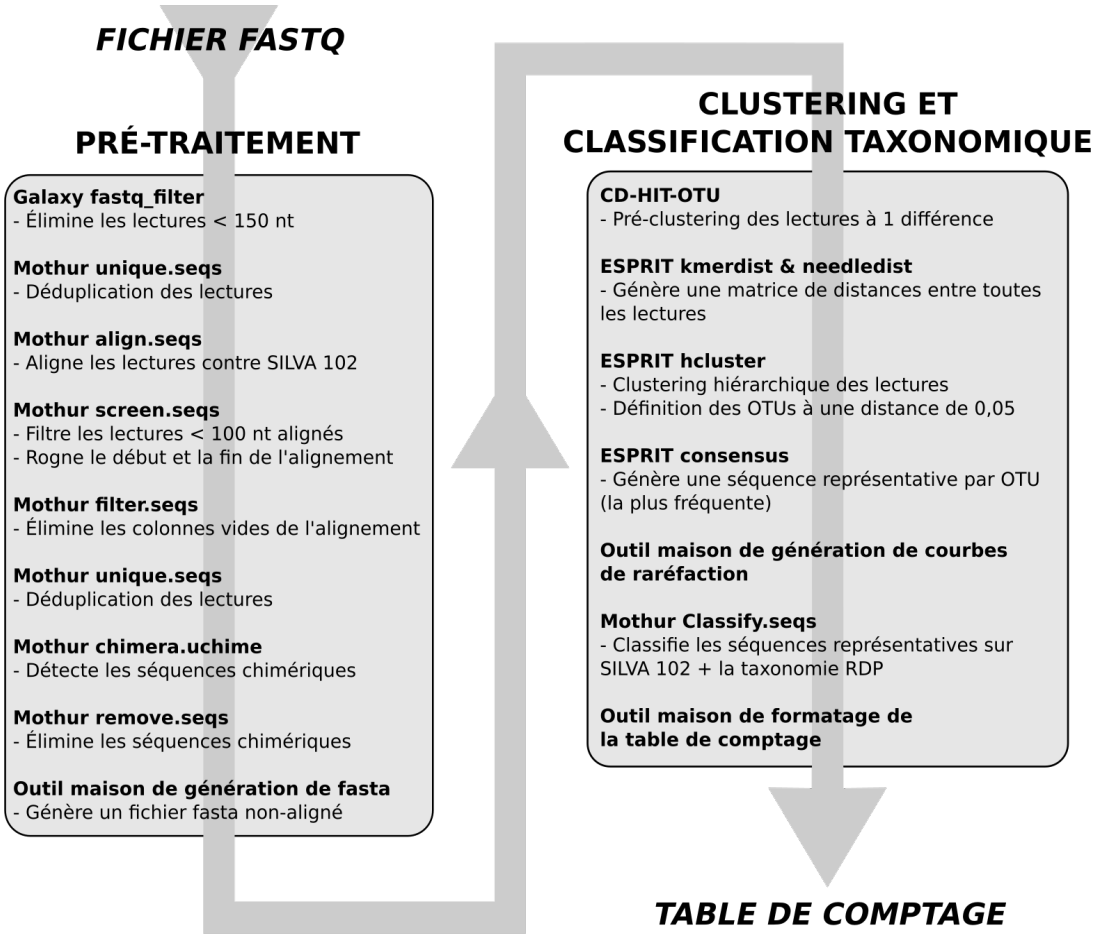


Figure 2.2 : Pipeline d'analyse métagénétique PEGASE v1 (pour un échantillon).

Le pipeline PEGASE v1 se déroule en deux étapes principales : l'étape de pré-traitement des lectures, puis l'étape de *clustering* en OTUs et de classification taxonomique. L'étape de pré-traitement des lectures, utilisant mothur ainsi que des scripts maison, filtre les données brutes afin de minimiser les erreurs générées par le séquençage Ion Torrent PGM. Le script Galaxy fastq_filter élimine les lectures plus courtes que 150 nucléotides et/ou contenant de longs homopolymères (> 10 nt). Les lectures sont ensuite dédupliquées (mothur unique.seqs) et alignées sur la banque GOLD (mothur align.seqs), et celles dont l'alignement est inférieur à 100 nucléotides (probablement des contaminants ou des lectures de trop basse qualité) sont éliminées (mothur screen.seqs). Ce dernier outil rogne également le début et la

fin de l'alignement pour conserver uniquement les positions couvertes à au moins 90 % des lectures. L'outil de détection de chimères UCHIME [Edgar *et al.* 2011] intégré à mothur est également utilisé par défaut pour éliminer les séquences chimériques. Une fois les lectures filtrées, celles-ci sont à nouveau dédupliquées pour éliminer les éventuelles lectures identiques après rognage de l'alignement. La quantité de lectures éliminées à chaque étape du pré-traitement du jeu de données *test_bio* est référencée dans le Tableau 2.3.

	Nombre de lectures restantes	**Pourcentage cumulé de lectures éliminées**
Données initiales	71 468	
Galaxy fastq_filter	55 030	23 %
mothur unique.seqs	44 863	37 %
mothur align.seqs / screen.seqs	13 986	80 %
mothur unique.seqs	11 985	83 %
mothur chimera.uchime	11 985	83 %

Tableau 2.3 : Impact des différentes étapes de pré-traitement sur la quantité de lectures initiales de l'échantillon test_bio.

La grande quantité de lectures éliminées ici est normale : seules les séquences bactériennes alignées sur la banque GOLD ont été conservées (les lectures issues des amplicons eucaryotes sont ainsi éliminées). En outre, les lectures dupliquées sont supprimées (leur quantité est sauvegardée dans un fichier intermédiaire afin de pouvoir être prise en compte dans les résultats). À noter qu'aucune lecture chimérique n'a ici été détectée.

Une fois les lectures filtrées, elles sont regroupées en OTUs dans la deuxième étape du pipeline. Le pipeline privilégie une approche de *clustering de novo* par *clustering* hiérarchique, solution plus exacte qu'un *clustering* glouton par centroïdes, mais aussi la plus gourmande en temps et en mémoire (voir Chapitre 1, Section 1.4.1). Afin de réduire la quantité de lectures à partitionner, et donc les temps d'analyse, une étape de pré-*clustering* a été intégrée afin de regrouper les lectures n'ayant qu'un nucléotide de différence

avant l'étape réelle de *clustering*. Le logiciel choisi pour l'étape de pré-*clustering* est CD-HIT-OTU [Li *et al.* 2012], lui-même un pipeline automatisant l'exécution successive de différents modules du logiciel CD-HIT. CD-HIT-OTU débute par un nouveau filtrage sur les lectures, avec un profil de paramétrage adapté au pyroséquençage 454, dont les données sont jugées proches du séquençage Ion Torrent PGM. Les lectures de longueur supérieure à la longueur médiane des lectures sont tronquées à cette longueur médiane (afin d'éliminer les nucléotides de trop basse qualité en 3' pour les longues lectures). Les lectures dont la taille est inférieure à 80 % de la longueur médiane des lectures sont éliminées. CD-HIT-OTU contient une étape d'élimination des chimères qui a été désactivée, car elle était déjà intégrée dans le pipeline PEGASE lors du pré-traitement des lectures. CD-HIT-OTU effectue ensuite une déduplication des lectures avant d'y appliquer un pré-*clustering* à un seuil donné. Ce seuil, par défaut à 97 %, a été modifié de manière arbitraire par les développeurs du pipeline PEGASE pour regrouper des lectures séparées d'une différence ponctuelle maximum. Enfin, les OTUs ne contenant qu'une seule lecture (appelés singletons) sont éliminés. La quantité de lectures éliminées à chaque étape du pré-*clustering* du jeu de données filtré *test_bio* est référencée dans le Tableau 2.4.

	Nombre de lectures restantes	**Pourcentage cumulé de lectures éliminées**
Données après pré-traitement	11 985	
Filtrage	6 824	43 %
Déduplication	6 640	45 %
Pré-clustering	4 113	66 %
Élimination des singletons	836	93 %

Tableau 2.4 : Impact des différentes étapes du pré-clustering CD-HIT-OTU sur la quantité de lectures pré-traitées de l'échantillon test_bio.

L'évaluation de cette étape de pré-*clustering* a révélé sa mauvaise intégration au pipeline PEGASE v1. En effet, sur ce jeu de données, le pré-*clustering* élimine de nombreux singletons (3277 lectures). Cette étape fait sens

lorsque CD-HIT-OTU est utilisé comme méthode de génération d'OTUs à un seuil de similarité plus faible (typiquement 97 %) ; les singletons représentent dans ce cas les lectures trop divergentes et/ou trop sous-représentées pour être validées, et il convient ainsi de les éliminer. Par contre, l'utilisation de CD-HIT-OTU comme étape de pré-*clustering* n'est pas adaptée : un seuil de *clustering* bien plus stringent (tel que celui utilisé dans le pipeline PEGASE v1) génère logiquement de nombreux singletons, qui sont simplement des lectures isolées qui n'ont pas été regroupées par le pré-*clustering*. Ces lectures ne doivent en aucun cas être éliminées.

Une mise en garde est également à émettre sur le rognage des lectures en 3' effectué au début de CD-HIT-OTU, dans sa propre étape de pré-traitement. Ce rognage a été mis en place pour éliminer les nucléotides de fin de lecture, qui sont généralement de moins bonne qualité pour l'Ion Torrent PGM. Toutefois, le séquençage de l'échantillon *test_bio* a été effectué sur une librairie bidirectionnelle, afin d'obtenir un séquençage de bonne qualité du début et de la fin de l'amplicon (voir Chapitre 1, Section 1.2.3, Figure 1.4). Durant le premier pré-traitement des lectures effectué par le pipeline PEGASE, celles séquencées dans le sens 3'-5' de l'amplicon ont été alignées de façon inverse et complémentaire sur la banque de séquences de référence. Elles sont ainsi inversées au sens du séquençage dans le fichier FASTA utilisé par CD-HIT-OTU. Ces lectures, rognées à leur extrémité 3' de ce fichier, sont en réalité rognées du côté 5' séquencé, donc du côté de meilleure qualité. Ainsi, l'étape de pré-*clustering* ajoutée au pipeline PEGASE v1 n'a pas été adaptée à l'approche de séquençage bidirectionnel développée par la plate-forme, nous avons donc décidé de l'éliminer du pipeline.

Après l'étape de pré-*clustering*, le logiciel ESPRIT [Sun *et al.* 2009] est utilisé pour effectuer un *clustering* hiérarchique moyen des lectures en OTUs, en se basant sur le calcul d'une matrice de distances entre lectures. Toutes les étapes d'ESPRIT ont été intégrées au pipeline avec les paramètres par défaut. ESPRIT génère des tables d'OTUs à différents seuils d'identité : le seuil choisi

par les développeurs du pipeline pour interpréter les résultats est un seuil de similarité de 95 %. ESPRIT sélectionne ensuite la lecture la plus abondante de chaque OTU comme étant sa séquence représentative. Cette étape soulève un problème majeur dans le pipeline PEGASE : en effet, puisque toutes les lectures ont été dédupliquées, chaque lecture est unique dans chaque OTU. La séquence représentative de chaque OTU est ainsi en réalité sélectionnée aléatoirement parmi toutes les lectures qui le composent, et ne peut donc pas être qualifiée de représentative.

La lecture de référence est ensuite comparée à la banque GOLD par le script *classify.seqs* intégré à mothur, pour être identifiée. Sur le jeu de données *test_bio*, le pipeline a généré 16 OTUs :

- 8 OTUs *Bordetella pertussis ;*
- 6 OTUs *Bacillus subtilis ;*
- 3 OTUs *Escherichia coli ;*
- 2 OTUs *Salmonella typhimurium ;*
- 1 OTU *Yersinia pseudotuberculosis*.

En alignant entre elles les lectures appartenant à deux OTUs différents assignés à la même espèce, on constate qu'elles contiennent un taux d'erreur élevé, causant une variation de séquence de plus de 5 % avec leur amplicon d'origine. Le seuil de similarité de 95 % semble ainsi mal défini pour prendre en compte la variabilité induite par les erreurs de séquençage. En outre, certains OTUs contiennent en réalité des séquences chloroplastiques, issues des plantes initialement présentes dans le mélange, et amplifiées par le couple d'amorces destiné à amplifier l'ADNr 16S bactérien. Ces séquences chloroplastiques sont suffisamment proches des séquences de la banque GOLD pour être mal classifiées comme étant des espèces bactériennes.

2.1.3 - Conclusion

L'utilisation d'un échantillon artificiel simplifié a permis de relever plusieurs points critiques du pipeline PEGASE v1, à savoir la mauvaise application d'une étape de pré-*clustering*, et la mauvaise sélection d'une séquence de référence d'un OTU lorsqu'il est composé de lectures dédupliquées. Toutefois, l'utilisation de données biologiques, même issues d'un échantillon artificiel, introduit trop de variabilité pour précisément optimiser le pipeline d'analyse. En effet de nombreuses variables biologiques (contamination chloroplastique, proportions initiales des différents génomes, erreurs d'amplification et/ou de séquençage, ...) sont difficiles à maîtriser *in vitro*. Afin d'avoir une plus grande maîtrise des données en entrée de pipeline, un nouvel échantillon artificiel *test_silico* a été créé, cette fois-ci *in silico*.

2.2 - Évaluation du pipeline PEGASE v1 sur une communauté microbienne artificielle générée *in silico*

2.2.1 - Description du jeu de données

CuReSim [Caboche *et al.* 2014] est un logiciel qui a été développé sur la plate-forme, dans le but de simuler des lectures de séquençage à partir de séquences de référence. Ce logiciel avait la particularité, à l'époque de cette évaluation, d'être le seul proposant un modèle de lectures Ion Torrent PGM. Initialement développé pour modéliser du *whole genome shotgun*, son auteur l'a adapté à notre contexte amplicon pour que les lectures simulées commencent obligatoirement par une des deux extrémités des amplicons afin de simuler un séquençage bidirectionnel (contrairement à une simulation WGS où les lectures sont choisies aléatoirement sur la séquence de référence). CuReSim a été appliqué sur les 5 séquences d'amplicons 400(V4-V5) de la banque GOLD, dans un premier temps sans simulation d'erreurs, et paramétré pour générer des lectures de taille moyenne de 200 nucléotides (similairement au jeu de données réel précédent). Les lectures composant l'échantillon artificiel *test_silico* sont

ainsi des fragments exacts des amplicons de la banque GOLD. Des lectures aléatoires (~ 5 %) ont également été intégrées à l'échantillon, afin d'ajouter du bruit dans les données et d'évaluer la capacité du pipeline PEGASE v1 à l'éliminer.

La description de la composition de l'échantillon *test_silico* est détaillée dans le Tableau 2.5.

Organisme	% de lectures	Nombre de lectures simulées	Nombre de lectures aléatoires
Bacillus subtilis	3 %	150	7
Bordetella pertussis	15 %	750	37
Escherichia coli	25 %	1250	62
Salmonella typhimurium	50 %	2500	125
Yersinia paratuberculosis	7 %	350	17
Total	100 %	5000	246 (en plus)

Tableau 2.5 : Composition de l'échantillon test_silico

2.2.2 - Résultats de l'analyse

L'analyse de ce jeu de données *test_silico* a été effectuée par le pipeline PEGASE v1 avec la banque complète SILVA cette fois-ci, comme représenté en Figure 2.2, afin de se mettre dans un contexte d'analyse réel. L'étape de pré-*clustering* a en outre été éliminée, puisqu'elle induisait des erreurs dans les résultats. Cette analyse de l'échantillon simulé *test_silico* a généré 4 OTUs, de composition suivante :

- *Bacillus subtilis* : 15 %;
- *Bordetella pertussis* : 21 %;
- *Salmonella typhimurium* : 45 %;
- Y*ersinia pseudotuberculosis* : 19%.

En analysant la composition des lectures dans les OTUs, on constate que l'ensemble des séquences aléatoires a été correctement éliminé. Toutefois, les OTUs contiennent un mélange de lectures issues des différentes espèces

initiales. Leur assignation taxonomique étant basée sur la sélection d'une séquence aléatoire dans l'OTU, elle n'est en aucun cas représentative de leur composition. Cette mauvaise répartition des lectures dans les OTUs peut être causée par le choix d'un seuil d'identité trop large (95 %) pour distinguer correctement les différents organismes.

L'utilisation d'un seuil d'identité plus stringent (96 %) permet de retrouver un découpage en OTUs correspondant à la composition réelle, où chaque OTU contient cette fois-ci uniquement les lectures de l'espèce à laquelle il est correctement identifié. Toutefois, comme montré par le Tableau 2.6, les proportions initiales des lectures sont loin d'être retrouvées, et le nombre total de lectures identifiées (447) est très inférieur au nombre de lectures initial (5000).

OTU	Nombre de lectures initiales	Nombre de lectures identifiées par le pipeline
Bacillus subtilis	150	68
Bordetella pertussis	750	93
Escherichia coli	1250	98
Salmonella typhimurium	2500	103
Yersinia paratuberculosis	350	85
Total	5000	447

Tableau 2.6 : Résultats de l'analyse de l'échantillon test_silico par le pipeline PEGASE v1 sans étape de pré-clustering, à un seuil d'identité de 96 %.

En évaluant les différents fichiers intermédiaires générés par le pipeline, il s'avère que les séquences qui ont été initialement dédupliquées (par mothur lors du pré-processing, puis par CD-HIT-OTU) ne sont pas correctement ré-injectées dans les OTUs dans le script final du pipeline, générant la table de comptages. En effet, seules les séquences issues de la première étape (mothur unique.seqs) ont été prises en compte. En corrigeant les calculs du nombre de lectures par OTUs (Tableau 2.7), les proportions initiales des différentes espèces sont correctement retrouvées. Les quelques lectures manquantes ont été

éliminées lors de la première étape de pré-traitement, car elles étaient trop courtes (< 150 nt). Ainsi, l'utilisation de données simulées sans erreurs, même peu complexes, a permis d'améliorer le pipeline PEGASE v1 en re-définissant le seuil de similarité permettant de délimiter les OTUs (96 %) et en corrigeant la prise en compte des lectures dédupliquées dans la génération de la table de comptages finale.

OTU	**Nombre de lectures initiales**	**Nombre de lectures identifiées par le pipeline après correction**
Bacillus subtilis	150	150
Bordetella pertussis	750	745
Escherichia coli	1250	1242
Salmonella typhimurium	2500	2486
Yersinia paratuberculosis	350	346
Total	5000	4969

Tableau 2.7 : Résultats de l'analyse de l'échantillon test_silico par le pipeline PEGASE v1 sans étape de pré-clustering, à un seuil d'identité de 96 %, après correction du comptage des lectures dédupliquées.

Afin d'observer le comportement du pipeline PEGASE sur des données Ion Torrent PGM tout en maîtrisant leur composition initiale, CuReSim a été utilisé sur les mêmes séquences d'amplicons mais cette fois-ci en incorporant un modèle de lectures représentatif de données Ion Torrent PGM (1 % de délétions, 0,5 % d'insertions, 0,5 % de substitutions, taille moyenne de lectures de 200 nucléotides, déviation standard de 20 nucléotides). L'analyse de ce nouvel échantillon, appelé *test_silico_PGM,* a engendré la génération d'un très grand nombre d'OTUs (110 à un seuil d'identité de 95 %, 261 à 96 %, et 597 à 97%). Cette grande quantité d'OTUs générés est causée par la présence d'erreurs dans les lectures, augmentant la dissimilarité entre celles appartenant à une même espèce. Toutefois, en cumulant les OTUs appartenant au même taxon, on retrouve la composition parfaite du jeu de données initial, comme montré dans le Tableau 2.8. Le nombre de lectures éliminées est plus conséquent car les

étapes de filtrage éliminent les lectures contenant trop d'erreurs.

OTU	Nombre de lectures initiales	Nombre de lectures identifiées par le pipeline
Bacillus subtilis	150	149
Bordetella pertussis	750	745
Escherichia coli	1250	1244
Salmonella typhimurium	2500	2486
Yersinia paratuberculosis	350	349
Total	5000	4373

Tableau 2.8 : Résultats de l'analyse de l'échantillon test_silico_PGM (dont les lectures ont été simulées avec un modèle d'erreurs et de taille Ion Torrent PGM) à un seuil d'identité de 96 %.

Un problème majeur identifié lors de l'utilisation de données simulant des erreurs de séquençage est le ralentissement du temps d'exécution d'ESPRIT, notamment durant la construction de la matrice de distances entre les séquences. En effet, sur l'échantillon *test_silico* sans erreurs, cette étape dure moins d'une minute, tandis qu'elle nécessite presque une heure sur les lectures simulées avec erreurs de l'échantillon *test_silico_PGM*. La présence d'erreurs cause une trop faible réduction des données dans les étapes de déduplication, et l'absence d'une étape de pré-*clustering* correctement intégrée ne permet pas non plus d'alléger la quantité de lectures à aligner.

2.3 - Améliorations du pipeline PEGASE

La réintégration d'une étape de pré-*clustering* aurait nécessité une plus longue période de tests et d'optimisation qui était impossible avec les délais imposés pour avoir un pipeline fonctionnel applicable en production. Pour réduire tout de même les temps d'analyse, ESPRIT a été remplacé par son successeur ESPRIT-Tree [Cai & Sun, 2011], allégeant les temps de calcul en promettant une qualité de résultats d'analyse similaire. En outre, le pipeline a été adapté pour pouvoir traiter simultanément tous les échantillons d'un même run de séquençage, là où il devait auparavant être exécuté échantillon par

échantillon. Enfin, une étape de regroupement des résultats en une même table de comptages, la conversion de ce fichier au format BIOM (*Biological Observation Matrix*, standard de représentation de tables de comptages [McDonald *et al.* 2012]) ainsi qu'une étape de normalisation de cette table ont été ajoutées à l'issue de ce pipeline. Toutes ces modifications ont été intégrées une nouvelle version du pipeline PEGASE, nommée v2, représentée dans la Figure 2.9.

Cette expertise du pipeline d'analyse PEGASE a permis d'identifier ses points critiques et d'optimiser son fonctionnement pour qu'il soit utilisable en production. Toutefois, son évaluation dans un délai restreint s'est limitée à de petits jeux de données non représentatifs de communautés réelles et complexes. En outre, ses performances n'ont pas été comparées avec des pipelines d'analyse publics, qui pourraient également servir de solution analytique pour la plate-forme. En effet, utiliser un pipeline public permettrait d'éviter la mise à jour manuelle et la maintenance d'un pipeline interne, tout en bénéficiant de recommandations d'usage adaptés à un contexte de séquençage d'Ion Torrent PGM. Ce premier travail a ainsi révélé la nécessité de mettre en place un protocole formel d'évaluation de pipelines d'analyses métagénétiques, afin de pouvoir mesurer précisément les performances de différents pipelines publics sur des jeux de données Ion Torrent PGM.

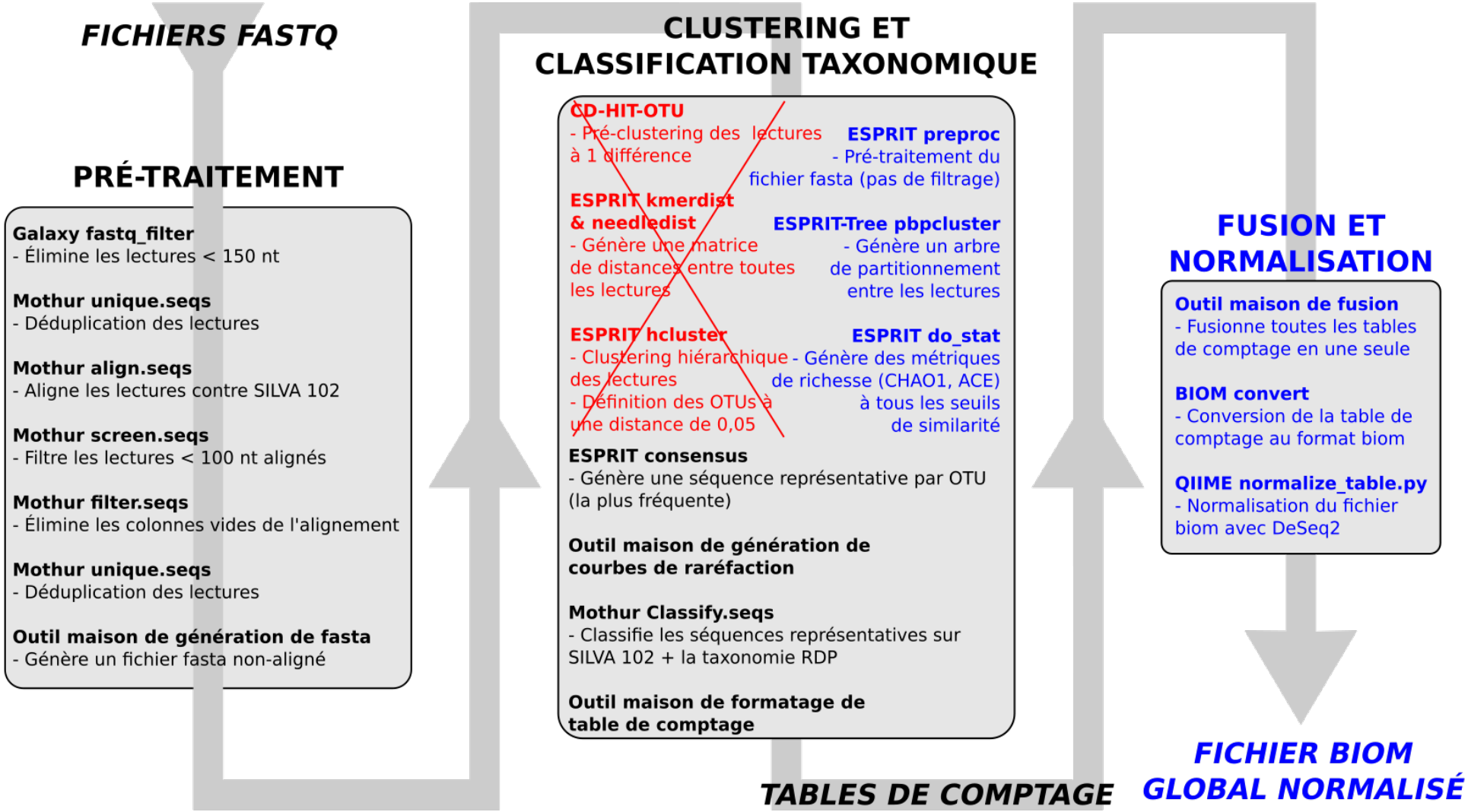


Figure 2.9 : Pipeline PEGASE v2 d'analyse métagénétique (pour plusieurs échantillons). Les étapes supprimées du pipeline PEGASE v1 sont indiquées en texte rouge, et les étapes ajoutées sont en texte bleu.

Chapitre 3 - Évaluation formelle de pipelines d'analyse de données métagénétiques

Les pipelines d'analyse de données métagénétiques sont spécifiquement développés pour lier les différentes étapes de l'analyse et automatiser leur exécution. Nombreux de ces logiciels sont disponibles, la plupart à code source ouvert dans un contexte académique, et à exécuter sur un serveur local. D'autres sont au contraire propriétaires, et/ou proposés sous la forme de services web distants. Chaque pipeline intègre ses propres recommandations d'étapes et algorithmes d'analyse comme comportement par défaut. De plus, plusieurs pipelines proposent à chaque étape la possibilité d'utiliser des algorithmes alternatifs, afin d'offrir un maximum de possibilités à l'utilisateur. Ce dernier doit ainsi avoir les compétences nécessaires pour évaluer si la solution par défaut correspond à son plan d'expérience, ou s'il doit basculer vers d'autres solutions d'analyse. Par exemple, l'utilisation d'une technologie de séquençage générant un taux élevé d'erreurs dans les lectures doit amener l'utilisateur à préférer des algorithmes moins sensibles à ces erreurs.

Il existe de nombreuses études comparant des algorithmes utilisés pour l'analyse de données métagénétiques, permettant d'estimer leurs forces et faiblesses dans différents contextes expérimentaux. Cependant, ces études comparatives se concentrent souvent sur une seule étape d'analyse, comme le pré-traitement des lectures [Bonder *et al.* 2012], le *clustering* d'OTUs [Westcott *et al.* 2015] ou l'assignation taxonomique [Bazinet *et al.* 2012, Garcia-Etxebarria *et al.* 2014]. Ces comparaisons utilisent généralement des données réelles et/ou simulées de séquençage Illumina [Sinclair *et al.* 2015] et 454 [D'Argenio *et al.* 2014] uniquement. Seules de rares études comparent les pipelines dans leur entièreté ; ces études présentent toutefois plusieurs limites. D'Argenio *et al.* ont évalué deux pipelines d'analyses métagénétiques sur des données réelles issues d'un séquençage 454 de microbiote intestinal humain.

Dans ce contexte, la composition réelle des microbiotes étudiés est inconnue : il est de ce fait impossible d'évaluer objectivement la qualité des résultats, n'ayant pas de référence à laquelle les comparer. Une autre étude [Lindgreen *et al.* 2016] a évalué la performance de différents pipelines sur des jeux de données artificiels, dont la composition est ainsi connue et maîtrisée. Ces jeux de données sont toutefois issus de simulations de séquençage métagénomique WGS, et non de métagénétique. Aucune étude n'a déjà comparé des méthodes d'analyses métagénétiques complètes (définies comme un pipeline et ses recommandations par défaut), ni n'a effectué une telle comparaison dans un contexte de séquençage de type Ion Torrent PGM. Dans ce chapitre, nous présentons la première contribution de la thèse : l'établissement d'un protocole d'évaluation afin de comparer des pipelines d'analyse complets dans un contexte de métagénétique bactérienne, et d'évaluer leur sensibilité à différentes variables propres au plan d'expérience.

Cette étude a fait l'objet d'une publication dans la revue PLOS ONE : Siegwald L, Touzet H, Lemoine Y, Hot D, Audebert C, Caboche S. Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics. PLOS ONE. Public Library of Science; 2017;12(1):e0169563.

3.1 – Mise en place d'un protocole d'évaluation de pipelines et contexte d'utilisation

3.1.1 - Jeux de données bactériens réels et simulés

De nombreux jeux de données issus de séquençages de communautés artificielles sont mis à disposition de la communauté afin de servir de jeux de données de référence [The Human Microbiome Consortium 2012, Singer *et al.* 2016]. Nous avons toutefois choisi d'utiliser des jeux de données simulés plutôt que de construire nos propres communautés bactériennes *in vitro*. En effet, le séquençage de microbiotes artificiels est tout de même sujet à des biais techniques pouvant introduire du bruit dans les données produites (par exemple les biais d'amplification). Ce bruit pourrait être à l'origine de variations

dans les résultats des pipelines, indépendamment des méthodes d'analyse qui y sont intégrées. Or, nous souhaitions évaluer la sensibilité des pipelines à différentes variables du plan d'expérience, et non à des biais techniques difficilement contrôlables et évaluables. En outre, nous voulions évaluer l'analyse de communautés artificielles riches (> 100 espèces) pour nous rapprocher de microbiotes réels d'intérêt, tels que la flore intestinale ou le microbiote du sol. Une telle richesse est très difficile et coûteuse à reproduire *in vitro*. Enfin, nous souhaitions évaluer l'impact de différents choix de plans d'expérience, ce qui aurait nécessité de nombreux séquençages distincts en utilisant des communautés artificielles.

Ces écueils sont évités en générant des jeux de données artificiels de séquençage par une approche *in silico*. En effet, la simulation permet d'obtenir un grand nombre de jeux de données précisément définis selon différentes conditions contrôlées, sans biais expérimentaux. Cette simulation, détaillée ci-dessous, a été menée en différentes étapes :

1. Définir la composition de plusieurs communautés bactériennes à simuler ;
2. Pour chaque communauté, récupérer le génome de référence de chaque organisme ;
3. Définir la proportion de chaque organisme dans chaque communauté, et le nombre de lectures à simuler par organisme selon un débit de séquençage donné ;
4. Extraire les amplicons d'intérêt de chaque génome, dans les proportions définies précédemment ;
5. Simuler un séquençage Ion Torrent PGM sur ces amplicons pour obtenir des lectures simulées.

La première étape de la simulation a ainsi été de définir la composition de plusieurs communautés bactériennes à simuler, sur base de la littérature. En 2007, Mavromatis *et al.* ont publié les jeux de données FAMeS dans l'idée de standardiser la comparaison d'outils d'assemblage et annotation de métagénomes [Mavromatis *et al.* 2007]. Ils ont sélectionné de façon aléatoire

des fragments issus de 113 génomes bactériens séquencés par la méthode de Sanger, et les ont répartis en trois jeux de données distincts, modélisant différentes complexités de communautés bactériennes. Cette notion de complexité n'est pas explicitement définie dans l'article. Toutefois, d'après l'appellation de ces communautés, on peut considérer que plus une communauté contient d'organismes majoritaires différents, plus elle est complexe. Une communauté est dite de faible complexité (LC, *low complexity*), avec un organisme largement majoritaire, de moyenne complexité (MC, *medium complexity*) avec quelques principaux organismes majoritaires, et de haute complexité (HC, *high complexity*) lorsqu'elle est hautement complexe sans organisme dominant. Ces jeux de données ne sont néanmoins pas simulés pour un contexte de séquençage haut-débit. La composition de ces microbiotes artificiels a été réutilisée dans différentes études [Mitra *et al*. 2010, Charuvaka & Rangwala 2011, Pignatelli & Moya 2011], qui se sont basées sur les génomes publiés des organismes listés pour simuler des lectures issues de séquençage à haut-débit de métagénomes WGS. Pignatelli & Moya ont publié dans leur étude une liste répertoriant pour les trois communautés LC, MC et HC l'ensemble des organismes sélectionnés, leur identifiant de génome NCBI ainsi que les proportions de chaque génome dans chaque communauté. Cette liste a été la base de la construction de nos propres jeux de données métagénétiques artificiels.

Les trois communautés artificielles ont une composition en organismes légèrement différente. Afin de pouvoir les comparer dans notre étude, tous les organismes ont été cumulés afin d'obtenir une liste totale de 122 organismes, tous présents dans LC, MC et HC. Cette liste a également été complétée par trois séquences de bactéries qui ne sont pas encore référencées dans les banques de séquences ribosomiques de référence, afin de pouvoir évaluer l'attribution taxonomique d'espèces « inconnues » :

- *Gloeobacter kilaueensis* : Cyanobactérie découverte en 2013, présente dans aucune banque spécifique, génome complet disponible ;

- *Tersicoccus phoenicis* : Actinomycète découverte en 2007 uniquement présent dans RDP, séquence 16S disponible ;
- *Eisenbergiella tayi* : Clostridiales découverte en 2013, présente dans aucune banque spécifique, séquence 16S disponible.

Au total, ces 125 organismes représentent 51 familles et 69 genres bactériens différents, dont le génome de référence a été téléchargé afin d'y simuler l'amplification de la cible (sauf pour *Tersicoccus phoenicis* et *Eisenbergiella tayi* qui n'ont pas encore été séquencées, et dont seule la séquence de l'ADNr 16S complet a été récupérée).

Afin d'évaluer l'impact du choix de la région cible d'intérêt, nous avons choisi deux régions différentes de l'ADNr 16S comme cible d'amplification (Figure 3.1) :

- 200(V3) est l'amplicon généré par les amorces *Probio_Uni* et *Probio_Rev* encadrant la région V3 de l'ADNr 16S (~188 nt). Ces amorces ont été designées et utilisées par Milani *et al.* dans le cadre d'une étude métagénétique de microbiote fécal par séquençage Ion Torrent PGM [Milani *et al.* 2013].
- 400(V4-V5) est l'amplicon généré par les amorces 519F et 907R [Weisburg *et al.* 1991], couramment utilisées sur la plate-forme PEGASE-biosciences, et encadrant les régions hypervariables V4 et V5 de l'ADNr 16S (~408 nt).

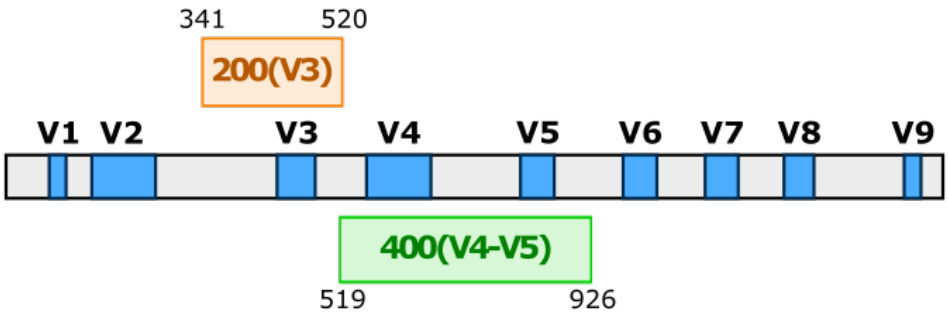


Figure 3.1 : Positions des deux amplicons 200(V3) et 400(V4-V5) sur l'ADNr 16S bactérien d'E. coli.

D'après le logiciel TestPrime [Klindworth *et al.* 2013], les couples d'amorces utilisés pour les amplicons 200(V3) et 400(V4-V5) capturent respectivement 82,1 % et 83,0 % des séquences présentes dans la banque SILVA SSU Ref 122. Sur notre liste de 125 séquences initiales, ces deux couples d'amorces sont parfaitement complémentaires à respectivement 102 (81,6 %) et 104 (83,2 %) séquences. Deux espèces (*Leuconostoc mesenteroides* et *Thermobifida fusca*) ne sont amplifiées que par les amorces 400(V4-V5). Cette observation met en évidence le biais majeur de toute étude métagénétique, qui est le choix d'amorces (Chapitre 1, Section 1.3.4) : il est impossible de garantir formellement leur universalité, ce qui induit la non-détection de certains organismes dans les microbiotes étudiés [Wang & Qian 2009]. En effet, les organismes non-amplifiés par PCR ne peuvent évidemment pas être séquencés, et par conséquent ne peuvent être présents dans les résultats finaux. Pour la construction de nos jeux de données artificiels, les séquences non sélectionnées par ces amorces ont été éliminées de la liste de séquences initiales, afin de ne pas introduire de biais de composition dans la simulation des jeux de données.

Nos communautés simulées étant légèrement différentes de celles décrites par Pignatelli & Moya (uniformisation des organismes, ajout de gènes absents des banques, élimination d'organismes non amplifiés), les généralisations suivantes ont été utilisées pour définir les proportions de chaque organisme dans les communautés LC, MC et HC :

- LC : 30 % de *Rhodopseudomonas palustris HaA2* (ID du génome dans Genbank : NC_007778.1), appartenant à la famille Bradyrhizobiaceae. Les autres organismes sont équitablement répartis ;
- MC : 20 % de *Rhodopseudomonas palustris HaA2* (NC_007778.1), *Rhodospirillum rubrum ATCC 11170* (NC_007643.1, famille Rhodospirillaceae), *Bradyrhizobium sp. BTAi1* (NC_009485.1, famille Bradyrhizobiaceae) et *Xylella fastidiosa M12* (NC_010513.1, famille Xanthomonadaceae). Les autres organismes sont équitablement répartis. À noter que la communauté MC contient 4 espèces dominantes de 4 genres différents, mais appartenant à seulement 3 familles distinctes ;
- HC : pas d'organisme dominant, tous sont équitablement répartis.

Ces proportions permettent d'estimer combien de lectures doivent être simulées par organisme, sur la base d'un débit de séquençage total. Pour les études métagénétiques classiquement effectuées sur la plate-forme, jusqu'à 96 échantillons sont séquencés par Ion Torrent PGM sur une puce Ion 318™, ce qui équivaut à environ 50 000 séquences par échantillon. Afin d'étudier l'impact du débit de séquençage sur les résultats d'analyse, nous avons décidé de varier le débit de séquençage dans nos données simulées. Pour chaque complexité (LC, MC et HC), 3 séquençages ont été simulés, chacun à une profondeur différente : 25 000 lectures, 50 000 lectures et 100 000 lectures.

De nombreux logiciels existent afin de simuler des sorties de séquençage haut-débit [Escalona *et al.* 2016] à partir de séquences génomiques. Grinder [Angly *et al.* 2012] a été utilisé dans notre contexte, car il est le seul à modéliser une approche amplicon en se basant sur la séquence dégénérée des amorces pour extraire les amplicons des séquences de génomes, à des proportions prédéfinies. Grinder permet également d'appliquer aux séquences extraites un modèle de lectures issues de différentes technologies de séquençage haut-débit. Ce logiciel ne propose toutefois pas de modèle de lectures Ion Torrent PGM, et a ainsi été utilisé uniquement pour extraire les amplicons des séquences de référence. CuReSim a été appliqué sur les amplicons générés par Grinder avec les paramètres par défaut, simulant des lectures Ion Torrent PGM (1 % de délétions, 0,5 % d'insertions, 0,5 % de substitutions, et une déviation standard de 20 nucléotides de la taille totale de l'amplicon).

36 jeux de données artificiels simulant trois communautés bactériennes de complexité différente, à trois couvertures différentes, sur deux amplicons distincts, avec et sans erreurs, ont ainsi été générés comme références dans notre protocole d'évaluation de pipelines d'analyses métagénétiques (Figure 3.2).

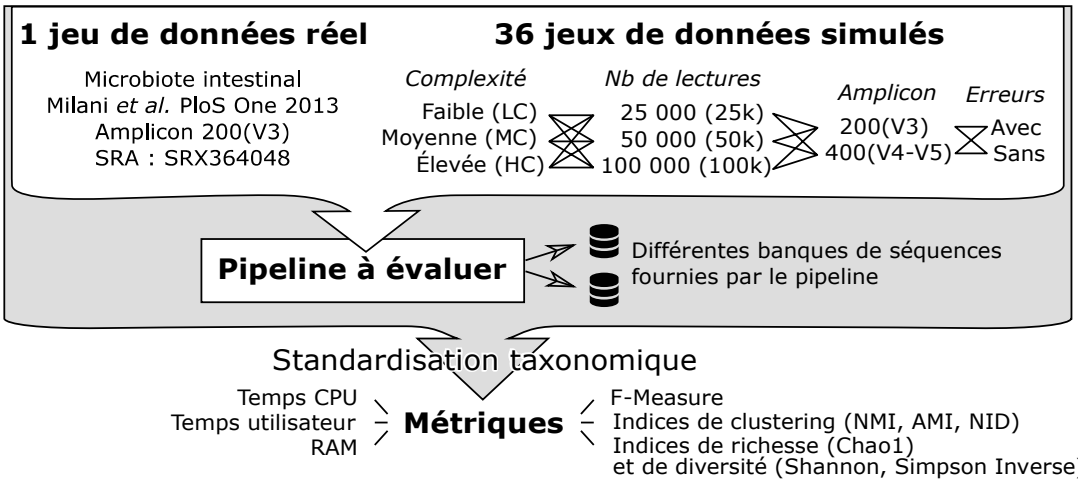


Figure 3.2 : Protocole d'évaluation de pipelines d'analyses métagénétiques (adapté de Siegwald et al. 2017).

Afin de valider les résultats d'analyse des jeux de données simulés, ce plan d'expérience a été complété par un jeu de données réel issu d'une étude métagénétique de microbiote intestinal humain par Ion Torrent PGM, utilisant les mêmes amorces que l'amplicon 200(V3) simulé. Ce jeu de données est public, archivé dans la banque SRA sous l'identifiant SRX364048, et contient 231 660 lectures. Ces données ont été nettoyées en utilisant cutadapt [Martin 2011] afin d'éliminer les séquences d'adaptateurs et les index de multiplexage des lectures.

3.1.2 - Choix de métriques d'évaluation adaptées

L'analyse des jeux artificiels de données par un pipeline permet de directement comparer ses résultats avec la composition initiale des jeux de données. Cette comparaison doit être effectuée en utilisant des métriques précises permettant d'évaluer la qualité des résultats et leur similarité avec les données de référence (Figure 3.2). De nombreuses études d'évaluation d'algorithmes utilisent dans un contexte similaire (comparaison de résultats à une référence) une mesure de précision et de rappel [Chen *et al.* 2013, Caboche *et al.* 2014, Kopylova *et al.* 2016]. Le rappel mesure la sensibilité du pipeline, en évaluant la proportion des lectures correctement identifiées parmi toutes les

lectures, tandis que la précision mesure la spécificité du pipeline, en évaluant parmi les lectures identifiées uniquement, lesquelles le sont correctement. Une lecture peut être catégorisée dans trois classes différentes : elle est un vrai positif (VP) à un niveau taxonomique donné si son assignation taxonomique à ce niveau est le même que la séquences de référence dont elle a été extraite. Une lecture est un faux positif (FP) si cette assignation est différente de la référence. Enfin, une lecture est considérée comme un faux négatif (FN) si elle est éliminée par le pipeline, ou annotée comme étant non classifiée (*unclassified*) à ce niveau taxonomique. Sur cette base, la précision et le rappel sont calculés de la façon suivante :

$$precision = \frac{VP}{VP + FP} \qquad recall = \frac{VP}{VP + FN}$$

La qualité globale d'un pipeline peut être alors évaluée grâce à la *F-mesure* (aussi appelée F-score), moyenne harmonique de la précision et du rappel. Bornée entre 0 et 1, plus la F-mesure est proche de 1, plus les résultats du pipeline sont proches de la composition réelle de l'échantillon simulé :

$$F - Mesure = 2 \times \frac{precision \times recall}{precision + recall}$$

Le calcul d'indices de richesse (Chao1) et de diversité (Shannon et Simpson inverse) permet également de comparer la performance des pipelines, ces indices étant connus pour les données simulées. Pour les pipelines *clustering-first*, ils peuvent être calculés immédiatement après l'étape de *clustering*, tout comme à différents niveaux taxonomiques après assignation taxonomique (les OTUs ayant la même assignation sont alors fusionnés en un seul OTU). Un pourcentage d'erreur de richesse estimée peut également être calculé, ce qui permet d'évaluer à quel degré l'estimation de richesse en résultat de pipeline est proche de la richesse réelle de l'échantillon simulé à un niveau taxonomique donné :

$$\%\, erreur\, Chao\,1 = \frac{Chao\,1\, estimé - Chao\,1\, réel}{Chao\,1\, réel} \times 100$$

Une forte similarité des indices de richesse et de diversité entre les résultats d'un pipeline et les données simulées n'est pas forcément un indicateur de la bonne capacité de ce pipeline à estimer ces indices. Par exemple dans la Figure 3.3, la composition des OTUs est différente entre les données initiales et le résultat de l'analyse ; toutefois, les indices sont identiques car ces indices sont indépendants de la manière dont les OTUs sont étiquetés.

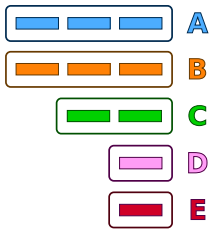
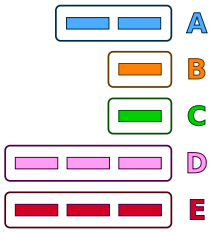
		 **Données initiales**	 **Résultat d'analyse**
RICHESSE	**Nombre d'OTUs**	5	5
	Chao1	5,5	5,5
	ACE	6,25	6,25
DIVERSITÉ	**Shannon**	2,17	2,17
	Simpson	0,24	0,24
	Simpson inverse	4,17	4,17

Figure 3.3 : Mesures de richesse et diversité identiques pour des résultats d'analyse différents des données initiales.

D'autres indices, nommés ici indices de *clustering*, permettent de répondre à ce problème en comparant le partitionnement des lectures dans différents OTUs à l'issue de l'analyse, et la répartition des lectures dans les taxons dont elles sont originaires [Vinh *et al.* 2010, Schmidt *et al.* 2015]. Les premières métriques développées dans ce but sont le RI (*rand index*, [Rand 1971]) et l'ARI (*adjusted rand index,* [Hubert & Arabie 1985]), qui observent la concordance et discordance de paires de séquences dans les partitions. Ces indices sont lourds à calculer pour nos jeux de données contenant de

nombreuses séquences, dont il faudrait évaluer toutes les paires. Ils sont en outre fortement biaisés envers les gros OTUs qui ont beaucoup plus de poids, car formant beaucoup plus de paires de séquences.

Certains indices issus de la théorie de l'information sont plus adaptés à notre contexte de comparaison, et se basent sur des mesures d'entropie entre les partitionnements. L'information mutuelle (MI, *mutual information*) calcule le recouvrement d'information entre deux partitions en se basant sur leur entropie respective. En considérant une partition A d'un nombre i OTUs de taille a_i (dans notre contexte le résultat d'un pipeline), et une partition B d'un nombre j d'OTUs de taille b_j (dans notre contexte le partitionnement initial des lectures) :

$$MI(A,B)=\sum_i \sum_j n_{ij} \log[\frac{n_{ij} n}{a_i b_j}]$$

où n_{ij} est le nombre de lectures présentes dans l'OTU i de A et l'OTU j de B, et n le nombre total de lectures.

L'information mutuelle normalisée (NMI, *normalized mutual information*, Fred & Jain 2003) est une version normalisée du MI qui le borne entre 0 et 1, ce qui permet de comparer cet indice entre différentes méthodes de partitionnement (dans notre contexte, entre différents pipelines par exemple). Plus le NMI est proche de 1, plus le partitionnement des lectures en OTUs par un pipeline est proche du partitionnement des lectures dans leurs taxons originels.

$$NMI=\frac{-2I(A,B)}{H(A)+H(B)}$$

où $H(A)$ et $H(B)$ sont les entropies respectives des partitions A et B, calculées de la façon suivante :

$$H(A)=-\sum_i \frac{a_i}{n}$$

Le NMI est toutefois dépendant du nombre de clusters, ce qui biaise la comparaison du NMI de deux méthodes de partitionnement générant un nombre de clusters différents. Ce biais a été corrigé par l'AMI (*adjusted mutual information* [Vinh et a. 2010]). L'AMI corrige le NMI par l'information mutuelle attendue pour deux *clusterings* aléatoires. Il se mesure entre -1 (discordance complète entre les partitionnements) et 1 (les deux partitionnements sont identiques) :

$$AMI = \frac{I(A,B) - E\{I(M)|a,b\}}{\sqrt{H(A)H(B)} - E\{I(M)|a,b\}}$$

où *E{I(M)|a,b}* est le MI moyen de tous les recouvrements de partitions théoriquement possibles entre A et B :

$$E\{I(M)|a,b\} = \sum_i \sum_j \sum_{n_{i,j}=(a_i+b_j-N)}^{min(a_i,b_j)} \frac{n_{i,j}}{n} \log\left(\frac{n_{i,j}}{a_i b_j}\right) \frac{a_i! b_j! (n-a_i)! (n-b_j)!}{n! n_{i,j} (a_i - n_{i,j})! (b_j - n_{i,j})! (n - a_i - b_j + n_{i,j})!}$$

L'AMI donne cependant de meilleurs résultats pour un nombre élevé de clusters, ce qui peut biaiser la comparaison de différentes méthodes de partitionnement. Une dernière correction a été proposée par le SMI (*Standardized Mutual Information*) [Romano *et al.* 2014], en corrigeant l'AMI par la variance du MI. Cette correction nécessite toutefois un calcul plus complexe et demande plus de ressources pour un partitionnement avec beaucoup de clusters.

Vinh *et al.* critiquent tous ces indices en leur reprochant le fait qu'ils ne représentent pas une distance entre les deux partitionnements [Vinh *et al.* 2010]. Ces auteurs proposent en remplacement le NID (*normalized information distance*), borné entre 0 et 1. À l'inverse des autres indices, plus le NID est faible, plus les deux partitionnements sont proches. Pour unifier son interprétation avec les autres métriques, nous avons privilégié l'emploi de l'indice 1-NID.

$$NID = \frac{1 - MI(A,B)}{max[H(A), H(B)]}$$

Le protocole d'évaluation calcule ainsi le NMI, l'AMI et l'indice 1-NID pour comparer le partitionnement des résultats avec celui des données initiales.

Outre la corrélation entre les résultats d'un pipeline et la taxonomie réelle des données simulées, ce protocole d'évaluation inclut également le temps d'exécution du pipeline (temps utilisateur et CPU) ainsi que la mémoire vive utilisée (RAM), afin d'estimer les ressources nécessaires à l'analyse. Notons que ces métriques ne peuvent être évaluées pour les services web qui délocalisent les calculs du client vers un serveur fermé. Dans cette situation, seul le temps de rendu des résultats peut être mesuré.

3.1.3 - Contexte d'utilisation du protocole d'évaluation

L'ensemble des jeux de données présentés en section 3.1.1 et des métriques présentées en section 3.1.2 compose un cadre de comparaison complet et robuste couvrant différents contextes d'expérience (Figure 3.2). Ce protocole a été utilisé pour évaluer 6 pipelines décrits dans le Tableau 3.4 : 3 pipelines de la catégorie *clustering-first*, et 3 pipelines de la catégorie *assignment-first*. mothur [Schloss *et al.* 2009] et QIIME [Caporaso *et al.* 2010] sont deux pipelines *clustering-first* en ligne de commande. BMP [Pylro *et al.* 2014] propose des étapes d'analyse optimisées pour un contexte Ion Torrent, en se basant sur QIIME et UCLUST dont les paramètres ont été adaptés pour mieux correspondre à cette technologie. One Codex [Minot *et al.* 2015] est un service web *assignment-first*, et kraken [Wood *et al.* 2014] et CLARK sont des outils *assignment-first* en ligne de commande.

Notre intérêt est d'évaluer ces pipelines sans *a priori*, dans un contexte non-expert : ainsi, chaque pipeline a été exécuté en suivant au maximum ses recommandations par défaut. L'objectif d'une telle étude est d'évaluer les performances des pipelines *assignment-first* dans un contexte métagénétique Ion Torrent PGM, ce qui n'a jamais été fait, et d'évaluer les pipelines sous différentes contraintes expérimentales dans un contexte d'utilisation par défaut.

Les pipelines en ligne de commande (mothur, QIIME, BMP, kraken et CLARK) ont été exécutés sur un même serveur, avec deux processeurs Intel Xeon E5-2470 et 192 GB de mémoire vive (RAM). Les étapes d'analyse parallélisables ont été réparties sur 32 cœurs. Les recommandations de pré-traitement des lectures proposées par certains pipelines ont été limitées afin de pouvoir évaluer l'impact des erreurs de séquençage dans les résultats : le filtrage et débruitage des lectures n'a pas été exécuté, sauf pour BMP qui nécessite une taille de lectures fixe (ce qui n'est pas le cas pour des lectures Ion Torrent). BMP utilise en effet le logiciel UPARSE pour son étape de *clustering* d'OTUs, nécessitant que les lectures aient toutes la même taille pour un fonctionnement optimal [Edgar 2013]. La détection des chimères a également été désactivée, puisque ce phénomène n'a pas été simulé, et que tous les pipelines n'intègrent pas une telle étape.

À l'issue des analyses, les OTUs ne contenant qu'une seule lecture (appelés singletons) ont été éliminés et les lectures les composant ont été considérés comme des faux négatifs. En effet, les pipelines *clustering-first* éliminent les singletons dans leurs recommandations (QIIME a un minimum de taille d'OTU de 2 lectures, BMP supprime les singletons, et mothur contient une étape de sous-échantillonnage qui *« supprime quelques OTUs qui n'ont pas assez de séquences »*, typiquement les singletons). Les singletons ont été également éliminés des résultats des pipelines *assignment-first* dans le but d'être dans un contexte d'étude similaire, puisqu'il n'existe pas encore de standard d'utilisation de cette catégorie de pipelines dans un contexte métagénétique.

	Clustering-first				***Assignment-first***		
	mothur	***QIIME***		***BMP***	***kraken***	***CLARK***	***One Codex***
Version	1.35.1	SortMeRNA + Sumaclust 1.9.0	UCLUST 1.9.0	Dez. 2014	0.10.5-beta	1.1.2	open beta
Banque par défaut	SILVA 119 (mothur S)	Greengenes 13.8 (QIIME SS GG)	Greengenes 13.8 (QIIME U GG)	Greengenes 13.8 (BMP GG)	Minikraken 20141208	RefSeq 71 adaptation	OneCodex 28k propriétaire (One Codex OC)
Autres banques évaluées	Greengenes 13.8	SILVA 119 (QIIME SS S)	SILVA 119 (QIIME U S)	SILVA 119 (BMP S)	NA	NA	RefSeq 65 (One Codex RS) SILVA 119 (One Codex S)
Interface	Ligne de commande locale						Serveur web
Référence	[Schloss *et al.* 2009]	[Caporaso *et al.* 2010]		[Pylro *et al.* 2014]	[Wood *et al.* 2014]	[Ounit *et al.* 2015]	[Minot *et al.* 2015]
Nombre de citations (avant 2017)	5 932	6 131		30	271	55	2

Tableau 3.4 : Description des 6 pipelines évalués, et abréviation de chaque pipeline en gris.

Chaque pipeline a été exécuté en suivant au mieux les recommandations de ses auteurs pour des données Ion Torrent. mothur (version 1.35.1) fournit des recommandations pour le séquençage 454 que nous avons suivies, les deux technologies étant très proches, et les recommandations Ion Torrent étant hors ligne au moment de notre étude. Ces dernières nous ont été transmises ultérieurement par l'auteur de mothur ; elles diffèrent des recommandations 454 uniquement dans l'étape de filtrage des lectures qui a été désactivée dans notre évaluation. QIIME (version 1.9.0) a été utilisé avec le script pick_open_reference_otus.py, en utilisant les logiciels alternatifs open-source SortMeRNA (Kopylova, 2012) et Sumaclust (Mercier, 2013) qui sont prévus comme algorithmes par défaut dans les prochaines versions de QIIME. L'approche de *clustering* UCLUST (non libre) par défaut a également été exécutée, mais n'est mentionnée dans ce texte que lorsqu'elle présente des différences significatives de résultats. BMP (version Dez. 2014) est le seul pipeline proposant des recommandations spécifiques aux analyses de métagénétique ADNr 16S sur la technologie Ion Torrent, utilisant à la fois UCLUST et quelques étapes de QIIME. Tous les pipelines *clustering-first* ont été exécutés sur la banque SILVA 119 (par défaut pour mothur), et Greengenes 13.8 (par défaut pour QIIME et BMP).

Les pipelines *assignment-first* ont été exécutés avec leurs paramètres par défaut. kraken (version 0.10.5-beta) a été utilisé avec la banque Minikraken (version 20141208, sous-ensemble de 10 000 *k-mers* sélectionnés aléatoirement dans RefSeq), CLARK (version 1.1.2) utilise la banque de *k-mers* bactériens extraits de RefSeq et discriminants au niveau du genre, et One Codex (un service propriétaire intégrant RefSeq) a également été utilisé avec sa propre banque de référence propriétaire (One Codex 28k) et une intégration expérimentale de SILVA 119.

L'ensemble des commandes utilisées pour l'exécution de chaque pipeline est décrit en Annexe 2.

3.1.4 - Standardisation taxonomique des résultats entre pipelines

Afin de pouvoir comparer les résultats entre eux en utilisant différents pipelines et banques de séquences, nous avons souhaité pouvoir les convertir dans la taxonomie du NCBI. C'est en effet la taxonomie qui a été utilisée pour annoter les séquences sur lesquelles les amplicons et lectures ont été simulés. En outre, il s'agit de la taxonomie utilisée par tous les pipelines *assignment-first*. Nous pensions pouvoir utiliser les dictionnaires fournis par les banques Greengenes et SILVA pour convertir leurs assignations taxonomiques dans la taxonomie du NCBI. En effet, Greengenes et SILVA fournissent des fichiers de correspondance entre leurs séquences et les séquences de Genbank. Il est ainsi aisé, à partir des séquences de ces banques de récupérer la fiche Genbank et donc la taxonomie associée. Cette solution s'est toutefois avérée impossible à intégrer en interprétant les résultats d'analyse : les pipelines *clustering-first* renvoient en résultat l'annotation taxonomique de chaque OTU, et non l'identifiant de la (ou des) séquence(s) de la banque associée(s). Ces banques ne fournissent pas de méthode pour convertir leurs propres annotations taxonomiques dans la taxonomie du NCBI, nous avons ainsi du développer notre propre méthode de conversion.

Pour ce faire, le TaxID (identifiant taxonomique du NCBI) de chaque OTU a été récupéré à différents niveaux taxonomiques à partir de son annotation, en parcourant le fichier XML renvoyé par le script e-search de E-utilities [Sayers 2013] accessible à l'adresse suivante :

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=taxonomy&term=*TAXON*[All%20Names]&rank=*RANG*&retmode=xml&rettype=full

où *TAXON* est le nom du taxon, et *RANG* est le rang taxonomique associé. Les assignations taxonomiques ambiguës ont été vérifiées manuellement et corrigées si possible (par exemple dans la taxonomie du NCBI, *Paracoccus* est un genre bactérien (TaxID 265) mais aussi d'insecte (TaxID 249411)). Si la

correction est impossible (par exemple un taxon n'ayant pas d'équivalent dans la taxonomie du NCBI), l'OTU correspondant est considéré comme étant un faux négatif, puisque tous les taxons des séquences de référence dont ont été tirées les lectures simulées ont une taxonomie bien définie dans le NCBI.

L'évaluation des pipelines a été effectuée à deux niveaux taxonomiques classiquement utilisés en métagénétique : la famille et le genre (abrégés respectivement F et G dans le texte). Des niveaux taxonomiques plus larges (ordre, classe, …) sont trop généralistes et ne discriminent pas suffisamment les résultats entre pipelines pour évaluer leur performance.

3.1.5 - Disponibilité du protocole d'évaluation

L'ensemble des jeux de données ainsi que leur composition détaillée (identifiants de génomes, positions des amplicons, nombre de lectures par amplicons, …), tout comme les lignes de commande exécutées pour l'évaluation de chaque pipeline, ont été publiés sur le site Internet de la plate-forme : http://www.pegase-biosciences.com/metagenetics/

3.2 – Évaluation des pipelines selon différentes variables du plan d'expérience

L'exécution de tous les pipelines sur tous les jeux de données et toutes les banques proposées a généré au total 481 résultats d'analyse. L'ensemble des métriques d'évaluation a été calculé pour chacun de ces résultats d'analyse, et peut se trouver en données supplémentaires de l'article Siegwald *et al.* (PLOS ONE, 2017). Nous y ferons référence dans la suite du manuscrit en tant que « données supplémentaires ». Ces mesures ont permis de précisément évaluer le comportement chaque pipeline face aux variations de conditions expérimentales simulées.

3.2.1 - Impact du changement de région amplifiée

Nous avons vu dans le Chapitre 1 (Section 1.3.4) l'importance du choix de la région génomique ciblée dans une étude métagénétique, et des amorces associées. Ces choix sont en effet une source connue de biais et de variations dans les résultats [Mao *et al.* 2012]. Les jeux de données simulés sur les amplicons 200(V3) et 400(V4-V5) ont permis d'évaluer le comportement des pipelines sur ces deux amplicons, afin d'évaluer dans quelle mesure le changement de cible peut avoir un impact dans les résultats. Cette évaluation a été effectuée sur les jeux de données HC, puisqu'ils n'induisent pas de biais de composition (tous les taxons y étant également représentés). Dans cette comparaison, l'impact du changement d'amplicon ne peut être dissocié de l'information contenue dans les deux régions distinctes, les deux éléments étant intrinsèquement liés. Ainsi, les variations de résultats entre les deux amplicons sont indissociablement liées à la fois au changement de taille et au changement de région.

La Figure 3.5 représente l'impact du changement d'amplicon sur la F-mesure pour chaque pipeline sur les jeux de données Ion Torrent simulés de débit 50k. Pour rappel, plus la F-mesure est proche de 1, plus les résultats d'analyse sont proches de la composition réelle de l'échantillon simulé. Cette métrique prend en compte la précision et le rappel ; nous pouvons observer dans les données supplémentaires que pour tous les pipelines, la précision est toujours plus élevée que le rappel. Ceci implique que tous les pipelines privilégient la spécificité des résultats à leur sensibilité.

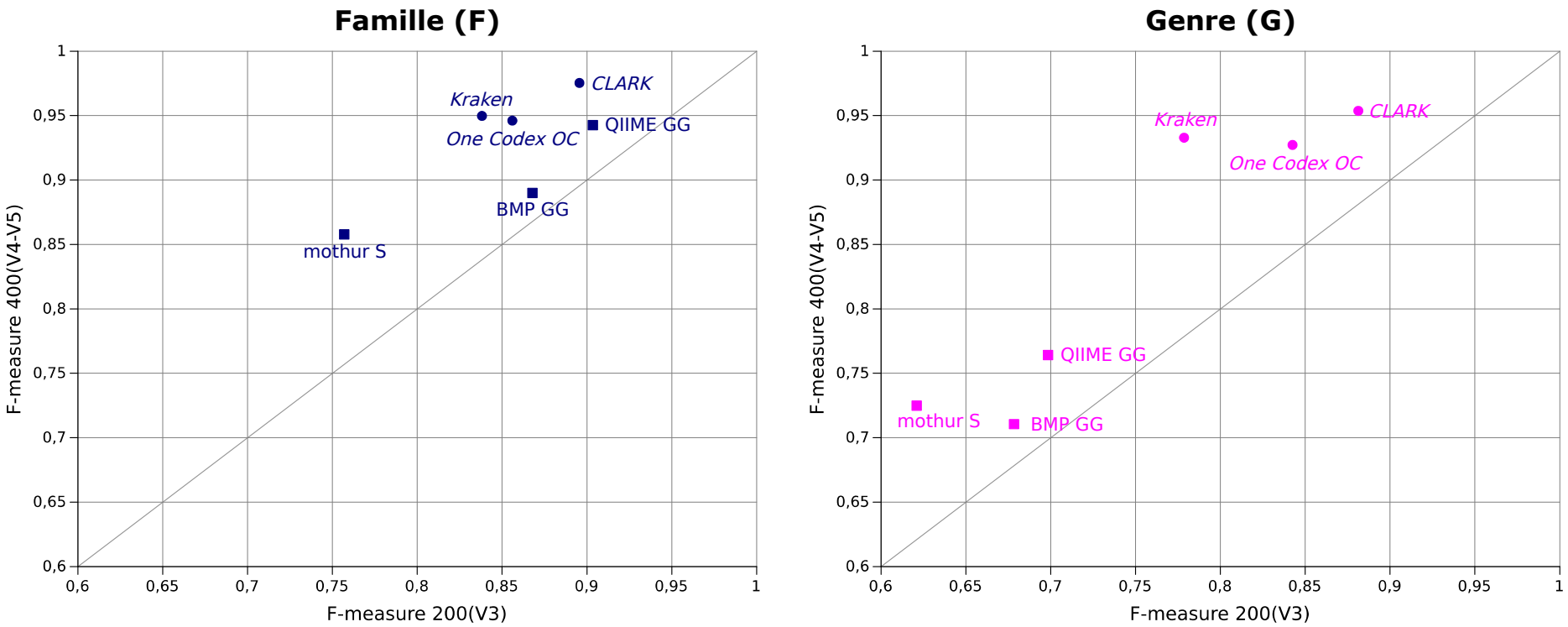


Figure 3.5 : Comparaison des F-mesures entre les amplicons 200(V3) et 400(V4-V5) au niveau de la famille (à gauche) et du genre (à droite) sur le jeu de données HC 50k avec simulation de lectures Ion Torrent (adapté de Siegwald et al. 2017)

Le changement de résolution taxonomique (de la famille au genre) affecte principalement les résultats des pipelines *clustering-first*, causant une chute de F-mesure entre 10 et 20 %, induite par une perte de précision et de rappel. Les pipelines *assignment-first* sont plus stables entre les deux niveaux taxonomiques, particulièrement pour l'amplicon 400(V4-V5).

Nous constatons avec surprise que le changement d'amplicon n'impacte pas drastiquement les résultats (dans la Figure 3.5, la F-mesure de tous les pipelines est relativement proche de la diagonale). mothur, kraken, CLARK et One Codex sont tout de même plus affectés par ce changement, gagnant environ 10 % de F-mesure au niveau de la famille en utilisant l'amplicon 400(V4-V5). Ce gain est causé par une augmentation du rappel (de 15 % F, 18,2 % F et 11 % F pour mothur, kraken et CLARK respectivement), sauf pour One Codex. L'augmentation de F-mesure pour ce dernier est causée à la fois par une augmentation de précision (de 9,6 % F) et du rappel (de 8,5 % F). QIIME n'est que peu impacté par le changement d'amplicon. En effet, l'utilisation de l'amplicon 400(V4-V5) cause une augmentation de F-mesure de moins de 4 % seulement pour ce pipeline au niveau de la famille. Cette augmentation est principalement causée par l'identification des deux familles bactériennes uniquement amplifiées par le couple d'amorces 400(V4-V5). BMP est le pipeline le moins impacté par le changement d'amplicon, probablement à cause de son étape de rognage des lectures, éliminant 25 % des nucléotides terminaux, ce qui cause une perte importante de l'information ajoutée par l'amplicon 400(V4-V5) plus long.

Avec des amplicons sans erreurs (voir données supplémentaires), l'amélioration des résultats en utilisant l'amplicon 400(V4-V5) est comme attendu beaucoup plus importante pour tous les pipelines. Ainsi, en présence d'erreurs, l'ajout de davantage de bases discriminantes dans l'amplicon 400(V4-V5) semble contre-balancé par la présence de nucléotides erronés dans les lectures. Il est ainsi essentiel d'évaluer la robustesse de chaque pipeline en présence d'erreurs de séquençage.

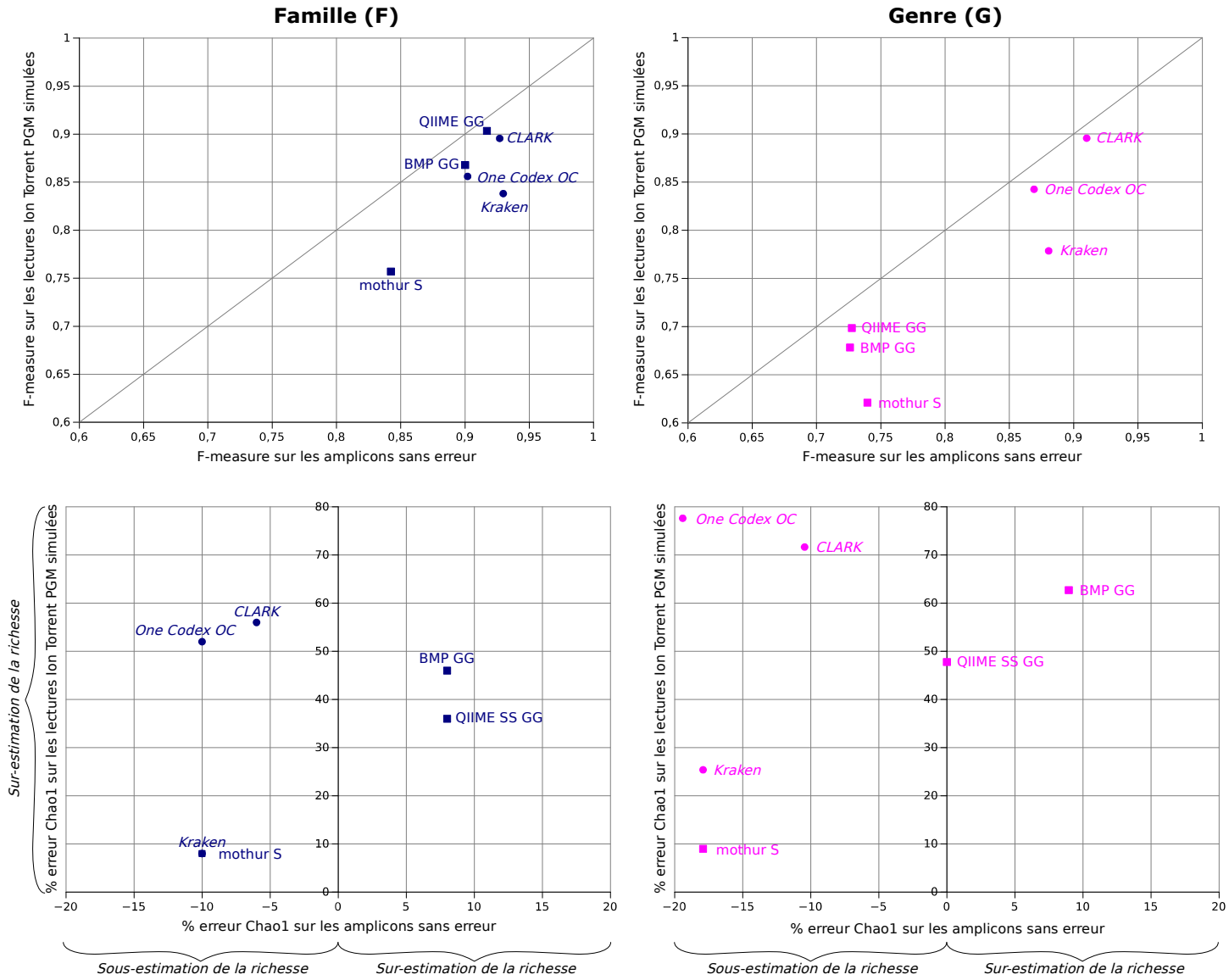


Figure 3.6 : Comparaison des F-mesures (haut) et du pourcentage d'erreur de richesse (bas) pour le jeu de données 200(V3) HC 50k avec et sans simulation de séquençage Ion Torrent PGM (adapté de Siegwald et al. 2017)

3.2.2 - Impact des erreurs de séquençage

Certaines technologies de séquençage haut-débit, telles que l'Ion Torrent PGM, génèrent des lectures présentant un taux d'erreurs non négligeable qui pourraient causer des erreurs dans les résultats d'analyse. Pour évaluer ce phénomène, les performances des pipelines ont été comparées sur les jeux de données avec et sans erreurs.

La Figure 3.6 (haut) représente les F-mesures des jeux de données 200(V3) 50k avec et sans simulation d'erreurs, au niveau de la famille et du genre. Comme constaté précédemment, tous les pipelines ont une F-mesure plus élevée au niveau de la famille et, comme attendu, la F-mesure chute en présence d'erreurs. Que ce soit avec ou sans erreurs, tous les pipelines présentent une F-mesure considérée comme acceptable (> 0,75) au niveau de la famille, même si mothur est proche de cette limite. Toutefois, seuls les pipelines *assignment-first* restent à un tel niveau de F-mesure au niveau du genre ; les pipelines *clustering-first* ont tous un rappel bien plus faible à ce niveau taxonomique. Cette chute de rappel est probablement causée par une mauvaise définition des OTUs au niveau du genre par ces pipelines (phénomène qui sera confirmé par les indices de *clustering* dans la section 3.2.5).

En présence d'erreurs, les pipelines sont affectés par une chute de rappel (jusqu'à 15,6 % G pour mothur). Ceci signifie qu'une lecture avec erreurs n'est préférentiellement pas classée, plutôt qu'assignée à un mauvais taxon. mothur et kraken sont les pipelines les plus affectés par la présence d'erreurs, avec une chute de F-mesure de 8,5 % F et 9,1 % F respectivement. QIIME est le pipeline le moins sensible aux erreurs de séquençage, avec une chute de F-mesure sous 3 % au niveau de la famille et du genre (principalement causée par une baisse de rappel). Mis à part pour One Codex, la précision des pipelines reste à peu près stable en présence d'erreurs (< 2,5 % de chute F & G). Il arrive même que la précision augmente légèrement (~1,5 % à 2 % F pour mothur et QIIME). Ce phénomène est dû à l'ajout d'erreurs dans certaines lectures qui sont déjà mal

assignées sans erreurs (faux positifs). Ces lectures avec erreurs engendrent une assignation taxonomique de qualité trop faible pour être validée par les pipelines, qui n'arrivent pas à les classifier, et les considèrent ainsi comme des faux négatifs – d'où une amélioration de la précision. One Codex est le seul pipeline ayant plus de faux positifs en présence d'erreurs (chute de précision d'environ 12-13 % F & G). Cette chute n'impacte pas significativement la F-mesure, puisque la chute de précision pour One Codex est associée à une augmentation du rappel. Pour ce pipeline, certaines lectures initialement non-classifiées (faux négatifs) deviennent mal identifiées (faux positifs) en présence d'erreurs. Avec erreurs, CLARK présente une chute de précision d'environ 6 % au niveau de la famille, réduite à 2 % au niveau du genre, ce qui démontre sa plus grande robustesse à ce niveau taxonomique plus fin.

Les erreurs de séquençage affectent également l'estimation de la richesse (Figure 3.6, bas). Les indices de richesse ont été calculés après assignation taxonomique (étape que nous avons nommée la fusion taxonomique des OTUs : tous les OTUs assignés au même taxon sont fusionnés et comptés comme une seule entité) afin de pouvoir être comparables entre pipelines *clustering-first* et *assignment-first*. Sans erreurs, QIIME et BMP sont les seuls pipelines qui surestiment la richesse, entre 8 % et 10 % d'erreur de Chao1 au niveau de la famille. À l'inverse, mothur, kraken et One Codex sous-estiment tous la richesse dans les mêmes proportions (-10 % F d'erreur). QIIME se rapproche beaucoup de la richesse réelle au niveau du genre (proche de 0 sur l'axe des abscisses dans la Figure 3.6 bas). QIIME est en outre le seul pipeline à mieux estimer la richesse au niveau du genre qu'au niveau de la famille ; la surestimation de richesse y est compensée par la chute de résolution au niveau du genre.

En présence d'erreurs, tous les pipelines surestiment la richesse. En effet, l'ajout d'erreurs de séquençage induit la formation d'un plus grand nombre de petits OTUs affectant le calcul de Chao1. Dans ce cas, kraken et mothur sont les pipelines les moins affectés au niveau de la famille et du genre (environ le même pourcentage d'erreur de richesse que sans erreurs de

séquençage), suivis par QIIME (augmentation de 36 % F d'erreur de richesse) et BMP (46 % F d'augmentation d'erreur de richesse). One Codex et CLARK sont les pipelines les plus sensibles aux erreurs en termes de richesse, atteignant autour de 50-55 % F et 70-80 % G d'erreur de Chao1 en présence d'erreurs de séquençage. Ce taux très élevé est causé par la mauvaise identification de beaucoup de petits taxons (< 5 lectures par taxon) pour ces pipelines en présence d'erreurs. Le classement des pipelines est similaire au niveau du genre, avec une augmentation du pourcentage d'erreur de richesse de la famille au genre (1 % pour kraken, plus de 25 % pour One Codex).

Comme attendu, la présence d'erreurs de séquençage dans les lectures cause une chute de F-mesure et une surestimation de la richesse, mais pas dans les mêmes proportions selon le pipeline utilisé. Ces résultats sont confirmés par l'interprétation des jeux de données de débit 25k et 100k (voir données supplémentaires).

3.2.3 - Impact du débit de séquençage

Les performances des pipelines ont ensuite été évaluées lorsqu'ils doivent gérer des quantités de lectures différentes – en utilisant les jeux de données simulés à différents débits de séquençage (25k, 50k et 100k). Pour les lectures sans erreurs, la richesse et la F-mesure restent stables peu importe le débit : puisque la composition bactérienne simulée ne varie pas entre les trois débits, les amplicons issus d'une même séquence de référence sont identiques peu importe le débit. Leur quantité n’impacte ainsi pas la qualité des résultats. À l'inverse, les pipelines sont très sensibles à la variation de débit en présence d'erreurs, surtout dans l'estimation de la richesse des communautés, comme présenté dans la Figure 3.7.

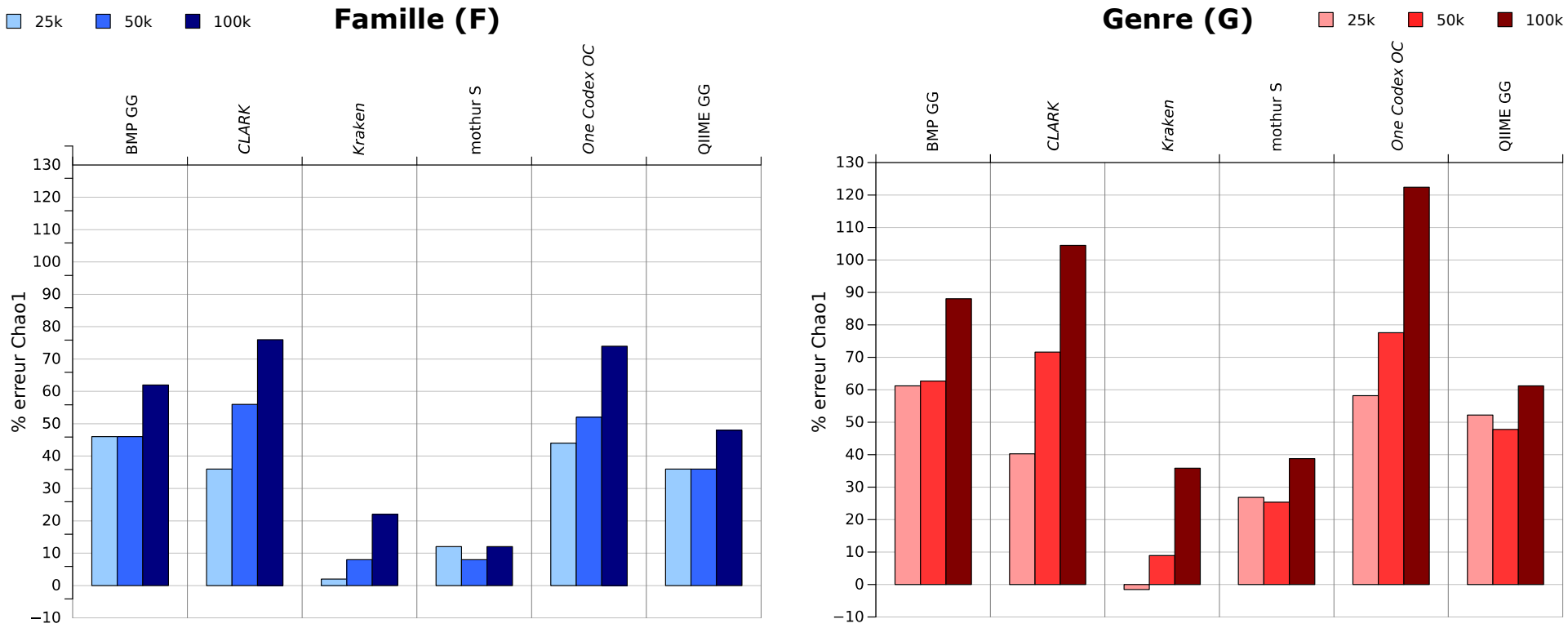


Figure 3.7 : Comparaison du pourcentage d'erreur de Chao1 sur le jeu de données 200(V3) HC avec simulation de lectures Ion Torrent PGM à trois débits différents (25k, 50k et 100k) au niveau de la famille et du genre, après fusion taxonomique des OTUs (adapté de Siegwald et al. 2017).

Comme observé dans la partie précédente, tous les pipelines surestiment la richesse en présence d'erreurs. Nous pouvons toutefois observer deux comportements bien distincts selon la catégorie des pipelines étudiés. Les pipelines *assignment-first* (kraken, One Codex et CLARK) augmentent le pourcentage d'erreur de richesse lorsque le débit augmente également. Ces pipelines sont les plus impactés par le changement de débit de séquençage, avec un pourcentage d'erreur au niveau de la famille d'au moins 20 % pour le jeu de données 25k, 30 % pour le jeu de données 50k et 40 % pour le jeu de données 100k, et plus encore au niveau du genre. One Codex et CLARK sont les plus sensibles à l'augmentation de débit, avec plus de 100 % de pourcentage d'erreur de Chao1 au niveau du genre pour le jeu de données 100k. La qualité de l'étape d'assignation taxonomique de ces pipelines est fortement dépendante du nombre de *k-mers* erronés dans les lectures, qui augmentent à plus haut débit. Cette surestimation est amoindrie pour kraken, probablement à cause de la banque réduite de *k-mers* qu'il utilise, les lectures avec beaucoup d'erreurs ayant plus de chances de ne pas être classifiées (faux négatifs), que d'être mal classifiées (faux positifs créant de nouveaux OTUs donc une augmentation de la richesse).

Les pipelines *clustering-first* (BMP, QIIME et mothur) sont plus stables entre les débits 25k et 50k. QIIME et BMP ne surestiment la richesse que pour le jeu de données 100k, et QIIME est plus proche de la richesse réelle que BMP au niveau de la famille et du genre. mothur est le pipeline le moins sensible à une variation de débit, même s'il augmente son pourcentage d'erreur de richesse de 100 % G entre les jeux de données 25k et 100k, et presque aucune variation au niveau de la famille. Les pipelines *clustering-first* semblent moins sensibles à la variation de débit, même s'ils surestiment tout de même drastiquement la richesse des échantillons ; en effet, l'étape de *clustering* (regroupant les lectures similaires jusqu'à un certain seuil d'identité) tout comme l'étape de fusion taxonomique (regroupant les OTUs ayant la même assignation taxonomique) minimisent l'impact de l'augmentation de débit sur l'estimation de richesse.

Il est important de souligner que les métriques de richesse et de diversité sont généralement calculées avant assignation taxonomique dans la plupart des études métagénétiques ; ces valeurs ont ainsi été générées avant fusion taxonomique pour les pipelines *clustering-first*, après *clustering* des lectures à un seuil de 97 % (Tableau 3.8).

	Chao1			**Simpson inverse**		
	25k	**50k**	**100k**	**25k**	**50k**	**100k**
BMP GG	1580	2780	4731	234,71	278,83	294,10
mothur S	1395	2655	4662	118.60	125,67	130.35
QIIME U GG	1140	1780	2764	149.27	151,73	154.11
QIIME SS GG	696	885	1256	130,31	129,33	129,09

Tableau 3.8 : Comparaison des indices de richesse Chao1) et de diversité (Simpson inverse) pour les pipelines clustering-first avant fusion taxonomique, sur le jeu de données 200(V3) HC avec simulation de séquençage Ion Torrent PGM, à trois débits différents (25k, 50k, 100k).

Pour les pipelines *clustering-first*, la surestimation de richesse et de diversité est de dix à cent fois plus importante avant fusion taxonomique. En effet, la présence d'erreurs dans les lectures augmente la distance entre celles issues du même amplicon originel. Ce bruit génère ainsi de nombreux petits OTUs isolés, augmentant l'estimation de richesse et de diversité. Pour tous ces pipelines, l'indice de Simpson inverse varie peu entre les trois débits (< 10 %), sauf pour BMP. Au contraire, l'indice Chao1 augmente fortement à plus haut débit, au moins de 80 % entre le jeu de données 25k et 100k. BMP est le pipeline le plus affecté en termes de richesse (augmentation d'environ 200 % de Chao1 entre 25k et 100k) et de diversité (augmentation de 25 % de l'indice de Simpson inverse). L'estimation de richesse faite par mothur est également affectée à des débits plus élevés (augmentation de 234 % de Chao1), mais sa variation de diversité reste sous 10 %. Le Tableau 3.8 confirme également que la version de QIIME utilisant SortMeRna et Sumaclust est plus performante que QIIME UCLUST, avec la plus petite variation de richesse entre les trois débits

(augmentation tout de même conséquente de 80 % de Chao1 entre 25k et 100k) et de diversité (diminution de 1 % de l'indice de Simpson inverse). Les nouveaux algorithmes de *clustering* utilisés dans QIIME permettent en effet une meilleure estimation de la richesse [Kopylova *et al.* 2016], d'où leur meilleure stabilité face à une variation de débit. Nous confirmons bien ici à quel point utiliser un seuil de similarité fixe pour représenter des unités taxonomiques est trompeur, surtout pour des lectures contenant des erreurs : le *clustering* semble bien plus guidé par les erreurs (d'où l'apparition de nombreux petits OTUs impactant la richesse et la diversité) que les similarités entre lectures.

La richesse et diversité sont généralement estimées avant fusion taxonomique dans la plupart des études métagénétiques utilisant un pipeline *clustering-first*. Il est toutefois essentiel de savoir que ces métriques sont largement surestimées dans un contexte de séquençage avec erreurs, et qu'elle ne peuvent par conséquent pas être évaluées comme quantités absolues représentatives du nombre et de la proportion de taxons présents dans un échantillon donné. Toutefois, l'étude de la variation de richesse et diversité entre différents groupes d'échantillons peut prendre sens dans une étude comparative, puisqu'avec un plan d'expérience adapté, les biais techniques seront relativement identiques pour tous les échantillons analysés. La variation significative de richesse et de diversité entre différents groupes d'échantillons pourra alors être interprétée selon les variables observées.

La variation de débit n'impacte significativement la F-mesure pour aucun des pipelines. Même si des débits plus élevés induisent l'identification de nombreux petits OTUs erronés, causant une augmentation de la richesse estimée, ces taxons ne concernent qu'une petite proportion des lectures, qui ont un impact négligeable sur la F-mesure. Nous avons néanmoins souhaité évaluer si une variation de composition plus importante pouvait avoir un impact sur la qualité des résultats des différents pipelines.

3.2.4 - Impact de la complexité des microbiotes

Sans jamais être clairement définie dans la littérature, la notion de complexité d'un microbiote peut être associée aux proportions des organismes qui le composent. En effet, un microbiote dit de faible complexité est principalement composé d'une ou deux espèces largement majoritaires, tandis qu'un microbiote dit de haute complexité est composé d'un grand nombre d'espèces équitablement représentées. Le respect de ces proportions dans les résultats d'analyse est crucial afin de pouvoir évaluer la structure et donc le fonctionnement du microbiote étudié. L'estimation de richesse et de diversité de chaque pipeline a été évaluée en analysant des échantillons de différentes complexités (LC contenant 30 % d'une espèce, MC contenant 4 espèces majoritaires à 20 % chacune, et HC contenant toutes les espèces dans les mêmes proportions). Le Tableau 3.9 présente les résultats de cette estimation pour chaque pipeline, avant et après fusion taxonomique pour les pipelines *clustering-first*.

	Avant fusion taxonomique			**Après fusion taxonomique (famille, valeur attendue = 50)**		
	LC	MC	HC	LC	MC	HC
BMP GG	2184	1066	2780	67	66	73
CLARK				73	64	78
kraken				53	49	54
mothur S	2111	1167	2655	54	52	54
One Codex OC				72	58	76
QIIME U GG	1370	713	1780	69	66	72
QIIME SS GG	773	422	885	67	60	68

Tableau 3.9 : Valeurs de Chao1 avant fusion taxonomique pour les pipelines clustering-first, et au niveau de la famille après fusion taxonomique pour tous les pipelines, à trois différentes complexités (LC, MC et HC, toutes composées de 50 familles bactériennes à proportions variables), sur le jeu de données 200(V3) 50k avec simulation de lectures Ion Torrent.

Comme observé précédemment, tous les pipelines ont tendance à surestimer la richesse sur les lectures simulées, peu importe la complexité de l'échantillon initial. Nous remarquons cependant que le nombre de familles est plus proche de la réalité avec le jeu de données MC, où la grande majorité des lectures est contenue dans trois familles principales. Cette tendance est accentuée pour les pipelines *clustering-first* avant fusion taxonomique, où la richesse est surestimée de deux à trois fois plus pour HC que pour MC. Une explication possible à ce phénomène est que les OTUs sont plus difficiles à délimiter lorsqu'ils sont nombreux et petits, comme c'est le cas pour LC et HC. Pour les données amplicon sans erreurs, la variation de complexité n'impacte pas significativement l'estimation de richesse.

La Figure 3.10 représente les proportions des dix familles majoritaires dans les résultats après fusion taxonomique pour tous les pipelines, aux trois niveaux de complexité LC, MC et HC. L'indice de *clustering* 1-NID est également représenté pour évaluer le partitionnement des lectures en OTUs. En effet, même si un pipeline a une F-mesure faible (à cause de nombreux faux positifs par exemple), une valeur de 1-NID élevée indique qu'il délimite tout de même correctement les OTUs. Les indices de *clustering* sont plus difficiles à interpréter au niveau du genre qu'au niveau de la famille à cause du nombre plus élevé de faux négatifs (OTUs non classifiés) à ce niveau ; les faibles valeurs de 1-NID au niveau du genre sont généralement causées par ces nombreuses lectures non-classifiées.

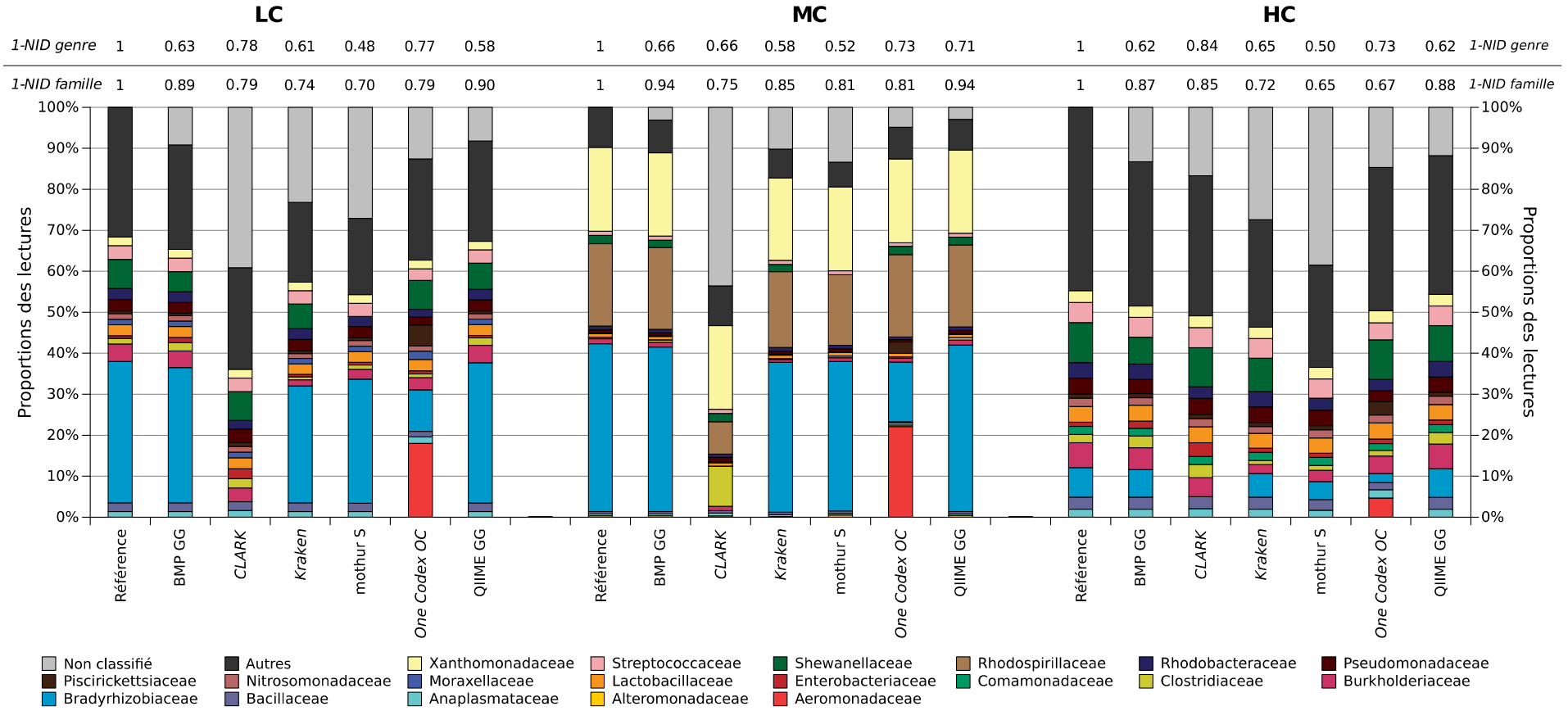


Figure 3.10 : Proportions des 10 familles majoritaires par pipeline pour les jeux de données LC, MC et HC 200(V3) 50k avec simulation de lectures Ion Torrent PGM, et indice de clustering 1-NID correspondant (calculé après fusion taxonomique) au niveau de la famille et du genre (adapté de Siegwald et al. 2017).

Pour le jeu de données de faible complexité (LC), tous les pipelines *clustering-first* réussissent à identifier la famille dominante, Bradyrhizobiaceae (en bleu dans la Figure 3.10). Ceci n'est pas le cas pour tous les pipelines *assignment-first* : One Codex présente une composition majoritaire différente de celle attendue, car de nombreuses lectures Bradyrhizobiaceae sont assignées comme étant de la famille Aeromonadaceae (en rouge dans la Figure 3.10), cette dernière étant plus représentée dans la banque One Codex 28k. Cependant, One Codex est l'un des pipelines délimitant au mieux les OTUs au niveau du genre, selon sa valeur de 1-NID de 0,77. Son faible taux de faux négatifs à ce niveau taxonomique contre-balance en effet les mauvaises proportions retrouvées. CLARK n'identifie pas du tout la famille Bradyrhizobiaceae, car les lectures correspondantes partagent trop de *k-mers* communs avec d'autres familles. Ces observations s'appliquent également pour le jeu de données MC, composé de quatre espèces dominantes regroupées en trois familles. Tous les pipelines reconnaissent correctement deux des trois familles (Xanthomonadaceae et Rhodospirillaceae). À nouveau, One Codex et CLARK n'identifient pas correctement la famille Bradyrhizobiaceae, composée de deux espèces dominantes dans cet échantillon. kraken est le seul pipeline *assignment-first* à correctement identifier cette famille.

Dans l'ensemble, les pipelines *clustering-first* générant le profil taxonomique le plus proche de la réalité sont QIIME et BMP, identifiant correctement les taxons majoritaires avec des valeurs d'indices de *clustering* élevés (> 0,86). Ces pipelines étant déjà les plus performants pour les jeux de données les plus complexes (HC), ils n'ont aucun mal à délimiter correctement les OTUs de jeux de données moins complexes. mothur est le pipeline le plus loin de la réalité pour tous les jeux de données, car il a le taux le plus élevé de lectures non classifiées (LC ~ 27 %, MC ~ 13 %, HC ~ 38%). De plus, mothur n'identifie pas la famille Shewanellaceae avec la banque SILVA, ce pipeline identifiant au mieux ces lectures au niveau de l'ordre.

Ces résultats prouvent qu'un microbiote dit de faible complexité n'est pas forcément synonyme d'une analyse facilitée. En effet, le jeu de données LC semble de faible complexité car il contient un seul organisme majoritaire ; il engendre toutefois des résultats moins bons que le jeu de moyenne complexité (MC), car il contient une quantité plus élevée de petits taxons, d'où une valeur de 1-NID plus faible pour la majorité des pipelines. La complexité d'une communauté semble plus justement définie par sa diversité : plus elle est élevée, plus la communauté sera complexe à analyser. La diversité initiale du jeu de données simulé LC est en effet plus élevée que la diversité du jeu de données MC. Ces observations ont été validées sur les jeux de données 25k et 100k ainsi qu'avec les autres indices de *clustering* (NMI et AMI). Ces analyses montrent donc que tous les pipelines ont plus de mal à délimiter correctement beaucoup de petits taxons dans une communauté très hétérogène (HC) que dans une communauté plus homogène composée de quelques taxons majoritaires (MC). Seul CLARK semble délimiter correctement les taxons sur le jeu de données HC, particulièrement au niveau du genre (1-NID = 0,86).

Cette variation de performances entre les pipelines peut être causée par leurs recommandations d'usage, différant par les algorithmes intégrés et les banques de séquences de référence recommandées. Changer ces dernières dans l'analyse permet d'évaluer leur importance pour chaque pipeline.

3.2.5 - Impact de la banque de séquences de référence

Que ce soit pour les pipelines *clustering-first* ou *assignment-first*, le choix d'une banque de séquences de référence apparaît comme un élément évidemment essentiel. Toutefois, différents pipelines recommandent l'utilisation de différentes banques de référence. Le choix d'une banque de référence est ainsi souvent guidé par le choix d'un pipeline. L'intégration de banques alternatives à celles recommandées par défaut nous a permis d'étudier l'impact du changement de banque de référence dans l'analyse. À noter que nos jeux de données simulés n'ont pas été créés dans l'optique d'évaluer la variation

de composition des banques utilisées : en effet, les taxons sélectionnés sont suffisamment bien décrits dans toutes les banques pour ne pas impacter significativement les métriques d'évaluation d'une banque à l'autre. Toutefois, toutes les banques ne sont pas intégrées de la même façon dans tous les pipelines ; ainsi, pour un même pipeline, changer de banque de référence peut tout de même avoir un impact sur les résultats.

Chaque pipeline propose une banque de séquences de référence par défaut. mothur recommande d'aligner les lectures en utilisant SILVA, et de classifier les OTUs en utilisant la taxonomie RDP. À l'inverse, QIIME et BMP préfèrent utiliser la banque Greengenes et la taxonomie qui y est associée. Les banques de séquences utilisées avec les pipelines *clustering-first* peuvent limiter la résolution taxonomique des résultats ; en effet, la banque SILVA intégrée dans mothur n'est pas annotée au niveau de l'espèce [Schloss 2014a], et la banque Greengenes intégrée dans QIIME contient moins de 7 % de séquences annotées au niveau de l'espèce. Pour les pipelines *assignment-first*, la plupart des banques proposées par défaut sont des variantes de RefSeq. One Codex a été évalué à la fois en utilisant RefSeq et One Codex 28k (par défaut). Cette dernière banque inclut les *k-mers* présents dans RefSeq, ainsi que 22 710 génomes supplémentaires. One Codex a également intégré une version beta de la banque SILVA. Les auteurs de kraken proposent pour une utilisation routinière l'emploi d'une banque de référence réduite, Minikraken, pour diminuer les ressources nécessaires à l'analyse. Cette banque contient 10 000 *k-mers* sélectionnés aléatoirement dans RefSeq. Minikraken a été utilisé dans notre évaluation, car il s'agit de la seule banque directement mise à disposition par les auteurs de kraken, et serait le choix par défaut d'un utilisateur ayant à sa disposition des ressources informatiques limitées. Enfin, CLARK fournit à ses utilisateurs un exécutable pour leur permettre de construire leur propre banque de *k-mers* à un niveau taxonomique donné, à partir de génomes bactériens, viraux, humains ou d'une sélection prédéfinie issus de RefSeq. Les banques générées par CLARK ont la particularité de ne contenir que des *k-mers* discriminant les taxons à un

niveau taxonomique donné. CLARK a été exécuté sur une banque de *k-mers* discriminants au niveau du genre, extraits des séquences bactériennes de RefSeq.

La Figure 3.11 représente la variation des résultats de chaque pipeline selon la banque de référence utilisée. Pour les pipelines *clustering-first*, le changement de banque n'impacte pas significativement la F-mesure. Toutefois, il affecte fortement l'estimation de richesse, particulièrement pour QIIME et BMP. Pour ces pipelines, utiliser SILVA au lieu de Greengenes (par défaut) double au minimum l'estimation de richesse au niveau de la famille, même si SILVA permet d'améliorer la précision des résultats. mothur est plus robuste au changement de banque, principalement à cause de ses taux de rappel plus faibles (les lectures faux-négatif ne sont identifiées avec aucune des banques).

Pour les pipelines *assignment-first*, kraken présente le taux d'erreur de richesse le plus faible, tout comme la meilleure précision. Il montre toutefois un taux de rappel plus faible, probablement causé par le nombre plus faible de *k-mers* présents dans Minikraken. CLARK, utilisant une banque complète de RefSeq ainsi qu'un algorithme de classification taxonomique plus précis, a une meilleure F-mesure (due à un meilleur rappel) que kraken. Toutefois, ses résultats semblent plus bruités puisqu'il surestime plus la richesse à cause de nombreux petits taxons mal identifiés, d'où une précision plus faible que kraken. One Codex est le seul pipeline dont l'utilisation de la banque par défaut ne donne pas les meilleurs résultats. La banque One Codex 28k n'améliore pas les résultats comparativement à RefSeq, bien au contraire : la surestimation de richesse est bien plus grande, et la F-mesure est plus faible (aux niveaux de la famille et du genre).

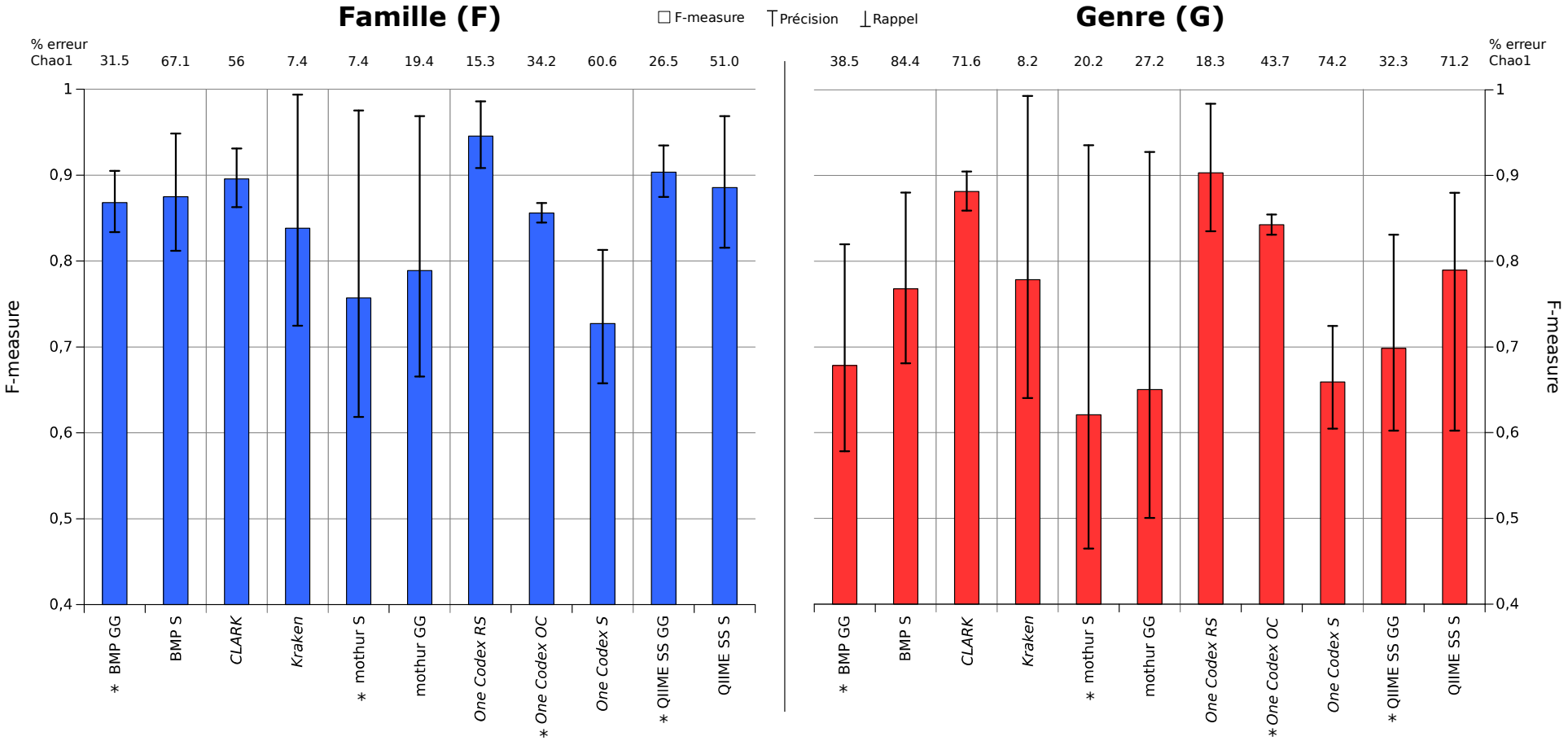


Figure 3.11 : F-mesure et pourcentage d'erreur de richesse après fusion taxonomique pour chaque pipeline sur le jeu de données 200(V3) 50k HC avec simulation de lectures Ion Torrent PGM, en utilisant différentes banques de référence (les banques recommandées par défaut pour chaque pipeline sont annotées avec une astérisque) (adapté de Siegwald et al. 2017).

Ces résultats peuvent indiquer que les génomes de référence ajoutés dans One Codex 28k, non validés dans RefSeq, causent une chute de précision. L'intégration expérimentale de SILVA dans One Codex aurait pu permettre d'évaluer les performances des pipelines *assignment-first* sur une banque dédiée à la cible étudiée ; toutefois, les résultats d'analyse sont bien moins bons que ceux générés par One Codex en utilisant une banque complète (RefSeq ou dérivée). Après discussion avec les développeurs de One Codex, il s'avère que leur intégration de SILVA n'avait pas été curée à l'époque de nos tests : de nombreuses séquences de références y étaient annotées comme « environnementales » ou « non cultivées ». Les lectures assignées à ces taxons ne pouvaient de ce fait pas être discriminées, et causent ainsi une chute de F-mesure et une mauvaise estimation de la richesse.

Les résultats de cette partie montrent que les pipelines *clustering-first* recommandent des banques de séquences en accord avec les algorithmes qui y sont intégrés par défaut, afin de minimiser au maximum les erreurs dans les résultats. Les pipelines *assignment-first*, quant à eux, sont très sensibles à l'annotation des séquences dans leurs banques de référence.

Nos jeux de données artificiels contenaient des organismes non présents dans les banques spécifiques d'ADNr (Chapitre 3, Section 3.1.1), afin d'évaluer le comportement des pipelines face à des séquences non référencées. Toutefois, la proportions de lectures concernées était trop faible pour impacter significativement les mesures de richesse, diversité, *clustering* et la F-mesure. Le tableau de l'Annexe 3 présente à quel organisme (au niveau du genre et de la famille) les lectures concernées ont été assignées par chaque pipeline. Ces informations ne permettent pas de conclure quant à la capacité de chaque pipeline à appréhender des séquences non référencées dans les banques de référence. Nos jeux de données ne sont pas adaptés à une telle évaluation ; il aurait été nécessaire d'y ajouter une proportion plus importante de séquences aléatoires pour évaluer leur effet sur les pipelines. Une telle évaluation serait intéressante pour étudier le comportement des pipelines face à microbiotes dits

« exotiques », dont les organismes sont pour la plupart inconnus (par exemple dans l'étude de grands fonds marins [Danovaro *et al.* 2010]).

3.2.6 - Validation sur un jeu de données réel

Un jeu de données réel contient du bruit inhérent aux biais techniques de préparation de librairies et de séquençage, qui ne peut être simulé. Afin de valider les conclusions tirées de l'interprétation des résultats d'analyse de données simulées, un jeu de données réel a été analysé avec les 6 pipelines. Ce jeu de données de microbiote intestinal humain (identifiant SRA : SRX364048) contient 231 660 lectures issues de l'amplicon 200(V3). Ce jeu de données contient plus de bruit (taille moyenne des lectures : 151 nt, déviation standard : 45nt) avec une qualité globale plus faible que notre jeu de données 200(V3) 100k simulé (taille moyenne des lectures : 185 nt, déviation standard : 11nt). La Figure 3.12 représente les résultats d'analyse obtenus sur le jeu de données réel avec les différents pipelines et banques de référence associées.

Pour tous les pipelines, nous observons une quantité de lectures non classifiées bien plus importante pour le jeu de données réel, pouvant être causée par le bruit expérimental, tout comme par la présence d'organismes dans l'échantillon qui ne sont pas présents dans les banques de référence. Dans l'article associé à ce jeu de données, les analyses ont été effectuées en utilisant QIIME UCLUST. Les auteurs ont décrit la présence de trois familles majoritaires (Lachnospiraceae, Ruminococcaceae et Veillonellaceae) dans le groupe d'échantillons dont a été tiré ce jeu de données. Seul One Codex SILVA, au profil de résultat bien différent, n'a pas retrouvé ces familles, ce qui confirme la mauvaise intégration de cette banque dans ce pipeline. La famille Oscillospiraceae est uniquement retrouvée par les pipelines *assignment-first* utilisant des banques non-spécifiques à l'ADNr 16S ; les autres pipelines identifient ces mêmes lectures comme appartenant à la famille Ruminococcaceae. Sans connaissance *a priori* de l'échantillon, il est impossible de déterminer à quelle famille ces lectures appartiennent réellement.

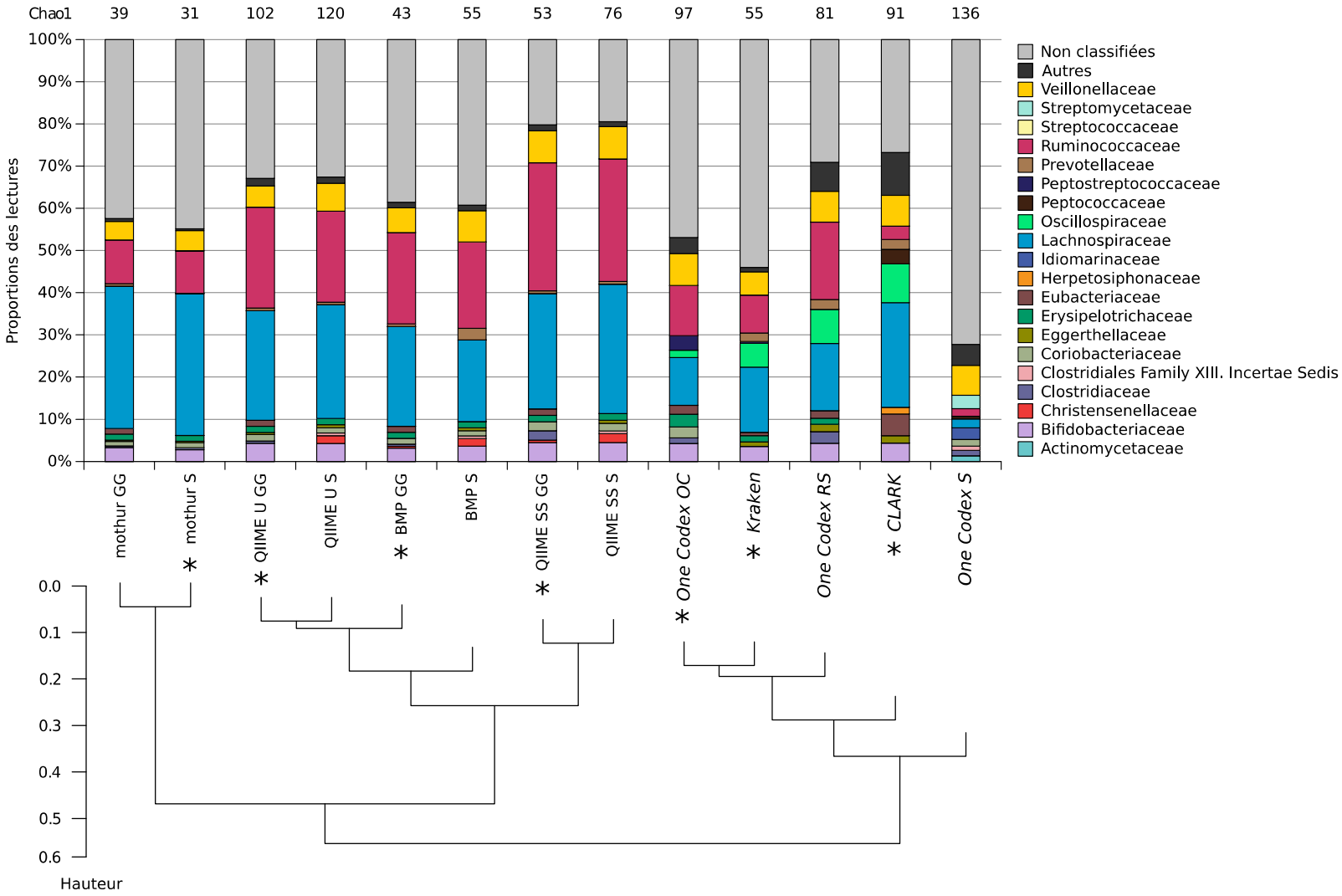


Figure 3.12 : Proportions des 10 familles majoritaires retrouvées par pipeline dans un jeu de données réel, et indices de richesse Chao1 (calculés après fusion taxonomique) au niveau de la famille. Au-dessous, clustering hiérarchique de tous les pipelines à partir de la quantité de lectures par famille par pipeline (en excluant les lectures non classifiées). Les pipelines sont annotés par une astérisque lorsqu'ils ont été exécutés avec leur banque de référence recommandée par défaut (adapté de Siegwald et al. 2017).

Un *clustering* hiérarchique par lien moyen de tous les pipelines a été effectué en se basant sur un calcul de distance Euclidienne à partir de la quantité de lectures par famille par pipeline (en excluant les lectures non classifiées) (Figure 3.12 bas). Les pipelines *clustering-first* sont tous regroupés sur une même branche du *clustering* hiérarchique. QIIME U, QIIME SS et BMP sont regroupés puisqu'ils utilisent des étapes d'analyse similaires, et le même algorithme de classification taxonomique des OTUs. QIIME SS est le pipeline *clustering-first* le plus récent, ayant le taux de lectures non-classifiées le plus faible tout comme l'indice de richesse estimée le plus bas parmi les trois pipelines. Ceci confirme ses bonnes performances déjà observées sur nos jeux de données simulés. mothur est également regroupé avec les pipelines *clustering-first* dans le *clustering* hiérarchique ; son fort taux de lectures non classifiées corrobore le faible rappel observé pour les données simulées, et explique sa faible estimation de richesse. Nous observons que la richesse estimée est plus faible pour les pipelines *clustering-first* en utilisant la banque de séquences recommandée par chacun d'eux, comme nous l'avons déjà observé sur les données simulées. Par contre, le jeu de données réel révèle un autre phénomène important lié aux banques de référence : l'importance de la taxonomie qui y est associée. Par exemple, tous les pipelines utilisant Greengenes identifient une proportion de lectures comme appartenant au genre *Eubacterium* et à la famille Eubacteriaceae, qui semble toutefois absente de SILVA. En réalité, dans SILVA, certaines espèces d'*Eubacterium* sont classées dans la famille Erysipelotrichaceae. Une certaine expertise de la taxonomie associée aux organismes d'intérêt retrouvés dans les résultats est ainsi essentielle à une interprétation avisée des résultats.

Les pipelines *assignment-first* sont tous regroupés par le *clustering* hiérarchique. CLARK est le pipeline capable d'identifier le plus grand nombre de lectures, particulièrement au niveau du genre (voir données supplémentaires). One Codex et kraken montrent des comportements similaires, ayant le même algorithme d'alignement de *k-mers* et d'assignation

taxonomique. One Codex OC identifie moins de lectures à cause de la banque de séquences One Codex 28k, comme montré dans la section précédente. L'utilisation de la version expérimentale de SILVA dans One Codex provoque à nouveau une grande quantité de lectures non identifiées dans les résultats (> 60 %), à cause du nombre trop élevé de lectures non classifiées dans la banque. One Codex RS a le moins de lectures non identifiées, grâce à la banque de référence RefSeq qui est curée et bien annotée. kraken produit une grande quantité de lectures non assignées, qui sont identifiées à des niveaux taxonomiques moins détaillés (ordre), à cause de l'utilisation d'une banque réduite (Minikraken) ne contenant pas assez de *k-mers* discriminants. Ce jeu de données a été ré-analysé avec kraken en utilisant la banque RefSeq complète ; les résultats générés étaient très proches de One Codex RS, prouvant à nouveau l'importance d'utiliser des banques de référence exhaustives et bien annotées pour les méthodes *assignment-first*.

Une analyse de beta-diversité a été exécutée entre les résultats de tous les pipelines sur les données réelles, en supposant que chaque pipeline puisse être considéré comme un contexte de biodiversité différent. La Figure 3.13 présente l'analyse en coordonnées principales résultante, qui confirme le *clustering* hiérarchique des pipelines observés en Figure 3.12.

Enfin, un test non-paramétrique de Mann-Whitney-Wilcoxon [Wilcoxon 1945, Mann-Whitney 1947] a été utilisé pour comparer les indices de diversité et de richesse entre les deux catégories de pipelines, *clustering-first* et *assignment-first*, qui étaient déjà ségréguées par le *clustering* hiérarchique. La diversité estimée était significativement plus élevée pour les pipelines *assignment-first* ($p < 0{,}005$ pour les indices de Shannon et Simpson inverse), validant la ségrégation des catégories des pipelines en fonction de la diversité.

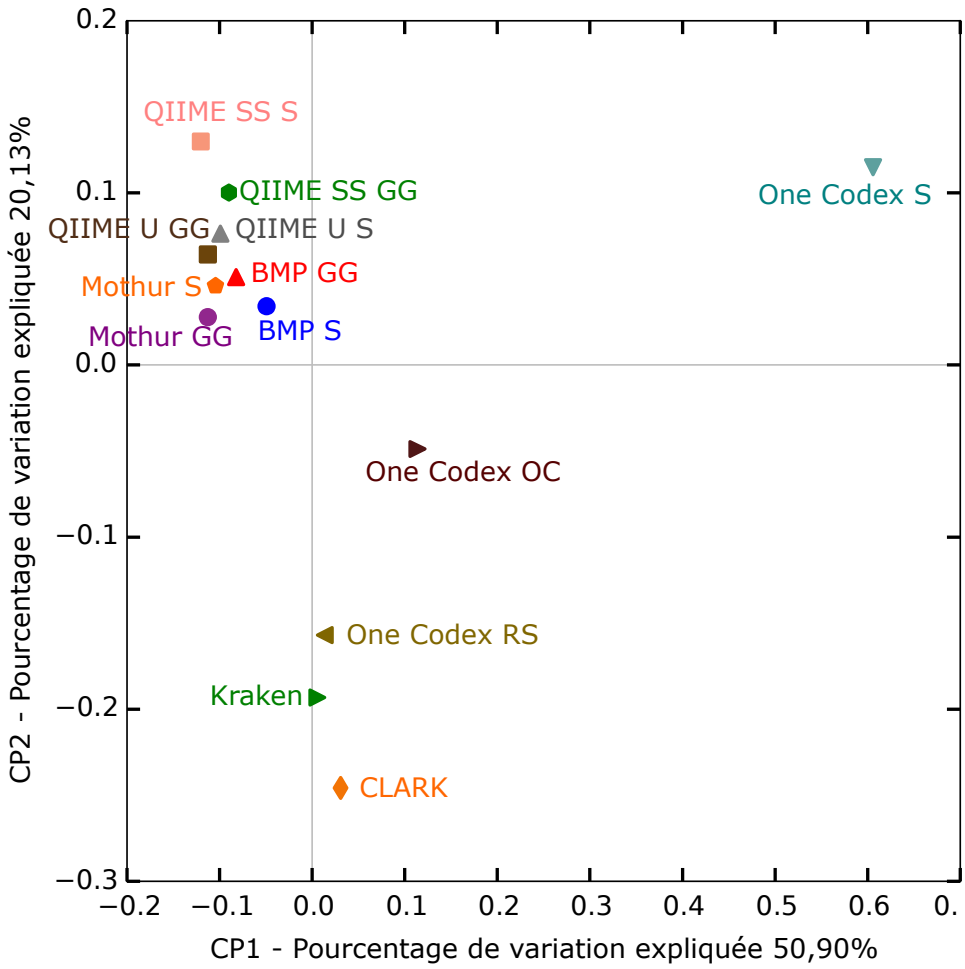


Figure 3.13 : Analyse en coordonnées principales (CP) de la beta-diversité entre tous les pipelines (adapté de Siegwald et al. 2017).

Cependant, aucune différence significative n'a été remarquée entre les deux catégories pour les mesures de richesse. Ceci peut s'expliquer par la plus grande sensibilité des indices de richesse au bruit dans les données, plus présent dans les données réelles, et géré de manière très différente selon les pipelines comme montré dans les sections précédentes.

L'analyse du jeu de données réel permet de confirmer plusieurs observations effectuées précédemment sur les jeux de données simulés. Nous pouvons ainsi conclure que nos simulations modélisent correctement des jeux

de données réels, même si ces derniers contiennent plus de bruit expérimental et possiblement une proportion plus importante de taxons non représentés dans les banques de référence.

3.2.7 - Temps d'exécution et ressources requises par chaque pipeline

Les ressources informatiques nécessaires à l'analyse sont des éléments importants à prendre en compte dans le montage du plan d'expérience, tout comme le temps d'analyse requis. Un pipeline générant d'excellents résultats ne pourra être utilisé dans toutes les conditions s'il nécessite une infrastructure informatique conséquente ou s'il demande des temps d'exécution de l'ordre de plusieurs jours. Nous avons mesuré dans cette étude les ressources (mémoire & temps CPU) utilisées par les pipelines QIIME, BMP, mothur, CLARK et kraken ont été mesurées, ainsi que le temps total d'exécution (temps utilisateur). Cette évaluation n'a pu être effectuée sur One Codex, qui fonctionne uniquement en tant que web service. La Figure 3.14 représente le pic d'utilisation de mémoire vive de chaque pipeline pour trois jeux de données différents, ainsi que le temps utilisateur et le temps CPU nécessaires pour les analyser.

kraken est de loin le pipeline le plus rapide et l'un de ceux nécessitant le moins de mémoire vive, utilisant moins de 500 MB de mémoire pour tous les jeux de données et s'exécutant en une seconde seulement, grâce à la banque réduite Minikraken. Par défaut, la banque de *k-mers* de kraken n'est pas pré-chargée dans la mémoire. Un tel pré-chargement est possible en utilisant l'option --preload, ce qui diminue encore plus le temps d'exécution de kraken, mais nécessite une quantité de mémoire vive au moins équivalente à la taille de la banque utilisée. Ce compromis doit être envisagé en utilisant la banque complète RefSeq, qui augmente considérablement les temps d'exécution. CLARK a été évalué comme étant plus rapide et demandant moins de mémoire que kraken [Ounit *et al.* 2015] puisque sa banque de référence ne contient que les *k-mers* discriminants à un niveau taxonomique donné.

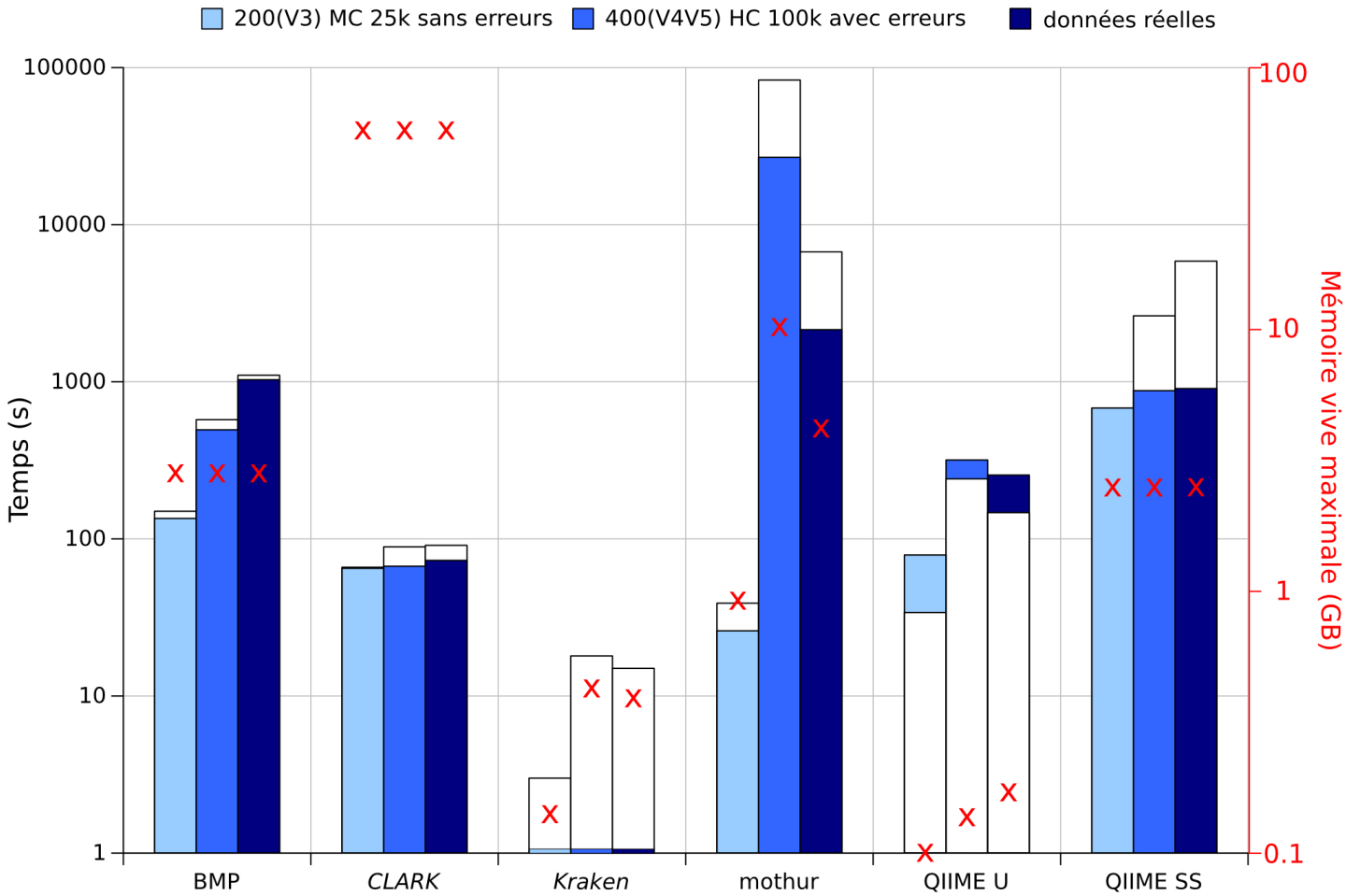


Figure 3.14 : Histogramme du temps utilisateur (couleur) et du temps CPU (blanc) et pic d'utilisation de mémoire vive (croix rouges) pour chaque pipeline et l'analyse de trois jeux de données différents (adapté de Siegwald et al. 2017).

Dans notre contexte, CLARK utilise les *k-mers* discriminants extraits de RefSeq dans sa totalité : cette approche demande ainsi plus de ressources (jusqu'à 160 Gb de mémoire vive) que la banque réduite proposée par kraken avec Minikraken (< 1 Gb de mémoire vive). Utiliser une plus petite banque de *k-mers* permet de maintenir la précision des résultats tout en consommant moins de mémoire vive, mais une banque complète de *k-mers* permet un meilleur rappel, comme vu dans les sections précédentes. CLARK propose un algorithme alternatif nécessitant moins de mémoire vive (CLARK-l) qui n'a toutefois pas été testé lors de notre étude. Comme kraken, CLARK permet également de précharger la banque de *k-mers* en mémoire, ce qui diminue les temps d'exécution mais augmente la mémoire vive nécessaire.

QIIME U est le pipeline consommant le moins de mémoire vive (au plus 150MB), variant très peu d'un jeu de données à l'autre. C'est également le pipeline *clustering-first* le plus rapide, s'exécutant en moins de 5 min pour le jeu de données réel. Nous avons constaté avec surprise que l'étape de *clustering* de QIIME U, pourtant parallélisée, nécessitait plus de temps utilisateur que de temps CPU, ce qui est peut-être causé par une étape non-optimisée de regroupement des résultats générés par chaque CPU. BMP ne nécessite pas d'alignement des lectures à une banque de référence, ce qui devrait le rendre plus rapide. Toutefois, le pipeline intègre par défaut une étape de génération d'arbre phylogénétique qui, quant à elle, nécessite un alignement des lectures à une banque de référence. Cette étape rallonge le temps d'exécution, mais est optionnelle si elle n'est pas requise par l'utilisateur. QIIME U et BMP utilisent tous les deux UCLUST, dont la version gratuite est compilée pour une architecture 32-bits uniquement, et est limitée à un maximum d'utilisation de 4GB de mémoire vive (2GB pour Windows). Cette limite n'a pas été atteinte dans notre étude, mais peut être un facteur limitant dans d'autres contextes d'analyse. La version 64-bits de UCLUST lève cette limitation, mais nécessite une licence payante. L'équipe de développement de QIIME cherche à rendre ses sources complètement ouvertes, d'où l'intégration de nouveaux algorithmes de

clustering open-source comme SortMeRNA, Sumaclust et Swarm dans la version 1.9. QIIME SS nécessite plus de mémoire (environ 2,5 GB pour tous les jeux de données) et de temps d'exécution que son prédécesseur, mais génère de meilleurs résultats comme démontré précédemment.

mothur était le pipeline le plus variable dans cette étude, ce qui apparaît également dans les ressources qu'il consomme selon différents jeux de données. Son étape limitante est l'interprétation de la matrice de distance entre lectures, nécessitant le plus de mémoire (de 1,4 GB pour LC 25k 200(V3) jusqu'à 45,5 GB pour MC 100k 400(V4-V5)); plus la matrice est grande, plus il faut de mémoire vive et de temps pour l'interpréter. Pour le jeu de données le plus complexe, le temps utilisateur atteint plus de 7 heures d'analyse. L'espace disque disponible est également un facteur à prendre en compte pour ce pipeline, dont les fichiers intermédiaires ont requis environ 16 GB d'espace disque pour le jeu de données HC 100k 400(V4-V5) avec simulation de lectures Ion Torrent PGM, à cause du très lourd fichier contenant la matrice de distances entre les lectures. Ce parti pris a été justifié par l'auteur de mothur, qui encourage les utilisateurs à utiliser des technologies de meilleure qualité, qui sont plus compatibles avec son pipeline [Schloss 2014b].

L'utilisation de services web tels que One Codex épargne à l'utilisateur la prise en compte des infrastructures informatiques requises dans le choix de sa méthode d'analyse. Toutefois, ces approches présentent aussi certaines limites. Tout d'abord, elles requièrent une connexion à Internet stable et un débit conséquent, puisque toutes les données brutes issues du séquençage doivent y être téléversées. Pour cette étude, l'ensemble des jeux de données à analyser représente 1,62 GB de données qui ont dû être transférées à des serveurs externes pour pouvoir être analysées par le service web One Codex. De plus, l'utilisation d'un service externe soulève la question de la gestion de la propriété et confidentialité de ces données. En outre, ce service autorise un maximum de 5 envois de fichiers simultanés, et des fichiers de 5 GB maximum, ce qui peut également être un facteur limitant. Par contre, One Codex a analysé tous les jeux

de données en quelques secondes, ce qui le rend aussi performant en termes de temps d'exécution que les autres pipelines *assignment-first*. Cette performance est contre-balancée par le temps humain nécessaire pour envoyer les fichiers bruts et récupérer les résultats d'analyse, même si une API (interface applicative de programmation) est en cours de développement.

3.3 – Bilan de l'étude et perspectives

Notre construction d'un protocole d'évaluation de pipelines d'analyse métagénétique a permis l'évaluation de 6 pipelines d'analyse, 3 *clustering-first*, et 3 *assignment-first*. Nous avons pu montrer que tous les pipelines sont sensibles à différentes variations du plan d'expérience. Ces résultats sont résumés dans la Figure 3.15.

Concernant les pipelines *clustering-first*, mothur est celui ayant montré les moins bonnes performances dans notre contexte d'étude. Il a en effet été développé afin d'analyser des données de séquençage d'excellente qualité, contenant ainsi un faible taux d'erreurs. Ce pipeline n'est de ce fait pas adapté à des technologies induisant un taux non négligeable d'erreurs dans les lectures, comme les technologies 454, Ion Torrent PGM, et d'Oxford Nanopore Technologies. BMP est le seul pipeline développé en prenant en compte les spécificités des erreurs induites par cette technologie dans les lectures ; ses résultats ne sont toutefois pas les meilleurs parmi les pipelines *clustering-first*. De meilleurs résultats ont été générés par QIIME, plus particulièrement en utilisant les nouveaux algorithmes qui y ont été intégrés. Ces derniers réduisent la surestimation de richesse, ce problème étant bien connu pour les approches *clustering-first* où la richesse est généralement estimée avant l'assignation taxonomique, et ce de façon d'autant plus importante pour les lectures comprenant des erreurs, comme montré dans notre étude. Ces erreurs réduisent la similarité des lectures appartenant à un même taxon, et rendent d'autant plus caduque l'utilisation d'un seuil de similarité fixe (habituellement 97 %) dans le *clustering* des lectures.

	Variables		Clustering-first			Assignment-first		
			mothur	QIIME	BMP	Kraken	One Codex	CLARK
Échantillon	Complexité		✓	✓		✓		✓
Séquençage	Locus				✗		✓	
	Débit (richesse)			✓	✓			✗
	Erreurs	F-measure		✓	✓			
		Richesse	✓			✓	✗	✗
Pipeline	Banque par défaut			✓	✓		✓	✓
	Temps d'exécution		✓			✓	✗	✗
	Paramétrages		✓	✓	✓			✓
	Convivialité		✗			✓	✓	✓
Résultats	Niveau taxonomique		✗	✗	✗			✓

Figure 3.15 : Résumé de la performance de chaque pipeline (avec sa banque de référence par défaut) en variant différents paramètres du plan d'expérience. Les disques de couleur représentent l'impact de ces paramètres sur les résultats pour chaque catégorie de pipeline (croix rouge = impact négatif, coche verte = robustesse, pas de symbole = pas d'impact notable) (adapté de Siegwald et al. 2017).

De nouveaux algorithmes de *clustering* comme SWARM contournent ce problème en n'utilisant pas de seuil de similarité fixe entre lectures, et commencent à être intégrés à des pipelines utilisés dans la communauté scientifique (par exemple QIIME *de novo*).

Cette étude a également confirmé que les pipelines *assignment-first* sont utilisables dans un contexte métagénétique, et sont tout aussi performants pour des technologies de séquençage induisant des erreurs dans les lectures. Ces pipelines ont en outre l'avantage d'être beaucoup plus rapides que les pipelines *clustering-first* (de 10 à plus de 10 000 fois, comme montré dans la Figure 3.14). De plus, de nombreux développements de ces algorithmes ont été récemment publiés, comme l'utilisation de *k-mers* espacés permettant d'améliorer les résultats [Břinda *et al.* 2015, Ounit & Lonardi 2016]. Ces pipelines sont toutefois fortement dépendants de la banque de référence qu'ils utilisent, qui n'est souvent pas adaptée à l'étude d'un locus défini. L'adaptation de ces pipelines à des analyses métagénétiques nécessiterait le développement de banques de *k-mers* spécifiques aux séquences cibles. Il est enfin important de rappeler que malgré les résultats prometteurs de ces pipelines dans un contexte métagénétique, ils sont sévèrement limités dans l'étude d'un microbiote exotique peu décrit dans les banques de référence, et pour lequel les approches *clustering-first de novo* restent une référence pour identifier les unités taxonomiques distinctes, même sans pouvoir les annoter.

Notre protocole d'évaluation de pipelines, comprenant des jeux de données simulés et réel ainsi que des métriques d'évaluation adaptées, est à ce jour le seul protocole publié permettant d'évaluer un pipeline d'analyses métagénétiques dans son entièreté, selon différents critères : sa précision et son rappel (F-mesure), son estimation de la richesse/diversité, et son partitionnement des lectures en différentes unités taxonomiques (indices de *clustering*). Il a permis une évaluation de pipelines d'analyses courants dans un contexte Ion Torrent PGM, révélant des comportements propres à différentes variables du plan d'expérience. En outre, ce protocole a été utilisé afin de valider

pour la première fois la performance de pipelines *assignment-first* pour l'analyse de données de métagénétique. Ce protocole a été mis à disposition de la communauté afin qu'il puisse être utilisé dans d'autres contextes de séquençage et d'analyse. Les jeux de données simulés peuvent également être utilisés comme échantillons contrôle dans une analyse métagénétique afin de mieux cerner les biais potentiellement présents dans les résultats d'analyse, et de permettre de mieux prendre en compte ces biais dans leur interprétation.

Chapitre 4 - Impact de la variation de pipeline d'analyse dans les conclusions d'une étude métagénétique du microbiote intestinal humain

4.1 - Contexte biologique

L'étude du microbiote intestinal est une des premières applications de la métagénétique bactérienne en recherche en santé humaine : en effet, plus de 1300 articles ont été publiés ces quinze dernières années sur ce sujet, chiffre augmentant de façon exponentielle depuis 2006 (soit depuis l'avènement du séquençage haut-débit). Plus de 1000 espèces bactériennes différentes ont été décrites dans le microbiote intestinal humain, représentant plus de 3 millions de gènes [Qin *et al.* 2010]. Sa composition et son activité métabolique ont été associées avec de nombreuses pathologies [Guinane & Cotter 2013] ; certaines sont des maladies inflammatoires chroniques de l'intestin (dites MICI) comme la maladie de Crohn [Tedjo *et al.* 2016] ou la rectocolite hémorragique [Ohjusa & Koido 2015], d'autres ont été associées au diabète et à l'obésité [Baothman *et al.* 2016]). Plus contre-intuitivement, le microbiote intestinal a également été lié à l'asthme et certaines formes d'allergies [Fujimura & Lynch 2015], tout comme des maladies neurodégénératives comme la maladie d'Alzheimer [Hill *et al.* 2014], et la maladie de Parkinson [Scheperjans *et al.* 2015], ou encore les troubles du spectre autistique [Fangiola *et al.* 2016]. L'observation d'une dysbiose, c'est-à-dire d'une altération de la composition des communautés microbiennes intestinales commensales [Normand *et al.* 2013] est désormais envisagée comme marqueur diagnostique de plusieurs de ces pathologies [Da Silva 2014]. Il est toutefois difficile de déterminer si la variation de microbiote observée est une cause ou une conséquence de la maladie.

Une dysbiose intestinale peut également avoir une origine infectieuse. Parmi les parasites intestinaux, on peut citer par exemple *Cryptosporidium* spp. et *Entamoeba histolytica* dont la pathogénicité n'est pas remise en cause.

Par contre, l'impact clinique d'autres protistes intestinaux comme *Blastocystis* spp., dont la prévalence est élevée dans une large proportion de la population, reste quant à lui plus incertain [Lukeš *et al.* 2015]. En effet, ce parasite peut coloniser l'intestin humain durant de larges périodes sans pour autant causer de symptômes cliniques [Scanlan *et al.* 2014] ; sa pathogénicité est ainsi difficile à déterminer, et pourtant d'un enjeu important au vu de sa prévalence. Afin de mieux appréhender l'impact clinique d'un tel protiste sur la santé humaine, le laboratoire Biologie et Diversité des Pathogènes Eucaryotes Émergents (BDPEE) de l'Institut Pasteur de Lille s'est associé à la plate-forme PEGASE-biosciences pour étudier l'impact de *Blastocystis spp.* sur la composition et la diversité des communautés bactériennes intestinales [Audebert *et al.* 2016]. Cette étude, démarrée avant le début de ce projet de thèse, a consisté à séquencer le microbiote intestinal bactérien de 48 patients colonisés par *Blastocystis* et de 48 patients non-colonisés par ce protiste, par une approche métagénétique ADNr 16S sur Ion Torrent PGM, suivi d'une analyse des données par le pipeline PEGASE v2 (voir Chapitre 2, Section 2.3, Figure 2.9). Elle a permis de mettre en évidence que la colonisation d'un individu par *Blastocystis* est associée à une diversité du microbiote intestinal plus élevée (Figure 4.1), ce qui est une observation opposée à une dysbiose habituellement constatée dans le cas de maladies intestinales inflammatoires ou métaboliques, et révélée par une diminution de la diversité [Winter *et al.* 2014]. En outre, les patients colonisés par *Blastocystis* présentaient une plus grande abondance de certaines familles bactériennes, comme Ruminococcaceae et Prevotellaceae, tandis que leur taux d'Enterobacteriaceae était plus faible. Ces tendances correspondent à la signature attendue d'un microbiote intestinal correspondant à un intestin sain [Hamer *et al.* 2008, Brestoff *et al.* 2013, Tan *et al.* 2014].

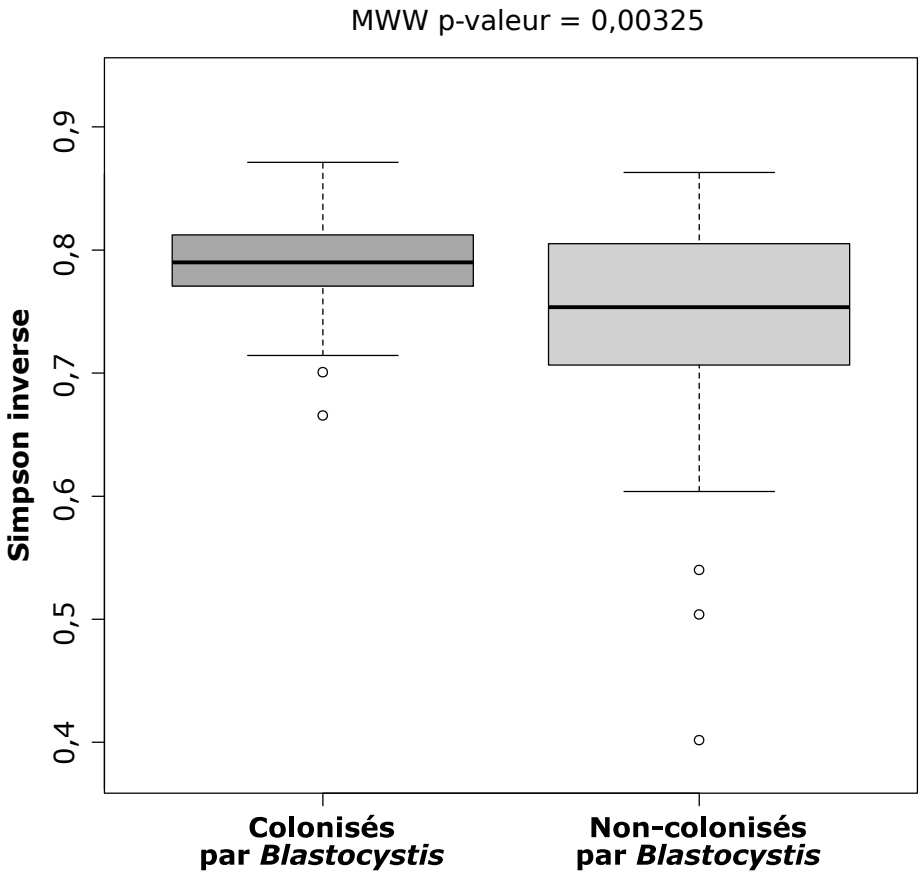


Figure 4.1 : Boîtes à moustaches de la diversité (Simpson inverse) des deux groupes de patients, colonisés ou non par Blastocystis. La différence entre les deux groupes a été testée par test de Mann-Whitney-Wilcoxon [Audebert et al. 2016].

Dans le contexte de l'expertise du pipeline PEGASE v2 menée dans le Chapitre 2, le traitement de ces données réelles par ce pipeline a été évalué étape par étape. Cette évaluation a permis la publication de ces données, et leur analyse par d'autres pipelines afin de comparer les résultats générés.

4.2 - Description des données de séquençage et analyses associées.

Nous avons soumis au journal Scientific Data un article complémentaire à l'étude initiale, mettant à disposition de la communauté scientifique l'ensemble des méthodes utilisées et des données générées. La description

détaillée des protocoles techniques utilisés, du prélèvement des échantillons aux résultats, en passant par la préparation de librairies, le séquençage et l'analyse des données, permet en effet une reproductibilité parfaite de l'étude initiale. La publication et description des jeux de données bruts issus du séquençage Ion Torrent PGM tout comme des jeux de données générés tout au long de l'analyse fournit également de nombreuses ressources à la communauté pour l'optimisation de différentes étapes d'analyse. Enfin, la publication des résultats bruts d'analyse primaire et de la démarche d'analyse secondaire permettent la reproduction des résultats de l'analyse originelle, et l'évaluation d'autres méthodes d'analyse secondaire.

4.2.1 - Collecte des échantillons

Pour cette étude, une enquête épidémiologique a tout d'abord été effectuée dans 11 hôpitaux français différents sur un total de 788 patients atteints de différentes pathologies, avec et sans symptômes gastro-intestinaux [El Safadi *et al.* 2016]. Pour chaque patient, un échantillon de selles a été récolté et testé pour la présence de *Blastocystis* par RT-PCR [Poirier *et al.* 2011], après extraction d'ADN. 143 échantillons sur 788 (18,1 %) ont été testés positifs à la présence de *Blastocystis*. Une première étape a été de réduire le nombre d'échantillons à 48 échantillons positifs à *Blastocystis*, et 48 échantillons négatifs, afin d'atteindre un total de 96 échantillons. Pour éviter un biais de sélection dans le choix des échantillons, des analyses statistiques ont été menées sur les variables cliniques et environnementales associées à chaque patient, afin de lui associer un score de risque de colonisation par *Blastocystis* [Audebert *et al.* 2016]. Les 24 patients ayant le score le plus élevé et les 24 patients ayant le score le plus bas ont été sélectionnés, dans deux populations différentes, pour atteindre un total de 96 échantillons répartis entre 4 groupes. Les groupes 1 et 2 étaient composés de patients colonisés par *Blastocystis*, tandis que ceux des groupes 3 et 4 ne l'étaient pas. Les groupes 1 et 3 rassemblaient des patients présentant des facteurs de risque de colonisation bas, tandis que les groupes 2 et 4 rassemblaient des patients présentant des facteurs de risque élevé (sur la base

de variables cliniques et environnementales). Le fichier associant le nom des échantillons à leur groupe, et le statut de colonisation par *Blastocystis* est disponible dans la section dédiée à ce chapitre sur la page Internet associée au projet de thèse : http://www.pegase-biosciences.com/2013-0920.

4.2.2 - Séquençage métagénétique d'ADNr 16S

La cible d'ADNr 16S 400(V4-V5) a été amplifiée pour l'ADN extrait de chaque échantillon, en utilisant les amorces sens 519F (CAGCMGCCGCGGTAATAC) et anti-sens 926R (CCGTCAATTCMTTTGAGTTT), par Fusion-PCR permettant d'ajouter aux amplicons les adaptateurs de séquençage et les index propres à chaque échantillon (voir Chapitre 1, Sections 1.2.1 et 1.2.3). Cette librairie de 96 échantillons a été séquencée par Ion Torrent PGM sur la plate-forme PEGASE-biosciences, en utilisant une puce Ion 318™ Chip et un kit de séquençage Ion PGM™ 400 Sequencing Kit.

4.2.3 - Description des données générées

Une première analyse du nombre de lectures par échantillon à l'issue du séquençage a révélé 3 échantillons aberrants qui ont été éliminés : les échantillons 50 et 63 à cause de leur faible nombre de lectures (< 10 000), et l'échantillon 18 suite à une erreur technique. En ne prenant pas en compte ces échantillons éliminés, le séquençage a généré un jeu de données d'environ 1,1 Gigabases, composé de 3 962 103 lectures, avec une moyenne de 42 603 lectures par échantillon, ayant un mode de taille de 403 nucléotides, et une taille moyenne de 272 nucléotides (Figure 4.2). La Figure 4.3 montre que les scores de qualité des lectures brutes (en moyenne Q27) est plutôt stable le long des lectures, et au-dessus de la qualité moyenne Q20 habituellement reportée pour un séquençage Ion Torrent PGM. Même la fin des lectures, généralement de moins bonne qualité, est toujours de l'ordre de Q20. Cette bonne qualité générale des lectures a permis l'utilisation des lectures complètes pour l'analyse bioinformatique, sans étape de rognage des lectures selon un seuil de qualité.

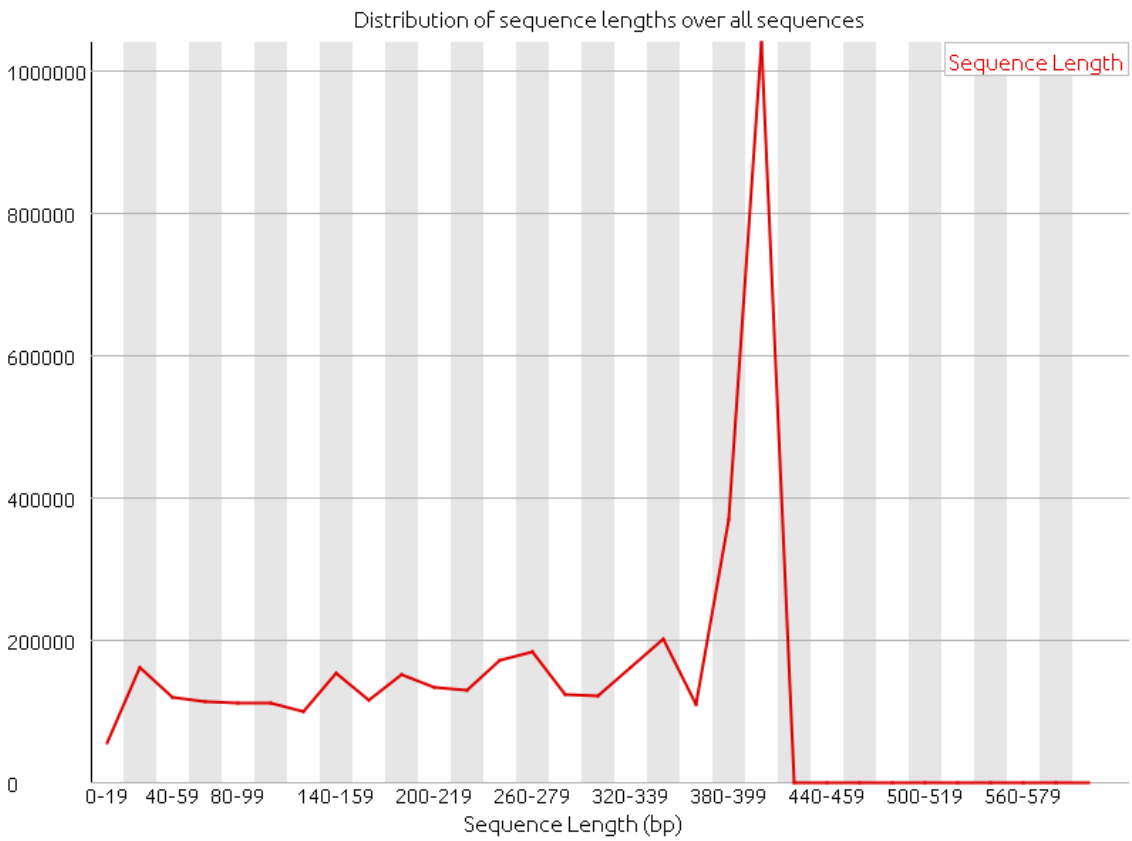


Figure 4.2 : Distribution de la longueur des lectures sur l'ensemble des lectures issues du séquençage, après élimination des index 18, 50 et 63. Taille moyenne : 272 nucléotides, mode = 403 nucléotides. Graphique généré par FastQC [Andrews 2010].

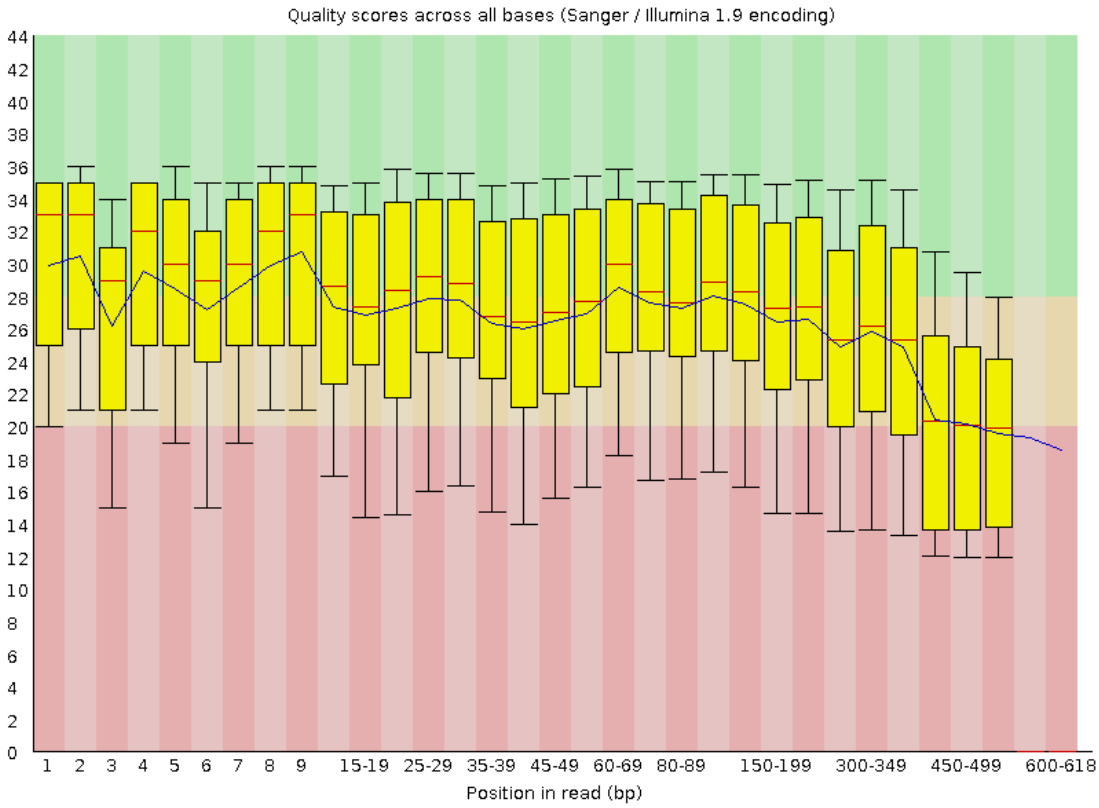


Figure 4.3 : Boîtes à moustaches du score qualité à chaque nucléotide des lectures. La ligne rouge et la ligne bleue représentent les valeurs de qualité médiane et moyenne respectivement. La boîte jaune représente l'écart interquartile, et les moustaches inférieure et supérieure représentent les valeurs adjacentes à 10 % et 90 % respectivement. Graphique généré par FastQC [Andrews 2010].

Une analyse de la variance (ANOVA) a été menée entre les quatre groupes de patients pour comparer le nombre moyen de lectures par échantillon et leur taille moyenne, représentés dans le Tableau 4.4. Aucune différence significative n'a été observée entre les quatre groupes pour ces deux variables (Nombre moyen de lectures moyen par index : p-valeur = 0,47 et taille moyenne des lectures : p-valeur = 0,78). Il n'y a ainsi aucun biais de séquençage observable selon le groupe d'échantillons considéré.

	Patients colonisés par *Blastocystis*		**Patients non-colonisés par *Blastocystis***	
	Groupe 1	**Groupe 2**	**Groupe 3**	**Groupe 4**
Nombre moyen de lectures par échantillon	43 058,00	44 150,04	41 374,65	41 862,54
Taille moyenne des lectures (nt)	270,00	273,96	275,13	271,62

Tableau 4.4 : Taille moyenne et nombre de lectures par échantillon sur les données de séquençage (après élimination des 3 échantillons aberrants 18, 63 et 50 appartenant aux groupes 1, 2 et 3 respectivement).

4.2.4 – Analyse des données

L'analyse de tous les échantillons a été effectuée en utilisant le pipeline PEGASE v2 (Chapitre 2, Section 2.3, Figure 2.9), après élimination des 3 échantillons aberrants 18, 63 et 50 appartenant aux groupes 1, 2 et 3 respectivement. Pour rappel, ce pipeline *clustering-first* fonctionne en trois étapes distinctes : un pré-traitement des lectures, un *clustering* des lectures en OTUs avec assignation taxonomique pour chaque OTU, et une fusion et normalisation des tables de comptage. L'ensemble des données intermédiaires générées par le pipeline d'analyse est disponible sur le site Internet de la plate-forme : http://www.pegase-biosciences.com/collaborations/blastocystis-gut-microbiota

L'étape de pré-traitement a éliminé 31,27 % des lectures : 856 053 lectures car elles étaient de taille inférieure à 150 nt et/ou avaient des homopolymères de taille supérieure à 10 nt, et 391 786 lectures car elles avaient moins de 100 nt alignés avec la banque bactérienne SILVA version 102 (correspondant à des séquences contaminantes ou chimériques). Au final, 2 742 108 lectures ont été conservées après cette étape de pré-traitement.

La deuxième étape du pipeline d'analyse était un *clustering* des lectures en OTUs, à l'issue duquel une table de comptages des lectures par OTU a été générée pour chaque échantillon. Ces tables de comptage sont des fichiers tabulés TSV (*tab-separated values*) résumés dans le Tableau 4.5, où chaque ligne représente un OTU, décrit par 4 colonnes :

- le nom de la lecture représentative de l'OTU (sélectionnée comme étant la plus abondante) ;
- le nombre de lectures associées à cet OTU ;
- le nom de la lecture représentative de l'OTU (information identique à la première colonne) ;
- le taxon auquel a été assigné cette lecture (et par extension l'OTU).

La troisième étape du pipeline a fusionné ces 93 tables d'OTUs sur la base de leur annotation taxonomique pour générer une table d'OTUs globale, en utilisant un script maison. Cette table de comptages globale est un fichier TSV dont chaque ligne représente un taxon, et chaque colonne un échantillon. Ce fichier décrit au total 474 taxons, leur annotation taxonomique, et le nombre de lectures attribuées à chaque taxon par échantillon. Finalement, cette table de comptages a été convertie en un fichier BIOM global (version 2.1.4), généré par la commande biom convert. Les caractéristiques de ce fichier sont résumées dans le Tableau 4.6.

Origine	Type de fichier	Nombre de fichiers	Nombre d'échantillon par fichier	Nombre moyen de lectures par OTU	Nombre moyen de taxons uniques par échantillon	Ratio moyen entre le nombre de taxons uniques et le nombre total d'OTUs
Pipeline PEGASE v2	TSV	93	1	25 (ET : 258,60)	1184 (ET : 846,00)	0,77

Tableau 4.5 : Résumé des tables de comptage à l'issue de la deuxième étape d'analyse. TSV = tab separated values, ET = écart-type.

Origine	Type de fichier	Nombre d'échantillons	Nombre d'observations (taxons)	Nombre total de lectures	Densité du tableau (fraction de valeurs non-nulles)	Nombre moyen de lectures par échantillon	Nombre médian de lectures par échantillon	Colonne additionnelle : méta-données
Table de comptages fusionnée	BIOM	93	474	2 742 108	0,187	29 485 (ET : 5 739)	29 419	taxonomie

Tableau 4.6 : Résumé du fichier BIOM global brut. ET = écart-type.

Origine	Type de fichier	Nombre d'échantillons	Nombre d'observations (taxons)	Nombre total de valeurs normalisées	Densité du tableau (fraction de valeurs non-nulles)	Nombre moyen de valeurs normalisées par échantillon	Nombre médian de valeurs normalisées par échantillon	Colonne additionnelle : méta-données
Fichier BIOM global	*BIOM*	*93*	*405*	*33 624*	*0,217*	*361 (ET : 138)*	*353*	*taxonomie*

Tableau 4.7 : Résumé du fichier BIOM global normalisé. ET = écart-type.

La normalisation de cette table de comptages au format BIOM a été effectuée par DESeq2 (intégré dans QIIME, et configuré pour remplacer les valeurs négatives après normalisation par des valeurs nulles). Cette normalisation permet de comparer les échantillons entre eux en évitant de recourir à une raréfaction des données (voir Chapitre 1 Section 1.5.1). L'information taxonomique associée à chaque taxon étant éliminée par DESeq, elle a été ré-intégrée au fichier résultat en utilisant la commande biom add metadata. Le fichier final BIOM normalisé est décrit dans le Tableau 4.7, et a été publié comme fichier final de l'analyse.

4.2.5 - Disponibilité des données

L'ensemble des lectures brutes issues du séquençage au format FASTQ a été déposé au sein de Sequence Read Archive (SRA) du NCBI, sous l'identifiant de projet PRJNA342805. Ces données sont publiées sous la forme d'un fichier FASTQ par index. Les données intermédiaires d'analyse, décrites dans les Tableaux 4.5, 4.6 et 4.7, ont été mises à disposition sur le site Internet PEGASE-biosciences :

http://www.pegase-biosciences.com/collaborations/blastocystis-gut-microbiota

Le tableau 4.8 récapitule l'ensemble des données mises à disposition de la communauté scientifique en association à cette étude :

Fichiers de données	Format	Figures / Tableaux descriptifs	Accessibilité des données	Nb de fichiers
Lectures brutes	FASTQ	Figures 4.2 & 4.3 Tableau 4.4	Projet PRJNA342805 dans la banque SRA du NCBI	96
Tables de comptages	TSV	Tableau 4.5	Site PEGASE-biosciences	93
BIOM global brut	BIOM	Tableau 4.6	Site PEGASE-biosciences	1
BIOM global normalisé	BIOM	Tableau 4.7	Site PEGASE-biosciences	1

Tableau 4.8 : Résumé des différentes données publiées.

La publication de ces données et des méthodes employées, du prélèvement d'échantillon à une table de comptages normalisée, est une ressource précieuse mise à disposition de la communauté. Elle garantit en effet la reproductibilité de l'étude cas/témoin initiale, mais permet aussi à d'autres utilisateurs et développeurs de ré-utiliser ces données pour évaluer et optimiser différentes étapes d'analyse. Les fichiers intermédiaires d'analyse (tables de comptage de taxons, fichier BIOM global avant et après normalisation) peuvent être ré-utilisés pour des analyses statistiques complémentaires, afin par exemple de comparer différentes méthodes de validation d'études cas/témoin.

4.3 - Évaluation de l'impact du changement de pipeline sur l'interprétation biologique des résultats

L'évaluation d'un outil analytique se fait habituellement en comparant ses résultats à un résultat de référence. Par exemple, dans un contexte métagénétique, la performance d'un pipeline d'analyse peut être évaluée en comparant ses résultats à la composition réelle du jeu de données analysé (que ce soit un jeu de données artificiel généré *in silico* ou une communauté artificielle *in vitro*), comme nous l'avons fait dans les Chapitres 2 et 3. Toutefois, cette évaluation est restreinte à à l'interprétation d'un résultat issu de données artificielles et décontextualisé, puisque ne correspondant pas à un plan d'expérience spécialement monté pour répondre à une problématique biologique donnée. Dans une étude cas/témoin réelle, les pipelines d'analyse sont utilisés pour traiter un grand nombre d'échantillons, dont les résultats sont soumis à comparaison afin d'en évaluer les différences, par exemple entre deux conditions biologiques données. Dans ce contexte, l'interprétation des résultats ne se fait pas sur la composition absolue d'un seul microbiote, mais sur la comparaison relative entre différents groupes d'échantillons, représentatifs de différentes conditions biologiques d'intérêt. Pour que cette comparaison soit robuste, il est nécessaire d'intégrer un relativement grand nombre

d'échantillons par condition dans le plan d'expérience, afin de gommer la variabilité inter-individuelle. Les résultats sont en outre interprétés après une analyse secondaire statistique, utilisée pour distinguer une différence potentiellement intéressante d'une différence hasardeuse ou causée par un certain bruit dans les données. Aucune étude n'a déjà évalué l'impact du pipeline d'analyse métagénétique utilisé sur les résultats d'une étude comparative réelle. Dans le cas d'une étude cas/témoin réelle, comme celle décrite précédemment, un changement de pipeline d'analyse peut-il entraîner un changement de paradigme dans les conclusions biologiques associées à la différence observée entre cas et témoin ?

4.3.1 - Plan d'expérience analytique

Afin de répondre à cette question, l'ensemble des données de l'étude clinique présentée dans la section précédente a été ré-analysé (sans les échantillons aberrants index 18, 50 et 63) en utilisant quatre pipelines déjà expertisés dans le Chapitre 3, comme présenté dans la Figure 4.9 : deux pipelines *clustering-first* (mothur et QIIME SortMeRNA+Sumaclust) et deux pipelines *assignment-first* (kraken et CLARK). Ces pipelines ont été utilisés sur l'ensemble des jeux de données initiaux (décrits dans les Figures 4.2 et 4.3 et le Tableau 4.4 précédents), avec la banque de séquences de référence recommandée par défaut par chaque pipeline (SILVA pour mothur, Greengenes pour QIIME, Minikraken pour kraken et RefSeq bactérien pour CLARK), en exécutant les commandes décrites en Annexe 2. Comme déjà constaté dans le Chapitre 3 Section 3.2.7, les pipelines nécessitent un temps d'analyse plus ou moins conséquent pour générer l'ensemble des résultats. Les 93 échantillons ont été traités en plus de 24 h pour mothur, 1 h pour QIIME, 5 min pour CLARK et 1 s pour kraken, confirmant les disparités importantes entre les temps d'exécution des différents pipelines.

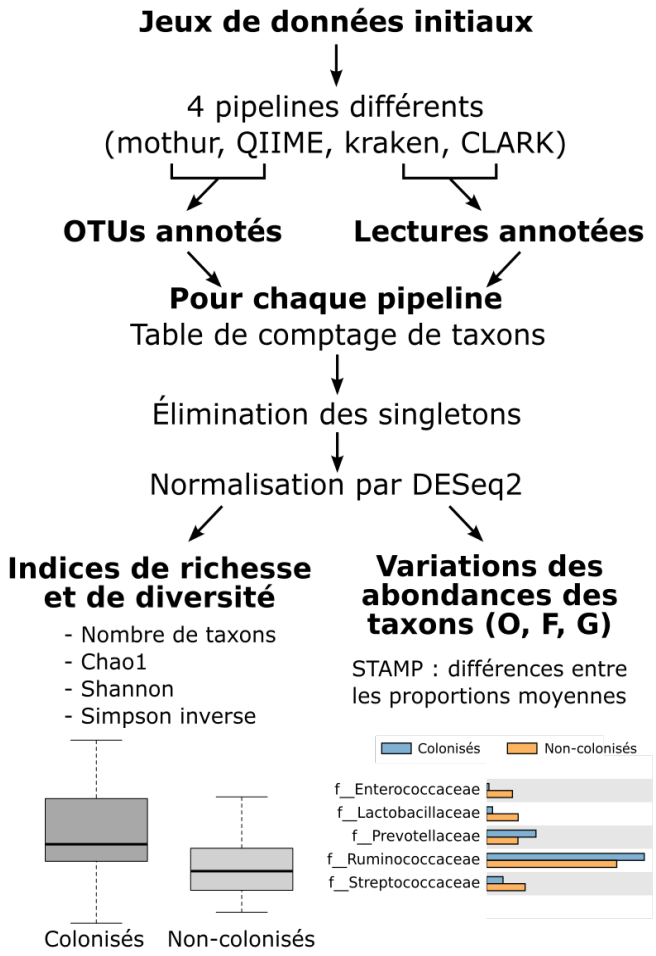


Figure 4.9 : Plan d'expérience analytique. O=ordre, F=famille, G=genre. STAMP est un logiciel d'analyse statistique secondaire de données métagénétiques.

Chaque pipeline a généré en résultat une table globale de comptages de taxons pour tous les échantillons. Cette table a été interprétée en suivant les méthodes utilisées dans l'étude originelle, afin d'appliquer la même méthode d'interprétation des résultats entre tous les pipelines. Les singletons ont tout d'abord été éliminés. Chaque table de comptages a ensuite été normalisée par DeSeq2. Pour chaque échantillon, les indices de richesse (Chao1) et de diversité (Shannon et Simpson inverse) ont été calculés à partir des comptages normalisés, après fusion taxonomique pour les pipelines *clustering-first*, comme cela avait été effectué dans l'étude initiale. La différence de ces indices

entre les deux groupes d'échantillons (patients colonisés ou non par *Blastocystis*) a été testée par un test de Mann-Whitney-Wilcoxon [Wilcoxon 1945, Mann-Whitney 1947], entre le groupe d'échantillons de patients colonisés par *Blastocystis*, et celui de patients non-colonisés par le protiste. Les tables de comptage normalisées ont également été interprétées pour chaque pipeline par le logiciel STAMP [Parks *et al.* 2014] afin d'évaluer les variations des abondances des taxons à différents niveaux taxonomiques (ordre, famille et genre). La représentation choisie dans STAMP pour évaluer les résultats est la différence des proportions moyennes de chaque taxon entre les deux groupes d'échantillons. Seuls les taxons ayant une différence de proportions moyennes supérieure à 1 % ont été retenus. Pour ces taxons, cette différence a été testée par un test de Student non-paramétrique [White *et al.* 2009], avec une correction de Benjamini-Hochberg [Benjamini & Hochberg, 1995], comme utilisés dans l'étude initiale.

4.3.2 - Interprétation des variations de richesse et de diversité selon le pipeline utilisé

Les mesures de richesse et de diversité permettent d'avoir une première image des microbiotes d'intérêt après analyse, en résumant sa composition pour la richesse, associée aux proportions des taxons identifiés pour la diversité. La Figure 4.10 représente l'estimation de l'indice de richesse Chao1 et de l'indice de diversité Simpson inverse générés pour l'ensemble des échantillons analysés par les 4 pipelines évalués (le nombre d'OTUs et l'indice de Shannon, non représentés ici, suivent les mêmes tendances).

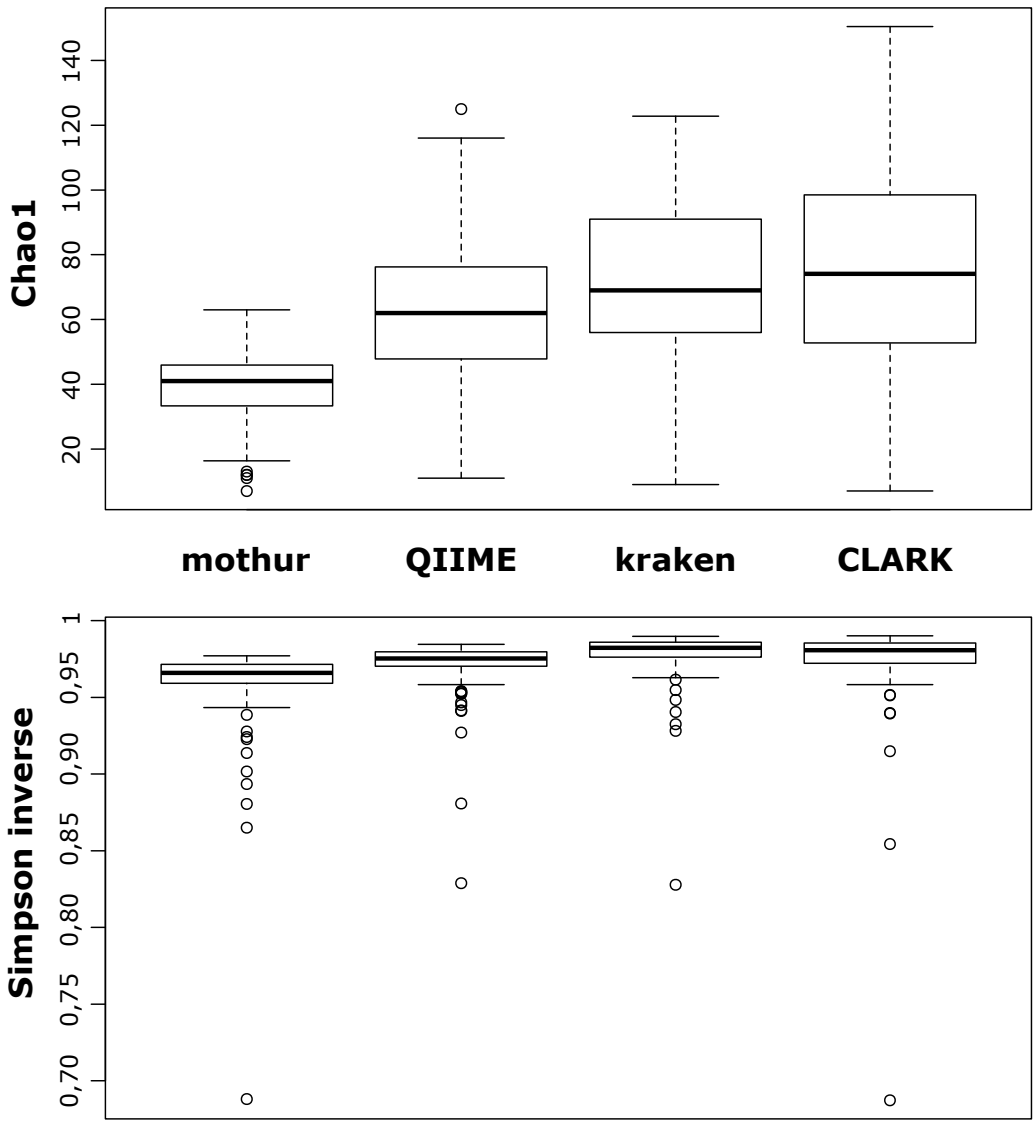


Figure 4.10 : Indices de richesse Chao1 et de diversité Simpson inverse pour les 93 échantillons (tous groupes confondus), selon 4 pipelines différents.

	Chao1	**Simpson inverse**
mothur	41,000 (EI : 12,542)	0,966 (EI : 0,012)
QIIME	62,000 (EI : 28,494)	0,975 (EI : 0,009)
kraken	69,000 (EI : 35,000)	0,982 (EI : 0,010)
CLARK	74,120 (EI : 45,730)	0,980 (EI : 0,013)

Tableau 4.11 : Valeurs médianes (écart interquartile EI entre parenthèses) des indices de richesse Chao1 et de diversité Simpson inverse pour les 93 échantillons (tous groupes confondus), selon 4 pipelines différents.

Comme nous l'avions observé sur données simulées (Chapitre 3 Section 3.2.3 Tableau 3.8), l'indice de richesse Chao1 est fortement variable d'un pipeline à l'autre, car corrélé à la façon dont chaque pipeline gère les erreurs et le débit de séquençage. Par exemple, comme représenté dans le Tableau 4.11, mothur estime à nouveau ici une richesse plus faible que les autres pipelines (Chao1 médian = 41, écart interquartile=12,542), comportement déjà observé dans l'utilisation de données réelles avec ce pipeline (Chapitre 3, Section 3.2.6). CLARK au contraire est le pipeline estimant la richesse la plus élevée (Chao1 médian = 74,120, écart interquartile = 45,730), comme déjà observé dans la même section précédente (Chapitre 3, Section 3.2.6). Nous confirmons ici sur des données réelles que même avec un grand nombre d'échantillons considéré et après fusion taxonomique, la mesure de richesse peut varier du simple au double selon le pipeline analytique utilisé. Les indices de diversité semblent également varier entre les quatre pipelines (Simpson inverse médian = 0,966 et écart interquartile = 0,012 pour mothur, Simpson inverse médian = 0,982 et écart interquartile = 0,010 pour kraken), bien qu'ils soient plus difficiles à interpréter car ils combinent le nombre de taxons retrouvés et leurs proportions respectives.

Cette variation de richesse et de diversité d'un pipeline à l'autre peut être mise en parallèle avec la proportion des lectures identifiée par chaque pipeline. En effet, la Figure 4.12 confirme des comportements déjà observés précédemment : contrairement à CLARK, mothur identifie le moins de lectures, notamment au niveau du genre (38,34 % pour mothur, 83,51 % pour CLARK), ce qui peut expliquer sa faible estimation de richesse et de diversité. Dans ce contexte, nous observons que plus un pipeline identifie de lectures, plus il estime une richesse élevée. Sans échantillon de référence, il est impossible de déterminer si cette augmentation de richesse s'approche de la réalité, ou si elle est causée par l'augmentation de lectures erronées qui impactent fortement l'estimation de richesse (comme démontré dans le Chapitre 3 Section 3.2.2).

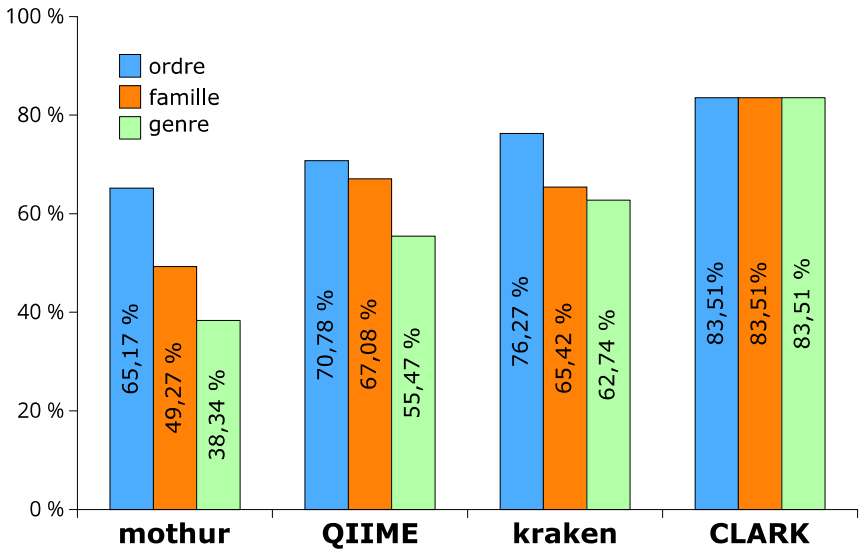


Figure 4.12 : Proportions de lectures assignées pour chaque pipeline à différents niveaux taxonomiques.

Afin de déterminer si ces différences de richesse et de diversité entre pipelines sont significatives, un test de Kruskall-Wallis a été utilisé pour comparer leur distribution sur tous les échantillons entre les 4 pipelines. Ce test s'est avéré significatif pour les deux indices (p-valeur < 2.2e-16 pour Chao1 et Simpson inverse), indiquant qu'au moins un des pipelines divergeait des autres dans leur estimation. Un test de Mann-Whitney-Wilcoxon a été appliqué entre toutes les paires de pipelines pour chaque indice (Tableaux 4.13 et 4.14). Seuls kraken et CLARK rendent un test de Mann-Whitney-Wilcoxon non-significatif, pour l'estimation de richesse comme de diversité (p-valeur Chao1=0,3446, p-valeur Simpson inverse = 0,2981), révélant que la distribution de richesse et de diversité entre tous les échantillons est similaire pour ces deux pipelines. Ainsi, on peut confirmer que l'estimation de richesse et de diversité sur l'ensemble des échantillons n'est pas similaire entre les pipelines, sauf entre kraken et CLARK. Ces deux derniers pipelines *assignment-first* sont en effet les pipelines les plus proches entre eux, fonctionnant avec une méthode similaire et sur la même taxonomie, même si kraken identifie moins de lectures (voir Figure 4.12) à cause de sa banque réduite Minikraken.

	QIIME	kraken	CLARK
mothur	5,933e-15	2,200e-16	2,200e16
	QIIME	2,882e-02	3,068e-03
		kraken	0,345

Tableau 4.13 : p-valeurs des tests de Mann-Whitney-Wilcoxon effectués sur la distribution de richesse Chao1 entre tous les échantillons pour chaque paire de pipelines.

	QIIME	kraken	CLARK
mothur	4,426e-11	2,200e-16	4,425e-14
	QIIME	8,485e-08	2,007e-04
		kraken	0,3446

Tableau 4.14 : p-valeurs des tests de Mann-Whitney-Wilcoxon effectués sur la distribution de diversité Simpson inverse entre tous les échantillons pour chaque paire de pipelines.

4.3.3 - Comparaison des indices de richesse/diversité entre les deux groupes de patients selon le pipeline utilisé

La conclusion principale de l'étude initiale était d'associer la colonisation d'un individu par *Blastocystis* à une richesse et diversité du microbiote intestinal plus élevées, représentatives d'un intestin sain. Même si elle est variable d'un pipeline à un autre dans l'absolu, l'interprétation des indices de diversité générés par les quatre pipelines sur les mêmes jeux de données en comparant les deux groupes de patients (colonisés ou non par *Blastocystis*) amène à la même conclusion (Tableaux 4.15 et 4.16, Figures 4.17 et 4.18). Pour les quatre pipelines, la richesse et la diversité du microbiote bactérien intestinal est plus élevée en présence de *Blastocystis*, confortant la conclusion de l'étude initiale.

Nous confirmons que les indices de richesse et de diversité ne peuvent être interprétés de façon absolue, puisqu'ils sont beaucoup trop variables d'un pipeline à un autre, particulièrement sur des données avec erreurs (comme nous l'avions déjà montré sur données simulées Chapitre 3 Section 3.2.2). Par contre, ces indices sont de bons estimateurs de différences potentielles entre deux groupes d'échantillons, puisque dans notre exemple, leur variation entre les

deux groupes de patients est confirmée peu importe le pipeline utilisé, et même si ces pipelines ont une estimation absolue différente de ces indices. Toutefois, que ce soit pour la diversité comme pour la richesse, les valeurs extrêmes mesurées dans chaque groupe sont très variables d'un pipeline à un autre (comme observé dans les Figures 4.17 et 4.18). Ces différences, dues à la variabilité de l'estimation de ces indices selon le pipeline utilisé, confirment l'importance du montage du plan d'expérience. Il est nécessaire d'avoir un nombre suffisamment conséquent d'échantillons par groupe étudié et d'une population homogène au sein de ces groupes pour pouvoir interpréter la variation de ces indices entre deux conditions de façon robuste.

	Patients colonisés par *Blastocystis*	**Patients non-colonisés par *Blastocystis***	**p-valeur du test de Mann-Whitney-Wilcoxon entre les deux groupes**
mothur	34,500 (EI : 13,708)	43,425 (EI : 7,268)	5,562e-05
QIIME	52,615 (EI : 28,883)	72,150 (EI : 23,353)	6,096e-04
kraken	59,000 (EI : 32,000)	80,000 (EI : 36,500)	2,185e-04
CLARK	63,333 (EI : 37,448)	89,183 (EI : 36,564)	1,128e-04

Tableau 4.15 : Valeurs médianes de la richesse Chao1 estimée par chaque pipeline pour les deux groupes de patients (EI=écart interquartile), et p-valeurs des tests de Mann-Whitney-Wilcoxon effectués sur les distributions de richesse entre les deux groupes.

	Patients colonisés par *Blastocystis*	**Patients non-colonisés par *Blastocystis***	**p-valeur du test de Mann-Whitney-Wilcoxon entre les deux groupes**
mothur	0,961 (EI : 0,017)	0,969 (EI : 0,007)	3,089e-04
QIIME	0,972 (EI : 0,018)	0,978 (EI : 0,006)	5,418e-04
kraken	0,978 (EI : 0,012)	0,984 (EI : 0,007)	2,223e-04
CLARK	0,975 (EI : 0,019)	0,984 (EI : 0,007)	6,989e-05

Tableau 4.16 : Valeurs médianes de la diversité Simpson inverse estimée par chaque pipeline pour les deux groupes de patients (EI=écart interquartile), et p-valeurs des tests de Mann-Whitney-Wilcoxon effectués sur les distributions de richesse entre les deux groupes.

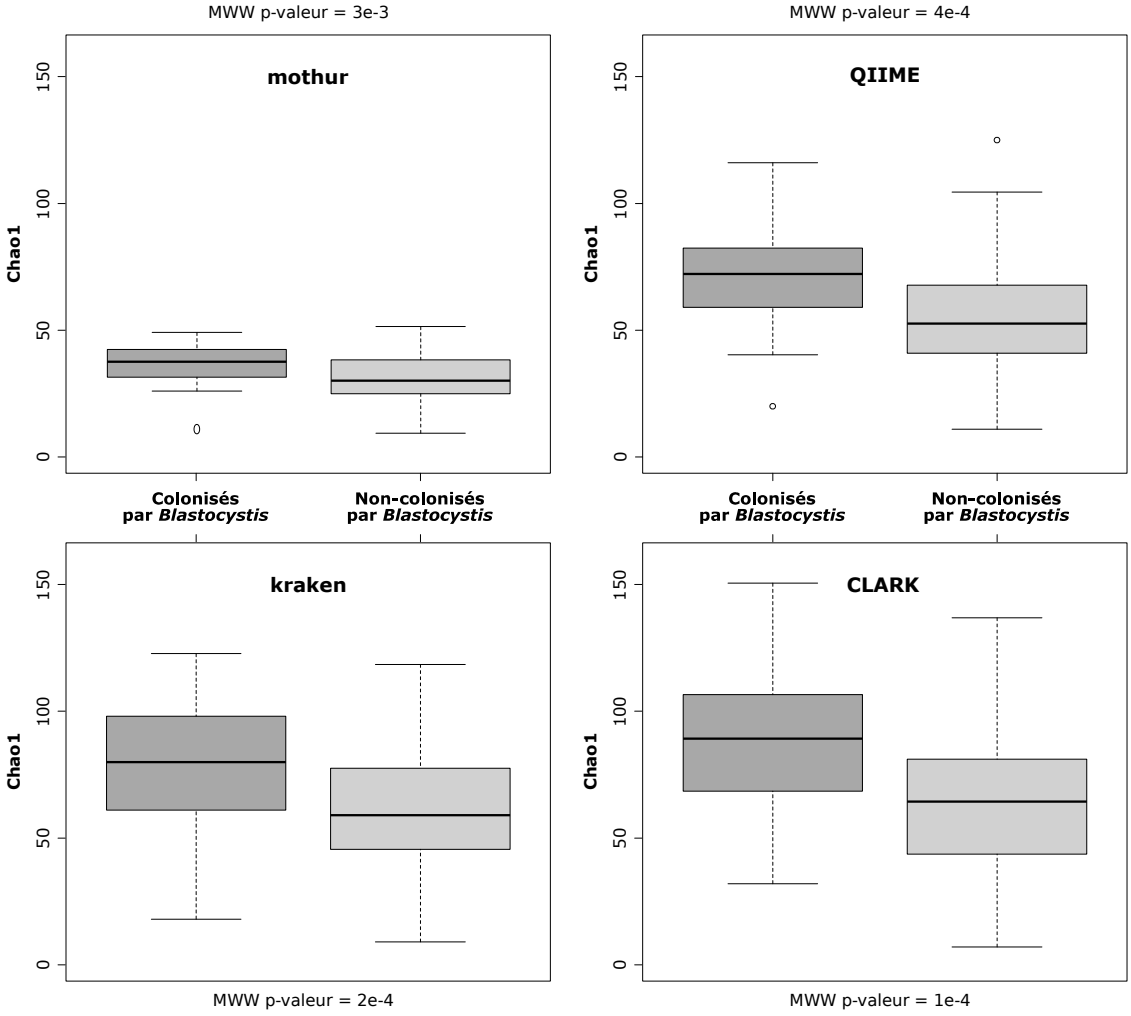


Figure 4.17 : Boîtes à moustaches de la richesse (Chao1) au niveau de la famille des deux groupes de patients, colonisés ou non par Blastocystis, pour tous les pipelines. La différence entre les deux groupes a été testée par test de Mann-Whitney-Wilcoxon (MWW).

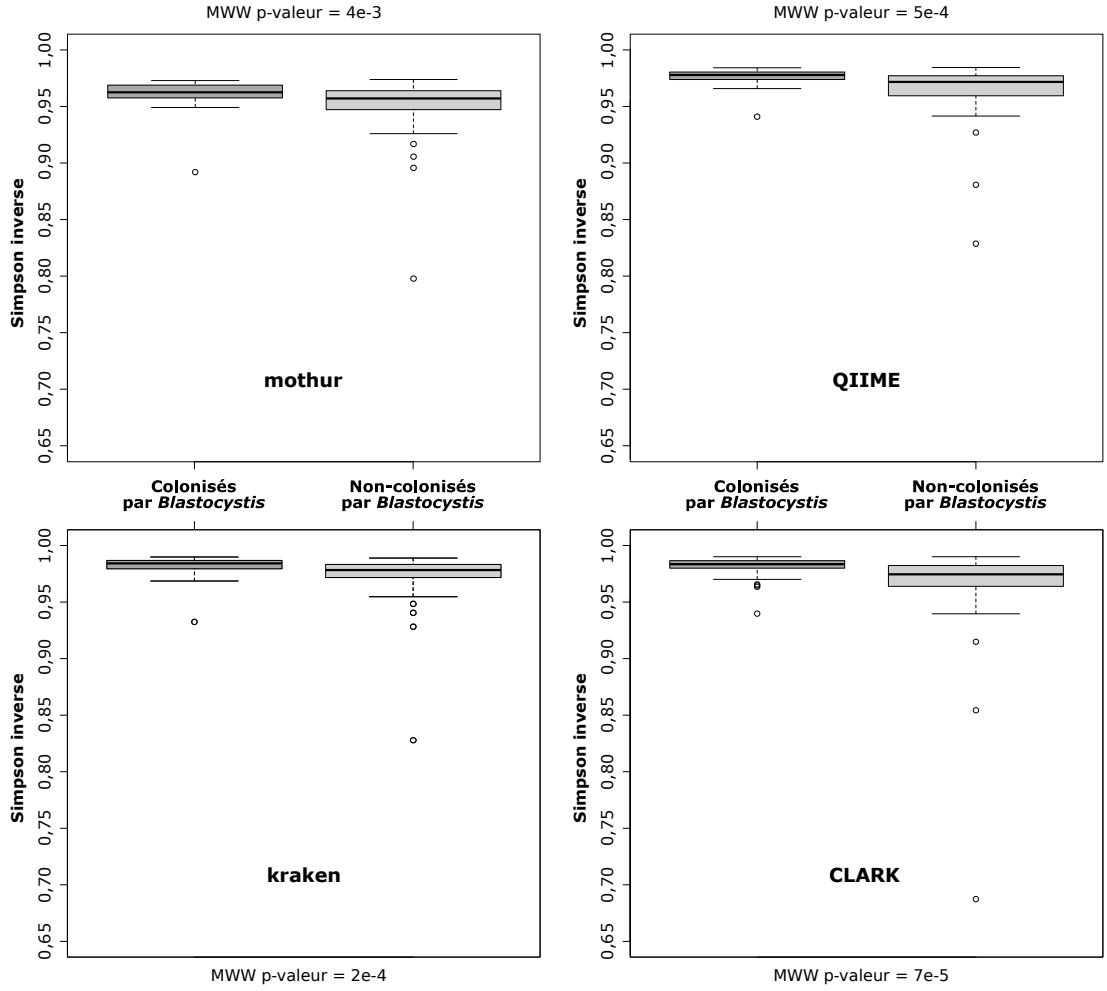


Figure 4.18 : Boîtes à moustaches de la diversité (Simpson inverse) au niveau de la famille des deux groupes de patients, colonisés ou non par Blastocystis, pour tous les pipelines. La différence entre les deux groupes a été testée par test de Mann-Whitney-Wilcoxon (MWW).

4.3.4 - Interprétation des variations de composition révélées par différents pipelines entre les deux groupes de patients

L'étude initiale de comparaison de flore intestinale bactérienne entre patients colonisés par *Blastocystis* et patients non colonisés avait révélé des variations significatives de proportions de différents taxons entre les deux groupes. Nous avons montré précédemment que l'estimation de richesse et diversité peut être variable d'un pipeline à un autre, laissant entendre que chaque pipeline ne détermine pas la même quantité de taxons dans les mêmes proportions. Cette variation impacte-t-elle les conclusions biologiques tirées de la comparaison de la composition taxonomique des deux groupes d'échantillons, selon le pipeline d'analyse utilisé ?

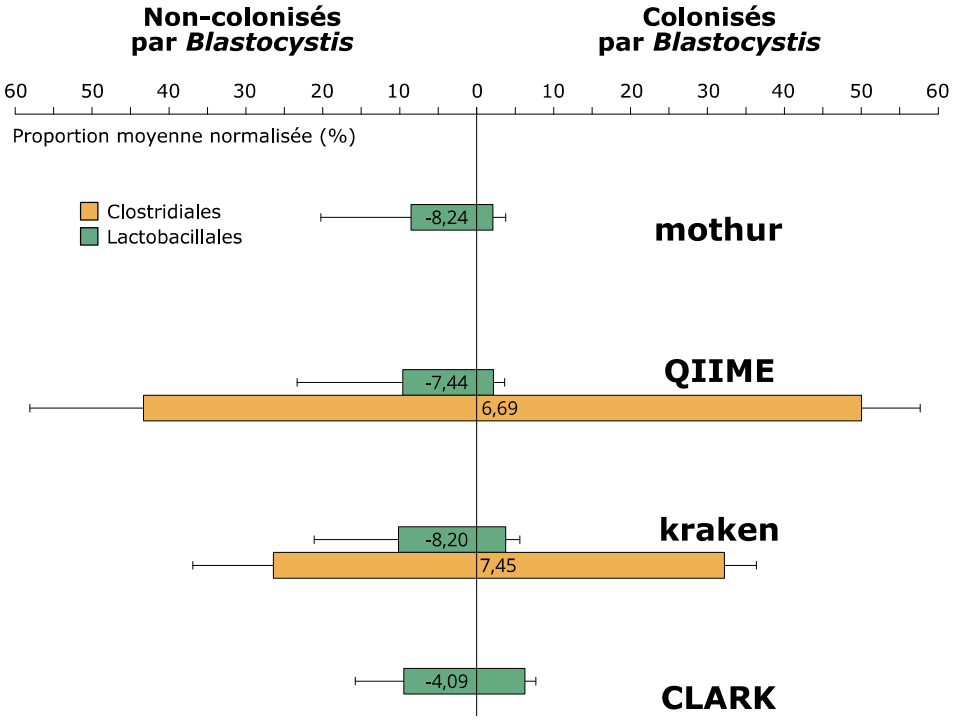


Figure 4.19 : Proportions moyennes de chaque ordre significativement différent entre les deux groupes d'échantillons. Les valeurs représentées sur chaque barre sont la différence de moyenne entre les deux groupes (seuls sont représentés les taxons pour lesquels le test de Student non-paramétrique avec correction de Benjamini-Hochberg entre les proportions moyennes des deux groupes a une q-valeur < 0,05, et dont la taille d'effet est supérieure à 1 %).

Dans l'étude initiale, deux ordres bactériens avaient été considérés comme significativement variables entre les deux groupes de patients : les Clostridiales étaient plus abondantes dans les patients colonisés par *Blastocystis*, tandis que les Lactobacillales étaient plus abondantes chez les patients non-colonisés. La Figure 4.19 représente les ordres bactériens différentiellement observés entre les deux groupes, selon les 4 pipelines.

Les Lactobacillales sont significativement plus abondantes chez les patients non colonisés pour les quatre pipelines, comme mentionné dans la publication originelle, mais seuls QIIME et kraken révèlent une augmentation significative de Clostridiales chez les patients colonisés par *Blastocystis*. Les Clostridiales contiennent de nombreuses familles comme Ruminococcaceae, Eubacteriaceae et Lachnospiraceae, dont la présence majoritaire est associée à un intestin sain [Biddle *et al.* 2013]. L'augmentation significative de cet ordre constatée par QIIME et kraken chez les patients colonisés par *Blastocystis* renforce l'association entre ce protiste et un intestin sain. Cette information n'est pas révélée par mothur et CLARK, qui détectent certes des Clostridiales dans les deux groupes d'échantillons, mais dont la variation n'est pas assez significative pour être représentée. Ainsi, dès le niveau taxonomique de l'ordre, on observe déjà que différents pipelines peuvent mener à des conclusions biologiques différentes.

Cette variation est renforcée au niveau de la famille, comme représenté dans la Figure 4.20. Dans la publication originelle, les Ruminococcaceae et Prevotellaceae étaient prévalentes dans les patients colonisés, tandis que les familles Enterococcaceae, Streptococcaceae, Lactobacillaceae et Enterobacteriaceae étaient plus abondantes dans les patients non colonisés. Cette signature de familles bactériennes différentiellement observées entre les deux groupes semble à première vue très variable selon le pipeline utilisé. Enterobacteriaceae, dont le bloom est un marqueur d'un intestin dysbiotique [Winter & Baümler 2014] n'est significativement variable pour aucun pipeline

entre les deux groupes.

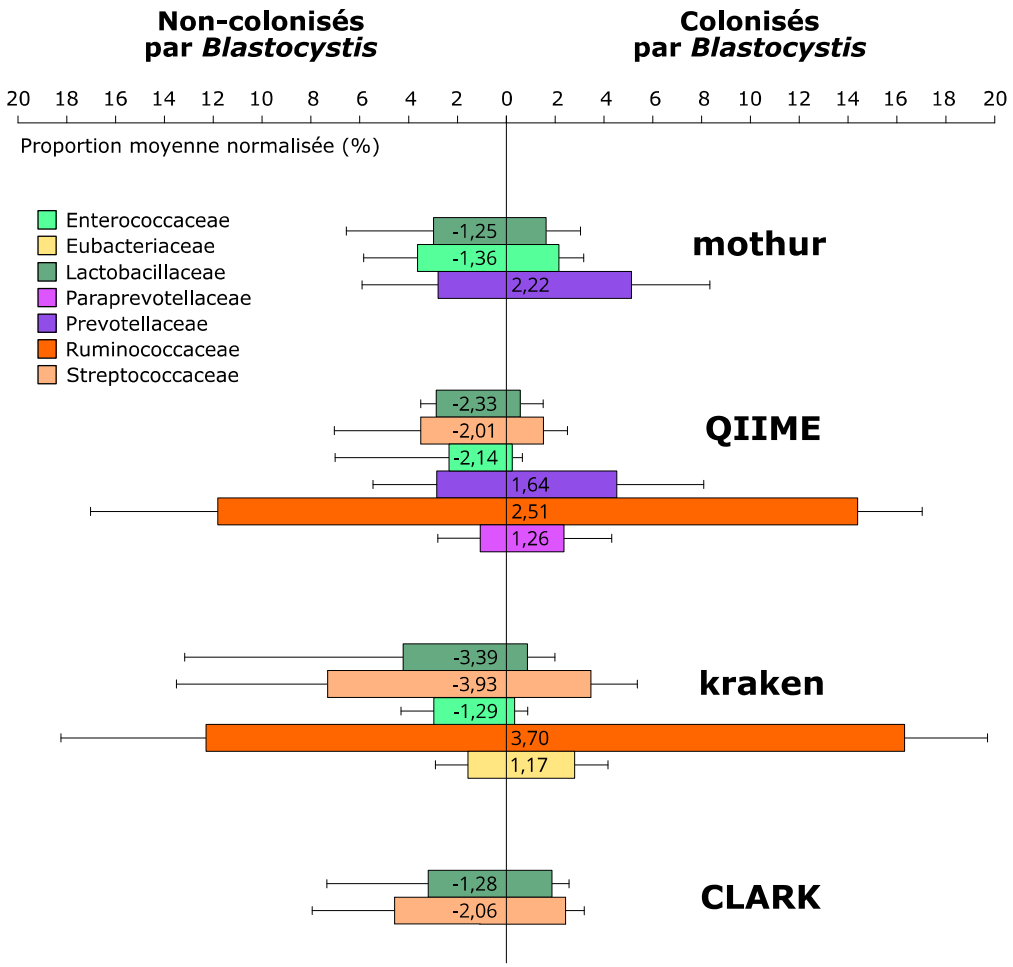


Figure 4.20 : Proportions moyennes de chaque famille significativement différente entre les deux groupes d'échantillons. Les valeurs représentées sur chaque barre sont la différence de moyenne entre les deux groupes (seuls sont représentés les taxons pour lesquels le test de Student non-paramétrique avec correction de Benjamini-Hochberg entre les proportions moyennes des deux groupes a une q-valeur < 0,05, et dont la taille d'effet est supérieure à 1 %).

QIIME est le pipeline le plus proche des résultats originels, identifiant les autres familles différentiellement observées. À l'inverse, mothur et CLARK n'identifient respectivement que 3 et 2 des familles précédemment citées, avec une différence relativement basse entre les deux groupes. Ces deux pipelines semblent ainsi ne pas révéler de variation majeure entre les deux groupes ;

malgré une richesse et diversité significativement plus élevées pour les patients colonisés par *Blastocystis*, ces pipelines ne permettent pas d'identifier de familles bactériennes majeures associées à cette variation. kraken identifie plus de familles différentiellement observées que CLARK, ce qui paraît contre-intuitif puisqu'il intègre ici une banque de référence plus réduite (Minikraken). En réalité, l'utilisation d'une banque réduite diminue la précision du pipeline (comme montré dans le Chapitre 3 Section 3.2.5) : ainsi, kraken donne une image plus grossière de chaque groupe, ce qui accentue leurs différences, et peut ainsi expliquer l'observation de plus de variations significatives entre leurs compositions. À noter que kraken est le seul pipeline à révéler une abondance d'Eubacteriaceae significativement plus élevée chez les patients colonisés par *Blastocystis*, et associée à un intestin sain.

Il est important de mentionner qu'aucun pipeline ne donne de résultats contradictoires, que ce soit dans la comparaison d'indices de richesse et de diversité, tout comme dans la comparaison de composition. La différence significative observée par deux pipelines entre les deux groupes pour un même taxon est toujours observée dans le même sens (par exemple dans la Figure 4.20, Lactobacillaceae et Streptococcaceae sont toujours plus abondants chez les patients non-colonisés par *Blastocystis*, tandis que Prevotellaceae est toujours plus abondant chez les patients colonisés par ce protiste). En outre, chez les patients colonisés par *Blastocystis,* aucun pipeline n'observe de familles bactériennes typiquement liées aux dysbioses associées aux maladies inflammatoires chroniques de l'intestin (bloom d'Enterobacteriaceae et diminution des Clostridiales).

Cette variation de résultats entre les pipelines semble de prime abord encore plus marquante au niveau du genre (Figure 4.21), où les analyses menées par mothur ne permettent pas d'observer de différence significative entre les deux groupes, tandis que QIIME semble bien plus éclater les différences observées.

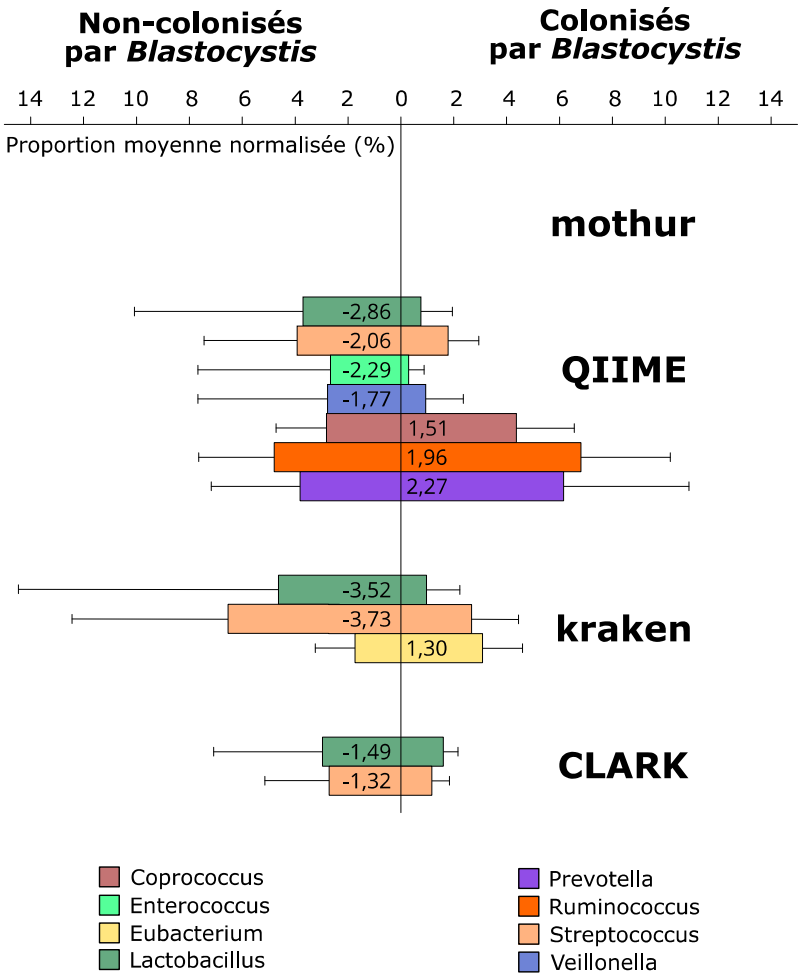


Figure 4.21 : Proportions moyennes de chaque genre significativement différent entre les deux groupes d'échantillons. Les valeurs représentées sur chaque barre sont la différence de moyenne entre les deux groupes (seuls sont représentés les taxons pour lesquels le test de Student non-paramétrique avec correction de Benjamini-Hochberg entre les proportions moyennes des deux groupes a une q-valeur < 0,05, et dont la taille d'effet est supérieure à 1 %).

La publication originelle mentionnait les genres *Acetanaerobacterium*, *Acetivibrio*, *Coprococcus*, *Faecalibacterium*, *Hespellia*, *Oscillibacter*, *Papillibacter*, *Sporobacter*, *Prevotella, Roseburia* et *Ruminococcus* comme étant plus abondants chez les patients colonisés par *Blastocystis*. Ici, seul QIIME retrouve une différence significative pour trois des genres

précédemment cités (*Coprococcus, Prevotella* et *Ruminococcus*). Ainsi, chaque pipeline donne une image apparemment fortement différente des genres variant significativement entre les deux groupes. Ces variations peuvent être induites par les différentes banques de séquences et taxonomies utilisées, connues pour varier de façon plus importante au niveau du genre. En outre, la capacité non égale des pipelines à identifier correctement les taxons à un tel niveau de résolution, particulièrement en utilisant une technologie avec erreurs, peut appuyer ce phénomène d'autant plus que la taille de l'amplicon utilisée pour réaliser cette identification peut être faible (voir Chapitre 3 Section 3.2.1). Au vu de ces résultats, il semble délicat d'affirmer des conclusions biologiques robustes sur la nature des taxons différentiellement observés au niveau du genre. Ces conclusions, même appuyées par des analyses statistiques valides, peuvent s'avérer inconsistantes et provoquer des inférences fausses. Malgré cela, ce sont des résultats souvent mentionnés dans la littérature, car ils peuvent être perçus comme plus éloquents d'un point de vue biologique : par exemple, on aura tendance à mettre en avant une variation de *Clostridium,* voire du pathogène *Clostridium difficile*, qui aura bien plus d'impact en termes de conclusions biologiques associées qu'une variation de leur famille parente Clostridiaceae, contenant également de nombreux taxons non-pathogènes. Ce biais de perception pousse de nombreuses études à approfondir leurs conclusions au niveau du genre, sans les modérer au regard du manque de robustesse de ce type d'analyses que nous pouvons clairement observer dans la Figure 4.21.

4.4 - Conclusion

Dans ce contexte d'étude, il est impossible de déterminer quel pipeline s'approche le plus de la réalité ; tout au plus peut-on mettre en parallèle leurs résultats avec d'autres propriétés inhérentes à chaque pipeline (par exemple mothur n'identifie pas de genre différentiellement observé entre les deux groupes, probablement car il est le pipeline le moins performant à ce niveau taxonomique, et qu'il identifie le moins de lectures initialement). Tous les

pipelines confirment la conclusion de l'étude initiale, selon laquelle la colonisation par *Blastocystis* ne semble pas pas associée à une dysbiose intestinale, et est au contraire associée à une richesse et diversité du microbiote bactérien intestinal plus élevé. Toutefois, différents pipelines rendent compte de plus ou moins d'observations de composition significativement différentes entre les deux groupes de patients pour renforcer cette conclusion, et associer la signature du microbiote bactérien chez les patients colonisés par *Blastocystis* à une signature de microbiote d'intestin sain. Ainsi, la variation d'interprétation biologique selon le pipeline utilisé ne se situe ainsi pas dans le sens de ce message, mais dans le nombre d'observations relevées pour le supporter.

La métagénétique a été popularisée comme étant un outil permettant d'accéder à l'ensemble des organismes en présence dans un échantillon. Par extension, cet outil est souvent utilisé à tort comme un moyen d'identification précis et absolu des organismes associés aux conditions étudiées. Nous démontrons ici que différents pipelines d'analyse révèlent un niveau de variation différent de ces organismes entre deux groupes d'échantillons, particulièrement à un niveau de résolution taxonomique détaillé. Ces variations, même si elles ne sont jamais contradictoires pour cette étude, renforcent le risque que le bioanalyste soit sélectif sur ses taxons d'intérêt, par biais de confirmation d'hypothèse. L'outil métagénétique doit être avant tout considéré comme un outil de profilage, adapté à révéler des différences entre deux conditions par l'utilisation d'indices dont la variation est robuste peu importe le pipeline d'analyse utilisé, comme les indices de richesse et de diversité.

L'évaluation des différences détaillées de compositions des microbiotes étudiés doit être effectuée avec un certain recul, en gardant à l'esprit les biais possibles induits par les outils d'analyse et les banques de séquences utilisés. L'emploi de deux pipelines d'analyse, aux approches et banques différentes, peut d'ailleurs être envisagée afin de corroborer les conclusions tirées des résultats des deux approches.

Chapitre 5 - Harpon : Design *de novo* d'amorces dégénérées à façon selon un microbiote d'intérêt

5.1 - Problématique du design d'amorces dans un contexte métagénétique

Comme évoqué dans l'état de l'art (Chapitre 1 Section 1.3.4) ainsi que démontré avec des données simulées (Chapitre 3 Section 3.2.1), le choix d'un *locus* cible pour une étude métagénétique, et d'un couple d'amorces permettant de l'amplifier, n'est pas trivial. Ce choix doit être effectué en considérant trois critères : la compatibilité physico-chimique des amorces utilisées, la capacité de ces amorces à capter un ensemble d'organismes d'intérêt, et la variabilité de l'amplicon qu'elles encadrent entre ces différents organismes.

5.1.1 – Critères généraux de design d'amorces

Certains critères inhérents à la séquence des amorces et aux principes même de la PCR sont à privilégier pour assurer une amplification conforme aux trois critères précédemment cités. Par exemple, plus une amorce est longue, plus elle sera spécifique d'un locus donné, moins elle aura de probabilité de capter une exhaustivité de taxons sur ce locus. En outre, elle nécessitera probablement une température de fusion élevée. Cette dernière est également impactée par le taux de nucléotides G et C constitutifs de l'amorce : plus ce taux est élevé, plus la température de fusion est élevée, et plus l'amorce s'hybridera de façon forte à sa cible en conditions salines (cations) et de température d'hybridation données. Toutefois, une température de fusion trop élevée peut générer du bruit et inhiber le fonctionnement de l'ADN polymérase. Dieffenbach *et al.* recommandent une taille d'amorce d'environ 20 nucléotides pour une amplification optimale, avec un taux de GC d'environ 50 %, ce qui garantit une température de fusion entre 56 et 62°C [Dieffenbach *et al.* 1993]. De nombreux autres critères peuvent être évalués, tels que la propension d'une amorce à s'hybrider sur elle-même (diminuant son efficacité), le nombre de polynucléotides, la complexité linguistique de l'amorce ...

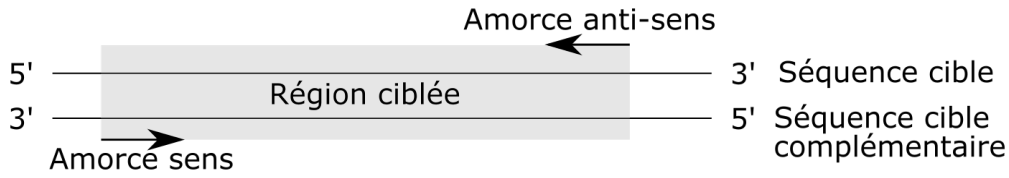


Figure 5.1 : Amorce sens et anti-sens encadrant une séquence cible d'intérêt.

Les amorces fonctionnent par couple dans une même réaction d'amplification : l'une dite sens, s'hybridant sur le brin complémentaire de la séquence cible, et l'autre dite anti-sens, s'hybridant sur la séquence cible (Figure 5.1). Pour une amplification optimale, les amorces d'un même couple doivent être compatibles entre elles, en ayant une température de fusion similaire et en n'étant pas complémentaires l'une de l'autre.

5.1.2 – Critères de design d'amorces dans un contexte métagénétique

Dans notre contexte métagénétique, le couple d'amorces sélectionné doit amplifier les cibles d'un maximum d'organismes constitutifs du microbiote d'intérêt. Cette universalité est l'idéal visé, essentiel pour limiter voire éviter un biais de sélection d'organismes, afin d'assurer que les résultats donneront une image représentative du microbiote initial. Chaque amorce doit ainsi être designée sur une région du *locus* d'intérêt qui est conservée chez tous ces organismes. Cette région est en général identifiée à partir d'un alignement multiple de la séquence de ce même *locus* pour une sélection de différents organismes jugés comme étant représentatifs du microbiote, ou d'intérêt pour l'étude. Une région conservée peut toutefois contenir certaines positions variables (en gras dans la Figure 5.2..a). Ainsi, si l'on choisit une amorce sur cette région en utilisant uniquement la séquence des nucléotides majoritaires, cette amorce ne captera pas l'ensemble des séquences (Figure 5.2..b). L'ajout de nucléotides dégénérés à certaines positions de l'amorce peut permettre d'augmenter le nombre de séquences captées (Figure 5.2.c). Ces nucléotides dégénérés correspondent à plusieurs bases possibles à une même position,

symbolisées par le code IUPAC dans la séquence de l'amorce (dans la Figure 5.2, S correspond aux bases C et G tandis que Y correspond aux bases C et T). Une amorce dégénérée est ainsi en réalité un mélange de plusieurs oligomères, de séquences quasi-identiques, sauf aux positions dites dégénérées. À noter que le nombre de dégénérescences d'une amorce (appelé ici *d*) ne correspond pas au nombre de positions dégénérées, mais au nombre d'oligomères différents générés à partir de cette amorce (dans la Figure 5.2.c, l'amorce a deux positions dégénérées, et a un nombre de dégénérescences $d = 4$).

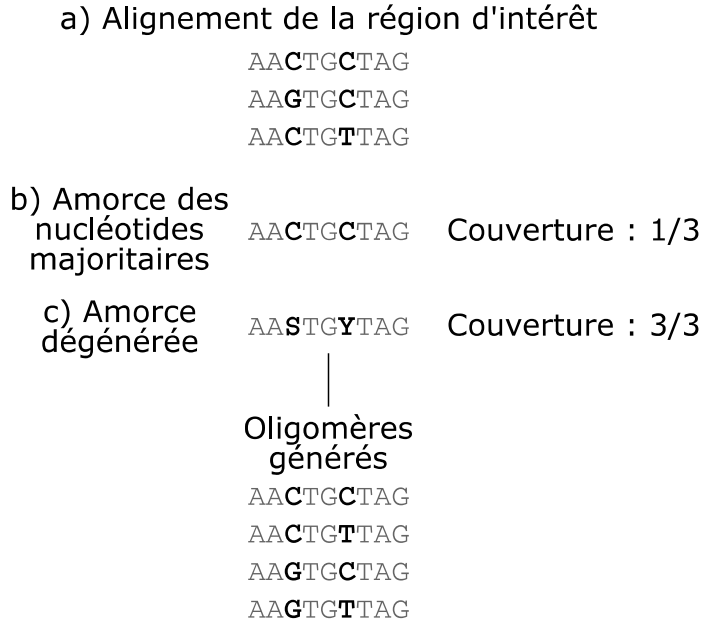


Figure 5.2 : Différences entre une amorce générée sur une fenêtre d'un alignement par les nucléotides majoritaires (a) et les nucléotides dégénérés aux positions variables en gras(b). Dans le code IUPAC, S correspond à C ou G et Y à C ou T.

Le design d'une d'amorce dégénérée à partir d'un alignement multiple de séquences correspond en réalité à deux problèmes distincts en informatique, tous deux NP-complets [Linhart & Shamir, 2002] :

- MD-DPD (*Minimum Degeneracy Degenerate Primer Design*) : recherche d'une amorce de taille fixe captant toutes les séquences de l'alignement en minimisant *d*, le nombre de dégénérescences ;

- MC-DPD (M*aximum Coverage Degenerate Primer Design*) : recherche d'une amorce de taille fixe captant un maximum des séquences de l'alignement avec un nombre maximum de dégénérescences d_{max} fixé.

MD-DPD optimise un nombre de dégénérescences pour que l'amorce s'aligne sur toutes les séquences de l'alignement, ce qui peut conduire à un nombre élevé de dégénérescences dans le cas d'un grand nombre de séquences considérées en entrée. À l'inverse, MC-DPD impose un nombre maximum de dégénérescences et va maximiser le nombre de séquences capturées.

5.1.3 – Solutions existantes

La plupart des logiciels de design d'amorces utilisent une approche exacte MD-DPD, en se basant sur la séquence consensus de l'alignement. Par exemple, primaclade [Gadberry *et al.* 2005], PriSM [Yu *et al.* 2011] et easyPAC [Rosenkranz 2012] utilisent une approche naïve en considérant tous les *k-mers* de la séquence consensus comme une amorce potentielle, qui est ensuite validée sur l'alignement complet. Toutefois, l'utilisation d'une séquence consensus génère souvent un nombre de dégénérescences excessif : il suffit en effet qu'une seule séquence diffère du nucléotide majoritaire à une position de l'alignement pour que cette position soit dégénérée. Or, plus une amorce sera dégénérée, moins l'amplification sera efficace à cause du grand nombre d'oligomères présents, augmentant le risque que l'amorce s'hybride sur d'autres régions que celle souhaitée (on parle alors d'aspécificité). Le nombre de dégénérescences doit ainsi être limité, ce qui nous impose de traiter le problème MC-DPD pour designer des amorces captant un maximum de séquences dans l'alignement tout en ayant un nombre de dégénérescences maximal pré-établi.

PrimerProspector [Walters *et al.* 2011] est un logiciel de design d'amorces développé pour un contexte métagénétique, et permettant le design de couples d'amorces dégénérées en intégrant RDP Classifier sur une banque de référence pour évaluer le niveau d'identification taxonomique des amplicons générés à partir des amorces designées. Toutefois, PrimerProspector génère des

amorces de taille fixe uniquement, et insère des dégénérescences dans ses amorces dès que les nucléotides concernés dépassent un certain pourcentage de représentation à cette position, sans limite de nombre de dégénérescences. Ainsi, sur une amorce de 15 nucléotides et avec un pourcentage de représentation de 50 %, on peut obtenir un nombre de dégénérescences allant de 1 à 32 768 ; PrimerProspector ne répond ainsi pas à la problématique MC-DPD. En outre, PrimerProspector utilise une ancre de 5 nucléotides conservés pour designer ses amorces, cette ancre correspondant à leur extrémité 3'. Ce fonctionnement se base en effet sur le critère communément utilisé selon lequel il est préférable d'éviter l'insertion de positions dégénérées dans les trois nucléotides en 3' des amorces, afin de renforcer leur spécificité. Une ancre conservée de 5 nucléotides, augmentant la taille de la région conservée par rapport à cette recommandation, empêche la possibilité de dégénérescences dans 10 nucléotides d'une amorce si on envisage de pouvoir l'utiliser comme amorce sens et comme amorce anti-sens. En recensant les amorces courramment utilisées dans la littérature pour des études métagénétiques, on constate que plusieurs d'entre elles ne remplissent pas ce critère : par exemple, 1/6^{e} des amorces évaluées et plus de la moitié des amorces prédites par Wang & Qian [Wang & Qian 2009] contiennent des positions dégénérées dans les 5 nucléotides en 5' et/ou en 3' des amorces. De même pour presque 2/3 des amorces évaluées par Teske & Sørensen [Teske & Sørensen 2008]. Dans notre contexte, nous souhaitions pouvoir générer des amorces pouvant contenir des dégénérescences en 5' et en 3', ce qui n'est pas possible avec PrimerProspector.

L'algorithme HYDEN [Linhart & Shamir, 2002] est le premier à avoir répondu au problème MC-DPD, en proposant deux approches heuristiques indépendantes, l'une dite d'expansion, l'autre dite de restriction. L'algorithme d'expansion (Figure 5.3) identifie dans un premier temps, dans chaque fenêtre de taille fixe de l'alignement, la séquence composée du nucléotide majoritaire à chaque position, qui devient l'amorce initiale. Afin d'augmenter le nombre de séquences captées par l'amorce, l'algorithme considère ensuite à chaque

position le second nucléotide majoritaire ; les positions pour lesquelles ce nucléotide est le plus élevé sont dégénérées, jusqu'à ce que le nombre maximum de dégénérescences d_{max} fixé soit atteint. À l'inverse, l'algorithme de restriction (non représenté ici) part d'une amorce totalement dégénérée sur chaque fenêtre, et supprime les nucléotides minoritaires à différentes positions jusqu'à atteindre le nombre maximum de dégénérescences d_{max} fixé.

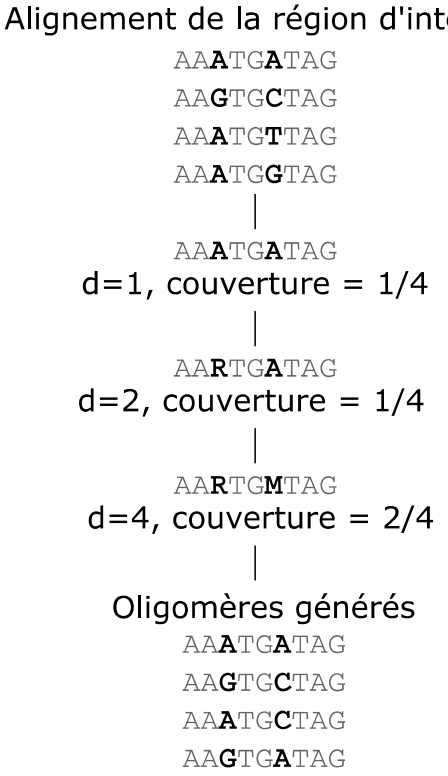


Figure 5.3 : Algorithme d'expansion de génération d'amorce par HYDEN, pour une taille d'amorce k=9 et un nombre maximum de dégénérescences d_{max}=4.

Hugerth *et al.* ont démontré que ces deux heuristiques ne produisent pas les solutions optimales [Hugerth *et al.* 2014], comme montré dans l'exemple simplifié de la Figure 5.3 : en effet, l'approche d'expansion tout comme de réduction génèrent l'amorce AARTGMTAG (captant la moitié des séquences), alors que l'amorce AAATGNTAG serait une meilleure solution dans ce cas, puisqu'elle capte 3/4 des séquences. Ainsi, inclure des positions dégénérées en

se basant sur le nombre décroissant de nucléotides majoritaires ne permet pas toujours de générer une solution optimale. Pour répondre à ce problème, Hugerth *et al.* ont développé DegePrime [Hugerth *et al.* 2014], qui applique l'algorithme d'expansion de HYDEN en modifiant l'incorporation de dégénérescences. Les positions dégénérées sont ajoutées aléatoirement dans l'amorce initiale (jusqu'à atteindre d_{max}), et le nombre de séquences captées par la nouvelle amorce est évalué. Cette étape est répétée n fois (par défaut n=100). Sur toutes ces répétitions, l'amorce captant le plus de séquences est retenue. Cette incorporation aléatoire génère toutefois trop de solutions différentes possibles, nécessitant un n très élevé pour trouver une solution optimale. Pour éviter cet écueil, les dégénérescences ajoutées sont sélectionnées avec une probabilité proportionnelle à leur fréquence. Ainsi, une position dégénérée captant un grand nombre de séquences aura plus de probabilité d'être incorporée dans les amorces designées DegePrime. À noter que les *gaps* dans l'alignement sont considérés comme une absence d'information ; DegePrime estime ainsi strictement qu'une amorce ne peut pas s'ancrer sur une région de l'alignement contenant un *gap*.

Outre le fait de devoir capter un maximum de séquences de l'alignement, les amorces designées doivent également encadrer une région la plus variable possible – c'est-à-dire dont la séquence sera la plus différente possible d'un taxon à un autre – puisque c'est sur ces différences que reposent les approches analytiques pour discriminer les taxons en présence. Peu de logiciels de design d'amorces existent actuellement qui prennent en compte cette variabilité de l'amplicon. PriFi [Fredslund *et al.* 2005] a été développé dans ce but, mais utilise une approche MD-DPD, et n'évalue pas la variabilité des amplicons générés. Jaric *et al.* ont publié une solution théorique à ce problème [Jaric *et al.* 2013], mais se basent sur une séquence de référence pour limiter le nombre de solutions ; une correspondance avec les co-auteurs de cet article nous a en outre confirmé l'indisponibilité du code associé à leur solution.

Nous avons développé Harpon afin de répondre à cette problématique de design d'amorces dans un contexte métagénétique, en prenant en compte trois critères : les paramètres physico-chimiques inhérents aux couples d'amorces, le taux de séquences d'intérêt qu'elles captent, et la variabilité des amplicons générés.

5.2 - Méthodes intégrées à Harpon

L'objectif principal de Harpon est, à partir d'un alignement multiple de séquences d'intérêt, de générer un ensemble de couples d'amorces compatibles, s'hybridant sur un maximum de ces séquences, et amplifiant les régions de l'alignement les plus variables possibles. L'algorithme de Harpon repose sur trois étapes. Tout d'abord, l'alignement est analysé afin d'identifier les régions les plus conservées. Puis, les amorces sont générées sur ces régions de façon à capturer un maximum des séquences. Enfin, des couples d'amorces sont validés en prenant en compte leur compatibilité pour la réaction d'amplification, et l'intervalle de taille d'amplicon choisi par l'utilisateur. Un calcul d'entropie et de distance topologique permet d'évaluer la variabilité, donc l'intérêt potentiel, de l'amplicon généré par chaque couple d'amorces.

5.2.1 Analyse de l'alignement

La Figure 5.4 représente les différentes étapes d'analyse de l'alignement initial. Tout d'abord, le nombre d'occurrences du nucléotide majoritaire à chaque position de l'alignement est calculé (Figure 5.4.b, les gaps sont ignorés). Puis, une fenêtre glissante de taille k (égale à la taille minimale d'amorce souhaitée, choisie par l'utilisateur) est utilisée afin d'identifier les régions conservées de l'alignement. Pour ce faire, la conservation moyenne cm pour chaque fenêtre est calculée, en faisant la moyenne du nombre d'occurrences du nucléotide majoritaire (hors gaps) à chaque position de la fenêtre. Une fenêtre est retenue si sa conservation moyenne cm se trouve à un certain pourcentage tp (*top percent*) de la conservation maximale trouvée dans l'alignement. Dans l'exemple de la Figure 5.4.c, l'alignement a une conservation moyenne cm

maximale de 4. Pour *tp* défini à 25 %, une fenêtre est retenue si sa conservation moyenne *cm* est supérieure ou égale à 3. Ensuite, les fenêtres contiguës sont fusionnées en régions conservées (Figure 5.4.d).

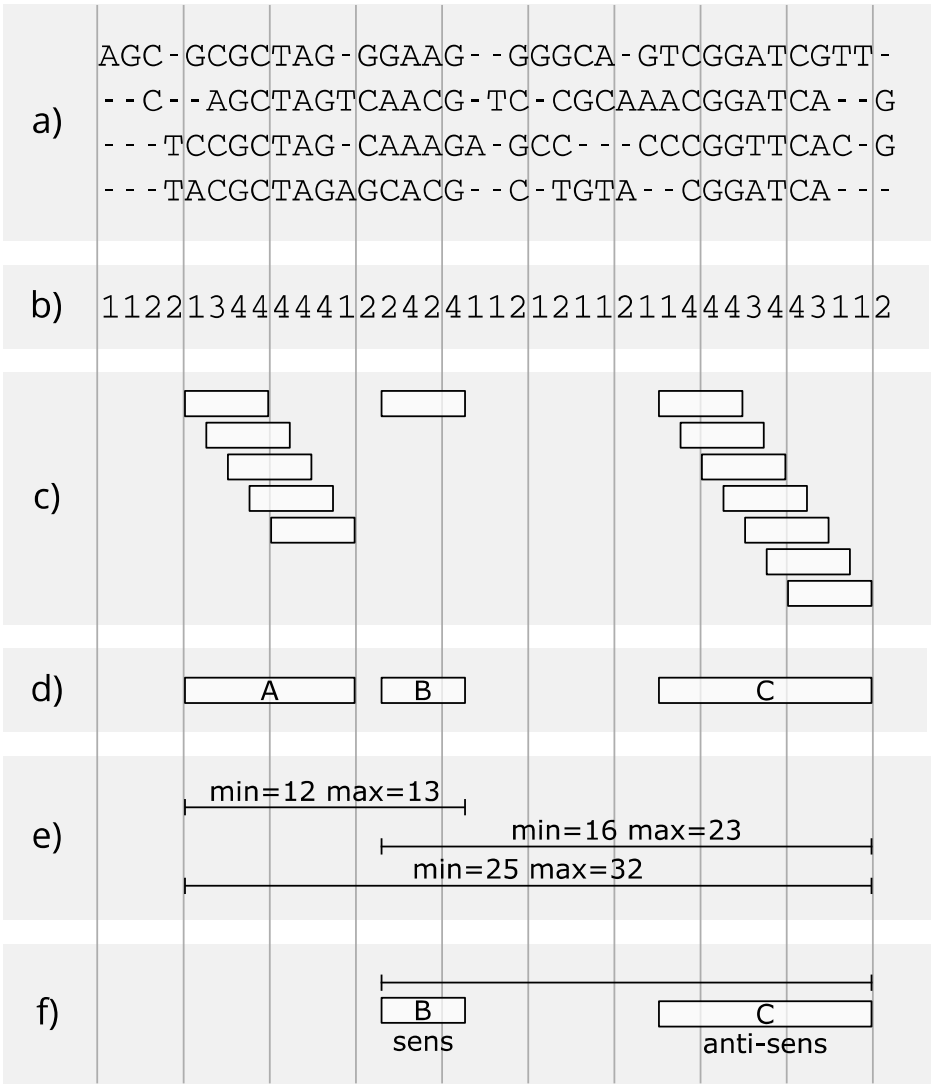


Figure 5.4 : Exemple d'analyse d'un alignement par Harpon, pour les paramètres k = 4, tp = 25, A_{min} = 15, A_{max} = 30.

a) Alignement initial
b) Calcul de la conservation des nucléotides majoritaires
c) Sélection des fenêtres (k = 4, tp = 25 %, soit cm ≥ 3)
d) Fusion des fenêtres contiguës en régions d'intérêt
e) Calcul des distances minimale (min) et maximale (max) entre toutes les paires de régions
f) Sélection des régions compatibles avec min > A_{min} et max < A_{max}

Un filtrage des régions conservées est ensuite effectué en fonction de la distance qui les sépare (Figure 5.4.e et 5.4.f). Pour chaque paire de régions conservées, la distance minimale (*min*) et maximale (*max*) des séquences de l'alignement (sans *gaps*) entre le début de la première région et la fin de la deuxième région est calculée (Figure 5.4.e). Si la plus petite distance entre deux régions est inférieure à la taille maximum d'amplicons A_{max} et si la plus grande distance entre deux régions est supérieure à la taille minimum d'amplicons A_{min} (valeurs choisies par l'utilisateur), alors ces deux régions sont sauvegardées (Figure 5.4.f, régions B et C).

À l'issue de cette étape, on obtient un ensemble de paires de régions d'intérêt de l'alignement, avec une région dite sens et l'autre région dite anti-sens. Ces régions peuvent maintenant être utilisées pour designer des amorces par une approche MC-DPD.

5.2.2 - Design d'amorces

DegePrime est utilisé sur chaque région d'intérêt, avec une fenêtre de taille k, et un nombre de dégénérescences maximum d_{max} choisi par l'utilisateur. DegePrime génère en fichier de sortie une liste d'amorces de taille fixe k, avec les informations représentées dans le Tableau 5.5.

Le code de DegePrime a été modifié pour également générer le champ SeqIds, correspondant à la liste des séquences où chaque amorce est retrouvée (en italique dans le Tableau 5.5). Cette information sera utilisée dans l'étape ultérieure de création de couples d'amorces.

Pos	Total Seq	Unique Mers	Entropy	Primer Deg	Primer Matching	PrimerSeq	*SeqIds*
0	28	11	3.16898695845306	8	19	TTAGCTAGTWGGTRRGG	*2;3;7;8;9;10;11;12;13;14;15;16;20;23;24;25;26;27*
1	28	11	3.16898695845306	8	19	TAGCTAGTWGGTRRGGT	*2;3;7;8;9;10;11;12;13;14;15;16;20;23;24;25;26;27*
2	28	11	3.16898695845306	8	19	AGCTAGTWGGTRRGGTA	*2;3;7;8;9;10;11;12;13;14;15;16;20;23;24;25;26;27*
3	28	11	3.16898695845306	8	19	GCTAGTWGGTRRGGTAA	*2;3;7;8;9;10;11;12;13;14;15;16;20;23;24;25;26;27*
4	28	17	3.83880436924461	8	14	CTAGTWGGTGRGGTAAM	*2;3;8;10;12;13;14;16;20;23;24;25;26;27*
5	28	15	3.65147892289331	8	15	TAGTWGGTGRGGTAAMG	*2;3;8;10;12;13;14;16;18;20;23;24;25;26;27*
6	28	15	3.65147892289331	8	15	AGTWGGTGRGGTAAMGG	*2;3;8;10;12;13;14;16;18;20;23;24;25;26;27*
7	28	13	3.45470124416735	8	19	GTWGGTGRGGTAAMGGC	*2;3;4;8;10;12;13;14;16;17;18;20;21;22;23;24;25;26;27*

Tableau 5.5 : Exemple de fichier de sortie DegePrime. Une modification du code originel permet d'ajouter, pour chaque amorce, la liste des séquences qu'elle capte (en italique).

- Pos : Position de l'amorce sur la région totale
- TotalSeq : Nombre de séquences sur la fenêtre
- UniqueMers : Nombre d'oligomères uniques (sans gaps) dans la fenêtre
- Entropy : Entropie de Shannon sur la fenêtre
- PrimerDeg : Nombre d'oligomères générés par l'amorce
- PrimerMatching : Nombre de séquences parfaitement captées par l'amorce
- PrimerSeq : Séquence de l'amorce
- SeqIds : Identifiants des séquences captées

DegePrime génère une liste d'amorces de taille fixe k. À partir de cette liste, nous souhaitions pouvoir générer des amorces de taille variable, entre k et $k_{max.}$; cette variation de taille permet de faire varier les paramètres physico-chimiques des amorces, afin d'augmenter le nombre d'amorces candidates et compatibles entre elles pour former des couples. L'évaluation de toutes les possibilités de taille d'amorce entre k et k_{max} nécessiterait d'exécuter DegePrime k_{max}-k+1 fois sur chaque fenêtre. Cette solution est plus exhaustive, mais augmenterait drastiquement le nombre d'exécutions de DegePrime et donc les temps d'exécution ; elle n'a ainsi pas été retenue.

Nous avons considéré la liste initiale des amorces générées sur une fenêtre k comme des ancres (en noir dans la Figure 5.6), qui vont être étendues en 3'. Pour chaque amorce candidate, si le nucléotide suivant sur l'alignement en 3' de cette amorce capte 100 % des séquences, alors une nouvelle amorce candidate (en gris dans la Figure 5.6) est créée, en ajoutant ce nucléotide à l'amorce candidate initiale. Cette action est répétée tant que les nouvelles amorces candidates sont de taille inférieure à k_{max}, et tant que le nucléotide ajouté en 3' capte 100 % des séquences. Ce seuil de 100 % permet de garantir que l'amorce allongée captera les mêmes séquences que son ancre, même si cette solution n'est pas la plus exhaustive.

Une solution intermédiaire, en cours d'intégration, est d'évaluer la possibilité d'ajouter une dégénérescence aux positions allongées pour toute amorce n'ayant pas un nombre de dégénérescences atteignant d_{max}.

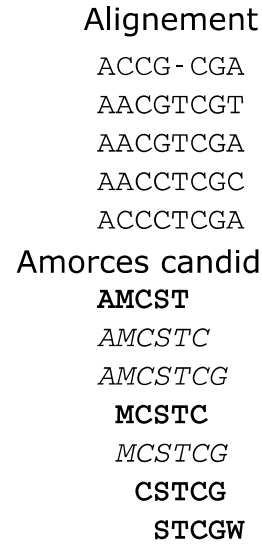


Figure 5.6 : Amorces générées à l'issue de DegePrime (en gras, k = 5) et allongées par Harpon (en italique).

Pour l'ensemble des amorces candidates générées, les propriétés suivantes sont récupérées du fichier de sortie de DegePrime (Les amorces allongées en italique dans la Figure 5.6 captant les mêmes séquences que leur amorce parente) :

- Séquence de l'amorce ;
- Pourcentage de séquences captées ;
- Liste des séquences parfaitement retrouvées par l'amorce.

Les propriétés suivantes sont calculées pour chaque amorce :

- Longueur maximale d'homopolymères ;
- Nombre de dégénérescences (= d'oligomères différents) ;
- Température de fusion T_m minimum, moyenne et maximum ;
- Pourcentage de GC moyen, minimum et maximum ;
- Complexité [Orlov & Potapov, 2004].

Si l'amorce est générée sur une région anti-sens, ces propriétés sont calculées à partir de la séquence complémentaire anti-sens de l'amorce. À l'issue de cette étape, on obtient une liste d'amorces candidates par région d'intérêt.

5.2.3 - Couples d'amorces compatibles

Un couple potentiel d'amorces est composé d'une amorce designée sur une région d'intérêt sens, et d'une amorce designée sur une région d'intérêt anti-sens. Pour chaque paire de régions d'intérêt précédemment définie (Figure 5.4.f), chaque couple potentiel d'amorces est évalué. Les couples d'amorces respectant les critères suivants sont conservés :

- Les amplicons générés par le couple d'amorce doivent avoir une taille minimale inférieure à A_{max} et une taille maximale supérieure à A_{min}
- Les amorces doivent avoir une température de fusion moyenne compatible (dont la différence est inférieure à *maxTmDiff*, choisi par l'utilisateur)

Les séquences captées par l'amorce sens d'un couple ne sont pas forcément captées par l'amorce anti-sens du même couple. Il est ainsi nécessaire de calculer le pourcentage de séquences captées à la fois par l'amorce sens et à la fois par l'amorce anti-sens. Si une des deux amorces capte 100 % des séquences, le taux de séquences captées par le couple d'amorces correspond à celui de l'autre amorce. Sinon, ce taux est la proportion de séquences à l'intersection des listes de séquences captées par chaque amorce (information ajoutée dans la sortie de DegePrime, en italique dans le Tableau 5.5). Si le taux de séquences capté par le couple d'amorces est supérieur ou égal à C % (seuil choisi par l'utilisateur), le couple est considéré comme un candidat.

Pour tous les couples candidats, l'entropie de Shannon est calculée sur l'alignement des séquences comprises entre les deux amorces Cette entropie permet d'évaluer la variabilité d'information de la région amplifiée : plus l'entropie est élevée, plus la région est variable et donc potentiellement intéressante dans notre contexte.

5.2.4 - Affichage des couples d'amorces candidats et évaluation de la variabilité des amplicons par calcul de distance topologique

Pour chaque paire de régions d'intérêt, les couples d'amorces validés sont affichés après tri décroissant des paramètres suivants :

- Taux de séquences captées par le couple d'amorces ;
- Taille moyenne de l'amplicon généré ;
- Température minimale de fusion ;
- Température moyenne de fusion ;
- Entropie de l'amplicon.

Les couples d'amorces considérés comme les meilleurs candidats sont ainsi ceux captant un maximum de séquences, générant l'amplicon le plus grand, ayant une température minimale de fusion la plus élevée, et encadrant une région la plus variable. Le nombre maximum d'amorces à afficher par couple de régions d'intérêt (par défaut, 10) peut être défini par l'utilisateur.

Comme défini précédemment, plus l'entropie d'une région est élevée, plus celle-ci est potentiellement informative. Une autre manière d'évaluer l'intérêt d'un couple d'amorces dans notre contexte est d'observer à quel point la région amplifiée est représentative de l'alignement total. Autrement dit, dans quelle mesure la région amplifiée permet-elle de ségréguer les amplicons aussi bien que les séquences complètes de l'alignement ? Pour répondre à cette question, on peut comparer l'arbre phylogénétique généré à partir de la région amplifiée à l'arbre phylogénétique généré par l'alignement des séquences complètes. Plus les deux arbres auront une topologie proche, plus la région amplifiée sera représentative de l'alignement global. Cette comparaison d'arbres a été intégrée à Harpon : pour chaque couple d'amorces affiché, un arbre phylogénétique est construit à partir de l'alignement de la région amplifiée par le logiciel FastTree 2.1.9, choisi pour sa rapidité d'exécution [Price *et al.* 2010]. Cet arbre doit ensuite être comparé à l'arbre phylogénétique généré par FastTree pour l'alignement complet. De nombreux outils permettent de

comparer la topologie de deux arbres phylogénétiques. Ces méthodes nécessitent toutefois que les arbres générés soient des arbres binaires stricts (chaque nœud interne doit avoir deux fils, Figure 5.7.a et 5.7.b) et qu'ils contiennent les mêmes séquences. Or, dans notre contexte, les amplicons ne concernent pas forcément les mêmes séquences, et ne sont souvent pas suffisamment informatifs pour les discriminer (Figure 5.7.c). Seul l'algorithme treeKO [Marcet-Houben & Gabaldón, 2011] permet la comparaison de la topologie de deux arbres en présence de ce type d'ambiguïtés. treeKO a ainsi été intégré à Harpon pour calculer une distance (appelée distance taxonomique) entre l'arbre généré par l'alignement global, et celui généré par l'alignement des amplicons. Plus cette distance est proche de 0, plus les deux arbres ont une topologie similaire (Figure 5.7.b). Afin de réduire les temps d'exécution de Harpon, la génération d'arbres et le calcul de distance topologique sont effectués uniquement pour les couples d'amorces affichés en résultat. Cette distance permet ainsi à l'utilisateur d'évaluer à quel point l'amplicon généré par un couple d'amorces est représentatif de l'alignement global donné en entrée.

Si l'utilisateur utilise un alignement de séquences d'ADNr 16S, il peut choisir d'afficher une information complémentaire, qui est la position des amorces sur la séquences d'ADNr 16S de référence d'*Escherichia coli*, ainsi que le nom des régions hypervariables amplifiées. Cette information complémentaire est générée par Harpon en exécutant un BLASTn de la séquence d'amorce sur la séquence de référence d'ADNr 16S d'*Escherichia coli*, et en évaluant la position des amorces par rapport aux positions des régions hypervariables définies par Baker *et al.* [Baker *et al.* 2003].

Un exemple de résultat final peut être observé dans la section dédiée à ce chapitre sur la page Internet associée au projet de thèse : http://www.pegase-biosciences.com/2013-0920.

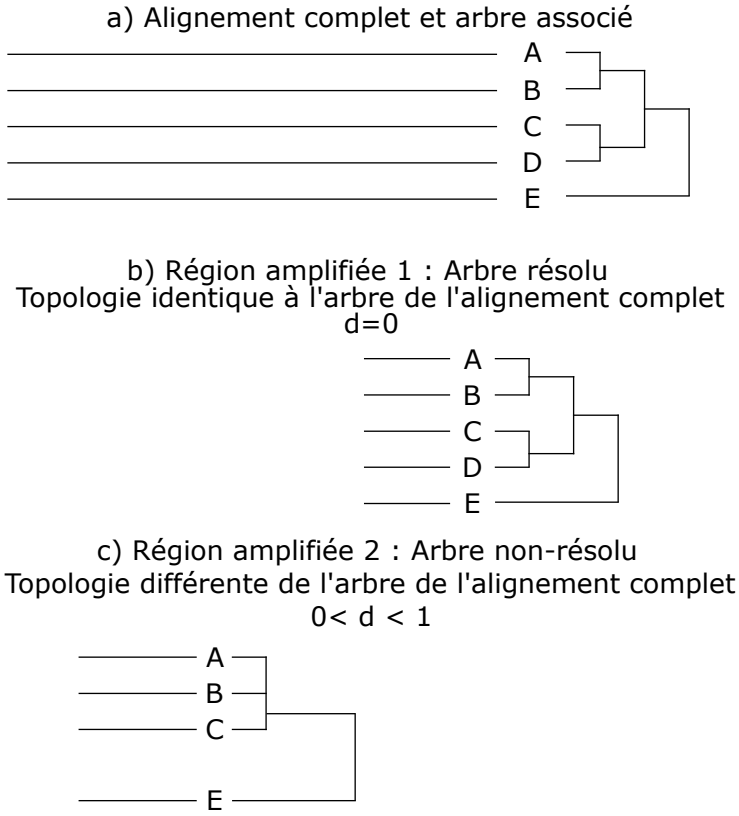


Figure 5.7 : Exemples d'arbres générés par différentes régions amplifiées, et distance d de l'arbre généré par l'alignement complet.

5.2.5 - Intégration et valorisation de Harpon

Harpon est un pipeline développé en Perl orienté objet, et s'appuie sur les librairies Bioperl 1.6 [Stajich *et al.* 2002]. Les modalités de distribution du logiciel sont en cours de discussion : en effet, Harpon peut être valorisé sous la forme d'une exclusivité de la plate-forme PEGASE-biosciences à créer des amorces les plus adaptées à un plan d'expérience. À l'inverse, une valorisation du logiciel et de son approche peut également être envisagée sous la forme d'une publication.

5.3 - Validation de Harpon sur différents contextes métagénétiques

L'ensemble des données utilisées et générées dans cette section sont disponibles sur la partie dédiée à ce chapitre de la page Internet associée au projet de thèse : http://www.pegase-biosciences.com/2013-0920.

5.3.1 - Validation de Harpon sur le design d'amorces ciblant deux régions de l'ADNr 16S bactérien

Harpon a été utilisé sur le jeu de données simulé décrit dans le Chapitre 3 Section 3.1.1, contenant 504 séquences d'ADNr 16S de 125 espèces bactériennes différentes. L'objectif était d'évaluer si Harpon permettait de retrouver des amorces similaires, voire plus performantes, que les couples d'amorces 200(V3) et 400(V4-V5). Pour ce faire, un alignement multiple de ces 504 séquences a tout d'abord été généré par Clustal Omega [Sievers *et al.* 2011] avec les paramètres par défaut (logiciel choisi car il générait l'alignement respectant au mieux les délimitations entre régions conservées et régions hypervariables). Cet alignement a été utilisé en entrée de Harpon, avec les paramètres suivants définis pour correspondre aux propriétés des amorces 200(V3) et 400(V4-V5) :

- Taille minimale d'amorce k : 15 nucléotides ;
- Taille maximale d'amorce k_{max} : 22 nucléotides ;
- Top percent tp : 5 % ;
- Seuil de conservation C : 5 % ;
- Nombre maximum de dégénérescences d_{max} : 4 ;
- Différence maximale de Tm moyenne entre deux amorces d'un même couple max$TmDiff$: 10°C ;
- Taille minimale d'amplicon A_{min} : 150 (200(V2) faisant en moyenne 188 nucléotides sur ce jeu de données) ;
- Taille maximale d'amplicon A_{max} : 410 (400(V4-V5) faisant en moyenne 408,08 nucléotides sur ce jeu de données).

Harpon identifie 1 421 amorces potentielles d'intérêt, pouvant former

8 008 couples différents. En filtrant les couples d'amorces pour ne garder que le meilleur couple par paire de régions conservées (selon l'ordre des critères définis Chapitre 5 Section 5.2.4), nous identifions 7 couples d'amorces candidats distincts, dont deux couples, appelés 16SA et 16SB, amplifiant les mêmes régions hypervariables que les couples 200(V3) et 400(V4-V5) respectivement. L'ensemble de ces couples d'amorces ainsi que leurs propriétés sont représentés dans les Tableaux 5.8 et 5.9.

Le couple d'amorces 16SA permet de capter 33 séquences (6,5 %) de plus dans le jeu de données artificiel que le couple 200(V3), en générant quasiment le même amplicon (décalé de 3 nucléotides), de variabilité comparable (distance topologique légèrement inférieure et entropie identique).

Le couple d'amorces 16SB, quant à lui, génère un amplicon plus court (380 nucléotides en moyenne) que l'amplicon 400(V4-V5) (408 nucléotides en moyenne) ; cette diminution de taille entraîne une diminution d'entropie, et une légère augmentation de distance topologique. L'amplicon ciblé est ainsi légèrement moins discriminant. Par contre, 16SB capte 48 séquences du jeu de données de plus que le couple 400(V4-V5), soit une augmentation de sensibilité de presque 10 %.

Nous pouvons ainsi valider que, sur le jeu de données utilisé, les couples d'amorces 16SA et 16SB captent plus de séquences que les couples 200(V3) et 400(V4-V5) respectivement, et génèrent des amplicons de variabilité similaire pour 16SA et légèrement plus faible pour 16SB.

Nom du couple d'amorces	Régions hypervariables encadrées	Taille moyenne	Écart type de taille	Nombre de séquences captées par le couple	Distance topologique	Entropie	Positions sur *E. coli*
16SA	V3	188,00	11,09	494 (98,02 %)	0,57	132,26	338-534
200(V3) (Probio_Uni-Probio_Rev)	V3	188,00	11,09	461 (91,47 %)	0,58	132,26	341-537
16SB	V4-V5	380,08	1,50	492 (97,62 %)	0,54	205,95	515-894
400(V4-V5) (519F-907R)	V4-V5	408,08	1,50	444 (88,10 %)	0,53	212,80	519-926

Tableau 5.8 : Couples d'amorces et leurs propriétés sur l'alignement des 504 séquences d'ADNr 16S d'intérêt.

Nom	Séquence	Taille	Tm moyen	Taille de l'homopolymère le plus long	Nombre de dégénérescences (= oligomères générés)
16SAF	ACTCCTRCGGGAGGC	15	53,22	3	2
16SAR	ACCGCGGCKGCTKGC	15	61,15	2	4
Probio_Uni	CCTACGGGRSGCAGCAG	17	58,33	6	4
Probio_Rev	ATTACCGCGGCTGCT	15	52,95	2	1
16SBF	GTGCMAGCMGCCGCGG	16	62,04	2	4
16SBR	CGTACTYCCCAGGYGG	16	53,19	4	4
519F	CAGCMGCCGCGGTAATAC	18	56,87	2	2
907R	CCGTCAATTCMTTTGAGTTT	20	50,08	3	2

Tableau 5.9 : Couples d'amorces et leurs propriétés, générées à partir de l'alignement de 504 séquences d'ADNr 16S d'intérêt.

Afin d'évaluer les performances des nouveaux couples d'amorces sur un jeu de données plus vaste, une PCR *in silico* a été effectuée avec les couples 16SA et 16SB sur la banque RDP par probe match [Cole *et al.* 2005], et sur la banque SILVA SSURef NR 128 par TestPrime 1.0 [Klindworth *et al.* 2013] (les deux logiciels étant intégrés sur les sites respectifs des banques utilisées). Les proportions de séquences bactériennes captées par les deux couples d'amorces ainsi que par 200(V3) et 400(V4-V5) sont présentées dans le Tableau 5.10. Cette évaluation a également été effectuée sur les séquences archées présentes dans ces banques de référence.

	% séquences amplifiées RDP (archées - bactéries)	**% séquences amplifiées SILVA (archées - bactéries)**
16SA	0,02 - 68,62	0,00 - 89,85
200(V3)	0,11 - 63,60	0,16 – 82,39
16SB	29,30 – 54,85	39,75 - 85,42
400(V4-V5)	0,31 - 50,12	0,00 - 79,45

Tableau 5.10 : Taux de séquences archées et bactériennes captées par les couples d'amorces dans les banques SILVA 128 et RDP 11.5.

16SA est le couple ayant les meilleures performances pour les deux banques de séquences, diminuant le taux de séquences archées captées de 9 % pour RDP à 16 % pour SILVA, et augmentant le taux de séquences bactériennes captées de 5 % et 7 % respectivement. En évaluant les amorces de façon individuelle, on remarque que l'amorce 16SAR est une variante de l'amorce E517F, décrite comme étant l'une des amorces bactériennes les plus universelles [Soergel *et al.* 2012]. E517F capte en effet 94,26 % des séquences bactériennes de SILVA (et 1,01 % d'archées) ; notre variante 16SAR en capte 95,53 % (et 58,10 % d'archées).

16SB augmente également le taux de séquences bactériennes captées, de 4 à 6 % pour les deux banques respectivement. Toutefois, 16SB augmente également fortement le taux de séquences archées amplifiées. Selon l'application souhaitée, cette présence d'archées peut être plus ou moins

souhaitée. À nouveau, en évaluant les amorces de façon individuelle, on remarque que que 16SBF est une variante de l'amorce U515F, également l'une des plus universelles [Soergel *et al.* 2012], et décrite pour capter également des séquences archées. Sur SILVA, U515F capte 93,61 % de séquences bactériennes (57,08 % d'archées), tandis que 16SBF capte 95,41 % de séquences bactériennes (57,44 % archées).

L'ensemble de ces amorces ont enfin été testées *in silico* sur le génome humain en utilisant MFE-primer2.0 [Qu *et al.* 2012] afin d'évaluer le risque d'interaction avec le génome hôte, dans un contexte d'étude de flores intestinale ou pulmonaire humaines, par exemple. Pour 16SA et 16SB, le risque d'amplifier du génome hôte contaminant est plus élevé que pour les couples 200(V3) et 400(V4-V5). En effet, l'augmentation de sensibilité des deux nouveaux couples d'amorces a entraîné une diminution de spécificité. L'intégration d'un module optionnel intégrant MFE-primer2.0 à Harpon est prévue, afin de mesurer le risque d'amplifier du génome contaminant pour que l'utilisateur puisse pondérer le classement des amorces par ce risque s'il le souhaite.

Les amorces largement utilisées dans la littérature pour amplifier l'ADNr 16S dans un contexte de métagénomique ciblée bactérienne sont déjà fortement optimisées pour capter un maximum de séquences les plus variables possibles avec un minimum de risque de contamination. Toutefois, Harpon a permis d'identifier deux couples d'amorces, 16SA et 16SB, augmentant le nombre potentiel d'organismes bactériens captés sur le jeu de données d'intérêt tout comme sur des banques de séquences de référence. Cette augmentation de sensibilité est contrebalancée par une perte de spécificité, notamment pour le couple 16SB (amplification d'archées et diminution de variabilité de l'amplicon). Ces amorces risquent en outre d'amplifier plus de séquences contaminantes si elles sont utilisées pour étudier un microbiote bactérien chez l'homme, et nécessitent ainsi d'être validées *in vitro*.

Cette première évaluation nous a permis de valider le principe de fonctionnement de Harpon, permettant d'identifier les couples d'amorces les plus optimaux possibles sur ce jeu de données de séquences d'ADNr 16S, en respectant certaines contraintes physico-chimiques. Nous avons, en outre, identifié des pistes d'amélioration sous la forme de modules complémentaires qui sont en cours de développement (évaluation du risque d'amplification de séquences contaminantes, évaluation de la performance des couples d'amorces sur une banque de référence, ...)

5.3.2 - Utilisation de Harpon pour un design d'amorces sur l'ADNr 16S d'un ensemble d'archées présentes dans la subsurface sédimentaire de fonds marins

Beaucoup d'amorces présentes dans la littérature ont été designées sur la base de banques de séquences de référence, cherchant à maximiser le nombre de séquences captées par ces amorces dans ces banques. Ce n'est toutefois pas parce qu'une amorce est la plus universelle possible sur une banque de référence, qu'elle sera adaptée à toute étude métagénétique. Par exemple, Teske & Sørensen ont publié en 2008 une évaluation de 10 amorces captant l'ADNr 16S d'archées, en étudiant leurs performances *in silico* sur des archées constitutives de la subsurface sédimentaire de fonds marins [Teske & Sørensen 2008]. Ce microbiome héberge en effet de nombreuses espèces d'archées dont la séquence d'ADNr n'est pas connue à ce jour. L'évaluation d'amorces communément utilisées sur des archées récemment découvertes dans ce milieu a permis d'évaluer le biais de sélection causé par ces amorces, qui pouvaient manquer jusqu'à 70 % des organismes en présence.

Pour mener cette évaluation, les auteurs ont constitué un jeu de données de 561 séquences d'ADNr 16S d'archées d'intérêt [Teske & Sørensen 2008, *Supplementary Material*], sur lequel la performance de 10 amorces issues de la littérature a été évaluée par PCR *in silico*. Ces amorces ne s'alignent pas sur de nombreuses séquences, allant de 5 % à 71 % de séquences manquées. Les amorces les plus performantes étaient PARCH519F et ARC806R, s'alignant sur

respectivement 91 et 95 % des séquences du jeu de données. Les auteurs indiquaient en conclusion de l'article que l'introduction de dégénérescences dans les amorces pouvait permettre d'augmenter ce pourcentage, mais au dépens de la spécificité d'amplification.

Nous avons souhaité réutiliser ces 561 séquences pour designer avec Harpon un couple d'amorces compatibles permettant d'en capter un maximum, encadrant la même région que PARCH519F et ARC806R. Cependant, l'alignement multiple de l'ensemble de ces 561 séquences était délicat, ces dernières étant de tailles très différentes et ne couvrant pas l'entièreté de l'ADNr 16S. Nous avons ainsi choisi d'utiliser comme alignement de référence celui d'un ensemble réduit de 61 de ces séquences, de taille supérieure à 600 nucléotides [Teske & Sørensen 2008, Figure 1]. Harpon a été exécuté dans un premier temps sur cet alignement réduit. Nous avons ensuite évalué la pertinence des meilleurs couples d'amorces générés sur l'ensemble des 561 séquences initiales, par PCR *in silico*.

Les 61 séquences du jeu de données réduit ont été alignées en utilisant Clustal Omega avec les paramètres par défaut. Cet alignement a été utilisé en entrée de Harpon, avec les paramètres suivants définis pour correspondre aux propriétés des amorces évaluées dans la publication de Teske & Sørensen, et pour générer des amplicons entre 100 et 400 nucléotides en moyenne :

- Taille minimale d'amorce k : 15 nucléotides ;
- Taille maximale d'amorce k_{max} : 25 nucléotides ;
- Top percent tp : 30 % ;
- Seuil de conservation C : 30 % ;
- Nombre maximum de dégénérescences d_{max} : 6 ;
- Différence maximale de Tm moyenne entre deux amorces d'un même couple *maxTmDiff* : 10° ;
- Taille minimale d'amplicon A_{min} : 100 ;
- Taille maximale d'amplicon A_{max} : 400.

Harpon identifie 1 569 amorces potentielles d'intérêt, pouvant former 855 couples différents, sur 4 paires de régions conservées. Le couple d'amorces captant la même région que PARCH519F et ARC806R est le couple nommé ArchA, représentés dans les Tableaux 5.11 et 5.12. Le pourcentage de séquences captées a été mesuré par PCR *in silico* sur l'ensemble des 561 séquences initiales, en utilisant le logiciel dispr [Scofield 2015].

ArchA est un couple d'amorces quasiment identique à PARCH519F-ARC806R, mais plus adapté aux séquences d'intérêt : il capte 4 séquences en plus (augmentation de la sensibilité), et augment l'entropie de la région amplifiée tout comme il diminue légèrement la distance topologique associée (augmentation de la spécificité).

Nom du couple d'amorces	Régions hypervariables encadrées	Taille moyenne	Écart type de taille	Nombre de séquences captées par le couple sur les 561 séquences	Distance topologique	Entropie	Positions sur *E. coli*
ArchA	V4	288,24	0,55	441 (78,60 %)	0,32	175,21	518-805
PARCH519F-ARC806R	V4	288,25	0,56	437 (77,90 %)	0,33	174,07	519-806

Tableau 5.11 : Couples d'amorces et leurs propriétés sur les 61 séquences d'ADNr 16S d'archées d'intérêt (sauf le nombre de séquences captées par le couple qui a été mesuré sur les 561 séquences initiales).

Nom	Séquence	Taille	Tm moyen	Taille de l'homopolymère le plus long	Nombre de dégénérescences (= oligomères générés)
ArchAF	BCAGCMGCCGCGGTA	15	57,94	2	6
ArchAR	GACTACMSGGGTATCTAATC	20	49,69	4	4
PARCH519F	CAGCMGCCGCGGTAA	15	55,57	2	2
ARC806R	GGACTACVSGGGTATCTAAT	20	51,85	5	6

Tableau 5.12 : Couples d'amorces et leurs propriétés, générées à partir de l'alignement de 61 séquences d'ADNr 16S d'archées d'intérêt.

Une amplification *in silico* a également été effectuée sur la banque RDP par probe match, et sur la banque SILVA SSURef NR 128 par TestPrime 1.0. Les résultats se trouvent dans le tableau 5.13 :

	% séquences amplifiées RDP (archées - bactéries)	**% séquences amplifiées SILVA (archées - bactéries)**
ArchA	74,17 - 4,74	85,33 - 6,91
PARCH519F-ARC806R	74,07 - 4,76	85,29 - 6,87

Tableau 5.13 : Taux de séquences archées et bactériennes captées par les couples d'amorces dans les banques SILVA 128 et RDP 11.5.

Ainsi, Harpon a permis de générer automatiquement le couple d'amorces candidat ArchA montrant des performances comparables au couple d'amorces PARCH519F-ARC806R pour l'étude d'archées présentes dans la subsurface sédimentaire de fonds marins.

5.3.3 - Utilisation de Harpon pour trouver des couples d'amorces adaptées à l'étude de la diversité fongique dans un contexte clinique et compatibles avec un projet de séquençage MiSeq paired-end 2x250 nt

L'étude métagénétique du microbiote fongique a historiquement été basée sur l'ADNr 18S, marqueur taxonomique de référence chez les eucaryotes. Ce n'est que récemment que les deux régions ITS (*Internal Transcribed Spacer*) ont été validées comme marqueurs taxonomiques chez les champignons, permettant de mieux discriminer leurs espèces que l'ADNr 18S [Schoch *et al.* 2012]. Les régions ITS1 et ITS2 sont des régions interstitielles hypervariables dans l'opéron ribosomique (Figure 5.14). Plusieurs couples d'amorces existent dans la littérature pour capter ces régions, notamment le couple d'amorces ITS1-ITS2 (encadrant la région ITS1) et le couple ITS3-ITS4 (encadrant la région ITS2) [White *et al.* 1990] Ces amorces sont communément utilisées dans les études de microbiotes fongiques environnementaux.

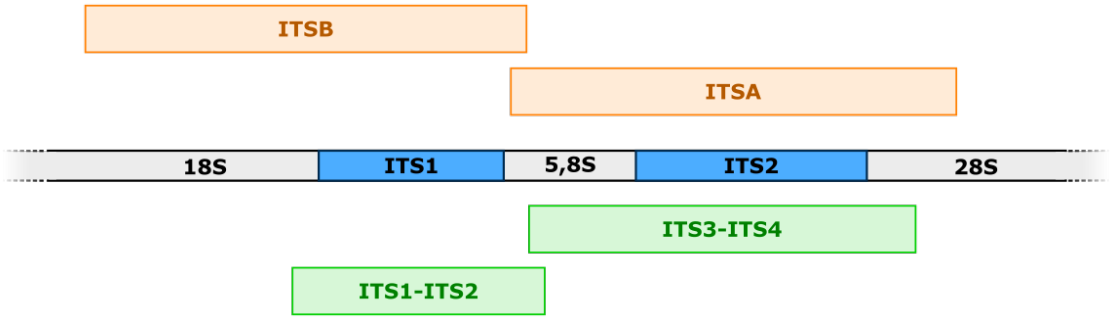


Figure 5.14 : Organisation de l'opéron ribosomique comprenant les régions hypervariables ITS1 et ITS2 (en bleu).

La plateforme PEGASE-biosciences travaille en collaboration avec l'équipe *Lung Infection and Innate Immunity* (LI3) de l'Institut Pasteur de Lille, afin d'étudier les microbiotes pulmonaires bactérien et fongique chez des souris modèles de maladies pulmonaires comme la Broncho-Pneumopathie Chronique Obstructive (BPCO). De précédentes études de métagénétique fongique menées sur la plate-forme ont révélé la non-applicabilité des amorces ITS1-ITS2 et ITS3-ITS4 dans ce contexte, puisque ces couples d'amorces n'amplifient pas certaines espèces d'intérêt clinique qui étaient détectées par qPCR dans les échantillons tel que, par exemple, *Aspergillus fumigatus* (responsable de plus de 80 % des aspergilloses humaines). Il était ainsi nécessaire de trouver de nouvelles amorces afin de pouvoir mener une étude métagénétique fongique sur ce type d'échantillons : ces amorces devaient capter une majorité de champignons d'intérêt clinique, tout en encadrant un amplicon suffisamment discriminant pour les distinguer lors d'une analyse métagénétique. Cet amplicon devait en outre être de taille compatible à un séquençage *paired-end* 2x250 nt Illumina MiSeq.

Une première idée a été d'utiliser la banque de séquences ISHAM ITS database. Cette banque recense plus de 3 600 séquences d'amplicons couvrant ITS1 et ITS2, représentant 535 espèces fongiques pathogènes pour les animaux et l'homme [Irinyi *et al.* 2015]. Une liste de 29 espèces de champignons d'intérêt dans un contexte clinique pour les maladies étudiées par l'équipe LI3 a été produite (Annexe 4). L'ensemble des séquences de ces organismes a été

téléchargé depuis la banque ISHAM ITS (version du 5 octobre 2016), et a été aligné par Clustal Omega. Ces séquences sont toutefois issues de différentes études, ayant utilisé différents couples d'amorces pour les générer. Ainsi, elles couvrent différentes zones de l'opéron ribosomique, leur alignement ne permet donc pas d'identifier des régions conservées pour l'ensemble des organismes d'intérêt.

Une autre approche a été envisagée, permettant de récupérer une région plus large que celles présentes dans la banque. Pour chaque espèce d'intérêt, une séquence consensus a été générée à partir d'un alignement Clustal Omega des amplicons téléchargés pour cette espèce. Cette séquence a ensuite été alignée sur les banques *refseq_genomic* et *whole-genome shotgun contigs* par BLASTn. Si une séquence cible a été identifiée dans ces banques avec un alignement de plus de 90 % avec la séquence consensus, et si cette séquence comporte au moins 500 nucléotides en amont et en aval de cette séquence consensus, alors la région dite étendue de cette séquence a été téléchargée (la région étendue est définie comme étant la séquence cible entre la position initiale de l'alignement – 500 et la position finale de l'alignement + 500). 14 séquences (marquées dans la colonne « Région génomique étendue » de l'Annexe 4) ont ainsi pu être récupérées sur les 29 espèces initialement d'intérêt. Ces séquences ont été alignées par Clustal Omega et utilisées en entrée de Harpon, avec les paramètres suivants :

- Taille minimale d'amorce k : 15 nucléotides ;
- Taille maximale d'amorce k_{max} : 25 nucléotides ;
- Top percent tp : 10 % ;
- Seuil de conservation C : 10 % ;
- Nombre maximum de dégénérescences d_{max} : 8 ;
- Différence maximale de Tm moyenne entre deux amorces d'un même couple *maxTmDiff* : 5° ;
- Taille minimale d'amplicon A_{min} : 300 ;
- Taille maximale d'amplicon A_{max} : 500.

Nom du couple d'amorces	Régions hypervariables encadrées	Taille moyenne	Écart type de taille	Nombre de séquences captées par le couple	Distance topologique	Entropie
ITSA	ITS2	434,77	29,02	13 (100,00 %)	0,091	390,88
ITSB	ITS1	469,62	21,09	13 (100,00 %)	0,000	377,40
ITS1-ITS2	ITS1	365,77	29,02	6 (46,15 %)	0,091	379,92
ITS3-ITS4	ITS2	252,46	20,86	11 (84,62 %)	0,091	316,86

Tableau 5.15 : Couples d'amorces et leurs propriétés sur l'alignement des régions génomiques encadrant les régions ITS1 et ITS2 de 13 espèces fongiques d'intérêt.

Nom	Séquence	Taille	Tm moyen	Taille de l'homopolymère le plus long	Nombre de dégénérescences (= oligomères générés)
ITSAF	TTBCCKCTTCACTCGCCG	18	57,64	3	6
ITSAR	AACAAYGGATCTCTTGGYTCY	21	54,72	2	8
ITSBF	RGARCCAAGAGATCCRTTGTT	21	54,71	2	8
ITSBR	CGTGCTGGGGATWGWSCATT	20	57,61	4	8
ITS1	TCCGTAGGTGAACCTGCGG	19	52,08	2	1
ITS2	GCTGCGTTCTTCATCGATGC	20	56,87	2	1
ITS3	GCATCGATGAAGAACGCAGC	20	56,87	2	1
ITS4	TCCTCCGCTTATTGATATGC	20	58,87	2	1

Tableau 5.16 : Couples d'amorces et leurs propriétés, générées à partir de l'alignement des régions génomiques encadrant les régions ITS1 et ITS2 de 13 espèces fongiques d'intérêt.

Harpon identifie 7 143 amorces potentielles d'intérêt, pouvant former 175 555 couples différents, sur 22 paires de régions conservées. Deux meilleurs couples d'amorces ont été sélectionnés, ITSA et ITSB, pour encadrer les régions hypervariables ITS2 et ITS1 respectivement (Figure 5.14, Tableaux 5.15 et 5.16).

Les couples d'amorces ITSA et ITSB captent la totalité des 13 séquences d'intérêt, et forment des amplicons plus grands que les couples d'amorces ITS1-ITS2 et ITS3-ITS4. Les quatre couples d'amorces ont ensuite été testés *in vitro* par PCR sur 3 espèces fongiques (*Absidia corymbifera*, *Candida albicans* et *Rhizopus microsporus*) ainsi que sur 2 contaminants potentiels (*Homo sapiens* et *Mus musculus*), et un témoin négatif NTC (*no template control*).

Les quatre couples d'amorces ont fait l'objet d'autant de réactions PCR sur les cinq matrices considérées et sur le témoin négatif, en utilisant le kit HotStart Taq Plus (Quiagen). Chaque mélange réactionnel d'un volume total de 50 µL était composé de:

- Tampon 10X 5,00 µL
- $MgCl_2$ (25 mM) 1,00 µL
- dNTP (25 mM) 0,50 µL
- Polymérase, HotStart Taq (5 U/µL) 0,50 µL
- amorce sens (10 µM) 5,00 µL
- amorce anti-sens (10 µM) 5,00 µL
- ADN matriciel (50 ng/µL) 2,00 µL
- eau 31,00 µL

Le programme de la réaction PCR a débuté par une activation de l'enzyme durant 5 min à 95 °C, suivie de 38 cycles de 15 sec de dénaturation à 94 °C, 20 sec d'hybridation des amorces à 50 °C (*ramping* bridé à 50 % équivalent à 3 °C/sec), et 45 sec d'élongation à 72 °C. Une ultime étape de terminaison a conclu la réaction pendant 1 min à 72 °C. Les produits PCR résultants ont ensuite été évalués par électrophorèse sur gel d'agarose 1,5 %, avec une migration à 120 V/60 mA, dont les résultats sont présentés dans la Figure 5.17.

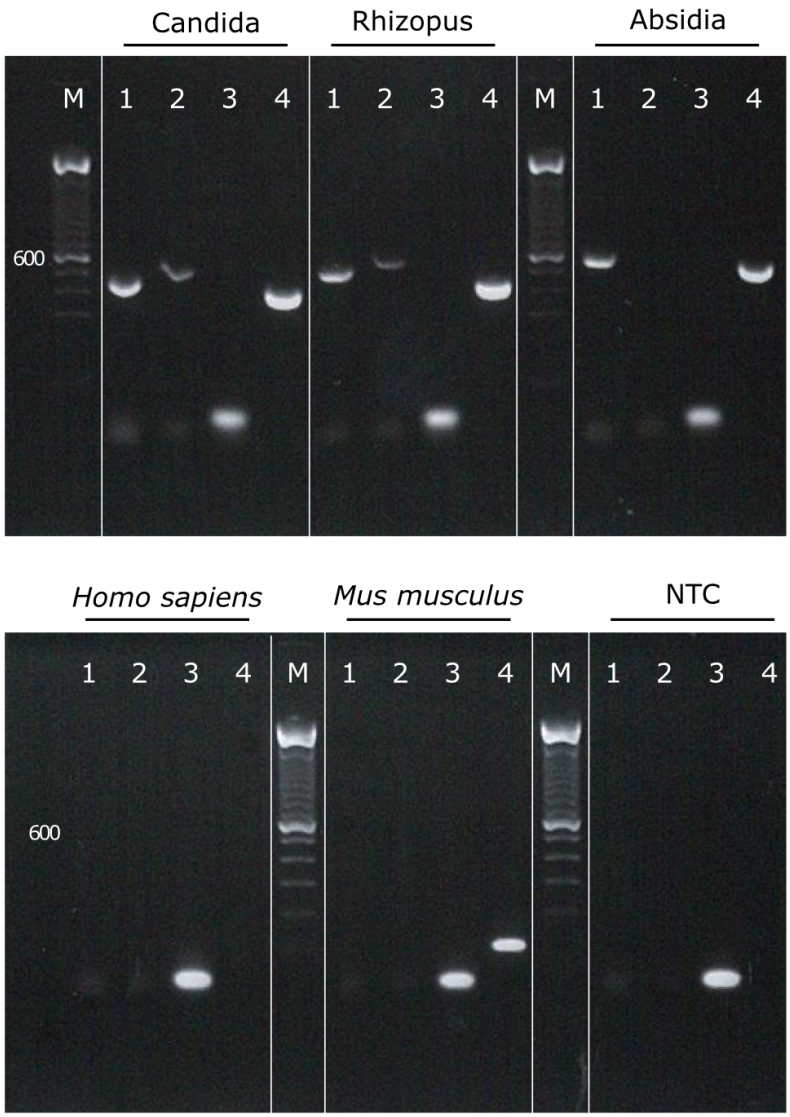


Figure 5.17 : Electrophoregrammes obtenus après PCR des quatre couples d'amorces (1=ITSA, 2=ITSB, 3=ITS1-ITS2, 4=ITS3-ITS4) sur Candida albicans, Rhizopus microsporus, Absidia corymbifera, Homo sapiens, Mus musculus et sans ADN matrice (NTP = no template control). M = marqueur de taille moléculaire (100 nt).

Ces gels révèlent que seuls les couples ITSA et ITS3-ITS4 amplifient les 3 espèces de champignons testées, dans une taille d'amplicon attendue (entre 400 et 500 nt). Le couple d'amorces ITSB n'amplifie pas *Absidia corymbifera* (qui n'avait pas été inclus dans la liste de séquences d'intérêt), tandis que le couple d'amorces ITS1-ITS2 n'amplifie aucune de ces 3 espèces, et génère en outre une forte proportion de dimères d'amorces. Le deuxième gel permet de constater que le couple ITS3-ITS4 amplifie de l'ADN contaminant de souris (produit PCR ~ 150 nt), ce qui cause problème dans un contexte d'étude se basant sur un modèle murin. Le couple d'amorces ITSA est ainsi le seul parmi les 4 testés qui répond à notre problématique initiale.

Les quatre couples d'amorces (ITSA, ITSB, ITS1-ITS2 et ITS3-ITS4) ont été utilisés sur plusieurs échantillons réplicats dans une étude pilote réelle actuellement en cours, pour étudier le microbiote fongique pulmonaire de souris par séquençage *paired-end* Illumina MiSeq. L'analyse de ces données permettra de comparer l'image obtenue des microbiotes de ces échantillons selon chaque couple.

5.4 - Conclusion et perspectives

Harpon est un logiciel de design d'amorces sur la base d'un alignement de séquences d'intérêt, dont l'innovation réside dans la sélection rapide (< 1 min pour chaque analyse décrite précédemment) de couples d'amorces compatibles avec une taille d'amplicon souhaitée, captant un maximum des séquences d'intérêt dans la limite d'un seuil de dégénérescences fixé, et dont les amplicons générés sont les plus variables possibles entre ces séquences. Ce logiciel est actuellement le seul existant cumulant tous ces critères et a été validé en comparaison avec des couples d'amorces issus de la littérature. Il a en outre été utilisé pour générer des couples d'amorces spécifiques à des champignons d'intérêt clinique. Ces amorces ont été intégrées à une étude pilote en cours afin d'être évaluées en conditions expérimentales réelles.

Ce logiciel se base sur un alignement multiple de séquences, comme un biologiste l'utiliserait s'il devait designer ses propres amorces. La qualité des résultats rendus dépend ainsi directement de la qualité de l'alignement initial, qui est de la responsabilité de l'utilisateur. En effet, une automatisation de l'alignement des séquences en entrée ne permettrait pas d'en garantir la qualité. Afin de s'affranchir de cette étape, une méthode de design d'amorces sans alignement préalable pourrait être envisagée, par l'identification de *k-mers* communs à toutes les séquences comme cela a été intégré dans le logiciel PriMux [Hysom *et al.* 2012]. Toutefois, PriMux ne répond pas au problème MC-DPD, puisqu'il designe un ensemble d'amorces captant un maximum des séquences d'intérêt. Ainsi, les dégénérescences permises par PriMux sont choisies par abondance majoritaire de *k-mers*, ce qui rejoint le problème de Hyden décrit Chapitre 5 Section 5.1.3 en choisissant la seule amorce de meilleure solution. Un développement possible de Harpon pour s'affranchir du besoin d'alignement multiple des séquences pourrait être de s'appuyer sur de récentes évolutions publiées sur la recherche de graines avec erreurs [Vroland *et al.* 2016] : dans un ensemble de séquences non-alignées, les *k-mers* majoritaires pourraient être envisagés comme des graines représentant des régions conservées d'intérêt, et les positions dégénérées ajoutées aux amorces pourraient être des positions d'erreurs dans ces graines. Ainsi, l'étape de détection de régions conservées dans Harpon pourrait être modifiée en s'appuyant sur ces développements récents pour être appliquée à un ensemble de séquences d'intérêt sans nécessité d’un alignement préalable.

Harpon n'a pas pour vocation de trouver les amorces les plus universelles possibles en général, mais de trouver celles qui ne manqueront pas d'amplifier un sous-ensemble d'organismes d'intérêt. Toutefois, un module complémentaire est actuellement en cours de développement afin de pouvoir tester les proportions de séquences captées par les amorces sur une banque de référence comme cela a été fait manuellement dans les sections précédentes. En outre, Harpon pourrait également être modifié pour ajouter la possibilité

d'intégrer des inosines (bases pouvant s'apparier à toute autre base, permettant d'éviter l'emploi d'une dégénérescence N) à certaines positions des amorces afin de diminuer le nombre de dégénérescences. Un autre module complémentaire est également prévu afin de vérifier l'amplification potentielle de génomes contaminants par les couples d'amorces d'intérêt ; les couples d'amorces générés pourraient être filtrés selon le risque d'amplifier des génomes contaminants.

Les pistes d'évolution de Harpon sont multiples : on peut par exemple envisager, à partir de la sélection d'un couples d'amorces d'intérêt, de générer un jeu de données métagénétique simulé selon un modèle de technologie de séquençage souhaité. Ce jeu de données pourrait servir d'échantillon artificiel à ajouter lors d'une analyse métagénétique, permettant sa validation, comme les jeux de données simulés que nous avons utilisés dans le Chapitre 3. Harpon peut également être adapté pour être utilisé dans des applications différentes des études métagénétiques. Par exemple, il pourrait être employé dans un protocole diagnostique de détection et/ou de typage bactérien/viral, en créant des amorces sur des régions conservées encadrant les régions discriminantes entre différentes souches d'intérêt.

Chapitre 6 - Recommandations d'analyse de données métagénétiques issues d'un séquençage de librairies bidirectionnelles Ion Torrent PGM

Suite aux travaux précédents, une nouvelle solution d'analyse de données métagénétiques a été intégrée sur la plate-forme PEGASE-biosciences en remplacement du pipeline PEGASE v2. Ce pipeline utilise des logiciels de source ouverte, et permet de traiter des données issues d'un séquençage multiplexé Ion Torrent PGM bidirectionnel. Il a été intégré sous la forme de plusieurs scripts Perl présentés dans la Figure 6.1, et disponibles sur la page Internet dédiée au projet de thèse :

http://www.pegase-biosciences.com/2013-0920

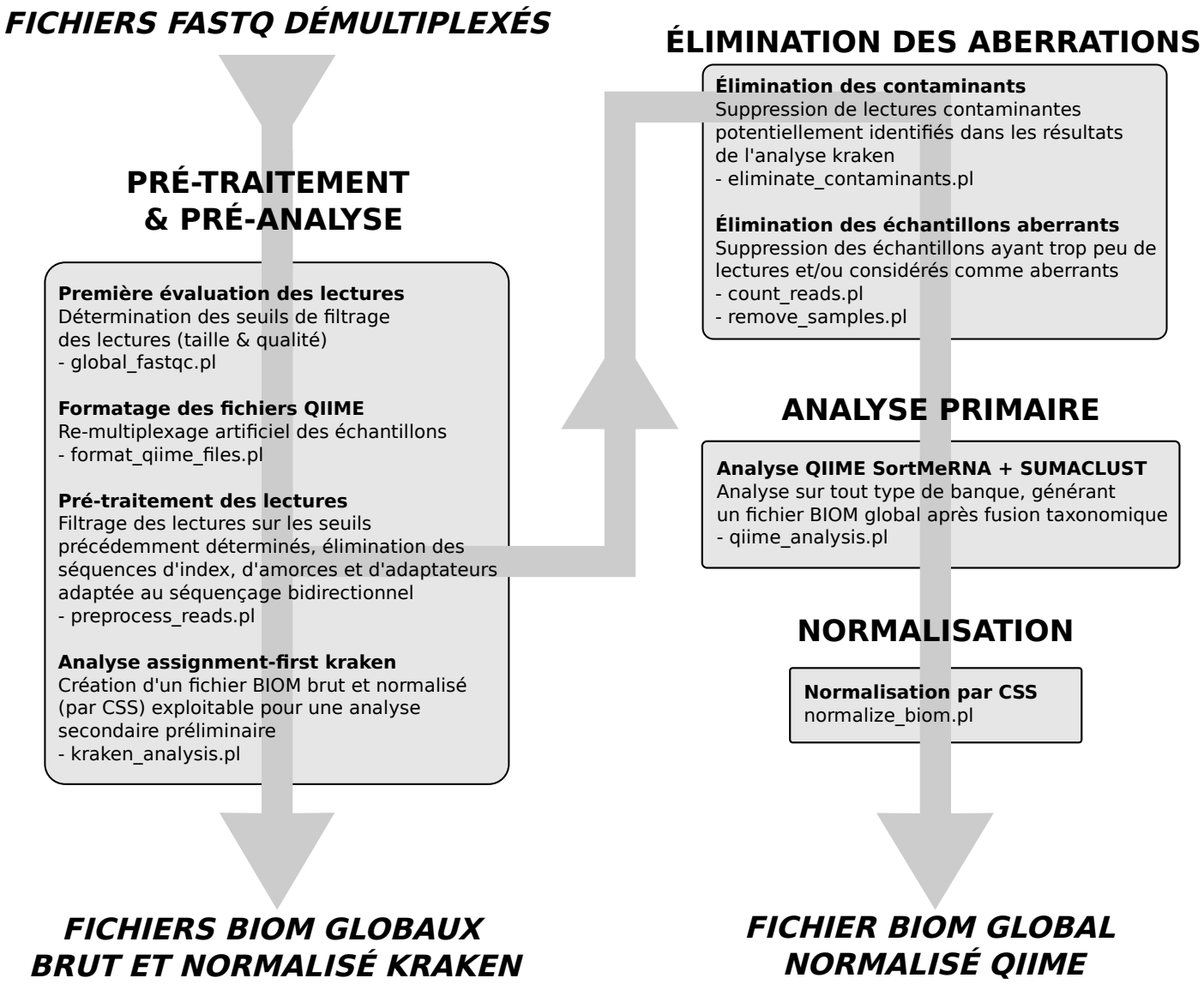


Figure 6.1 : Récapitulatif du pipeline d'analyse métagénétique PEGASE actuellement utilisé pour analyser des données métagénétiques Ion Torrent PGM (librairies bidirectionnelles).

6.1 - Première évaluation des lectures

Script :

```
global_fastqc.pl --dir dossier --min_proportion 0.3
--max_proportion 0.1
```

Arborescence générée :

```
- fastqc
   - Images
      sequence_length_distribution.png
      per_base_sequence_quality.png
-logs
      log_global_fastqc.txt
```

Une première évaluation de taille et de qualité des lectures peut être effectuée en exécutant le script *global_fastqc.pl*, ayant pour argument le dossier contenant les fichiers FASTQ des différents échantillons. Ce script se base sur FastQC afin de générer des graphiques représentatifs de l'ensemble de ces données, qui permettront de fixer des seuils qui seront utilisés dans le pré-traitement des lectures. Ce script génère le dossier *fastqc*, contenant un dossier *Images* qui contient l'ensemble de ces graphiques. Ce script génère également un fichier journal *logs/log_global_fastqc.txt*

Le graphique *sequence_length_distribution.png* représente la distribution de taille de toutes les lectures, dont le pic correspond normalement à la taille d'amplicon attendu (comme dans la Figure 6.2, entre 360 et 440 nt). Néanmoins, la présence d'un deuxième pic d'une amplitude de taille plus faible ou plus grande peut correspondre à une contamination. Avec la technologie Ion Torrent PGM, de nombreuses lectures sont plus courtes que le pic d'intérêt, et correspondent à des lectures fragmentées. À l'inverse, des lectures plus longues peuvent correspondre à des artefacts de séquençage.

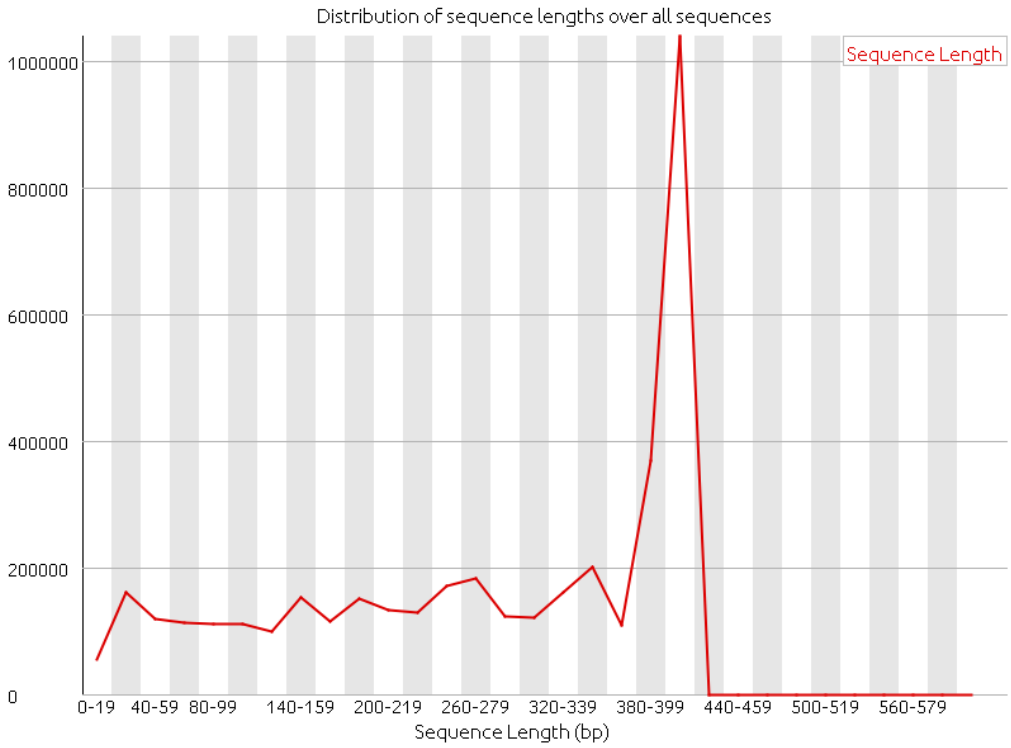


Figure 6.2 : Exemple de graphique sequence_length_distribution.png, représentant la distribution de la longueur des lectures sur l'ensemble des lectures issues du séquençage.

Une approche drastique de filtrage des lectures serait de définir un intervalle de taille de lecture correspondant uniquement au pic de taille d'intérêt. Nous déconseillons toutefois cette approche : en effet, elle peut éliminer une proportion très importante de lectures qui, même si elles sont fragmentées, contiennent de l'information biologique pouvant être exploitée par des approches *assignment-first* par exemple.

En effet, dans l'approche d'analyse *clustering-first open-reference* intégrée par la suite, les lectures conservées de taille inférieure à celle attendue formeront soit des OTUs à assignation taxonomique très peu détaillée (n'étant pas assez longues), soit à des OTUs *de novo* impossibles à classifier. Dans les deux cas, la génération de ces OTUs augmentera l'estimation de richesse et de diversité des microbiotes étudiés par échantillon. Toutefois, nous avons indiqué

dans le Chapitre 4 Sections 4.3.2 et 4.3.3 que ces métriques ne prennent sens que si elles sont interprétées de façon comparative entre groupes d'échantillons, et non de façon absolue. La génération de lectures étant considérée comme étant homogène pour tous les index sur un run de séquençage Ion Torrent [Singh *et al.* 2013], cette surestimation de richesse/diversité sera similaire pour tous les échantillons considérés (hors échantillons aberrants qui seront éliminés dans les étapes de pré-traitement) et donc n'impactera pas une étude comparative robuste.

Pour sélectionner les lectures d'intérêt, nous préférons ainsi utiliser un intervalle de taille plus large, permettant d'éliminer les artefacts les plus flagrants, et filtrer les lectures conservées par une approche plus fine ultérieure (qualité, présence d'homopolymères, détection de lectures contaminantes, …) décrite par la suite. Cet intervalle de taille est choisi de façon arbitraire comme incluant les lectures de taille minimale t_{min} correspondant par défaut à 30 % du pic de taille d'intérêt, et de taille maximale t_{max} correspondant par défaut au pic de taille +10 %. Par exemple dans la Figure 6.2, pour un pic à 410 nt, t_{min} serait estimé à 123 nt, et t_{max} à 451 nt. Ces seuils sont ajustables par les paramètres *--min_proportion* et *--max_proportion*. L'intervalle de taille correspondant est automatiquement calculé et proposé dans le fichier log/*log_global_fastqc.txt*. Sa valeur doit être validée au regard du graphique *sequence_length_distribution.png*.

Le graphique *per_base_sequence_quality.png* présente la qualité de chaque nucléotide sur l'ensemble des lectures. Il permet d'évaluer une éventuelle chute de qualité, typiquement observée en fin de lecture pour les lectures Ion Torrent PGM (par exemple à partir de la position 410 dans la Figure 6.3). On peut ainsi fixer un seuil de qualité minimale *minqualtrim* (par exemple Q20), qui sera utilisé dans le pré-traitement des lectures pour rogner les séquences.

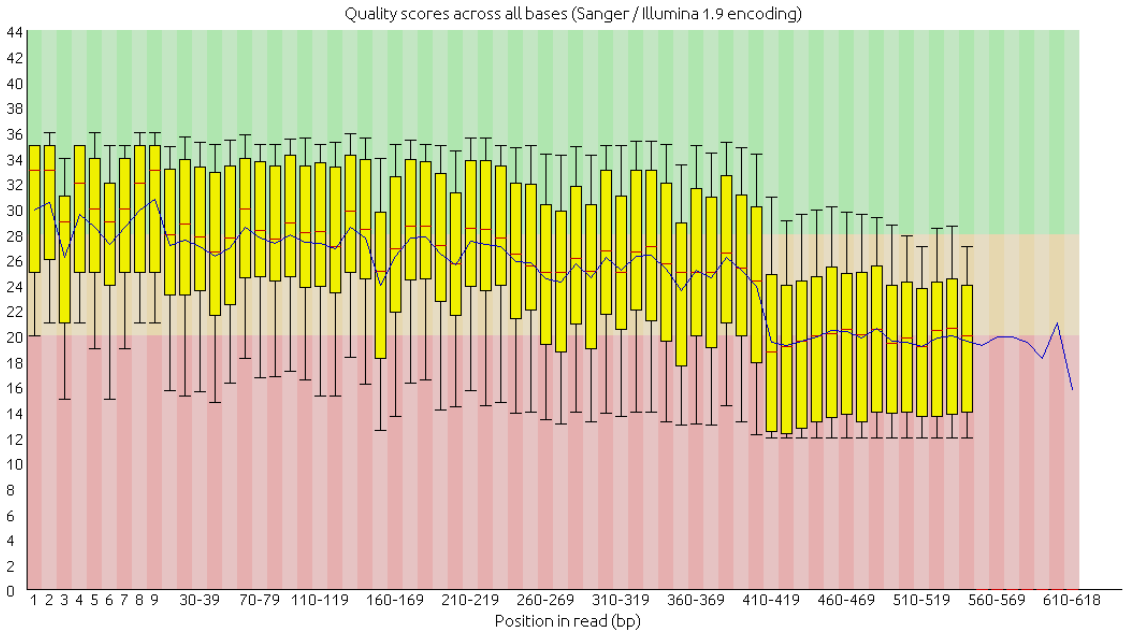


Figure 6.3 : Exemple de graphique per_base_sequence_quality.png, représentant des boîtes à moustaches du score qualité à chaque nucléotide des lectures. La ligne rouge et la ligne bleue représentent les valeurs de qualité médiane et moyenne respectivement. La boîte jaune représente l'écart interquartile, et les moustaches inférieure et supérieure représentent les valeurs adjacentes à 10 % et 90 % respectivement.

À l'issue de cette étape d'évaluation, on a ainsi potentiellement défini un seuil de taille de lecture minimale t_{min}, un seuil de taille de lecture maximale t_{max}, et/ou un seuil de qualité minimale *minqualtrim*. Ces seuils seront utilisés dans l'étape suivante de pré-traitement des lectures.

6.2 - Pré-traitement des lectures par QIIME

QIIME SortMeRNA + Sumaclust a été choisi pour effectuer l'analyse primaire des données, car il s'agissait du pipeline évalué au Chapitre 3 montrant les meilleures performances sur des données de séquençage avec erreurs, tout en permettant la prise en compte d'OTUs non classifiés dans l'estimation de richesse et de diversité. QIIME nécessite toutefois en entrée un seul fichier FASTA, comprenant l'ensemble des lectures indexées de tous les échantillons. Le script *format_qiime_files.pl* permet de générer ce fichier FASTA en ajoutant

artificiellement une séquence d'index à chaque lecture, sur la base des index utilisés lors du séquençage. Il prend en argument le dossier contenant les fichiers FASTQ à analyser, et ajoute à chaque lecture sa séquence d'index.

Script :

```
format_qiime_files.pl --dir dossier --primerF primer_sens
--primerR primer_antisens
```

Arborescence générée :

```
- qiime_preprocessing
   all_reads.fasta
   all_reads.qual
   mapping.txt
```

format_qiime_files.pl va ainsi générer le dossier qiime_preprocessing, contenant deux fichiers, *all_reads.fasta* et *all_reads.qual*, et un fichier *mapping.txt* nécessaire à QIIME pour pré-traiter les lectures. Ce fichier associe les séquences d'amorces et d'index aux noms des échantillons.

Script :

```
preprocess_reads.pl
--fasta qiime_preprocessing/all_reads.fasta
--qual qiime_preprocessing/all_reads.qual
--mapping qiime_preprocessing/mapping.txt
(--tmin tmin --tmax tmax --minqualtrim minqualtrim –
max_homopolymer max_homopolymers)
```

Arborescence générée :

```
- qiime_analysis
   seqs.fna
```

Le script *preprocess_reads.pl* va exécuter de façon automatique le scripts QIIME *split_libraries.py*. QIIME va reconnaître la séquence de l'index et de l'amorce sens en 5' pour les éliminer, et va reconnaître la séquence inverse complémentaire de l'amorce anti-sens en 3' des lectures afin d'éliminer tout ce

qui suit le début de cette amorce (l'adaptateur trP1 ayant été potentiellement séquencé si la lecture est plus longue que le locus d'intérêt). *preprocess_reads.pl* prend en entrée les fichiers *all_reads.fasta, all_reads.qual* et *mapping.txt* ainsi que les seuils optionnels t_{min}, t_{max}, et/ou *minqualtrim* définis à l'étape précédente. Une lecture est rognée si la qualité moyenne sur une fenêtre de 50 nucléotides chute sous *minqualtrim*. Les lectures peuvent également être filtrées par QIIME si elles contiennent un homopolymère plus long que *maxhomopolymer* (6 par défaut). Le script va générer en sortie un dossier *qiime_analysis*, contenant le fichier *seqs.fna* qui contiendra l'ensemble des lectures après filtrage, dont l'en-tête est modifiée afin de contenir le nom de l'échantillon dont chaque lecture est issue.

Nous n'avons pas inclus dans le pipeline d'étape de débruitage ni de correction d'erreurs (comme décrit dans le Chapitre 1 Section 1.3.5). En effet, aucune méthode existante à ce jour n'est adaptée à des données Ion Torrent PGM, dont le format de fichier brut est spécifique à cette technologie.

6.3 - Pré-analyse *assignment-first*

Une première analyse rapide globale des données est effectuée avec le pipeline *assignment-first* kraken sur la banque de *k-mers* Minikraken, afin d'obtenir une première image des échantillons en un minimum de temps. En effet, nous avons démontré dans le Chapitres 3 Section 3.2.7 qu'une analyse kraken sur cette banque de référence réduite est quasiment instantanée avec des ressources minimales, et permet d'observer les principaux éléments constitutifs d'un microbiote d'intérêt.

Script :

```
kraken_analysis.pl --fasta qiime_analysis/seqs.fna
(--db directory)
```

Arborescence générée :

```
- kraken_analysis
   results.txt
   kraken.biom
   kraken_normalized.biom
```

Le script *kraken_analysis.pl* permet d'exécuter cette analyse en prenant en argument le fichier *qiime_analysis/seqs.fna* précédemment généré. Ce script va générer en résultat dans le dossier *kraken_analysis* un fichier *results.txt* contenant l'assignation par kraken de toutes les lectures du fichier *seqs.fna*. Le script va ensuite générer un fichier BIOM *kraken.biom*, en séparant les lectures dans leur échantillon respectif sur la base de leur en-tête FASTA. Le fichier *kraken.biom* permet d'observer les proportions brutes des lectures entre chaque échantillon afin d'estimer les variations de quantités de lectures d'un échantillon à un autre. Un fichier *kraken_normalized.biom* normalisé par CSS (voir Chapitre 1 Section 1.5.2) est également généré, afin de pouvoir être interprété manuellement par des scripts R ou le logiciel STAMP pour une analyse secondaire préliminaire.

Cette première analyse peut par exemple permettre d'estimer si les données contiennent des séquences contaminantes, ou d'observer des phénomènes globaux comme une variation de diversité entre deux groupes d'intérêt. Cette approche permet également d'identifier des échantillons potentiellement aberrants, qui pourront être éliminés avant l'analyse *clustering-first*. Néanmoins, ces résultats ne tiennent pas compte d'organismes potentiellement non identifiés, et l'utilisation d'une banque de référence réduite donne une image grossière des résultats (comme vu dans le Chapitre 3 Section 3.2.5). Ainsi, cette analyse kraken est à mener à titre exploratoire et pour

obtenir un aperçu assez global des données, avant d'exécuter une analyse plus complète.

6.4 - Élimination des lectures contaminantes (optionnel)

Script :

```
eliminate_contaminants.pl --fasta qiime_analysis/seqs.fna
--kraken_results kraken_analysis/results.txt
--contaminants contaminants.txt
```

Arborescence générée :

```
- qiime_analysis
   seqs_no_contaminants.fna
- logs
   log_eliminate_contaminants.tsv
```

Si, lors de l'analyse précédente, des taxons contaminants ont pu être observés (par exemple des lectures issues du génome de l'hôte, de génomes chloroplastiques ou mitochondriaux), il est préférable d'éliminer les lectures concernées avant analyse. Pour ce faire, le script *eliminate_contaminants.pl* prend en entrée le fichier FASTA *qiime_analysis/seqs.fna* précédemment généré, le fichier de résultats bruts *kraken* (*kraken_analysis/results.txt*) ainsi qu'un fichier texte généré manuellement (*contaminants.txt*), contenant une annotation taxonomique par ligne, correspondant aux taxons à éliminer (cette annotation taxonomique doit être identique à l'annotation taxonomique du fichier *kraken.biom)*. Ce script identifie dans les résultats de kraken les lectures assignées aux taxons contaminants, et génère un fichier *qiime_analysis/seqs_no_contaminants.fna*, correspondant au fichier *seqs.fna* sans lectures contaminantes. Il génère également le fichier *logs/log_eliminate_contaminants.tsv* indiquant le nombre de lectures éliminées par échantillon.

Le modèle de génération de lectures chimériques par l'Ion Torrent PGM n'est à ce jour pas connu, aussi nous n'avons pas inclus au pipeline d'étape de détection de telles lectures. Toutefois, on pourrait envisager d'interpréter les résultats de kraken pour les détecter : en effet, une lecture chimérique est composée de *k-mers* appartenant à deux taxons différents. Cette signature pourrait être détectée dans les résultats de kraken, et donc utilisée pour filtrer les lectures. Ce module de détection de chimères serait ainsi un axe de développement possible pour ce pipeline.

6.5 - Élimination des échantillons aberrants (optionnel)

Script :

```
count_reads.pl
--fasta qiime_analysis/seqs(_no_contaminants).fna
```

Arborescence générée :

```
- logs
   log_count_reads.tsv
```

Certains échantillons doivent être éliminés de l'analyse, soit car ils ont été identifiés comme aberrants dans l'étape précédente, soit car ils contiennent trop peu de lectures. Le script *count_reads.pl* permet de compter le nombre de lectures par échantillons. Il prend en entrée le fichier *qiime_analysis/seqs.fna* (ou *qiime_analysis/seqs_filtered.fna* si l'étape de filtrage des lectures contaminantes a été réalisée). Il crée en sortie le fichier *logs/log_count_reads.tsv* composé de deux colonnes : le nom de l'échantillon, et le nombre de lectures dans cet échantillon.

Script :

```
remove_samples.pl
--fasta qiime_analysis/seqs(_no_contaminants).fna
--samples bad_samples.txt
```

Arborescence générée :

```
- qiime_analysis
   seqs_no_bad_samples.fna
```

Si certains échantillons sont à supprimer, le fichier *bad_samples.txt* doit être créé manuellement, contenant un nom d'échantillon à supprimer par ligne. Ensuite, le script *remove_samples.pl*, basé sur le script QIIME *filter_fasta.py*, doit être exécuté. Il prend en argument ce fichier *bad_samples.txt* et le fichier *qiime_analysis/seqs.fna* (ou *qiime_analysis/seqs_filtered.fna* si l'étape de filtrage des lectures contaminantes a été réalisée) d'où les lectures concernées seront éliminées. Ce script va générer en sortie le fichier *qiime_analysis/seqs_no_bad_samples.fna*

6.6 - Exécution de l'analyse QIIME

L'analyse QIIME SortMeRNA + Sumaclust est exécutée par le script *qiime_analysis.pl* à partir d'un fichier FASTA qui peut être *seqs.fna*, *seqs_no_contaminants.fna* ou *seqs_no_bad_samples.fna* selon les étapes précédentes effectuées. Cette analyse est effectuée par défaut sur la banque d'ADNr 16S Greengenes 13.8. Si une autre banque doit être utilisée (pour une analyse de métagénétique fongique par exemple), elle doit être indiquée au script *qiime_analysis.pl* par les paramètres --*reference_seqs* (fichier FASTA de la banque) et – reference_taxonomy (fichier de taxonomie de la banque).

Script :

```
qiime_analysis.pl
--fasta qiime_analysis/seqs(_no_contaminants ||
_no_bad_samples).fna
(--reference_seqs banque.fasta
--reference_taxonomy taxonomie.txt --min_otu_size 2)
```

Arborescence générée :

```
- qiime_analysis
   -  results
      merged_results.biom
```

Ce script exécute le script *pick_open_reference_otus.py* de QIIME, en utilisant les algorithmes SortMeRNA+Sumaclust, en autorisant un alignement des lectures de façon inverse complémentaire (nécessaire puisqu'on traite des lectures issues du séquençage de librairies bidirectionnelles), et en éliminant les OTUs singletons (ce paramètre peut être modifié en passant au script *qiime_analysis.pl* le paramètre *–min_otu_size,* avec la valeur de taille minimale d'OTU – par défaut 2).

Le script *qiime_analysis.pl* convertit également le fichier BIOM de résultat en fusionnant les OTUs ayant la même assignation taxonomique, ce qui génère en sortie le fichier BIOM global *qiime_analysis/merged_results.biom*

6.7 - Normalisation des données

Script :

```
normalize_biom.pl
--biom qiime_analysis/merged_results.biom
```

Arborescence générée :

```
- qiime_analysis
   - results
      merged_results_normalized.biom
```

Enfin, le script *normalize_biom.pl* prend en argument le fichier BIOM *qiime_analysis/merged_results.biom* et y appplique une normalisation CSS sur la base du script *normalize_table.py* de QIIME, générant le fichier final *qiime_analysis/merged_results_normalized.biom*.

6.8 - Conclusion

Ce pipeline d'analyse est à notre connaissance le premier pipeline d'analyse métagénétique utilisant une méthode *assignment-first* afin de renforcer les résultats issus d'une approche *clustering-first*. La pré-analyse *assignment-first* est d'intérêt multiple : elle permet tout d'abord d'obtenir une image générale très rapide des échantillons étudiés. Elle permet également la détection de lectures potentiellement contaminantes qui peuvent ainsi être éliminées. Elle peut aussi être à la base d'une nouvelle approche de détection de séquences chimériques. Enfin, le fichier BIOM normalisé généré par l'analyse kraken peut être utilisé afin de conforter des observations issues de l'analyse secondaire des résultats QIIME.

Ce pipeline est actuellement en cours de validation et d'intégration à un workflow Galaxy afin d'être utilisé de façon routinière sur la plate-forme. Un second pipeline est également en cours de développement selon la même approche afin de traiter des données *paired-end* Illumina MiSeq.

Conclusions & perspectives

« Nous nous noyons dans l'information mais sommes assoiffés de connaissances » [Naisbitt 1982]. John Naisbitt soulevait déjà en 1982 ce qui peut être considéré comme le défi majeur de la bioinformatique à l'ère du séquençage haut-débit : nous générons actuellement bien plus de données que ce que nous sommes capables d'interpréter. Ce projet de thèse s'inscrit dans la nécessité de renforcer le rôle de la bioinformatique comme interface de dialogue indispensable entre plusieurs disciplines, permettant une meilleure appréhension d'un tel volume de données afin d'en extraire des informations biologiques de façon robuste, pertinente et reproductible. Ce continuum interdisciplinaire est souvent revendiqué de manière incantatoire, mais il est encore trop rarement appliqué en réalité ; le transfert de connaissances et d'expertises de la bioinformatique vers la biologie, et réciproquement, reste actuellement toujours un point critique. Un exemple marquant de ce cloisonnement est BLAST [Altschul *et al.* 1990], qui est depuis plus de 25 ans l'outil de référence des biologistes pour comparer une séquence à une banque de référence, alors qu'il existe de nombreuses autres solutions bioinformatiques bien plus adaptées à différentes problématiques. Le statut même du bioinformaticien n'est pas toujours clairement caractérisé ; certains revendiquent la nécessité de le définir par son expertise des outils manipulés [Vincent & Charette 2015], d'autres souhaitant élargir cette définition afin d'encourager le décloisonnement des disciplines [Smith 2015].

Dans le contexte de la métagénétique, cette thèse a permis de renforcer les passerelles entre biologie humide et biologie sèche, en apportant plusieurs solutions concrètes d'amélioration des études que ce soit par accompagnement du montage du plan d'expérience biologique tout comme dans l'appréhension des méthodes d'analyse bioinformatiques. Ce projet a permis d'identifier et d'évaluer des solutions d'analyse hégémoniques tout comme des solutions émergentes, en les catégorisant sous les appellations *clustering-first* et *assignment-first*, et en les évaluant au regard de différents contextes biologiques grâce à l'établissement un protocole d'évaluation formel innovant.

Cette thèse a également révélé l'impact des pipelines d'analyse dans les conclusions d'un projet métagénétique répondant à une question biologique donnée. La surprenante variabilité des résultats d'un pipeline à un autre nous a permis d'émettre des mises en garde sur leur interprétation, afin qu'elle soit menée de façon critique non plus seulement au regard de la problématique biologique d'intérêt, mais également par rapport aux méthodes d'analyses utilisées. Enfin, le développement du logiciel Harpon a permis de fournir une solution bioinformatique innovante au biais majeur constitué par la sélection d'amorces de PCR pour capter le locus choisi chez tous les organismes d'intérêt, étape préalable à toute étude métagénétique. L'ensemble de ces travaux a pu être mis à profit de la plate-forme PEGASE-biosciences sous la forme de recommandations détaillées d'analyse de données métagénétiques.

De façon plus détaillée, une première expertise du pipeline d'analyse PEGASE développé sur la plate-forme PEGASE-biosciences a révélé la nécessité de mettre en place un protocole objectif et formel pour l'évaluation de pipelines d'analyse. Nous avons développé ce protocole, composé de données simulées et réelles accompagné de métriques adaptées, afin de mesurer l'impact de plusieurs variables liées aux plans d'expérience sur les résultats d'analyse, en particulier lorsque les données interprétées contiennent des erreurs de séquençage. Ce protocole a été utilisé pour comparer six pipelines d'analyse que nous avons définis par une nouvelle appellation, trois pipelines *clustering-first* et trois pipelines *assignment-first*. Nous avons observé avec surprise que l'effet des erreurs de séquençage a un impact plus élevé sur les résultats que le choix de différentes régions cibles amplifiées. En outre, en présence d'erreurs, l'augmentation du débit de séquençage engendre une surestimation de richesse pour tous les pipelines évalués, particulièrement pour les microbiotes de complexité élevée. Enfin, le choix de la banque de séquences de référence a de façon attendue un impact majeur sur l'estimation de la richesse pour les pipelines *clustering-first*, et sur la qualité des taxons identifiés pour les pipelines *assignment-first*. Cette étude a permis de valider pour la première fois l'utilisation de pipelines *assignment-first* dans un contexte métagénétique,

démontrant que leurs résultats ont une qualité comparable à des pipelines *clustering-first*, ceci en considérant des lectures comportant des erreurs de séquençage. Ce protocole d'évaluation a été publié, et peut être réutilisé par des bioinformaticiens et bioanalystes pour évaluer leurs solutions analytiques et pour accompagner le développement et paramétrage de nouvelles méthodes d'analyse.

Dans le contexte de cette évaluation, nous avons également mis à disposition de la communauté scientifique un ensemble de jeux de données réelles issus du séquençage et de l'analyse de microbiotes intestinaux humains de patients colonisés ou non par le protiste *Blastocystis*. Ces données nous ont permis de démontrer que l'utilisation de différents pipelines d'analyse peut amener à des conclusions biologiques variables, chaque pipeline interprétant différemment la richesse et la composition des microbiotes étudiés. Cependant, nous avons démontré que les indices de richesse et de diversité sont de bons estimateurs de différences potentielles entre groupes d'échantillons, sous réserve d'un nombre suffisant d'échantillons par groupe comparé, et d'une population homogène dans chaque groupe. La variation de l'image des microbiotes étudiés donnée par les différents pipelines nous a amenés à questionner l'utilisation courante de la métagénétique comme d'une méthode d'identification précise des organismes en présence. En effet, cette variation d'un pipeline à un autre renforce le risque que le bioanalyste soit sélectif sur ses taxons d'intérêt, les interprétant de façon biaisée selon sa problématique biologique sans mettre en perspective les résultats observés selon l'approche analytique utilisée. Nous avons ainsi révélé que l'outil métagénétique n'est robuste que s'il est utilisé comme outil de profilage, et non comme un outil absolu d'identification des micro-organismes en présence en l'absence d'autres méthodes de validation complémentaires.

Le développement de Harpon a permis de fournir une solution innovante de design d'amorces adapté à un contexte métagénétique. Ce logiciel identifie les couples d'amorces dégénérées compatibles entre elles et captant un nombre maximum de séquences d'intérêt à partir de leur alignement ; ces

couples d'amorces sont en outre designés pour générer un amplicon dont la séquence est la plus variable possible, et dont la taille est compatible avec un intervalle de taille choisi (afin de pouvoir correspondre aux prérogatives d'une technologie de séquençage par exemple). Harpon a été validé *in silico*, retrouvant des amorces couramment utilisées dans la littérature pour l'ADNr 16S bactérien et archée, ainsi que pour l'ITS. Ce logiciel a également été utilisé pour designer des amorces adaptées à l'étude d'un microbiote fongique dans un contexte clinique, pour lequel aucun couple d'amorces référencé dans la littérature à ce jour n'était adapté. Ces amorces ont été validées *in vitro*, et sont actuellement utilisées dans un projet pilote pour l'étude du microbiote fongique pulmonaire sur modèle murin, actuellement en cours de séquençage. Nous avons identifié de multiples possibilités de développement de Harpon, que ce soit dans son approche algorithmique tout comme par l'addition de modules complémentaires, qui permettront, à terme, de l'utiliser dans de nombreux contextes dépassant celui de la métagénétique. Harpon peut en effet répondre aux problématiques de design d'amorces à des fins diagnostiques ou de sous-typage (nécessité de designer un couple d'amorces captant un ensemble d'organismes et encadrant des positions variables permettant de les distinguer).

Enfin, l'expertise acquise sur l'ensemble de ces travaux a été mis à disposition de la plate-forme PEGASE-biosciences par la proposition de recommandations d'analyse adaptées au traitement de données métagénétiques issues du séquençage de librairies bidirectionnelles Ion Torrent PGM. Ce nouveau pipeline d'analyse utilise de façon innovante un pipeline *assignment-first* afin d'optimiser l'analyse *clustering-first* associée.

La métagénomique WGS est souvent mise en opposition à la métagénétique, par sa plus grande exhaustivité et son absence de problématiques liées au ciblage d'un locus d'intérêt. Nous considérons néanmoins que les deux approches n'ont pas la même application. La métagénomique WGS est avant tout d'un intérêt fonctionnel, permettant la découverte de voies métaboliques intervenant dans un microbiote d'intérêt. Toutefois, son utilisation à des fins d'étude de diversité et de composition

taxonomique nécessite des débits de séquençage conséquents et le développement de solutions analytiques encore émergentes ; ces deux points critiques verrouillent actuellement le transfert de la métagénomique WGS des laboratoires de recherche vers des applications cliniques et industrielles. Nous avons néanmoins démontré dans ces travaux de thèse que certaines approches bioinformatiques développées dans un contexte WGS peuvent être transférées à un contexte de métagénomique ciblée.

L'approche métagénétique permise par les séquenceurs de paillasse de seconde génération reste à ce jour la solution privilégiée pour l'étude de microbiotes d'intérêt dans des processus routiniers, d'où son application actuelle dans de nombreux contextes cliniques et industriels. Cette démocratisation a également déplacé la maîtrise de l'analyse du bioinformaticien vers le biologiste. Comme l'indiquait déjà Stuart Hurlbert en 1971, *« L'utilisation d'une approche mathématique ne doit pas contraindre un biologiste à être modeste sur sa capacité à faire preuve de discernement dans ses conclusions biologiques »* [Hurlbert 1971]. De ce point de vue, les solutions apportées par cette thèse fournissent aux bioanalystes des méthodes et garde-fous indispensables à une utilisation optimale de l'outil métagénétique, afin d'interpréter de façon robuste les données générées par les technologies actuelles comme les technologies futures.

Glossaire

ADNr 16S/18S : Gène codant pour la petite sous-unité ribosomique des procaryotes (16S) et eucaryotes (18S).

Biologie humide : Manipulations de matrice biologique *in vitro*.

Biologie sèche : Manipulations de données biologiques *in silico*.

Bloom : Développement rapide d'un ou plusieurs organisme(s) colonisant un milieu de façon prépondérante.

BIOM (*Biological Observation Matrix*) : Format de fichier destiné à représenter des tables de contingence entre échantillons biologiques.

Clustering : Partitionnement de données.

Code IUPAC (*International Union of Pure and Applied Chemistry)* : Représentation standardisée des nucléotides en code à une lettre.

Design d'amorce : Détermination de la séquence d'une amorce sur la base d'une ou plusieurs séquences d'intérêt.

Dysbiose : Altération de la composition des communautés microbiennes intestinales.

FASTA : Format de fichier texte contenant une ou plusieurs séquences (ici nucléotidiques). Chaque séquence est représentée sur deux lignes : une ligne en-tête descriptive, et une ligne contenant la séquence.

FASTQ : Format de fichier texte contenant une ou plusieurs séquences (ici nucléotidiques). Chaque séquence est représentée sur quatre lignes : une ligne en-tête descriptive, une ligne contenant la séquence, une ligne séparatrice (contenant un + et parfois l'en-tête descriptive à nouveau) et la séquence de score qualité correspondant au score qualité de chaque nucléotide de la séquence, encodé sur un seul caractère ASCII.

Flowgram : Fichier de sortie de séquenceur contenant le signal brut qui sera converti en séquence par l'étape de *basecalling*.

Gap, aussi appelé brèche : un ou plusieurs espaces successifs dans une même séquence d'un alignement de séquences.

Hit BLAST : Séquence d'une banque de référence sur laquelle s'aligne localement une séquence d'intérêt, selon des paramètres définis lors de l'exécution de l'alignement par le logiciel BLAST.

Index, aussi appelé code-barre : Oligomère artificiel d'une dizaine de nucléotides, unique pour un échantillon d'intérêt. Cet oligomère est présent en 5' d'une lecture de séquençage, permettant de reconnaître de quel échantillon cette lecture est issue sur la base de sa séquence.

ITS (*Internal Transcribed Spacer*) : Région interstitielle non-codante entre deux gènes ribosomiques.

LCA (*Lowest Common Ancestor*) : Méthode d'assignation taxonomique d'une lecture. Lorsque cette dernière peut être assignée à différents taxons, la

méthode LCA l'assigne à l'ancêtre commun le plus proche de ces taxons.

Lecture : Séquence de texte nucléotidique au format FASTQ issue d'un séquenceur.

Locus : Région génomique.

Marqueur taxonomique : Locus variable entre différents taxons, permettant de les discriminer sur la base de sa séquence.

Métadonnée : Information complémentaire à une donnée.

Métagénomique : Méthode d'étude du contenu génomique d'un échantillon par séquençage.

Métagénomique WGS (*Whole Genome Shotgun*) : Métagénomique par séquençage shotgun de l'ensemble des génomes en présence dans l'échantillon d'intérêt.

Métagénétique, aussi appelée métagénomique ciblée : Métagénomique par amplification préliminaire d'un *locus* d'intérêt (servant de marqueur taxonomique) avant séquençage.

MICI : Maladie inflammatoire chronique de l'intestin.

Microbiome : Milieu dans lequel existe un microbiote.

Microbiote : Ensemble des micro-organismes présents dans un milieu donné à un temps donné.

MPSS (*Massive Parallel Signature Sequencing*) : première génération de séquençage haut-débit.

NP-complet : Problème informatique ne pouvant peut être résolu en temps. polynomial qu'en utilisant une méthode non-déterministe.

OTU (*Operational Taxonomic Unit*, unité taxonomique opérationnelle) : Regroupement de lectures sur la base d'une similarité de séquence.

Paired-end : Technique de séquençage permettant de séquencer les deux extrémités d'un même fragment d'ADN, et de conserver l'information que les lectures résultantes sont appariées.

Pipeline : Enchaînement d'étapes d'analyse informatique.

Run : Processus médié par un séquenceur visant à séquencer un ADN matrice.

Shotgun : Séquençage de fragments aléatoires d'ADN.

Singleton : OTU ne contenant qu'une lecture.

Taxon : OTU annoté à un niveau de taxonomie donné.

Taxonomie : Science de la classification (ici, des organismes vivants).

Température de fusion (Tm) : Température à laquelle une molécule double brin d'ADN est à moitié désappariée.

TSV (*tab-separated values*) : Format de fichier dont les informations sont délimitées par des tabulations.

Workflow Galaxy : Pipeline d'analyse créé sur la plate-forme informatique Galaxy, permettant d'automatiser des traitements de fichiers et analyses bioinformatiques par interface graphique.

Bibliographie

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J. Bacteriol.* 186, 2629-35 (2004).

Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44, W3-W10 (2016).

Ahn, J.-H., Kim, B.-Y., Song, J. & Weon, H.-Y. Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *J. Microbiol.* 50, 1071-1074 (2012).

Aitchison, J. Reducing the dimensionality of compositional data sets. *J. Int. Assoc. Math. Geol.* 16, 617-635 (1984).

Allen, B., Kon, M. & Bar-Yam, Y. A New Phylogenetic Diversity Measure Generalizing the Shannon Index and Its Application to Phyllostomid Bats. *Am. Nat.* 174, 236-243 (2009).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410 (1990).

Amann, R. I., Ludwig, W. & Schleifer, K. H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* 59, 143-69 (1995).

Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* 11, R106 (2010).

Andrews, S. FastQC: A Quality Control tool for High Throughput Sequence Data. (2010). Disponible sur : http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Consulté le 1 février 2017).

Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P. & Tyson, G. W. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* 40, e94-e94 (2012).

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J. & Weightman, A. J. At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies. *Appl. Environ. Microbiol.* 71, 7724-7736 (2005).

Audebert, C. *et al.* Colonization with the enteric protozoa Blastocystis is associated with increased diversity of human gut bacterial microbiota. *Sci. Rep.* 6, 25255 (2016).

Baker, G. C., Smith, J. J. & Cowan, D. A. Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* 55, 541-555 (2003).

Balzer, S., Malde, K., Lanzen, A., Sharma, A. & Jonassen, I. Characteristics of 454 pyrosequencing data--enabling realistic simulation with flowsim. *Bioinformatics* 26, i420-i425 (2010).

Baothman, O. A., Zamzami, M. A., Taher, I., Abubaker, J. & Abu-Farha, M. The role of Gut Microbiota in the development of obesity and Diabetes. *Lipids Health Dis.* 15, 108 (2016).

Bashiardes, S., Zilberman-Schapira, G. & Elinav, E. Use of Metatranscriptomics in Microbiome Research. *Bioinform. Biol. Insights* 10, 19-25 (2016).

Bazinet, A. L. & Cummings, M. P. A comparative evaluation of sequence classification programs. *BMC Bioinformatics* 13, 92 (2012).
Begon, M., Harper, J. L. & Townsend, C. R. *Ecology : individuals, populations, and communities*. (Blackwell Science, 1986).
Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* 57, (1995).
Benson, A. K. *et al.* Microbial successions are associated with changes in chemical profiles of a model refrigerated fresh pork sausage during an 80-day shelf life study. *Appl. Environ. Microbiol.* 80, 5178-94 (2014).
Biddle, A., Stewart, L., Blanchard, J. & Leschine, S. Untangling the Genetic Basis of Fibrolytic Specialization by Lachnospiraceae and Ruminococcaceae in Diverse Gut Communities. *Diversity* 5, 627-640 (2013).
Birtel, J. *et al.* Estimating Bacterial Diversity for Ecological Studies: Methods, Metrics, and Assumptions. *PLoS One* 10, e0125356 (2015).
Bonder, M. J., Abeln, S., Zaura, E. & Brandt, B. W. Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* 28, 2891-2897 (2012).
Bowers, R. M. *et al.* Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* 16, 856 (2015).
Bozec, A., Le Roux, A. & Feurer, C. Analyse métagénomique de la dynamique de l'écosystème bactérien de la viande de porc biopréservée. in *16èmes Journées Sciences du Muscle et Technologies des Viandes* 15-16 (2016).
Bragg, L., Stone, G., Imelfort, M., Hugenholtz, P. & Tyson, G. W. Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nat. Methods* 9, 425-426 (2012).
Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* 27, 325-349 (1957).
Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18, 630-634 (2000).
Brestoff, J. R. & Artis, D. Commensal bacteria at the interface of host metabolism and the immune system. *Nat. Immunol.* 14, 676-684 (2013).
Břinda, K., Sykulski, M. & Kucherov, G. Spaced seeds improve k-mer-based metagenomic classification. *Bioinformatics* 31, 3584-3592 (2015).
Brooks, J. P. *et al.* The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* 15, 66 (2015).
Caboche, S., Audebert, C., Lemoine, Y. & Hot, D. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC Genomics* 15, 264 (2014).
Cai, L. *et al.* Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets. *PLoS One* 8, e53649 (2013).
Cai, Y. & Sun, Y. ESPRIT-Tree: Hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res.* 39, e95 (2011).
Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335-336 (2010).

Carlsen, T. *et al.* Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecol.* 5, 747-749 (2012).
Cato, M. P. & Varro, M. T. *On Agriculture*. (Traduit par W. D. Hooper, Harrison Boyd Ash. Loeb Classical Library 283. Cambridge, MA: Harvard University Press, 1934., 36av. J.-C.).
Chakravorty, S., Sarkar, S. & Gachhui, R. Identification of new conserved and variable regions in the 16S rRNA gene of acetic acid bacteria and acetobacteraceae family. *Mol. Biol.* 49, 668-677 (2015).
Chao, A. Species Estimation and Applications. *Encycl. Stat. Sci.* 7907-7916 (2004).
Chao, A. Nonparametric Estimation of the Number of Classes in a Population. *Scand. J. Stat.* 11, (1984).
Charuvaka, A. & Rangwala, H. Evaluation of short read metagenomic assembly. *BMC Genomics* 12 Suppl 2, S8 (2011).
Chen, W. *et al.* A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. *PLoS One* 8, e70837 (2013).
Choo, J. M., Leong, L. E. & Rogers, G. B. Sample storage conditions significantly influence faecal microbiome profiles. *Sci. Rep.* 5, 16350 (2015).
Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* 44, D67-72 (2016).
Clarridge, J. E. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews* 17, 840-862 (2004).
Clooney, A. G. *et al.* Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. *PLoS One* 11, e0148028 (2016).
Cochrane, G. *et al.* Sequence Database Collaboration, I. N. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 44, D48-D50 (2016).
Cole, J. R. *et al.* The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* 33, D294-6 (2005).
Cole, J. R. *et al.* Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633-D642 (2014).
D'Argenio, V., Casaburi, G., Precone, V., Salvatore, F. & Salvatore, F. Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. *Biomed Res. Int.* 2014, 325340 (2014).
Da Silva, K. Microbiota: Dysbiosis as a diagnostic. *Nat. Med.* 20, 348-348 (2014).
Danovaro, R. *et al.* Deep-Sea Biodiversity in the Mediterranean Sea: The Known, the Unknown, and the Unknowable. *PLoS One* 5, e11832 (2010).
Davenport, C. F. *et al.* Genometa - A Fast and Accurate Classifier for Short Metagenomic Shotgun Reads. *PLoS One* 7, e41224 (2012).
Delcenserie, V. *et al.* Microbiota characterization of a Belgian protected designation of origin cheese, Herve cheese, using metagenomic analysis. *J. Dairy Sci.* 97, 6046-6056 (2014).

DeSantis, T. Z. *et al.* Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069-5072 (2006).
Deusch, S., Tilocca, B., Camarinha-Silva, A. & Seifert, J. News in livestock research — use of Omics-technologies to study the microbiota in the gastrointestinal tract of farm animals. *Comput. Struct. Biotechnol. J.* 13, 55-63 (2015).
Dieffenbach, C. W., Lowe, T. M. & Dveksler, G. S. General concepts for PCR primer design. *PCR Methods Appl.* 3, S30-7 (1993).
ecSeq. Why does the per base sequence quality decrease over the read in Illumina? (2017). Disponible sur : http://ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina (Consulté le 1 février 2017).
Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194-2200 (2011).
Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996-998 (2013).
Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460-2461 (2010).
Edwards, R. A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7, 57 (2006).
El Safadi, D. *et al.* Prevalence, risk factors for infection and subtype distribution of the intestinal parasite Blastocystis sp. from a large-scale multi-center study in France. *BMC Infect. Dis.* 16, 451 (2016).
Escalona, M., Rocha, S. & Posada, D. A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat. Rev. Genet.* 17, 459-469 (2016).
Esling, P., Lejzerowicz, F. & Pawlowski, J. Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Res.* 43, 2513-2524 (2015).
Esposito, A., Kirschberg, M., HM, B. & RJ, M. How many 16S -based studies should be included in a metagenomic conference? It may be a matter of etymology. *FEMS Microbiol. Lett.* 351, 145-146 (2014).
Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.* 26, 541-7 (2008).
Filée, J., Tétart, F., Suttle, C. A. & Krisch, H. M. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc. Natl. Acad. Sci. U. S. A.* 102, 12471-6 (2005).
Franklin, R. E. & Gosling, R. G. Molecular configuration in sodium thymonucleate. *Nature* 171, 740-1 (1953).
Fredslund, J., Schauser, L., Madsen, L. H., Sandal, N. & Stougaard, J. PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. *Nucleic Acids Res.* 33, W516-W520 (2005).
Friedman, J. *et al.* Inferring Correlation Networks from Genomic Survey Data. *PLoS Comput. Biol.* 8, e1002687 (2012).

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150-3152 (2012).
Fujimura, K. E. & Lynch, S. V. Microbiota in allergy and asthma and the emerging relationship with the gut microbiome. *Cell Host Microbe* 17, 592-602 (2015).
Gadberry, M. D., Malcomber, S. T., Doust, A. N. & Kellogg, E. A. Primaclade--a flexible tool to find conserved PCR primers across multiple species. *Bioinformatics* 21, 1263-1264 (2005).
Ganda, E. K. *et al.* Longitudinal metagenomic profiling of bovine milk to assess the impact of intramammary treatment using a third-generation cephalosporin. *Sci. Rep.* 6, 37565 (2016).
Garcia-Etxebarria, K., Garcia-Garcerà, M. & Calafell, F. Consistency of metagenomic assignment programs in simulated and real data. *BMC Bioinformatics* 15, 90 (2014).
Gaspar, J. M. & Thomas, W. K. FlowClus: efficiently filtering and denoising pyrosequenced amplicons. *BMC Bioinformatics* 16, 105 (2015).
Gaspar, J. M. *et al.* Assessing the Consequences of Denoising Marker-Based Metagenomic Data. *PLoS One* 8, e60458 (2013).
Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. Metagenomic Assembly: Overview, Challenges and Applications. *Yale J. Biol. Med.* 89, 353-362 (2016).
Gori, F., Folino, G., Jetten, M. S. M. & Marchiori, E. MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. *Bioinformatics* 27, 196-203 (2011).
Gotelli, N. J. & Chao, A. in *Encyclopedia of Biodiversity* 195-211 (Elsevier, 2013).
Grimont, F. & Grimont, P. A. Ribosomal ribonucleic acid gene restriction patterns as potential taxonomic tools. *Ann. Inst. Pasteur. Microbiol.* 137B, 165-75 (1986).
Guinane, C. M. & Cotter, P. D. Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. *Therap. Adv. Gastroenterol.* 6, 295-308 (2013).
Guo, F., Ju, F., Cai, L. & Zhang, T. Taxonomic precision of different hypervariable regions of 16S rRNA gene and annotation methods for functional bacterial groups in biological wastewater treatment. *PLoS One* 8, e76185 (2013).
Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494-504 (2011).
Hamer, H. M. *et al.* Review article: The role of butyrate on colonic function. *Alimentary Pharmacology and Therapeutics* 27, 104-119 (2008).
Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, R245-9 (1998).
Harrison, J. G., Forister, M. L., Parchman, T. L. & Koch, G. W. Vertical stratification of the foliar fungal community in the world's tallest trees. *Am. J. Bot.* 103, 2087-2095 (2016).

Hiergeist, A., Reischl, U. & Gessner, A. Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability. *Int. J. Med. Microbiol.* 306, 334-342 (2016).
Hill, J. M., Bhattacharjee, S., Pogue, A. I. & Lukiw, W. J. The Gastrointestinal Tract Microbiome and Potential Link to Alzheimer's Disease. *Front. Neurol.* 5, 43 (2014).
Hooke, R. *Microscopium*. (In Early Science in Oxford, Vol. 8, p. 333. Oxford: R. T. Gunther, 1931., 1678).
Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* 2, 193-218 (1985).
Hugerth, L. W. *et al.* DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies. *Appl. Environ. Microbiol.* 80, 5116-23 (2014).
Hurlbert, S. H. The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology* 52, 577-586 (1971).
Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* 12, 1889-98 (2010).
Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* 17, 377-86 (2007).
Hyde, E. R. *et al.* The Living Dead: Bacterial Community Structure of a Cadaver at the Onset and End of the Bloat Stage of Decomposition. *PLoS One* 8, e77733 (2013).
Hysom, D. A. *et al.* Skip the Alignment: Degenerate, Multiplex Primer and Probe Design Using K-mer Matching Instead of Alignments. *PLoS One* 7, e34560 (2012).
IBM Research and Mars, Incorporated Launch Pioneering Effort to Drive Advances in Global Food Safety. (2015). Disponible sur : http://www-03.ibm.com/press/us/en/pressrelease/45938.wss (Consulté le 1 février 2017).
Irinyi, L. *et al.* 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. *Med. Mycol.* 53, 313-337 (2015).
Izsak, C. & Price, A. Measuring b-diversity using a taxonomic similarity index, and its relation to spatial scale. *Mar. Ecol. Prog. Ser.* 215, 69-77 (2001).
Jaccard, P. Distribution comparée de la flore alpine dans quelques régions des Alpes occidentales et orientales. *Bull. la Murithienne* 81-92 (1902).
Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*. 96, 3801-3806 (1999).
Ju, F. & Zhang, T. Experimental Design and Bioinformatics Analysis for the Application of Metagenomics in Environmental Sciences and Biotechnology. *Environ. Sci. Technol.* 49, 12628-12640 (2015).
Jünemann, S. *et al.* Bacterial Community Shift in Treated Periodontitis Patients Revealed by Ion Torrent 16S rRNA Gene Amplicon Sequencing. *PLoS One* 7, e41606 (2012).
Kembel, S. W. *et al.* Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Comput. Biol.* 8, e1002743 (2012).

Kim, M. *et al.* Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era. *Genomics Inform.* 11, 102 (2013).
Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* 66, 1328-33 (2000).
Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41, e1 (2013).
Knight, R. *et al.* Unlocking the potential of metagenomics through replicated experimental design. *Nat. Biotechnol.* 30, 513-20 (2012).
Koeppel, A. F. & Wu, M. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Res.* 41, 5175-5188 (2013).
Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211-3217 (2012).
Kopylova, E. *et al.* Open-Source Sequence Clustering Methods Improve the State Of the Art. *mSystems* 1, (2016).
Krause, L. *et al.* Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.* 36, 2230-2239 (2008).
Krebs, C. J. in *Ecological Methodology* 531-595 (1999).
Kumar, P. S. *et al.* Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing. *PLoS One* 6, e20956 (2011).
Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* 12, 118-123 (2010).
Laehnemann, D. *et al.* Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Brief. Bioinform.* 17, 154-179 (2016).
Lederberg, J. 'Ome sweet 'omics -- A genealogical treasury of words. (2001). Disponible sur http://www.the-scientist.com/?articles.view/articleNo/13313/title/-Ome-Sweet--Omics---A-Genealogical-Treasury-of-Words/ (Consulté le 1 février 2017).
Li, W., Fu, L., Niu, B., Wu, S. & Wooley, J. Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief. Bioinform.* 13, 656-668 (2012).
Lindgreen, S. *et al.* An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* 6, 19233 (2016).
Linhart, C. *et al.* The degenerate primer design problem. *Bioinformatics* 18, S172-S181 (2002).
Linhart, C. & Shamir, R. The degenerate primer design problem. *Bioinformatics* 18 Suppl 1, S172-81 (2002).
Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* 12, S4 (2011).
Liu, Z., DeSantis, T. Z., Andersen, G. L. & Knight, R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* 36, e120 (2008).

Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550 (2014).
Lowyick, A., Even, G., Blervaque, R., Merlin, S. & Audebert, C. A targeted metagenomic analysis pipeline dedicated to Ion Torrent PGM data. Dans *European Conference on Computational Biology* (2014).
Lozupone, C. A. & Knight, R. Species divergence and the measurement of microbial diversity. *FEMS Microbiol. Rev.* 32, 557-578 (2008).
Lukeš, J. *et al.* Are Human Intestinal Eukaryotes Beneficial or Commensals? *PLOS Pathog.* 11, e1005039 (2015).
Luo, C., Tsementzi, D., Kyrpides, N., Read, T. & Konstantinidis, K. T. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* 7, e30087 (2012).
Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2, e593 (2014).
Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Heal. Dis.* 26, (2015).
Mangiola, F. *et al.* Gut microbiota in autism and mood disorders. *World J. Gastroenterol.* 22, 361-8 (2016).
Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* 18, 50-60 (1947).
Mao, D.-P., Zhou, Q., Chen, C.-Y. & Quan, Z.-X. Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC Microbiol.* 12, 66 (2012).
Marcet-Houben, M. & Gabaldón, T. TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. *Nucleic Acids Res.* 39, e66 (2011).
Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376-80 (2005).
Marinier, E., Brown, D. G. & McConkey, B. J. Pollux: platform independent error correction of single and mixed genomes. *BMC Bioinformatics* 16, 10 (2015).
Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10 (2011).
Mashima, J. *et al.* DNA Data Bank of Japan. *Nucleic Acids Res.* 45, D25-D31 (2017).
Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11, 538 (2010).
Mavromatis, K. *et al.* Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods* 4, 495-500 (2007).
McDonald, D. *et al.* The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience* 1, 7 (2012).
McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610-618 (2012).
McMurdie, P. J. *et al.* Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput. Biol.* 10, e1003531 (2014).

Menzel, P. *et al.* Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7, 11257 (2016).
Mercier, C., Boyer, F., Bonon, A. & Coissac, É. Sumatra and Sumaclust: fast and exact comparison and clustering of sequences. in *SeqBio* 27-29 (2013).
Meyer, F. *et al.* The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9, 386 (2008).
MICROWINE - Microbial metagenomics and the modern wine industry. (2015). Disponible sur http://cordis.europa.eu/project/rcn/193985_fr.html (Consulté le: 1 février 2017).
Milani, C. *et al.* Assessing the Fecal Microbiota: An Optimized Ion Torrent 16S rRNA Gene-Based Analysis Protocol. *PLoS One* 8, e68739 (2013).
Minot, S. S., Krumm, N. & Greenfield, N. B. One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification. *bioRxiv* (2015).
Mitra, S., Schubach, M. & Huson, D. H. Short clones or long clones? A simulation study on the use of paired reads in metagenomics. *BMC Bioinformatics* 11, S12 (2010).
Monzoorul Haque, M., Ghosh, T. S., Komanduri, D. & Mande, S. S. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* 25, 1722-1730 (2009).
Naisbitt, J. *Megatrends: Ten new directions transforming our lives*. (Warner Books, 1982).
Navas-Molina, J. A. *et al.* Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol.* 531, 371-444 (2013).
Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933-2935 (2013).
Nilsson, R. H., Kristiansson, E., Ryberg, M., Hallenberg, N. & Larsson, K.-H. Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evol. Bioinform. Online* 4, 193-201 (2008).
Normand, S., Secher, T. & Chamaillard, M. La dysbiose, une nouvelle entité en médecine ? *médecine/sciences* 29, 586-589 (2013).
Nossa, C. W. *et al.* Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World J. Gastroenterol.* 16, 4135-44 (2010).
Nuzzo, R. Scientific method: Statistical errors. *Nature* 506, 150-152 (2014).
O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733-45 (2016).
Ohkusa, T. & Koido, S. Intestinal microbiota and ulcerative colitis. *J. Infect. Chemother.* 21, 761-768 (2015).
Orlov, Y. L. & Potapov, V. N. Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic Acids Res.* 32, W628-W633 (2004).
Ounit, R. & Lonardi, S. Higher classification sensitivity of short metagenomic reads with CLARK- *S. Bioinformatics* 32, 3823-3825 (2016).

Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 16, 236 (2015).
Pace, N., Stahl, D., Lane, D. & Olsen, G. Analyzing natural microbial populations by rRNA sequences. *ASM News* 51, 4-12 (1985).
Pallmann, P. *et al.* Assessing group differences in biodiversity by simultaneously testing a user-defined selection of diversity indices. *Mol. Ecol. Resour.* 12, 1068-1078 (2012).
Parks, D. H., Tyson, G. W., Hugenholtz, P. & Beiko, R. G. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30, 3123-4 (2014).
Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* 10, 1200-1202 (2013).
Pignatelli, M. & Moya, A. Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data. *PLoS One* 6, e19984 (2011).
Poinar, H. N. *et al.* Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA. *Science.* 311, (2006).
Poirier, P. *et al.* Development and evaluation of a real-time PCR assay for detection and quantification of blastocystis parasites in human stool samples: prospective study of patients with hematological malignancies. *J. Clin. Microbiol.* 49, 975-83 (2011).
Price, M. N. *et al.* FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* 5, e9490 (2010).
Pylro, V. S. *et al.* Data analysis for 16S microbial profiling from different benchtop sequencing platforms. *J. Microbiol. Methods* 107, 30-37 (2014).
Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59-65 (2010).
Qu, W. *et al.* MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer specificity. *Nucleic Acids Res.* 40, W205-W208 (2012).
Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590-6 (2013).
Quince, C., Lanzen, A., Davenport, R. J. & Turnbaugh, P. J. Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics* 12, 38 (2011).
Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* 66, 846-850 (1971).
Reeder, J. & Knight, R. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat. Methods* 7, 668-669 (2010).
Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4, e2584 (2016).
Romano, S., Bailey, J., Nguyen, V. & Verspoor, K. Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance. 1143-1151 (2014).
Rosen, M. J., Callahan, B. J., Fisher, D. S. & Holmes, S. P. Denoising PCR-amplified metagenome data. *BMC Bioinformatics* 13, 283 (2012).

Rosenkranz, D. & Rosenkranz. easyPAC: A Tool for Fast Prediction, Testing and Reference Mapping of Degenerate PCR Primers from Alignments or Consensus Sequences. *Evol. Bioinforma.* 8, 151 (2012).
Rossi-Tamisier, M., Benamar, S., Raoult, D. & Fournier, P.-E. Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species. *Int. J. Syst. Evol. Microbiol.* 65, 1929-1934 (2015).
Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463-7 (1977).
Santamaria, M. *et al.* Reference databases for taxonomic assignment in metagenomics. *Brief. Bioinform.* 13, 682-695 (2012).
Santori, G. Research papers: Journals should drive data reproducibility. *Nature* 535, 355-355 (2016).
Sayers, E. E-utilities Quick Start. (2008). Disponible sur : https://www.ncbi.nlm.nih.gov/books/NBK25500/ (Consulté le 1 février 2017).
Sboner, A., Mu, X., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biol.* 12, 125 (2011).
Scanlan, P. D. *et al.* The microbial eukaryote *Blastocystis* is a prevalent and diverse member of the healthy human gut microbiota. *FEMS Microbiol. Ecol.* 90, 326-330 (2014).
Scheperjans, F. *et al.* Gut microbiota are related to Parkinson's disease and clinical phenotype. *Mov. Disord.* 30, 350-358 (2015).
Schloss, P. D. & Handelsman, J. Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Appl. Environ. Microbiol.* 71, 1501-1506 (2005).
Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* 6, e27310 (2011).
Schloss, P. D. & Westcott, S. L. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* 77, 3219-26 (2011).
Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537-41 (2009).
Schloss, P. D. A High-Throughput DNA Sequence Aligner for Microbial Ecology Studies. *PLoS One* 4, e8230 (2009).
Schloss, P. D. Readme for the SILVA v119 reference files. (2014a). Disponible sur : http://blog.mothur.org/2014/08/08/SILVA-v119-reference-files/ (Consulté le 1 février 2017) .
Schloss, P. D. Why do I have such a large distance matrix? (2014b). Disponible sur : http://blog.mothur.org/2014/09/11/Why-such-a-large-distance-matrix/ (Consulté le 1 février 2017).
Schmidt, T. M., DeLong, E. F. & Pace, N. R. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* 173, 4371-8 (1991).
Schmidt, T. S. B., Matias Rodrigues, J. F. & von Mering, C. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ. Microbiol.* 17, 1689-1706 (2015).

Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U. S. A.* 109, 6241-6 (2012).
Schreiber, F., Gumrich, P., Daniel, R. & Meinicke, P. Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics* 26, 960-961 (2010).
Schulz, M. H. *et al.* Fiona: a parallel and automatic strategy for read error correction. *Bioinformatics* 30, i356-i363 (2014).
Scofield, D. G. Degenerate In-Silico PcR. (2015). Disponible sur : https://github.com/douglasgscofield/dispr/ (Consulté le 1 février 2017).
Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27, 379-423 (1948).
Siegwald, L. *et al.* Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics. *PLoS One* 12, e0169563 (2017).
Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539 (2011).
Simpson, E. H. Measurement of Diversity. *Nature* 163, 688-688 (1949).
Sinclair, L. *et al.* Microbial Community Composition and Diversity via 16S rRNA Gene Amplicons: Evaluating the Illumina Platform. *PLoS One* 10, e0116955 (2015).
Singer, E. *et al.* Next generation sequencing data of a defined microbial mock community. *Sci. Data* 3, 160081 (2016).
Singh, R. R. *et al.* Clinical Validation of a Next-Generation Sequencing Screen for Mutational Hotspots in 46 Cancer-Related Genes. *J. Mol. Diagnostics* 15, 607-622 (2013).
Sinha, R., Abnet, C. C., White, O., Knight, R. & Huttenhower, C. The microbiome quality control project: baseline study design and future directions. *Genome Biol.* 16, 276 (2015).
Smith, D. R. Broadening the definition of a bioinformatician. *Front. Genet.* 6, 258 (2015).
Soergel, D. A. W., Dey, N., Knight, R. & Brenner, S. E. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J.* 6, 1440-4 (2012).
Soergel, D. A. W., Dey, N., Knight, R. & Brenner, S. E. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J.* 6, 1440-4 (2012).
Soueidan, H. & Nikolski, M. Machine learning for metagenomics: methods and tools. (2015).
Stackebrandt, E. & Goebel, B. M. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int J Syst Bacteriol* 44, 846-849 (1994).
Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12, 1611-8 (2002).
Staley, J. T. & Konopka, A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* 39, 321-46 (1985).
Stämmler, F. *et al.* Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* 4, 28 (2016).

Sun, Y. *et al.* A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief. Bioinform.* 13, 107-121 (2012).
Sun, Y. *et al.* ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Res.* 37, e76-e76 (2009).
Tan, J. *et al.* in *Advances in immunology* 121, 91-119 (2014).
Tedjo, D. I. *et al.* The fecal microbiota as a biomarker for disease activity in Crohn's disease. *Sci. Rep.* 6, 35216 (2016).
Teske, A. & Sørensen, K. B. Uncultured archaea in deep marine subsurface sediments: have we caught them all? *ISME J.* 2, 3-18 (2008).
Thaiss, C. A., Zmora, N., Levy, M. & Elinav, E. The microbiome and innate immunity. *Nature* 535, 65-74 (2016).
The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* 486, 215-221 (2012).
Thompson, C. C. *et al.* Use of recA as an alternative phylogenetic marker in the family Vibrionaceae. *Int. J. Syst. Evol. Microbiol.* 54, 919-924 (2004).
Toribio, A. L. *et al.* European Nucleotide Archive in 2016. *Nucleic Acids Res.* 45, D32-D36 (2017).
Tsilimigras, M. C. B. & Fodor, A. A. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* 26, 330-335 (2016).
Vincent, A. T. & Charette, S. J. Who qualifies to be a bioinformatician? *Front. Genet.* 6, 164 (2015).
Vinh, N. X., Epps, J. & Bailey, J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res.* 11, 2837-2854 (2010).
Vos, M. *et al.* A Comparison of rpoB and 16S rRNA as Markers in Pyrosequencing Studies of Bacterial Diversity. *PLoS One* 7, e30600 (2012).
Vroland, C., Salson, M., Bini, S. & Touzet, H. Approximate search of short patterns with high error rates using the 01∗0 lossless seeds. *J. Discret. Algorithms* 37, 3-16 (2016).
Wagner, B. D., Robertson, C. E. & Harris, J. K. Application of two-part statistics for comparison of sequence variant counts. *PLoS One* 6, e20296 (2011).
Walters, W. A. *et al.* PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* 27, 1159-1161 (2011).
Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261-7 (2007).
Wang, Y. & Qian, P.-Y. Conservative Fragments in Bacterial 16S rRNA Genes and Primer Design for 16S Ribosomal DNA Amplicons in Metagenomic Studies. *PLoS One* 4, e7401 (2009).
Wasserstein, R. L. & Lazar, N. A. The ASA's Statement on p-Values: Context, Process, and Purpose. *Am. Stat.* 70, 129-133 (2016).
Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171, 737-8 (1953).
Weisburg, W. G., Barns, S. M., Pelletier, D. A. & Lane, D. J. 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* 173, 697-703 (1991).

Weiss, S. J. *et al.* Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. (2015).
Westcott, S. L. & Schloss, P. D. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 3, e1487 (2015).
White, B. A. in *PCR Protocols: A Guide to Methods and Applications* 15, 315-322 (Humana Press, 1990).
White, J. R. *et al.* Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Comput. Biol.* 5, e1000352 (2009).
Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bull.* 1, 80 (1945).
Wilke, A. *et al.* The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res.* 44, D590-D594 (2016).
Winter, S. E. & Bäumler, A. J. Why related bacterial species bloom simultaneously in the gut: principles underlying the 'Like will to like' concept. *Cell. Microbiol.* 16, 179-184 (2014).
Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5088-90 (1977).
Wommack, K. E., Bhavsar, J. & Ravel, J. Metagenomics: Read Length Matters. *Appl. Environ. Microbiol.* 74, 1453-1463 (2008).
Wood, D. E. & Salzberg, S. L. kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15, R46 (2014).
Wright, E. S., Yilmaz, L. S. & Noguera, D. R. DECIPHER, a Search-Based Approach to Chimera Identification for 16S rRNA Sequences. *Appl. Environ. Microbiol.* 78, 717-725 (2012).
Yang, B., Wang, Y. & Qian, P.-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* 17, 135 (2016).
Yilmaz, P. *et al.* The SILVA and « all-species Living Tree Project (LTP) » taxonomic frameworks. *Nucleic Acids Res.* 42, D643-8 (2014).
Yu, Q. *et al.* PriSM: a primer selection and matching tool for amplification and sequencing of viral genomes. *Bioinformatics* 27, 266-7 (2011).
Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* 8, 357-66 (1965).

Annexe 1 : Principe du séquençage Ion Torrent PGM

La préparation de la matrice de séquençage (Figure A1.1) se fait hors séquenceur, sur l'Ion OneTouch2, et est suivi d'un enrichissement en ISPs (*Ion Sphere Particles*). Chaque fragment de librairie va être amplifié à la surface d'une ISP, par une PCR en émulsion (emPCR) qui va recouvrir chaque ISP du même fragment d'ADN, afin d'augmenter le signal émis lors du séquençage afin qu'il puisse être détecté. Le mélange d'ISPs et de fragments d'ADN est effectué dans un ratio 1/1, il est néanmoins possible que plusieurs fragments différents soient amplifiés sur la même bille (appelées alors billes polyclonales). Ces configurations non désirées généreront des signaux non interprétables lors du séquençage, qui seront filtrés par le logiciel du séquenceur.

L'Ion Touch ES (Figure A1.2) est ensuite utilisé pour effectuer un enrichissement des ISPs d'intérêt par capture de l'amorce ePCR-A, liée à de la biotine, qui sera captée par des billes magnétiques recouvertes de streptavidine. Cette étape permet d'éliminer les ISPs sans fragments d'ADN amplifiés.

Avant d'effectuer le séquençage (Figure A1.3), l'Ion Torrent PGM est initialisé (étape « Auto pH ») de façon à ce que l'ensemble de ses réactifs soient à une valeur de pH autour de 7,8. Le support de séquençage est une puce contenant des millions de puits, existant en trois modèles selon les capacités de séquençage nécessaires (Chip 314™ ~ 40 Mb, Chip 316™ ~ 200 Mb, Chip 318™ ~ 1 Gb). Chacune des puces est déclinable en une v2, permettant d'utiliser la chimie permettant de générer des lectures plus longues (~ 400-450 nt). La puce de séquençage est tout d'abord chargée par la matrice de séquençage (ISPs recouverts de fragments d'ADN), ainsi que de polymérase et d'amorces de séquençage. Le séquençage correspond à la détection du signal émis lors de la réaction de polymérisation, induite par l'ajout séquentiel des nucléotides. Ce signal est généré par un changement de pH du milieu induit par l'émission d'ions H+ lors de la polymérisation.

Cette variation de pH est détectée au niveau de la couche mince au fond de chaque puits, suivant la technologie des semi-conducteurs. Les mesures de pH à chaque cycle de séquençage sont enregistrées dans un fichier nommé l'ionogramme.

Une fois le run de séquençage terminé (Figure A1.4), les ionogrammes sont transmis du séquenceur vers le serveur informatique qui lui est associé. Ce serveur va exécuter un algorithme de *basecalling*, convertissant les données des ionogrammes en séquences et qualités associées. Cette étape est suivie d'une étape de pré-traitement propre à la technologie, éliminant les lectures de mauvaise qualité ou celles qui ne correspondent qu'à la séquence de l'adaptateur. Les lectures de mauvaise qualité ou de trop petite taille sont également éliminés. À l'issue de cette analyse, le serveur génère un fichier FASTQ contenant l'ensemble des lectures séquencées, ainsi qu'un rapport résumant le déroulement du séquençage et de ces premières étapes analytiques.

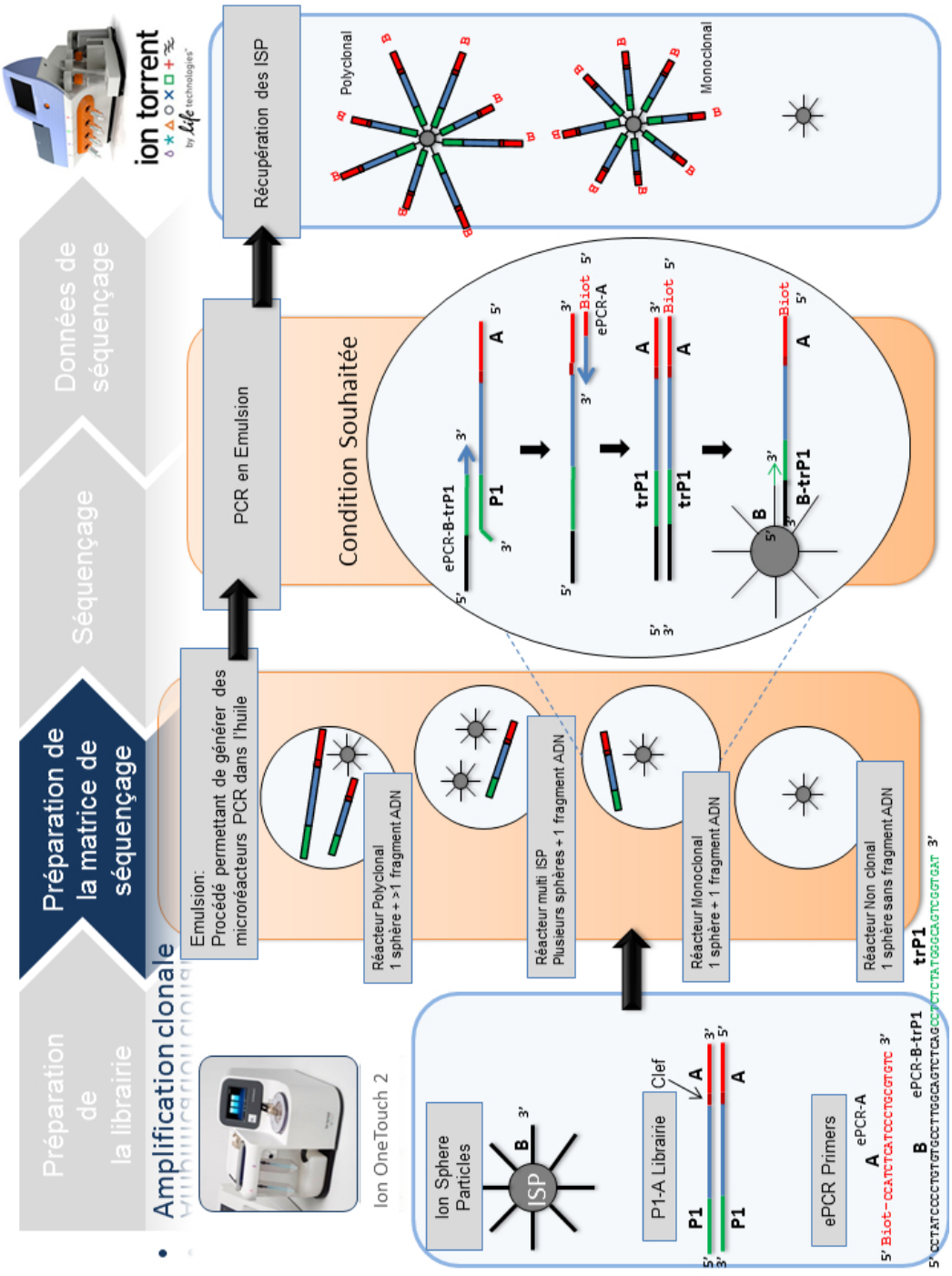


Figure A1.1 : Préparation de la matrice de séquençage (Source : http://www.biorigami.com/?p=4643)

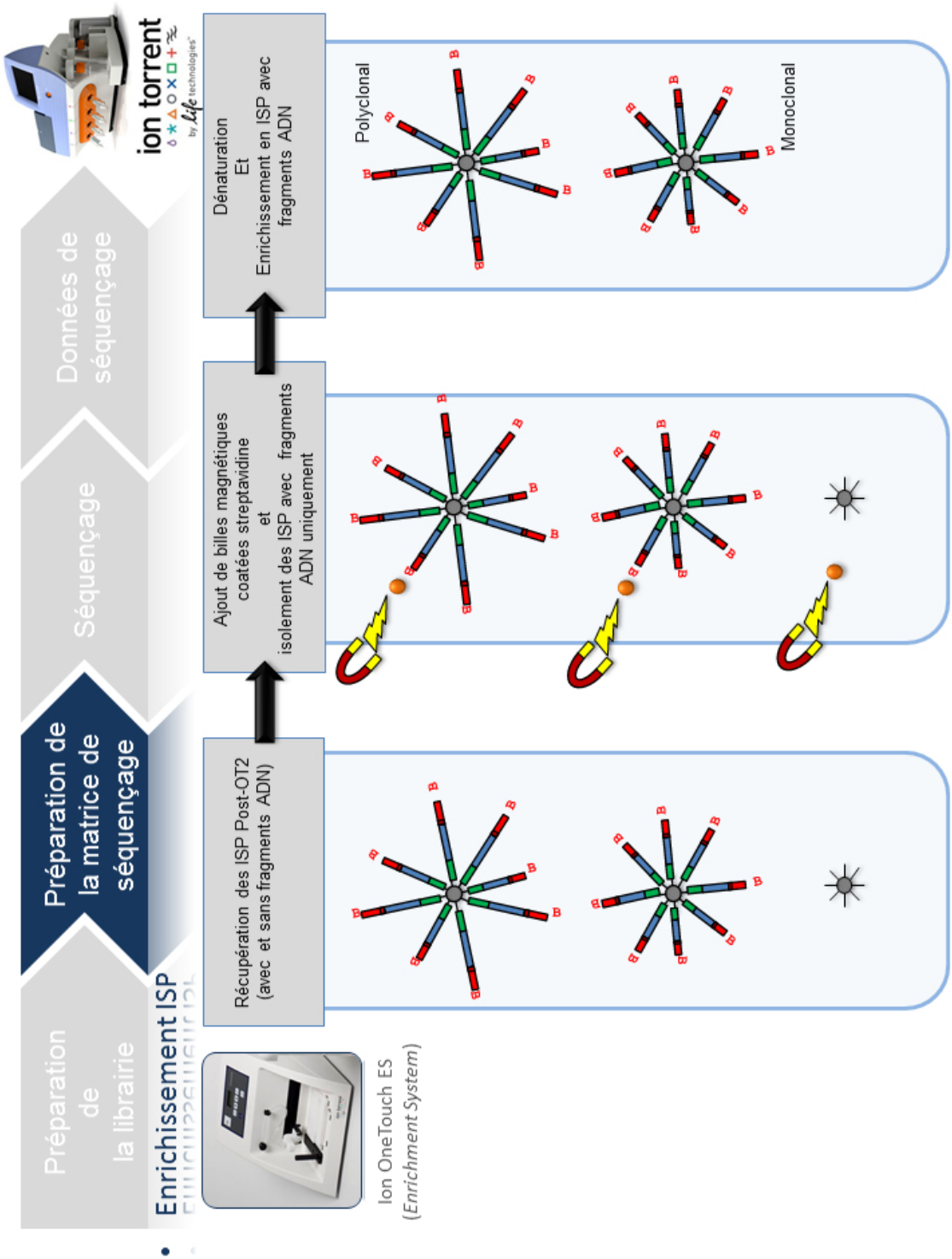


Figure A1.2 : Enrichissement en ISPs
(Source : http://www.biorigami.com/?p=4643)

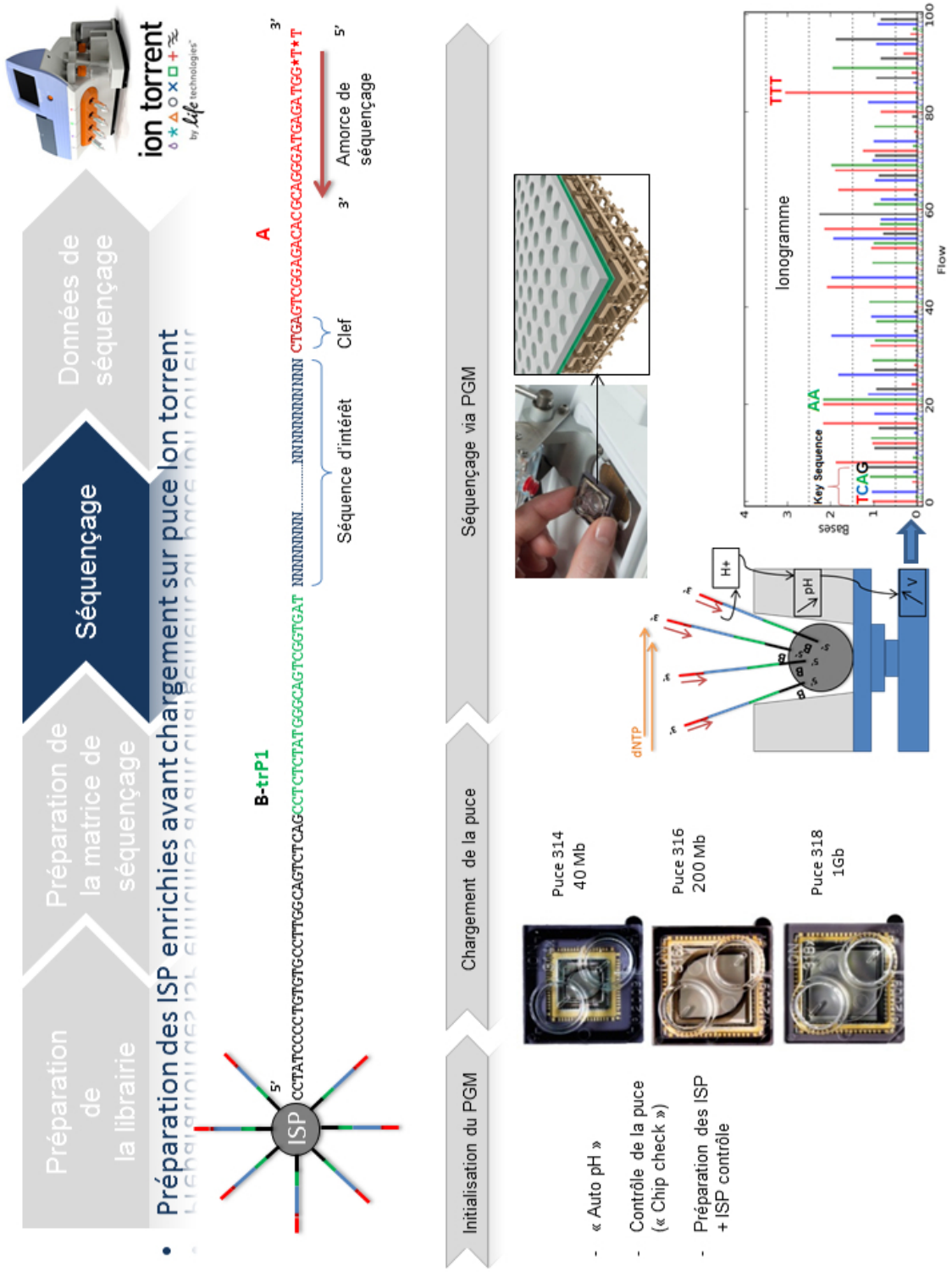


Figure A1.3 : Séquençage Ion Torrent PGM
(Source : http://www.biorigami.com/?p=4643)

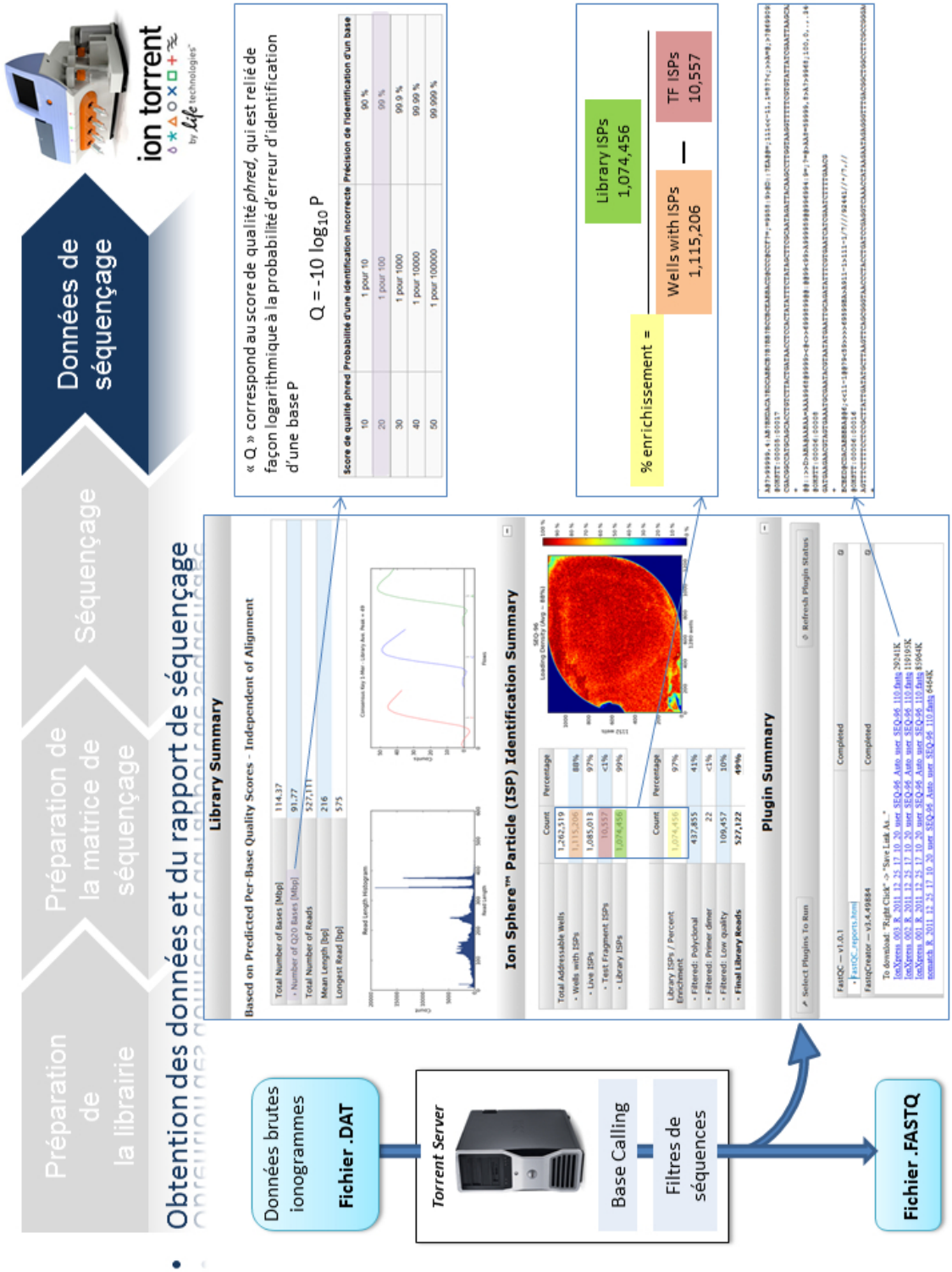


Figure A1.4 : Traitement du signal brut et génération des lectures (Source : http://www.biorigami.com/?p=4643)

Annexe 2 : Lignes de commande exécutées pour chaque pipeline

A2.1 - BMP

Commandes basées sur : http://www.brmicrobiome.org/#!16s-profiling-ion-torrent/cpdg
Logiciels & versions : usearch v7.0.1090, QIIME 1.9.0, uc2otutab.py (script UPARSE)

Variables :

- `$file = fichier FASTQ des lectures`
- `$cut = taille d'élagage des lectures (150 pour 200(V3) & réel, 300 pour 400(V4-V5)`
- `$ref_db = fichier FASTA de la banque de référence (séquences clusterisées à 97 %)`
- `$ref_aligned = fichier FASTA aligné de la banque de référence (séquences clusterisées à 97 %)`
- `$ref_taxo = taxonomie de la banque de référence`

Fichiers de sortie d'intérêt :

- `map.uc – Association entre les OTUs (colonne 9) et chaque lecture (colonne 8)`
- `output/otus_tax_assignments.txt – Assignation taxonomique (colonne 2) pour chaque OTU (colonne 1)`

Commandes :

01 – Filtrage qualité, élagage des lectures et conversion en FASTA

```
usearch -fastq_filter $file -fastq_trunclen $cut
-fastaout reads.fa
```

02 – Déreplication des lectures

```
usearch -derep_fulllength reads.fa -output derep.fa
-sizeout
```

03 – Tri des lectures par abondance et élimination des singletons

```
usearch -sortbysize derep.fa -output sorted.fa -minsize 2
```

04 – *Clustering* en OTUs en utilisant la méthode UPARSE

```
usearch -cluster_otus sorted.fa -otus otus.fa
```

05 – Alignement de chaque lecture initiale sur son OTU de référence

```
usearch -usearch_global reads.fa -db otus.fa -strand plus
-id 0.97 -uc map.uc
```

06 – Assignation taxonomique des OTUs en utilisant la méthode UCLUST de QIIME

```
assign_taxonomy.py –i otus.fa –o output –r $ref_db
–t $ref_taxo
```

07 – Alignement des lectures de référence de chaque OTU sur la banque Greengenes de référence par QIIME

```
align_seqs.py –i otus.fa –o rep_set_align –t $ref_aligned
```

08 – Filtrage des alignements par QIIME

```
filter_alignment.py –i rep_set_align/otus_aligned.fasta
–o filtered_alignment
```

09 – Construction d'un arbre phylogénétique des OTUs par QIIME

```
make_phylogeny.py
–i filtered_alignment/otus_aligned_pfiltered.fasta
–o rep_set.tre
```

10 – Génération d'une table d'OTUs

```
python uc2otutab.py map.uc > otu_table.txt
```

11 – Conversion de la table d'OTUs en fichier biom

```
biom convert –i otu_table.txt –o otu_table.biom
--table-type=\"OTU table\" --to-json
```

12 – Ajout de métadonnées (taxonomie) à la table d'OTUs

```
biom add-metadata –i otu_table.biom –o otu_table_tax.biom
--observation-metadata-fp output/otus_tax_assignments.txt
--observation-header OTUID,taxonomy,confidence
--sc-separated taxonomy --float-fields confidence
```

13 -Vérification de la table d'OTUs par QIIME

```
biom summarize-table –i otu_tabl
e_tax.biom –o results_biom_table
```

A2.2 - mothur

Commandes basées sur : http://www.mothur.org/wiki/454_SOP
Logiciels & versions : mothur 1.35.1

Variables :

```
$file = fichier FASTQ des lectures
$filename = nom du fichier FASTQ (sans l'extension)
$nb_threads = nombre de processus parallélisés (dans notre
cas, 32)
$ref_db = fichier FASTA de la banque de référence
(séquences clusterisées à 99 %)
$ref_aligned = fichier FASTA aligné de la banque de
référence (séquences clusterisées à 97 %)
$ref_taxo = taxonomie de la banque de référence
$rdp_fasta = fichier FASTA formatté par mothur du jeu de
données d'entraînement de RDP (dans notre cas,
trainset9_032012.pds.fasta)
$rdp_taxo = taxonomie associée au jeu de données
d'entraînement de RDP (dans notre cas,
trainset9_032012.pds.tax)
```

Fichiers de sortie d'intérêt :

```
$filename.final.an.0.03.fasta – Fichier FASTA de chaque
lecture avec son OTU associé (en-tête de chaque lecture
sous la forme : « >numero_de_la_lecture \t OTU »
$filename.final.an.0.03.cons.taxonomy – Assignation
taxonomique (colonne 3) pour chaque OTU (colonne 1)
```

Commandes :

01 – Conversion du FASTQ en FASTA

```
mothur "#fastq.info(fastq=$file, fasta=T)"
```

02 – Filtrage des lectures

```
mothur "#trim.seqs(fasta=$filename.fasta, pdiffs=2,
bdiffs=1, minlength=100, maxhomop=8,
processors=$nb_threads)"
```

03 – Résumé du fichier sortant

```
mothur "#summary.seqs(fasta=$filename.trim.fasta)"
```

04 – Déréplication des lectures

```
mothur "#unique.seqs(fasta=$filename.trim.fasta)"
```

05 – Alignement des lectures sur la banque de référence

```
mothur "#align.seqs(fasta=$filename.trim.unique.fasta,
flip=true, reference=$ref_aligned, processors=$nb_threads)"
```

06 – Résumé de l'alignement

```
mothur "#summary.seqs(fasta=$filename.trim.unique.align,
name=$filename.trim.names)"
```

07 – Elagage de l'alignement – Note : vérifier que le paramètre « optimize » est choisi correctement par mothur, en se basant sur le résumé de l'alignement précédents, et le remplacer par start=position de début et end=position de fin si nécessaire

```
mothur "#screen.seqs(fasta=$filename.trim.unique.align,
name=$filename.trim.names, optimize=start-end-minlength,
criteria=95, processors=$nb_threads)"
```

08 – Filtrage de l'alignement

```
mothur
"#filter.seqs(fasta=$filename.trim.unique.good.align,
trump=., vertical=T, processors=$nb_threads)"
```

09 – Déréplication des lectures

```
mothur
"#unique.seqs(fasta=$filename.trim.unique.good.filter.fasta
, name=$filename.trim.good.names)"
```

10 – Résumé du fichier final de lectures

```
mothur
"#summary.seqs(fasta=$filename.trim.unique.good.filter.uniq
ue.fasta, name=$filename.trim.good.names)"
```

11 – Pré-*clustering*

```
mothur
"#pre.cluster(fasta=$filename.trim.unique.good.filter.uniqu
e.fasta, name=$filename.trim.unique.good.filter.names,
diffs=2)"
```

12 – Création de nouvelles variables contenant les noms de fichiers afin de les simplifier

```
$preclustered_fasta="$filename.trim.unique.good.filter.uniq
ue.precluster.fasta"
$preclustered_names="$filename.trim.unique.good.filter.uniq
ue.precluster.names"
$taxonomy_file="$filename.trim.unique.good.filter.unique.pr
ecluster.xxx.taxonomy" # xxx doit être manuellement
remplacé par le nom correct du fichier, variable selon la
banque de séquences utilisée
```

13 – Classification de toutes les lectures pré-clusterisées

```
mothur "#classify.seqs(fasta=$preclustered_fasta,
name=$preclustered_names, template=$rdp_fasta,
taxonomy=$rdp_taxo, cutoff=80, processors=$nb_threads)"
```

14 – Elimination des séquences non bactériennes

```
mothur "#remove.lineage(fasta=$preclustered_fasta,
name=$preclustered_names, taxonomy=$taxonomy_file,
taxon=Mitochondria-Chloroplast-Archaea-Eukaryota-unknown)"
```

15 – Nouveau résumé des séquences restantes

```
mothur
"#summary.seqs(fasta=$filename.trim.unique.good.filter.uniq
ue.precluster.pick.fasta,name=$filename.trim.unique.good.fi
lter.unique.precluster.pick.names)"
```

16 – Renommage des fichiers

```
mv
$filename.trim.unique.good.filter.unique.precluster.pick.na
mes $filename.final.names
mv
$filename.trim.unique.good.filter.unique.precluster.pick.fa
sta $filename.final.fasta
mv $taxonomy_file $filename.final.taxonomy
```

17 – Calcul de la matrice de distances

```
mothur "#dist.seqs(fasta=$filename.final.fasta,
cutoff=0.20, processors=$nb_threads)"
```

18 - *Clustering* des séquences en OTUs

```
mothur "#cluster.split(column=$filename.final.dist,
name=$filename.final.names, cutoff=0.10,
processors=$nb_threads)"
```

19 – Classification des OTUs

```
mothur "#classify.otu(list=$filename.final.an.list,
name=$filename.final.names,
taxonomy=$filename.final.taxonomy, label=0.03)"
```

20 – Génération d'un fichier FASTA avec toutes les lectures et leur OTU associé

```
mothur "#bin.seqs(list=$filename.final.an.list,
fasta=$filename.fasta, name=$filename.final.names,
label=0.03)"
```

A2.3 - QIIME SortMeRNA + Sumaclust

Commandes basées sur :
http://qiime.org/scripts/pick_open_reference_otus.html
Logiciels & versions : QIIME 1.9.0

Variables :

```
$file = fichier FASTQ des lectures
$filename = nom du fichier FASTQ (sans l'extension)
$nb_threads = nombre de processus parallélisés (dans notre
cas, 32)
$ref_db = fichier FASTA de la banque de référence
(séquences clusterisées à 97 %)
$ref_taxo = taxonomie de la banque de référence
$outdir = dossier de sortie
```

Fichiers de sortie d'intérêt :

```
$outdir/final_otu_map_mc2.txt – OTUs (colonne 1, un OTU par
ligne) et tous les reads associés (toutes les autres
colonnes)
$outdir/uclust_assigned_taxonomy/rep_set_tax_assignments.tx
t – Assignation taxonomique      (colonne 2) pour chaque
OTU (colonne 1)
```

Commandes :

01 – Créer un fichier parameters.txt contenant les lignes suivantes :

```
assign_taxonomy:reference_seqs_fp $ref_db
assign_taxonomy:id_to_taxonomy_fp $ref_taxo
```

02 – Génération d'un fichier FASTA

```
convert_fastaqual_fastq.py –c fastq_to_fastaqual –f $file
–o .
```

03 – Exécution du pipeline de *clustering* d'OTUs *open-reference*

```
pick_open_reference_otus.py –p parameters.txt
–i $filename.fna –m sortmerna_sumaclust –a
––jobs_to_start $nb_threads –o $outdir –r $ref_db
```

A2.4 - QIIME UCLUST

Commandes basées sur :
http://qiime.org/scripts/pick_open_reference_otus.html
Logiciels & versions : QIIME 1.9.0

Variables :

```
$file = fichier FASTQ des lectures
$filename = nom du fichier FASTQ (sans l'extension)
$nb_threads = nombre de processus parallélisés (dans notre
cas, 32)
$ref_db = fichier FASTA de la banque de référence
(séquences clusterisées à 97 %)
$ref_taxo = taxonomie de la banque de référence
$outdir = dossier de sortie
```

Fichiers de sortie d'intérêt :

```
$outdir/final_otu_map_mc2.txt - OTUs (colonne 1, un OTU par
ligne) et tous les reads associés (toutes les autres
colonnes)
$outdir/uclust_assigned_taxonomy/rep_set_tax_assignments.tx
t - Assignation taxonomique (colonne 2) pour chaque OTU
(colonne 1)
```

Commandes :

01 – Créer un fichier parameters.txt contenant les lignes suivantes :

```
pick_otus:enable_rev_strand_match True
assign_taxonomy:reference_seqs_fp $ref_db
assign_taxonomy:id_to_taxonomy_fp $ref_taxo
```

02 – Génération d'un fichier FASTA

```
convert_fastaqual_fastq.py -c fastq_to_fastaqual -f $file
-o .
```

03 – Exécution du pipeline de *clustering* d'OTUs *open-reference*

```
pick_open_reference_otus.py -p parameters.txt -i
$filename.fna -a --jobs_to_start $nb_threads -o $outdir
-r $ref_db
```

A2.5 - kraken

Commandes basées sur : https://ccb.jhu.edu/software/kraken/
Logiciels & versions : kraken 0.10.5-beta

Variables :
```
$file = fichier FASTQ des lectures
$nb_threads = nombre de processus parallélisés (dans notre cas, 32)
$db = dossier de la banque de k-mers (dans notre cas, minikraken_20141208)
```

Fichiers de sortie d'intérêt :
```
output.txt – Attribution d'un taxid du NCBI (colonne 3) pour chaque lecture (colonne 2)
```

Commandes :
01 – Classification des lectures
```
kraken --db $db --fastq-input $file --threads $nb_threads > output.txt
```

A2.6 - CLARK

Commandes basées sur : http://clark.cs.ucr.edu/
Logiciels & versions : CLARK v1.1.2

Variables :
```
$file = fichier FASTQ des lectures
$nb_threads = nombre de processus parallélisés (dans notre cas, 32)
```

Fichiers de sortie d'intérêt :
```
output.txt – Attribution d'un taxid du NCBI (colonne 3) pour chaque lecture (colonne 1)
```

Commandes :
```
01 – Construction de la banque de k-mers discriminants (bactéries de RefSeq au niveau du genre)
set_targets.sh ref Bacteria --genus
02 – Classification des lectures
classify_metagenome.sh -O $file -R output.txt -n $nb_threads
```

A2.7 - One Codex

Site Internet : https://www.onecodex.com/
Ressources nécessaires : Connexion et navigateur Internet

Fichiers de sortie d'intérêt :
`Read-level data – Attribution d'un taxid du NCBI (colonne 2) pour chaque lecture (colonne 1)`

Actions :
S'inscrire sur www.onecodex.com et suivre les instructions d'envoi de fichiers (Menu « Upload / Import »). Les résultats sont visibles à partir du menu « Samples » : Cliquer sur l'échantillon d'intérêt, et télécharger le fichier de résultat brut en cliquant sur le bouton « Read-level data ».

Annexe 3 : Identification des espèces non référencées dans les banques de séquences d'ADNr.

	Gloeobacter kilaueensis		***Tersicoccus phoenicis***		***Eisenbergiella tayi***	
	Famille	Genre	Famille	Genre	Famille	Genre
BMP GG	VP	VP	VP	FP	VP	FP
BMP S	FN	FN	FN	FN	VP	FP
CLARK	VP	VP	FN	FN	VP	FN
kraken Minikraken	VP	VP	FN	FN	VP	FN
mothur GG	VP	VP	VP	FN	VP	FN
mothur S	FN	FN	FN	FN	VP	FP
One Codex RefSeq	VP	VP	FN	FN	FN	FN
One Codex OneCodex28k	VP	VP	FP	FP	FN	FN
One Codex SILVA	VP	VP	VP	VP	FN	FN
QIIME U S	FN	FN	VP	FN	VP	FN
QIIME U GG	VP	VP	VP	FP	VP	FN
QIIME SS S	FN	FN	FN	FN	VP	FN
QIIME SS GG	VP	VP	VP	FN	VP	FN

Tableau A3.1 : Identification par les pipelines évalués des lectures appartenant aux bactéries dont les espèces ne sont pas référencées dans les banques de séquences d'ADNr de référence, au niveau de la famille et du genre.

VP = Vrai Positif, FP = Faux Positif, FN = Faux Négatif.

Annexe 4 : Espèces de champignons d'intérêt.

Nom de l'espèce	Région génomique étendue:positions
Aspergillus fumigatus	gb|ABDB01000088.1:178485-180485
Aspergillus flavus	gb|ZDT01000286.1:910-2910
Aspergillus terreus	gb|LWBM01000021.1:1915980-1917980
Aspergillus niger	gb|MCQH01000015.1:248244-250244
Aspergillus nidulans	
Aspergillus ochraceus	
Scedosporium apiospermum	gb|JOWA01000106.1:250510-252510
Scedosporium aurantiacum	gb|JUDQ01000713.1:2733-4733
Scedosporium prolificans	
Scedosporium minutispora	
Fusarium oxysporum	gb|MALU01000072.1:750-2750
Fusarium virguliforme	gb|MADZ01000003.1:1230-3230
Fusarium verticillioides	gb|AAIM02000210.1:3250-5250
Scopulariopsis brevicaulis	
Scopulariopsis candida	
Acremonium strictum	
Paecilomyces variotii	
Paecilomyces lilacinus	
Mucor circinelloides	gb|AMYB01000003.1:504514-506614
Mucor racemosus	gb|JNEI01006051.1:537-2537
Mucor indicus	
Lichtheimia corymbifera	
Lichtheimia ramosa	
Lichtheimia ornata	
Rhizopus oryzae	gb|JNDX01002683.1:2455-4455
Rhizopus delemar	gb|AACW02000152.1:4235-6335
Rhizopus microsporus	
Rhizopus stolonifer	
Candida albicans	NW_139715.1:1275-2776

Tableau A4.1 : Nom des espèces de champignons d'intérêt, identifiant et positions de la séquence Genbank associée quand la région génomique étendue encadrant les régions ITS est disponible.