Université Claude Bernard – Lyon 1
EA 2160 Laboratoire Mer, Molécules, Santé (Université Le Mans)

Identification des micro-organismes et de leurs fonctions métaboliques dans un lac d'Oran par séquençage Shotgun métagénomique

Théophile BOYER
Master 2 Bioinformatique Moléculaire : Méthodes et Analyses

Juin 2021

Encadrantes :
Aurore Caruso EA 2160 Laboratoire Mer, Molécules, Santé
Nathalie Casse EA 2160 Laboratoire Mer, Molécules, Santé

Année Universitaire 2020 - 2021

EA 2160 Laboratoire Mer, Molécules, Santé (Université Le Mans)

Identification des organismes et fonctions métaboliques présents dans un lac d'Oran par séquençage Shotgun métagénomique

Théophile Boyer
Sous la tutelle de Aurore Caruso et Nathalie Casse

Résumé

Le lac Dayat Morsli est un lac d'eau salé à la périphérie d'Oran, en Algérie. Longtemps considéré comme première décharge de la ville, cet ancien lac pollué est aujourd'hui classé comme réserve naturelle et est au centre d'un programme de dépollution et revitalisation depuis le début des années 2000. L'étude des paramètres physico-chimiques du lac a permis d'établir la présence de deux saisons : l'une humide, de janvier à mai, et l'autre sèche, de juin à décembre. Cependant aucune étude sur les micro-organismes de lac n'a encore été réalisée dans cette région. Identifier les communautés présentes dans ce lac et comprendre leur rôle permettraient de compléter les connaissances sur ce milieu encore peu caractérisé et mieux appréhender sa préservation.

C'est ainsi qu'en 2018, deux échantillons d'eau du lac Dayat Morsli ont été séquencés par Shotgun métagénomique pour identifier les communautés microbiennes présentes à chaque saison. L'identification des différentes espèces a été effectuée grâce à deux méthodes complémentaires : les k-mer exacts et l'homologie des séquences ADN. Cette double approche a permis de classifier, avec robustesse, taxonomiquement la majorité des microorganismes. L'étude de l'abondance relative au rang domaine montre, pour les deux saisons, la présence d'une majorité de séquences bactériennes et une augmentation importante des archées lors de la saison sèche. Pour les deux échantillons, l'analyse des phyla montre des compositions similaires pour les archées et les bactéries mais divergentes chez les eucaryotes. L'étude a également été enrichie par une assignation fonctionnelle des gènes présents dans les séquences grâce à différentes bases de données métabolomiques. Les principales voies métaboliques identifiées ont également démontré une similitude entre les deux saisons, bien que plus faible lors de la saison sèche.

Pour l'ensemble de ces analyses un assemblage métagénomique des données a été nécessaire. Ce dernier a permis de reconstruire au mieux les organismes présents à chaque saison et d'analyser leurs fonctions.

Table des matières

Table des figures

Table des tableaux

Tableau des logiciels utilisés

Logiciels	**Version**
Bwa	0.7.17
FastQC	0.11.9
Kraken2	2.0.9beta
Quast	5.0.2
Megahit	1.2.9
R	4.0.3
Samtools	1.10
Seqkit	0.14.0
Squeezemeta	1.4.0beta1
Spades	3.14.1
Trimmomatic	0.39

Bases de données	**Date d'accession**
NCBI nr	25/09/20
NCBI nt	29/01/21
KEGG	25/09/20
PFAM	25/09/20
COG	25/09/20

Liste des abréviations

ADNg	Acide Désoxyribonucléique génomique
ARNr	Acide RiboNucléique ribosomique
CAMI	Critical Assessment of Metagenomic Interpretation
DDBJ	DNA Data Bank of Japan
EMBL	European Molecular Biology Laboratory
GC	Guanine Cytosine
IFB	Institut Français de Bioinformatique
LCA	Plus petit ancêtre commun / Lowest Common Ancestor
LUCAS	Land Use/Land Cover Area Frame Survey
MAGs	Génomes Assemblés de Métagénomes
NGS	Nouvelle génération de séquençage
ONT	Oxford Nanopore Technology
ORF	Open Reading Frame
PB	Paire de base

Remerciement

Je souhaite remercier chaleureusement madame Aurore Caruso et madame Nathalie Casse pour m'avoir accueilli dans le laboratoire, d'avoir eu la patience et l'envie de comprendre la bioinformatique (même si ce n'était pas toujours facile). Merci d'avoir suivi les péripéties et rebondissement de ce stage en cherchant toujours à me rassurer et à me faire aller de l'avant.

Merci Aurélie pour tes conseils toujours pertinents et même s'ils ne me plaisent pas tout le temps et que je n'aime pas les entendre, tu es et reste la meilleure.

Merci à Mariem, Julie, Maxime, Romuald, Gurvan et toutes les personnes avec qui j'ai partagé la vie de laboratoire pendant ce stage.

1. Introduction

1.1. Du simple séquençage à la métagénomique

Connaître la composition en acides nucléiques d'un génome a toujours été l'un des buts de la compréhension du vivant. Du séquençage de Sanger (Sanger et al., 1977), aux technologies actuelles de troisième génération apparues dans les années 2010 (Schadt et al., 2010), beaucoup de chemin a été parcouru. C'est l'essor des Nouvelles Générations de Séquençage (NGS) qui ont permis la démocratisation du séquençage et de la génomique de masse (Metker, 2010). En effet, elles permettent de réaliser du séquençage à très haut débit pour un prix et un intervalle de temps drastiquement réduit. Ces nouvelles technologies ont grandement accéléré la caractérisation de nombreux organismes qu'ils soient viraux, bactériens, archéens ou eucaryotes. Au fil des années, plusieurs générations de séquençage ont vu le jour (Heather et Chain, 2016). La première mise en place dans les années 70, lente et coûteuse, reposait sur l'addition de nucléotides dénaturés afin de produire des fragments de l'ordre du kilobase (kb) visualisables par électrophorèse. La deuxième génération apparaît vers 2005. Elle s'affranchit des nucléotides marqués pour leur préférer la luminescence. Tout comme la première génération, elle s'appuie sur une amplification de l'échantillon et produit une grande quantité de courtes séquences allant de 30 à 250 paires de bases (pb) selon les méthodes. Cependant, la technologie Illumina grâce à la longueur et la qualité de ses séquences, 0,1% d'erreur par base, s'est imposée petit à petit comme leader dans ce domaine (Pfeiffer et al., 2018). C'est dans les années 2010 que la troisième génération de séquençage voit le jour. Elle vise principalement à résoudre les problèmes liés à la petite taille des fragments générés par la génération précédente, particulièrement la résolution des séquences répétées dans un génome. Les technologies PacBio et Oxford Nanopore (ONT) actuellement sur le marché peuvent produire des séquences entre 10 et 100 kb mais possèdent un taux d'erreur très élevé autour de 15 % (Dohm et al., 2020).

A l'aide de ces technologies, les biologistes ont d'abord cherché à séquencer des organismes simples comme des bactéries. Le premier organisme entièrement séquencé fut *Haemophilus influenzae* en 1995 (Fleischmann et al., 1995). Un an plus tard, la première archée, *Methanocaldococcus jannaschii* fut complètement séquencée (Bult et al., 1996). Enfin, le premier organisme multi-cellulaire eucaryote à avoir été complètement séquencé fut *Caenorhabditis elegans* en 1998 (The *C. elegans* Sequencing, 1998). En accédant à la séquence, les chercheurs ont eu par extension accès aux voies métaboliques et enzymatiques mais aussi aux Open Reading Frame (ORFs) et aux transcripts. L'évolution des techniques biochimiques basées sur les nouvelles technologies de séquençage a vu l'émergence de nouveaux domaines d'étude comme la transcriptomique, l'épigénomique, la métabolomique, etc. Ces disciplines font partie du champ des "omiques" qui permettent l'étude de l'ensemble d'une composante d'un système.

En voulant s'intéresser à la biodiversité présente dans l'environnement, les chercheurs se sont rendus compte qu'une grande partie de cette dernière n'était pas caractérisée. Par exemple, il a été estimé que les bactéries cultivables ne représentaient que 1 % des bactéries présentes dans le sol (Dauga et al., 2005). Vouloir déterminer et étudier la composition en micro-organismes d'un environnement spécifique, a donné naissance à une nouvelle discipline "omique" : la métagénomique.

1.2. L'approche métagénomique

La métagénomique se définit comme l'étude de l'ensemble des micro-organismes présents dans un environnement. La première utilisation du terme "métagénomique"

remonte à 1998 (Handelsman et al., 1998). Mais c'est en 1996 que le premier séquençage métagénomique a été réalisé sur un échantillon d'un environnement marin (Stein et al., 1996). Leur approche via la création d'une librairie combinant *E. coli* et fosmide a bien évolué depuis.

Au cours des vingt dernières années, de nombreuses études ont été lancées pour permettre de mieux comprendre le monde microbien aussi bien à l'intérieur qu'à l'extérieur du corps humain. Ainsi des initiatives, comme le projet microbiome humain, ont permis de mettre en lumière l'extraordinaire hétérogénéité des micro-organismes présents dans nos intestins et leur rôle essentiel sur notre santé (Turnbaugh et al., 2007). D'autres projets ont mis en évidence la biodiversité en micro-organismes des sols via le projet Earth Microbiome (Gilbert et al., 2014) ou encore l'étude LUCAS, acronymes de Land Use/Land Cover Area Frame Survey, portée par la commission européenne (European Commission, 2016). L'étude des océans est aussi à l'honneur avec des programmes comme le projet global "ocean sampling" porté par la J. Craig Venter Institute (Rusch et al. 2007) ou plus récemment le projet "Tara" (Sunagawa et al. 2015). Avec la métagénomique, des espèces ou des fonctions métaboliques jusqu'alors inconnues apparaissent, démontrant encore une fois la complexité du monde vivant.

En métagénomique, il existe différentes approches pour caractériser un échantillon (Breitwieser et al., 2019) :

- la métagénomique ciblée, aussi appelée « metabarcoding », consiste à sélectionner un gène très conservé et à ne séquencer que ce dernier pour tous les organismes d'un même échantillon. Généralement, il est choisi l'ARNr 16S pour les bactéries ou l'ARNr 18S pour les eucaryotes. Avec cette approche, bien que l'entièreté des micro-organismes présents dans un échantillon ne soit pas visible (notamment les virus), il est possible, en fonction du fragment sélectionné, de réaliser une phylogénie très précise des espèces présentes. Et cela, sans exiger d'importantes ressources informatiques pour traiter les données.
- la métagénomique globale aussi appelée « Shotgun » métagénomique est le deuxième type d'approche pour l'analyse d'un échantillon. Contrairement au metabarcoding, cette technique vise à prendre l'intégralité des séquences nucléiques de l'échantillon, à les fragmenter puis à les séquencer. Cette technique a comme avantage de pouvoir étudier en même temps l'ensemble des génomes microbiens sans distinction, qu'ils soient eucaryotes, archéens, bactériens ou viraux.

Une fois le séquençage effectué, les lectures des séquences peuvent être analysées de deux façons : par assemblage ou par différence d'abondance des séquences.

La première méthode repose sur l'assemblage des lectures en contigs, soit des séquences génomiques continues et ordonnées. Le but de cet assemblage est de reconstruire *in silico* le génome d'origine d'un micro-organisme à partir des lectures des séquences. L'assemblage reste une étape compliquée lorsqu'il faut traiter un génome unique, mais l'est encore plus lorsqu'il s'agit de traiter un mélange de génomes. Cette difficulté provient notamment de la différence de profondeur de séquençage entre les génomes présents dans l'échantillon, et de la non clonalité de ces derniers. Des outils adaptés à la métagénomique ont ainsi été développés et ont été comparés grâce au challenge CAMI (Sczyrba et al., 2017). Les plus performants se basent sur des graphes De Bruijn via une approche par k-mer (Ghurye et al. 2016) comme MegaHit (Li et al., 2015) ou Metaspades (Nurk et al., 2017). Les contigs ainsi assemblés peuvent alors soit être directement utilisés pour réaliser un profilage taxonomique des organismes dans l'échantillon, soit servir pour une approche de binning (Mande et al. 2012).

L'assemblage généré pour des données métagénomiques est souvent fragmenté. Le binning a pour but de regrouper entre eux les contigs appartenant à des organismes proches. Les techniques de binning sont capables de regrouper des contigs, et ceci même pour des espèces rares. De plus, elles sont capables de récupérer des génomes partiels et presque complets provenant d'espèces non cultivables. Il existe plusieurs approches de binning basées sur des algorithmes d'apprentissages supervisés dits « taxonomie dépendant » ou non supervisés dits « taxonomie indépendant ».

La première requiert l'utilisation de bases de données et d'outils de classification pour regrouper les contigs.

La deuxième approche utilise la différence d'abondance des contigs entre échantillons pour les regrouper comme MBBC (Wang et al., 2015) ou AbundanceBin (Wu et Yu, 2011). D'autres outils reposent sur la teneur en GC (Guanine et Cytosine) ou la fréquence en x-nucléotides au sein du même échantillon comme MetaWatt (Strous et al., 2012) ou VizBin (Laczny et al.,2015). Cependant les meilleurs résultats s'obtiennent (Yue et al. 2020) avec des programmes mélangeant les deux approches précédentes c'est le cas de CONCOCT (Alneberg et al., 2014), MetaBat (Kang et al., 2019) ou GroopM (Lmelfort et al., 2014).

Les contigs assignés à un même taxon forment des génomes assemblés de métagénomes (MAGs).

Contigs ou MAGs peuvent ensuite être utilisés pour la détection des ORFs avec Prodigal pour les procaryotes (Hyatt et al., 2010) et Augustus pour les eucaryotes (Stanke et al., 2008). Après annotation via l'utilisation d'outils d'alignement comme DIAMOND (Buchfink et al., 2021) ou Blast (Altschul et al., 1990), les séquences sont annotées fonctionnellement grâce à une recherche homologique dans des bases de données protéiques spécifiques telles que KEGG (Kanehisa et al., 2016) ou RefSeq (O'Leary et al., 2016).

L'approche par contigs et binding s'oppose à une deuxième approche dite de « libre d'assemblage » qui ne nécessite que les lectures brutes (pas d'assemblage). Dans ce cas l'assignation taxonomique peut se faire directement à l'aide d'une référence et permet de réaliser un profil quantitatif de la répartition des communautés présentes dans l'échantillon (Figure 1). Différentes méthodes existent. Les séquences peuvent être alignées contre une base de données comme GenBank en utilisant un logiciel d'alignement type Blast. D'autres approches ont tenté d'éviter d'aligner l'ensemble des lectures contre des bases de données complètes, tâche nécessitant beaucoup de temps. L'utilisation de marqueurs spécifiques comme des gènes conservés avec MetaPhlAn2 (Beghini et al., 2020) ou mOTU (Milanese et al., 2019) permettent d'accélérer l'assignation taxonomique grâce à de plus petites bases de données tout en conservant une grande précision. Des approches par k-mers ont aussi été employées avec des logiciels comme Kraken2 (Wood et al., 2019) ou Clark (Ouinit et al., 2015).

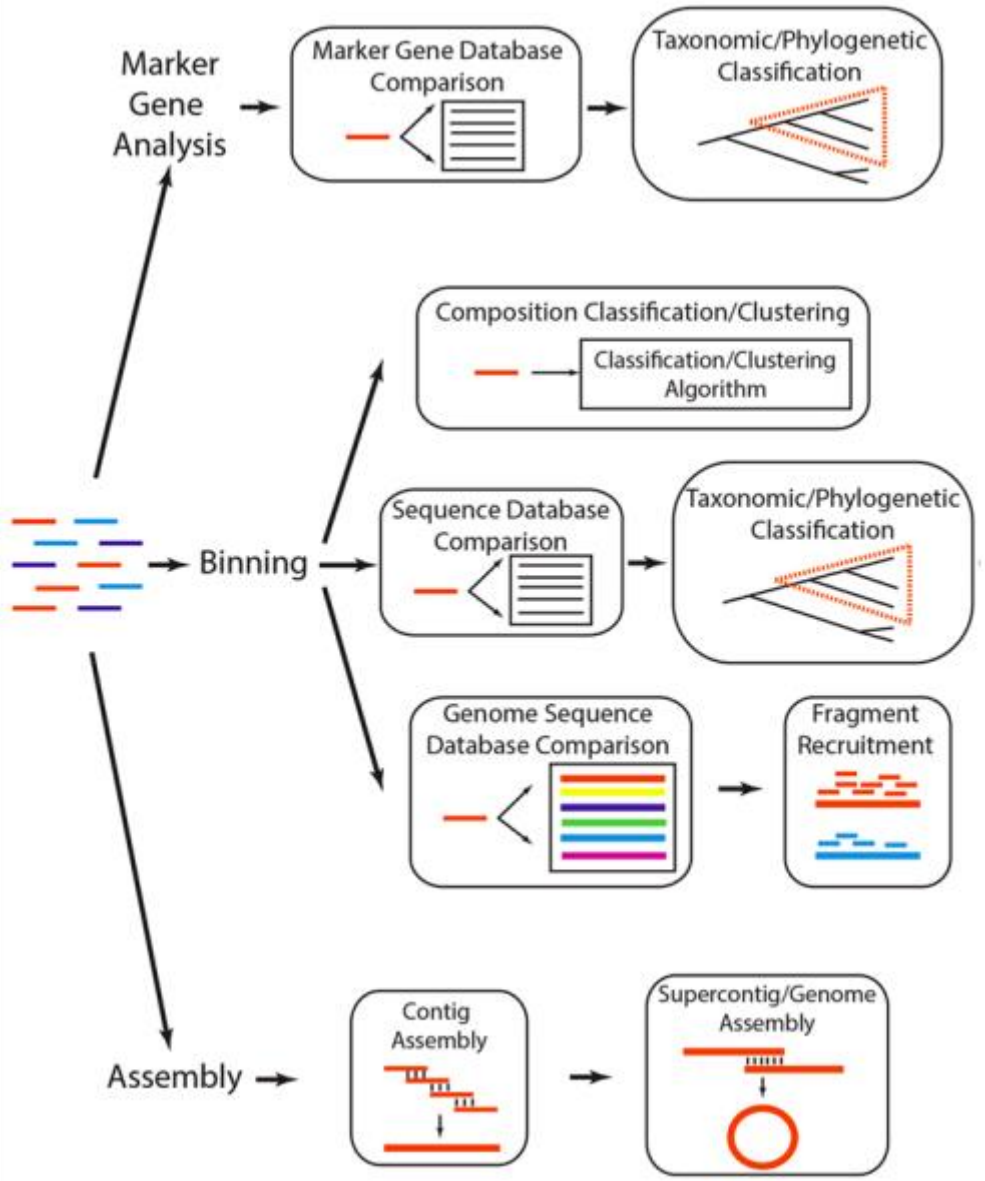

Figure 1 : modifiée d'après Sharpton (2014) : Différentes approches possibles par shotgun métagénomique : approche par libre assemblage avec des gènes marqueurs, approche par binning supervisée et non supervisée, approche par assemblage.

1.3. Objectif : caractérisation des micro-organismes d'un lac salé par métagénomique

La métagénomique est un domaine en plein essor et les données provenant d'échantillons d'eau de mer, d'océans, de rivières ou de lacs font l'objet de nombreuses publications ces dernières années (Grossert et al. 2019). Cependant l'identification des organismes reste complexe et repose sur des bases de données pour l'heure encore incomplètes, d'où la nécessité de caractériser des écosystèmes jusque-là encore inconnus.

L'objectif du présent stage est la caractérisation taxonomique et fonctionnelle de la population en micro-organismes d'un lac salé d'une superficie de 150 hectares, la Dayat Morsli ou Petit Lac d'Oran (Algérie) (Ben Bayer, 2019) durant deux saisons. Ce lac se trouve proche d'activités humaines qui ont entraîné la pollution de ses eaux pendant de nombreuses années. Avant d'être classé comme réserve naturelle par la municipalité et soumis à un programme de revitalisation et dépollution dans le début des années 2000. Le lac n'est depuis alimenté que par les eaux de pluie et le drainage des routes. En 2017 un projet visant à le réhabiliter pour en faire une zone de loisir a vu le jour.

Deux périodes ont été mises en évidence grâce aux paramètres physico-chimiques du lac (Ben Bayer et al. 2019), principalement expliquée par le niveau d'eau, le pH et la concentration en métaux lourds (Tableau 1). La saison humide, entre janvier et mai, contraste avec la période sèche, de juin à décembre, par un niveau d'eau ainsi qu'un pH plus élevé, mais aussi des variations dans la concentration en métaux lourds, particulièrement pour le cadmium et le plomb présents en concentration supérieure aux taux autorisés.

Tableau 1 : Paramètres physico-chimiques du lac Dayat Morsli en fonction des périodes sèches (juin à décembre) et humides (janvier à mai)

	Niveau d'eau (cm)	pH	Salinité (mg/L)	phosphates (mg/L)	Nitrate (mg/L)	Cadmium (ppm)	Cuivre (ppm)	Zinc (ppm)	Fer (ppm)	Plomb (ppm)
Période Humide	32± 12	9± 0,5	30± 1,7	4,4± 0,3	90 ± 7	0,07 ± 0,02	0,09 ± 0, 01	0,07 ± 0, 03	2,1 ± 0, 1	1,3 ± 0,2
Période Sèche	9,8± 3,2	8,3± 0,6	56	6	150	0,71 ± 0,05	0,2 ± 0,0 3	0,04 ± 0, 02	0,2± 0, 07	1,1 ± 0,2

2. Matériel et méthodes

L'ensemble des scripts développés au cours de ce stage est disponible à l'adresse suivante : http://pedago-service.univ-lyon1.fr:2325/thboyer/oran_lake_meta.

2.1. Matériel biologique

Les échantillons d'eau nécessaires à cette étude ont été collectés en triplicat sur sept sites autour du lac Dayat Morsli et ce chaque mois durant l'année 2017 (Figure 2). L'ADN génomique (ADNg) de chaque prélèvement a été extrait à l'aide du kit d'extraction du sol MACHEREY-NAGEL. Un traitement à la RNase a été effectué suivi d'une purification au Phénol-Chloroforme. La concentration et la qualité de l'ADNg (260/280) ont été estimées avec le Spectromètre NanoDrop (Thermo Scientific). Ces extractions et mesures ont été réalisées lors de la thèse de Wided Ben Bayer en 2018.

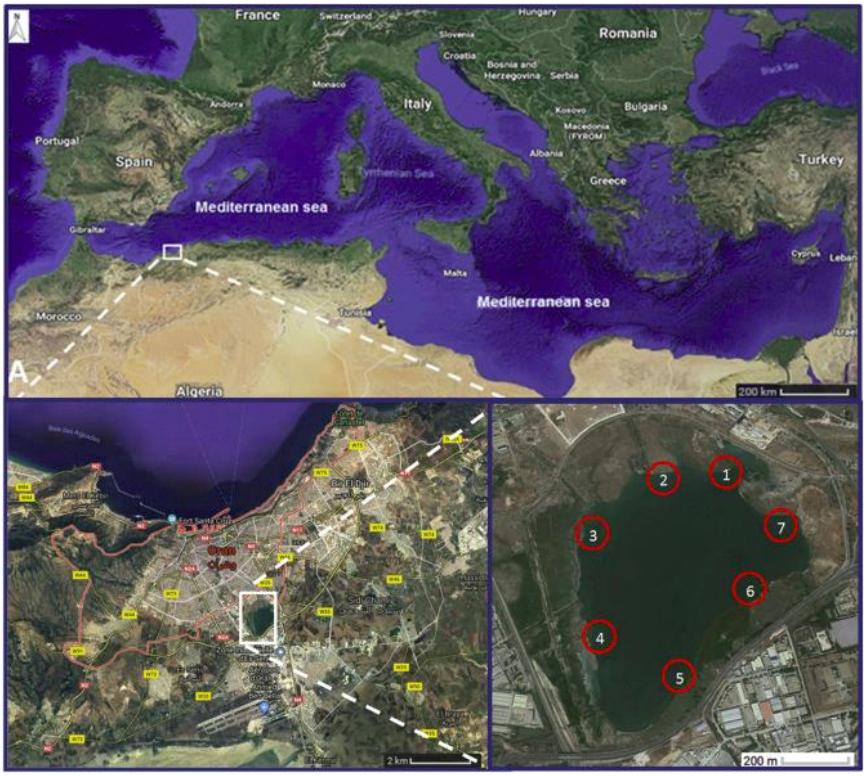

Figure 2 : d'après Ben Bayer et al. (2019) : Localisation du lac Dayat Morsli et des sites d'échantillonnages : A) La Dayat Morsli est située en Algérie, B) au sud-est d'Oran, sa superficie est de 150 hectares, C) Localisation sur le pourtour du lac des prélèvements d'eau mensuels ayant servi à extraire les ADNg totaux pour les analyses métagénomiques.

2.2. Données de séquençage

Dû à des concentrations en ADNg faibles, les prélèvements ont été regroupés en deux échantillons représentant la composition des communautés de la saison humide, échantillon H (janvier à mai) et de la saison sèche, échantillon S (juin à décembre). Le séquençage par shotgun métagénomique avec la méthode Illumina HiSeq 4000 (paired-end, 2x150 bp) a été réalisé par la plateforme Genewiz.

Le contrôle qualité des données a été réalisé par fastQC 0.11.9 (Figure 3). Les adaptateurs, les bases de mauvaise qualité et les lectures trop courtes ont été filtrés à l'aide du logiciel Trimmomatic 0.39. Une longueur minimale de 36 pb pour les lectures a été requise, les bases sur une fenêtre glissante de 4 ayant un score phred moyen inférieur à 25 ont été retirées ce qui équivaut à une probabilité d'erreur par base entre 99 % et 99,9 %.

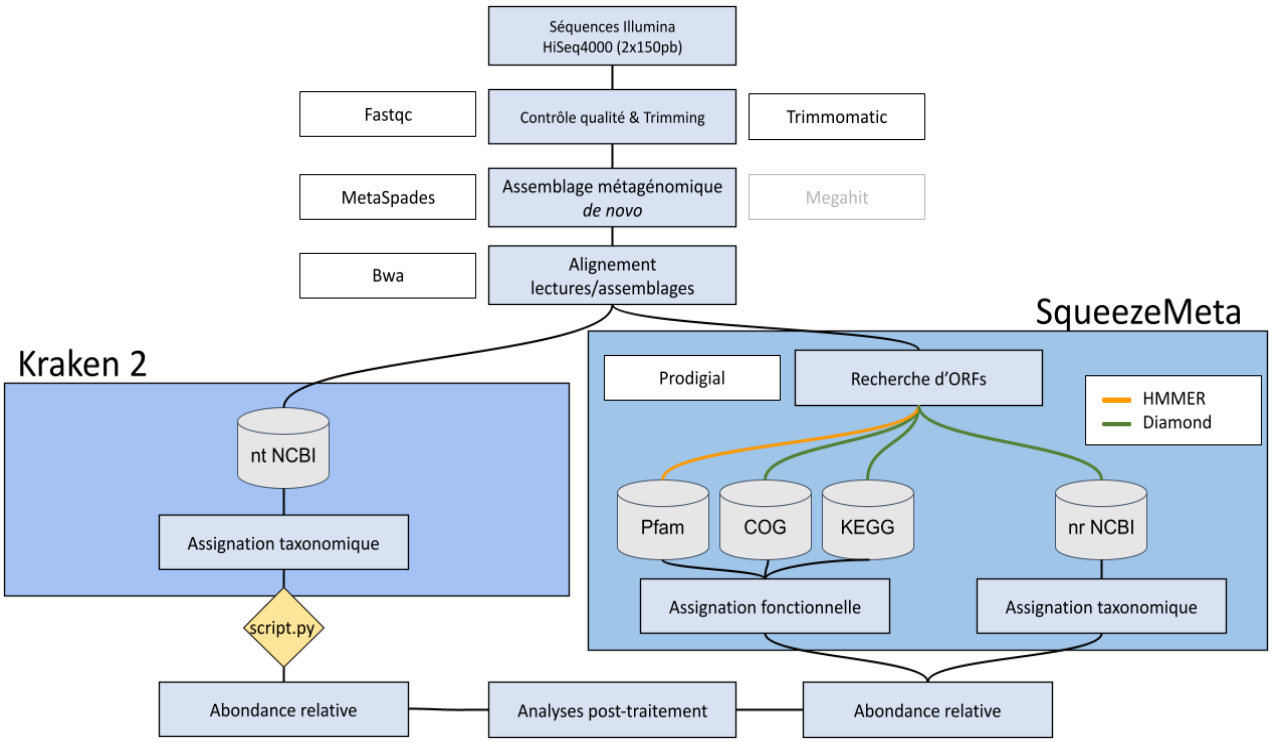

Figure 3 : Schéma général reprenant les différentes étapes du pipeline de traitement des données métagénomiques. L'analyse qualité des séquences est faite à l'aide de Trimmotic et de FastQC; L'assemblage métagénomique de novo repose sur la comparaison de deux logiciels : Metaspades et Megahit; L'alignement des lectures sur l'assemblage avec bwa permet d'obtenir la couverture de l'assemblage; Les contigs sont ensuite traités par deux logiciels différents : Kraken2 et Squeezemeta qui permettent d'obtenir l'assignation fonctionnelle et taxonomique; Après calcul de l'abondance, une analyse post-traitement est réalisée

2.3. Matériel informatique

L'ensemble des analyses réalisées durant ce stage ont été faites sur le core cluster de l'institut français de bioinformatique (IFB). Il fonctionne sous un environnement Linux et est géré par le système de gestion des ressources Slurm 19.05.8. Les différents logiciels présents sur le cluster ont été installés via le gestionnaire d'environnement Conda.

2.4. Outils pour l'assemblage *de novo*

L'assemblage des lectures *de novo* a été réalisé en comparant deux logiciels : Megahit 1.2.9 (Li et al., 2015) et Metaspades (Nurk et al., 2017) une extension de Spades 3.14.1. Megahit a été désigné comme l'un des meilleurs assembleurs métagénomiques (top 3) selon les comparatifs réalisés lors du CAMI challenge (Sczyrba et al., 2017).

Spades a été développé en premier lieu pour traiter les données bactériennes ou single-cell avec une profondeur de séquençage non uniforme. L'ajout d'un nouvel algorithme lui permet de traiter les données métagénomiques efficacement. De plus, son approche est connue pour générer des contigs de plus grande longueur que les autres assembleurs. Plusieurs paramètres ont été testés afin de choisir le meilleur assemblage, le choix de ces derniers est explicité dans la partie 3.2.

Le logiciel Quast 5.0.2 (Gurevich et al., 2013) a été utilisé pour s'assurer de la qualité des assemblages en générant les différentes métriques de qualité. L'alignement des lectures brutes sur les contigs s'est fait à l'aide du logiciel Bwa MEM 0.7.17 (Li., 2013). Samtools 1.10 (Li et al., 2009) avec l'option flagstats a été utilisé pour obtenir les statistiques liées au taux de couverture, c'est-à-dire la longueur totale séquencée rapportée à la taille du génome.

De plus, l'utilisation de la commande coverage après tri des alignements permet de détecter puis d'éliminer les contigs dont la profondeur moyenne est inférieure à 1X ou dont le nombre de lectures brutes alignées dessus est inférieur à 5. La profondeur étant le nombre de fois où une base est séquencée, si ce paramètre est inférieur à 1X alors il y a de forte probabilité que le contig généré ne soit qu'un artefact. Ne conserver que les contigs dont le nombre de lectures brutes alignées est supérieur à 5 permet d'éviter de garder de petits contigs avec une trop faible couverture.

La combinaison de ces deux paramètres permet d'enlever au maximum les artefacts, sans pour autant être trop strict pour conserver les séquences appartenant à des espèces rares (peu présentes dans les données). La filtration des séquences a été réalisée grâce à seqkit 0.14.0 (Shen et al., 2016).

2.5. Outils pour l'assignation taxonomique

2.5.1. Kraken

Kraken2 2.0.9beta (Wood et al., 2019) est un logiciel d'assignation taxonomique développé en premier lieu pour une utilisation sur des lectures courtes. Cependant, son utilisation sur des contigs a montré une amélioration de la précision, tout en maintenant un rappel similaire (Tran et Phan, 2020). La précision étant la capacité à trouver les taxons pertinents sur l'ensemble des taxons retrouvés ; le rappel correspond au ratio de taxons pertinent trouvé sur l'ensemble des taxons pertinents existants.

L'utilisation de Kraken2 sur contigs s'est faite avec la taxonomie générée avec la base de données nucléotidiques "nt" de NCBI téléchargée le 29/01/21. Cette base de données contient les séquences issues des divisions traditionnelles de GenBank, EMBL (European Molecular Biology Laboratory) et DDBJ (DNA Data Bank of Japan). Les divisions "bulk" (gss, sts, pat, est, htg) ainsi que wgs n'ont pas été incluses.

2.5.2 SqueezeMeta

Contrairement à Kraken2, SqueezeMeta 1.4.0beta (Tamames et Puente-Sánchez, 2019) est un pipeline d'analyse de données métagénomiques qui base son approche sur les séquences codantes présentes dans les contigs. SqueezeMeta assigne un taxon à un contig via plusieurs étapes (Figure 4).

Tout d'abord, Prodigal permet de prédire les gènes présents sur la séquence. Puis les gènes prédits sont comparés à l'aide de Diamond contre la base de données nr de Genbank téléchargée le 25/09/20. Toutes les solutions dont le son pourcentage d'identité est supérieur à 50 % et avec une e-value inférieure à 1e-03 sont retenues.

L'étape d'assignation taxonomique se fait tout d'abord sur les gènes. SqueezeMeta recherche le plus petit ancêtre commun (LCA - Lowest Common Ancestor) à partir des résultats de Diamond. Le LCA est le plus petit rang taxonomique commun à un ensemble de solutions. Le LCA s'applique à la solution optimale et à toutes les solutions dont le bit-score est d'au moins 80% et le score d'identité d'au plus 10% divergeant de la solution optimale.

Le bit-score est une matrice de score qui mesure la similarité entre deux alignements de séquences, plus ce score est grand, plus les séquences sont similaires.

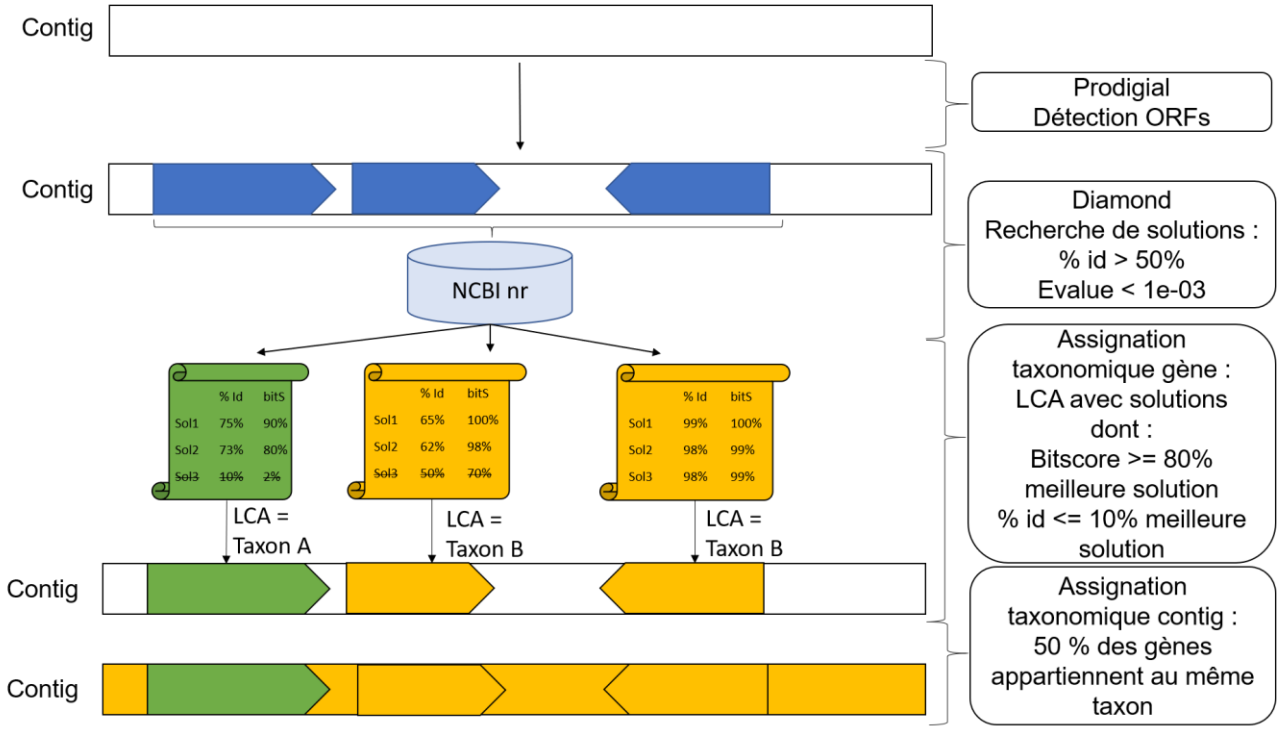

Figure 4 : Schéma explicatif de la méthode d'assignation taxonomique des contigs par Squeezemeta ; Prodigial permet de détecter les ORFs présents sur le contig ; Diamond effectue une recherche de solutions, seules celles passant le filtre sont conservées ; L'assignation taxonomique est réalisée grâce à la recherche du LCA sur les solutions dont le bitscore et le pourcentage d'identité sont proches de la meilleure solution ; L'assignation taxonomique du contig est réalisé si 50% des gènes appartiennent au même taxon

SqueezeMeta recherche le LCA pour un ensemble de solutions. Cependant, un taxon ne peut être attribué à un rang spécifique si un pourcentage d'identité précis n'est pas atteint. Il est de 85, 60, 55, 50, 46, 42, et 40 % pour respectivement : l'espèce, le genre, la famille, l'ordre, la classe et le phylum et le règne. Ce qui veut dire qu'une séquence ne pourra être classée si aucune des solutions trouvées par Diamond n'excède 40 % d'identité. Tous ces paramètres permettent d'éviter que l'assignation d'un gène ne se fasse avec des solutions de faible qualité et peu probables.
Vient ensuite l'étape d'annotation des contigs qui se fait par consensus de l'assignement fait sur les gènes. Si 50 % des gènes d'un contig appartiennent à un même taxon alors ce contig est classé comme tel. Ce pourcentage permet d'inclure d'éventuels transferts horizontaux sans pour autant modifier l'assignation.

2.6. Outils pour l'assignation fonctionnelle

SqueezeMeta, en plus de réaliser l'assignation taxonomique, permet d'obtenir une annotation fonctionnelle d'un échantillon. Cette partie de l'analyse est réalisée avec les gènes prédits par Prodigal et Diamond par homologie contre deux autres bases de données : KEGG (Kanehisa, 2000), la dernière version publique d'accès, afin d'obtenir les annotation KEGG ID et COG (Huerta-Cepas et al., 2019) pour les annotations COG. Toutes deux ont été téléchargées le 25/09/20. SqueezeMeta permet également une classification des gènes grâce à la base de données PFAM (Finn et al., 2014), téléchargées le 25/09/20, à l'aide du logiciel HMMER3 (Eddy, 2009).

2.7 Abondance

Pour Kraken, l'abondance relative est calculée à partir de l'alignement des lectures contre l'assemblage. L'abondance se calcule en utilisant la profondeur moyenne, et non via le nombre de lectures par contigs, afin de la normaliser. Afin d'obtenir l'abondance en associant la profondeur moyenne du contig à son rang taxonomique ainsi qu'à l'ensemble du taxon auquel il appartient, un script a été développé au cours du stage.

Squeezemeta réalise de son côté l'alignement des lectures simples sur l'assemblage à l'aide de Bowtie2. Cet alignement permet ensuite d'obtenir de la même manière l'abondance relative pour les gènes et les contigs.

2.8 Analyses post-traitement et représentation graphique

Les données sont ensuite traitées à l'aide de logiciel RStudio 4.0.3 et les packages ggplot2 et pavian pour l'analyse de la taxonomie issue de Kraken2. Pour l'exploration des données, Squeezemeta fournit le package SQMtools.

3. Résultats

3.1. Contrôle qualité et élagage des données de séquençage

Tableau 2 : Résultats obtenues à l'aide de Trimmomatic

Résultats Trimmomatic			
Échantillon H		Échantillon S	
Lectures appariées en entrées	216056171	Lectures appariées en entrées	139474651
Lectures appariées survivantes	**195126953**	**Lectures appariées survivantes**	**121725241**
Pourcentage lectures appariées survivantes	**90,31**	**Pourcentage lectures appariées survivantes**	**87,27**
Lectures avant survivantes seulement	14524512	Lectures avant survivantes seulement	129622653
Pourcentage lectures avant survivantes seulement	6,72	Pourcentage lectures avant survivantes seulement	9,29
Lectures inverses survivantes seulement	3546355	Lectures inverses survivantes seulement	1940999
Pourcentage lectures inverses survivantes seulement	1,64	Pourcentage lectures inverses survivantes seulement	1,39
Lectures éliminées	2858350	Lectures éliminées	2845757
Pourcentage lectures éliminées	1,32	Pourcentage lectures éliminées	2,04

Afin de s'assurer de la bonne qualité des lectures avant assemblage, un premier filtre est appliqué à l'aide de Trimmomatic. Pour le premier échantillon H, 9,89 % des lectures ne peuvent pas être utilisées pour l'assemblage. Pour l'échantillon S, 12,73 % des lectures sont écartées du futur assemblage (Tableau 2).

Le rapport généré par le logiciel FastQC permet d'estimer si la qualité des séquences obtenue avant et après utilisation de Trimmomatic est satisfaisante.

3.2 Comparaison des assemblages *de novo*

3.2.1 Megahit

Plusieurs paramètres ont été testés afin de sélectionner le meilleur assemblage obtenu avec Megahit. L'échantillon H étant le plus volumineux en termes de lectures et donc de taille de fichier, c'est ce dernier qui a été choisi pour réaliser les tests. Si son analyse se déroule sans erreur alors l'analyse de l'échantillon S le sera aussi.

Dans le but d'optimiser les étapes d'assignation taxonomique et fonctionnelle, différents paramètres ont été testés (Tableau 3). La priorité a été donnée à l'assemblage générant les plus grands contigs et en plus grand nombre. En effet, ces étapes sont plus

performantes avec des contigs de grande taille, soit une taille supérieure à 1000 pb (Tamames et al., 2019 ; Tran et Phan, 2020).

Tableau 3 : Liste des paramètres utilisés pour les différents assemblages en utilisant Megahit

Assemblages	Megahit_1	Megahit_2	Megahit_3	Megahit_4
Paramètres	--k-min=27 --k-max=91 --k-step=10 --min-contig-len 200	--k-min=21 --k-max=91 --k-step=10 --max-tip-len 200 --min-contig-len 300	--k-min=27 --k-max=141 --k-step=10 --min-contig-len 200	--k-min=21 --k-max=141 --k-step=10 --min-contig-len 200

Les paramètres de l'assemblage Megahit_4 correspondent aux paramètres par défaut du logiciel. L'augmentation du paramètre k-min à 27 pour l'assemblage Megahit_1 et 3 est recommandée pour diminuer la complexité du graphe de Bruijn pour des échantillons complexes. Les paramètres utilisés pour l'assemblage Megahit_2 reprennent les paramètres ayant produit les meilleurs résultats avec Megahit lors du CAMI challenge. L'option max-tip-len spécifiée permet d'enlever les "tips", des chaînes de nœuds non connectées à l'une de leur extrémité, inférieur à 200 pb. Cette option permet de simplifier le graphe. Enfin, les paramètres utilisés pour l'assemblage Megahit_1 reprennent ceux du CAMI challenge mais en augmentant k-min à 27. Le paramètre k-step = 10 est utilisé par défaut par le logiciel et permet de mieux prendre en compte les séquences avec une faible profondeur de séquençage.

Les paramètres utilisés avec Megahit_3 produisent les meilleurs résultats (Tableau 4). Cet assemblage crée le plus grand nombre de très grands contigs, supérieurs à 10 000 pb, tout comme le plus grand contig, 712 154 pb. Le nombre total de contigs supérieurs à 1 000 s'élève à 357 271, soit 369 de moins qu'obtenu avec l'assemblage en produisant le plus grand nombre : Megahit_1. Enfin, le N50 mesure la fragmentation du génome assemblé en contigs. Plus le N50 est élevé, moins le génome est fragmenté. Il est de 1 279 pour Megahit_3 soit la deuxième plus élevée.

Pour toutes ces raisons, l'assemblage réalisé avec les paramètres Megahit_3 est considéré comme le meilleur.

Tableau 4 : Résultats obtenus avec Quast pour les assemblages de Megahit avec différents paramètres. Les valeurs en gras représentent les critères optimaux recherchés pour chaque catégorie. L'assemblage numéro 3, présente les meilleurs résultats avec un nombre de grands contigs, supérieur à 1000 pb, plus élevé, un N50 très proche de celui obtenu avec Megahit_2, ainsi que le plus grand contig produit.

	Echantillon H							
	Megahit_1		Megahit_2		**Megahit_3**		Megahit_4	
Nb Contigs > 1000 pb	**357 640**		344 941		357 271		329 185	
Plus grand contig	615 340		412 670		**712 154**		691 541	
GC (%)	49,1		48,55		**49,14**		48,42	
N50 (Pour contigs > 500bp)	1255		**1280**		1279		1192	
Contigs < 1000	89,80 %	3 147 507	**86,46 %**	**2 201 713**	87,86 %	2 585 809	88,34 %	2 494 248
1000 < Contigs < 5000	9,59 %	**336 095**	**12,76 %**	325 070	11,38 %	335 021	11,01 %	310 728
5000 < Contigs < 10000	0,41 %	14 505	**0,54 %**	13 679	0,50 %	**14 812**	0,45 %	12 666
10000 < Contigs < 25000	0,16 %	5530	0,19 %	4910	**0,20 %**	**5808**	0,16 %	4604
25000 < Contigs < 50000	0,03 %	1077	0,04 %	925	**0,04 %**	**1129**	0,03 %	858
Contigs > 50000	0,01 %	433	0,01 %	357	**0,02 %**	**501**	0,01 %	329
Total Contigs	3505147		2546654		2943080		2823433	

3.2.2 Metaspades

Tableau 5 : Liste des paramètres utilisés pour les différents assemblages en utilisant Metaspades

Assemblages	Metaspades_1	Metaspades_2	Metaspades_3
Paramètres	k = [21,33,55]	k= [27,37,47,57,67,77,87,91]	k= [27,37,47,57,67,77,87,97,107,117,127]

Les paramètres sélectionnés (Tableau 5) essaient de reprendre ceux utilisées avec Megahit. Metaspades_1 correspond au programme par défaut. Metaspades_2 reprend les paramètres de Megahit_1 avec un k minimal de 27 et maximal de 91. Metaspades_3 reprend les paramètres de Megahit_3, soit un k minimal de 27 et maximal de 127 (maximum possible avec Metaspades).

Tableau 6 : Résultats obtenus avec Quast pour les assemblages de Metaspades avec différents paramètres. Les valeurs en gras représentent les critères optimaux recherchés pour chaque catégorie.

	Echantillon H					
	Metaspades_1		Metaspades_2		Metaspades_3	
Nb Contigs > 1000 pb	602 128		**771 628**		770 432	
Plus grand contig	661 746		844 155		**1 134 719**	
GC (%)	49,92		50,02		50,36	
N50 (Pour contigs > 500bp)	**1658**		1298		1333	
Contigs < 1000	94,78%	10 937 559	88,70%	6 059 354	88,06%	**5 681 312**
1000 < Contigs < 5000	4,79%	553 050	10,79%	**737 063**	**11,39%**	734 863
5000 < Contigs < 10000	0,27%	**31 426**	0,37%	25 275	**0,40%**	25 658
10000 < Contigs < 25000	0,11%	**12 999**	0,10%	7104	**0,12%**	7 482
25000 < Contigs < 50000	**0,03%**	**3098**	0,02%	1366	0,02%	1 516
Contigs > 50000	0,01%	**1555**	0,01%	820	0,01%	913
Total Contigs	11 539 687		6 830 982		6 451 744	

Les résultats obtenus avec l'assemblage Metaspades_3 sont globalement les meilleurs. Bien qu'il ne produise pas le meilleure N50 (Metaspades_1 : 1658), ni le plus grand nombre de grands contigs supérieur à 1000 pb (Metaspades_2 : 737 063). L'assemblage Metaspades_3 produit un plus grand nombre de très grands contigs que Metaspades_2 avec un N50 supérieur (1333) et cela, pour un assemblage général moins fragmenté (5 681 312 contigs inférieurs à 1000 pb) que Metaspades_1 et 2. Enfin, il génère le contig le plus grand (1 134 719 pb).
Pour toutes ces raisons, l'assemblage issus des paramètres de Metaspades_3 est retenu.

3.2.3 Megahit vs Metaspades

La comparaison entre les assemblages obtenus avec Megahit et Metaspades (Tableau 7) se base sur les mêmes critères que pour la comparaison des assemblages d'un même logiciel. Cependant, un critère important vient s'ajouter à cette comparaison : la conservation de l'information contenue dans les données brutes par l'assemblage. Plus le pourcentage de lectures alignées sur ce dernier est élevé, plus l'information génétique est conservée entre les contigs et les données brutes, et plus l'assemblage est représentatif des micro-organismes présents dans l'environnement. Avec un pourcentage d'alignement des lectures de 85 % ou plus, l'assignation taxonomique sera meilleure que si elle l'avait été avec les lectures seules (Tamames et al., 2019) et de ce fait capturera mieux la diversité présente dans les données.

Pour l'ensemble des paramètres, l'assemblage réalisé avec Metaspades est meilleur que celui réalisé avec Megahit. Le nombre de contigs supérieurs à 1 000 pb est 2,1 fois supérieur, particulièrement pour les contigs compris entre 1 000 et 5 000 pb (2,1X) et les contigs supérieurs à 50 000 pb (1,8X). La mesure de fragmentation du génome, le N50, est elle aussi plus élevée : 1 333 pour Metaspades contre 1 279. Enfin, le pourcentage total de lectures s'alignant sur l'assemblage réalisé par Metaspades est de 95,43 % contre 90,68 % pour Megahit. Cependant, l'assemblage généré par Metaspades produit 2,1 fois plus de contigs de petite taille, inférieure à 1 000 pb.

Tableau 7 : Comparaison des résultats obtenus avec Quast pour les meilleurs assemblages de Megahit et Metaspades. Les éléments en gras correspondent aux valeurs les plus pertinentes de leur catégorie. Metaspades obtient globalement de meilleures performances

	Megahit		Metaspades		Ratio Metaspades/Megahit
Nb Contigs > 1000 pb	357 271		**770 432**		2,1
Plus grand contig	712 154		**1 134 719**		1,6
GC (%)	49,14		**50,36**		/
N50 (Pour contigs > 500bp)	1279		**1333**		/
Contigs < 1000	**87,86%**	**2 585 809**	88,06%	5 681 312	2,1
1000 < Contigs < 5000	11,38%	335 021	11,39%	**734 863**	2,1
5000 < Contigs < 10000	0,50%	14 812	0,40%	**25 658**	1,7
10000 < Contigs < 25000	0,20%	5 808	0,12%	**7 482**	1,2
25000 < Contigs < 50000	0,04%	1 129	0,02%	**1 516**	1,3
Contigs > 50000	0,02%	501	0,01%	**913**	1,8
Total Contigs	100,00 %	2 943 080	100,00 %	6 451 744	2,2
	Statistiques d'alignement				
% mapped	90,68		**95,43**		/
% mapped properly paired	79,43		**85,83**		/
% mapped singleton	1,59		**0,99**		/

3.2 Affiliation Taxonomique

Préalablement à l'assignation taxonomique, les contigs, dont la profondeur et le nombre de lectures alignées n'était pas suffisant, ont été retirés. Cela représente 19 482 contigs pour l'échantillon H et 14 012 contigs pour l'échantillon S.

3.2.1 Kraken2

Pour l'échantillon H, un peu moins de la moitié des contigs a été classifiée par Kraken2, soit 54,06% des séquences contre 49,56% pour l'échantillon S (Tableau 8). Les contigs non classifiés sont des séquences pour lesquelles Kraken2 n'a pas pu trouver la présence de k-mer exact entre sa base de données et la séquence.

Tableau 8 : pourcentage de contigs classifiées et non classifiées pour les échantillons S et H. Kraken2 permet de classifier environ la moitié des contigs.

	Séquences classifiées		Séquences non classifiées	
	Nombres contigs	Pourcentage	Nombres contigs	Pourcentage
Échantillon H	3 488 085	54,06%	2 949 647	45.94%
Échantillon S	3 366 318	49,56%	3 406 771	50.44%

L'assignation taxonomique des contigs par Kraken2 montre que la majorité des séquences retrouvées dans les deux échantillons appartiennent au domaine des bactéries (Figure 5).

Le domaine eucaryote est le deuxième plus représenté et ce pour les deux saisons. Pour le domaine archée, un plus grand nombre de séquences lui est attribué pour la saison sèche que pour la saison humide. Les séquences virales sont quant à elles équivalentes entre les échantillons.

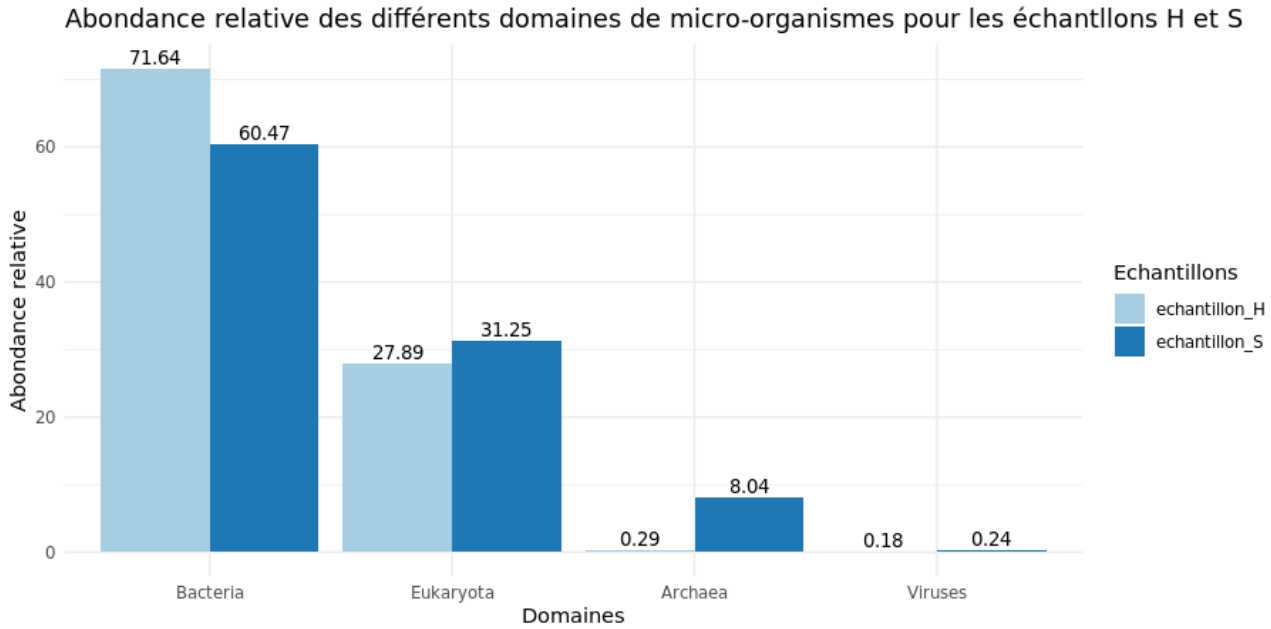

Figure 5 : Pourcentage de contigs assignés pour chaque échantillon en fonction du domaine taxonomique dans chaque échantillon.

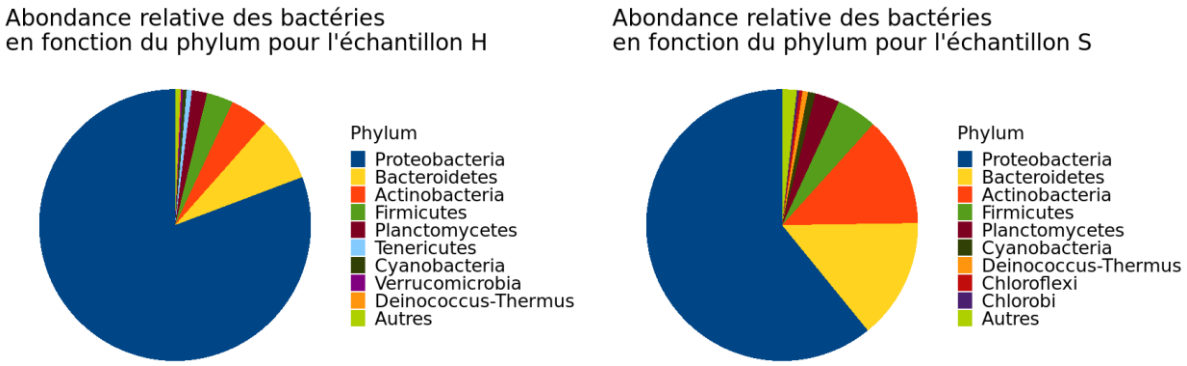

Figure 6 : Assignation taxonomique des contigs bactériens en fonction du phylum pour les échantillons H et S

Les principaux phyla bactériens (Figure 6) sont les mêmes entre les échantillons à l'exception des Tenericute et Verrucomicrobia plus présents lors de la saison humide et des Chloroflexi et Chlorobi plus présents durant la saison sèche. Néanmoins, l'abondance en Protéobactéries est plus importante pour l'échantillon H : 80,75 %; que dans l'échantillon S : 60,87 % (Tableau 9). L'abondance des bactéries de la catégorie "Autres" augmente dans l'échantillon S par rapport à l'échantillon H (1,75 % contre 0,58 %).

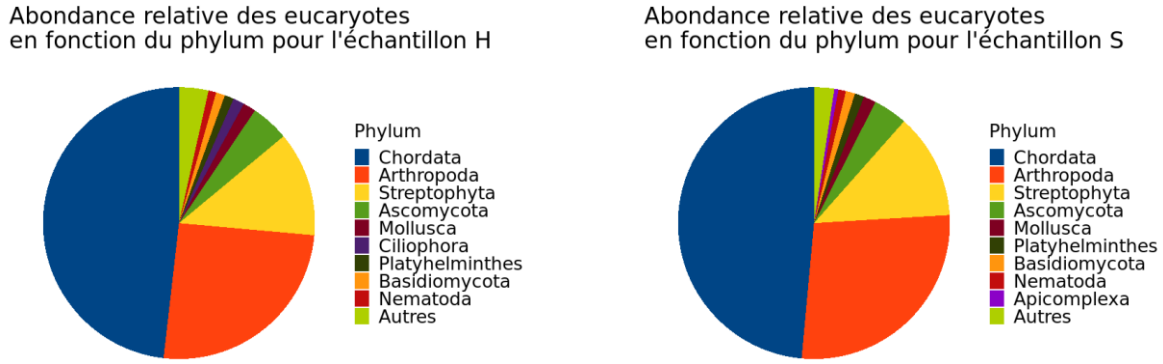

Figure 7 : Assignation taxonomique des contigs eucaryotes en fonction du phylum pour l'échantillon H et S

La composition des phyla eucaryotes est très similaire en termes d'abondance et de composition entre les deux échantillons (Figure 7). Les phyla majoritaires sont les Chordata (48,19 % et 48,55 %), Arthropoda (25,38 % et 27,4 %) et Streptophyta (12,46 % et 12,52 %) (Tableau 9). Ces trois phyla représentent respectivement 86,03 % des phyla eucaryotes de l'échantillon H et 88,87 % dans l'échantillon S.

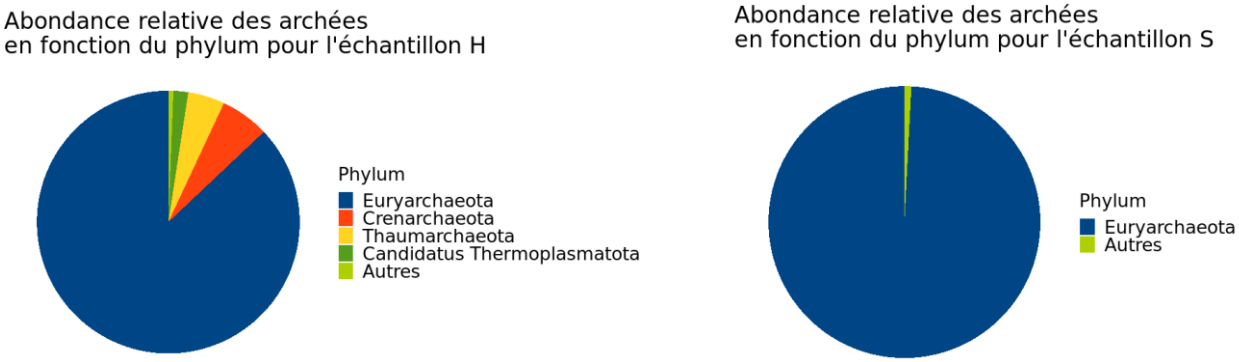

Figure 8 : Assignation taxonomique des contigs archéens en fonction du phylum pour les échantillons H et S

Durant la saison humide, la diversité des archées est caractérisée par 4 phyla principaux : Euryarchaeota, Crenarchaeota, Thaumarchaeota et Candidatus Thermoplasmatota (Figure 8). Durant la saison sèche, le phylum Euryarchaeota représente à lui seul 99,18 % des séquences archéennes (Tableau 9).

Tableau 9 : Récapitulatif de l'abondance et des pourcentage d'abondance obtenus pour les différents phyla avec Kraken2 pour les bactéries, archées et eucaryotes.

Bactéries	Échantillon H	Phylum	Proteobacteria	Bacteroidetes	Actinobacteria	Firmicutes	Planctomycetes	Tenericutes	Cyanobacteria	Verrucomicrobia	Deinococcus-Thermus	Autres
		Abondance	13 441 837	1 288 570	755 487	527 400	301 663	104 658	70 663	39 970	19 502	96 448
		Pourcentage	80,75	7,74	4,54	3,17	1,81	0,63	0,42	0,24	0,12	0,58
	Échantillon S	Phylum	Proteobacteria	Bacteroidetes	Actinobacteria	Firmicutes	Planctomycetes	Cyanobacteria	Deinococcus-Thermus	Chloroflexi	Acidobacteria	Autres
		Abondance	5 213 157	1 230 625	1 119 235	413 335	247 768	77 470	55 447	34 308	22 832	175 500
		Pourcentage	60,87	14,37	13,07	4,83	2,89	0,90	0,65	0,40	0,27	1,75
Eucaryotes	Échantillon H	Phylum	Chordata	Arthropoda	Streptophyta	Ascomycota	Mollusca	Cillophora	Platyhelminthes	Basidiomycota	Nematoda	Autres
		Abondance	2 532 219	1 333 369	654 861	237 655	79 696	70 605	57 576	56 414	48 346	183 697
		Pourcentage	48,19	25,38	12,46	4,523	1,517	1,344	1,096	1,074	0,9201	3,4959
	Échantillon S	Phylum	Chordata	Arthropoda	Streptophyta	Ascomycota	Mollusca	Platyhelminthes	Basidiomycota	Nematoda	Apicomplexa	Autres
		Abondance	1 894 218	1 068 884	488 590	157 973	57 178	44 199	42 945	34 685	19 211	93 638
		Pourcentage	48,55	27,4	12,52	4,049	1,465	1,133	1,101	0,889	0,4924	2,4

Archées	Échantillon H	Phylum	Euryarchaeota	Crenarchaeota	Thaumarchaeota	Candidatus Thermoplasmatota	Autres
		Abondance	59 407	4086	3090	1238	424
		Pourcentage	87,05	5,987	4,528	1,814	0,621
	Échantillon S	Phylum	Euryarchaeota	Autres			
		Abondance	1 219 138	10 079			
		Pourcentage	99,18	0,82			

3.2.2 SqueezeMeta

SqueezeMeta classifie taxonomiquement 69,79 % de contigs pour l'échantillon H et 81,43 % des contigs dans l'échantillon S (Tableau 10). Cependant, à partir du rang classe, plus de contigs sont classifiés pour l'échantillon H : 49,22 % des contigs; que pour l'échantillon S : 41,23 %. Les séquences non classifiées sont soit des contigs sans gènes, soit des séquences inconnues.

La disparité mesure pour un contig le nombre de paires de gènes appartenant à deux taxons différents au rang consensus, divisé par le nombre de paires possibles. La disparité mesurée est de 0,1 % des séquences pour les deux échantillons.

Tableau 10 : Répartition des séquences classifiées en fonction du rang pour les échantillons S et H et proportion de contigs ayant une disparité supérieure à 0.

	Echantillon H		Échantillon S	
	Nombre contigs	Pourcentage	Nombre contigs	Pourcentage
Contigs au rang domaine (k)	4 498 728	69,79	5 527 640	81,43
Contigs au rang phylum (p)	3 792 758	58,84	4 246 896	62,56
Contigs au rang classe (c)	3 172 578	49,22	2 798 716	41,23
Contigs au rang ordre (o)	2 550 388	39,56	1 968 225	28,99
Contigs au rang famille (f)	2 215 529	34,37	1 570 909	23,14
Contigs au rang genre (g)	1 297 597	20,13	965 646	14,23
Contigs au rang espèces (s)	494 711	7,67	416 389	6,13
Congruence	6 445 783	99,9	6 787 869	99,9
Disparité > 0	5962	0,1	4703	0,1

L'extrême majorité des séquences classifiées sont attribuables au domaine des bactéries. Elles représentent 95,84 % et 90,35 % des lectures, respectivement pour la période humide et sèche (Figure 9). Les eucaryotes représentent 3,98 % des séquences de l'échantillon H et 1,57 % de l'échantillon S. Les archées sont quasiment absentes de l'échantillon H, avec 0,13 % de séquences mais beaucoup plus présentes dans l'échantillon S avec 7,85 % des lectures. Enfin, les séquences virales sont présentes en faible proportion dans les deux échantillons.

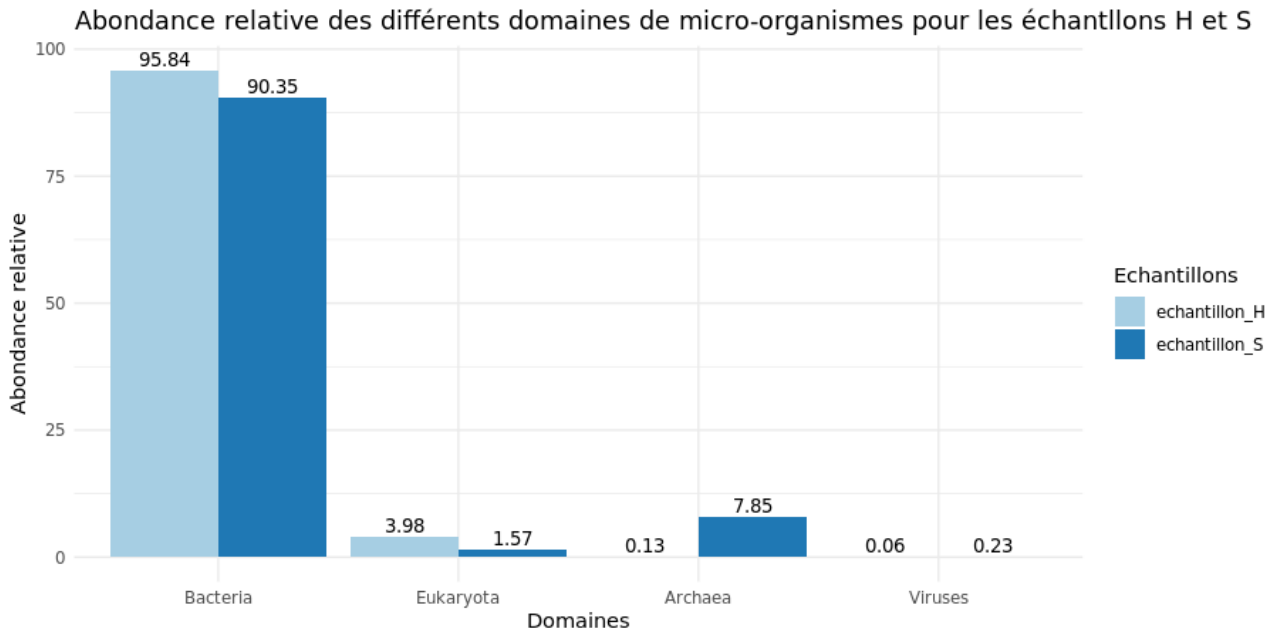

Figure 9 : Pourcentage de contigs assignés pour chaque échantillon en fonction du domaine taxonomique

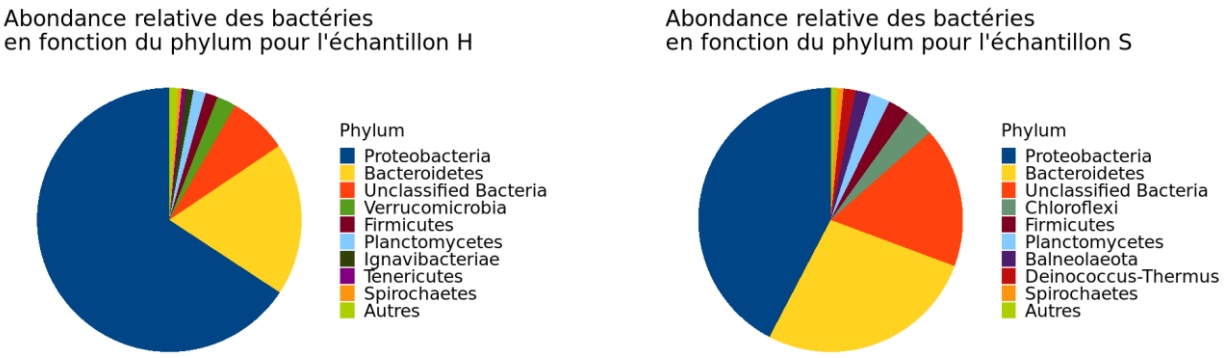

Figure 10 : Assignation taxonomique des contigs bactériens en fonction du phylum pour les échantillons H et S

Les phyla bactériens classifiés par Squeezemeta sont relativement similaires entre les échantillons (Figure 10). Cinq des six premiers phyla sont communs, seuls les Verrucomicrobia sont remplacés par les Chloroflexi dans l'échantillon S. Au total ces 6 premiers phila représentent la majorité de l'abondance avec respectivement 97,05 % et 95,18 % pour les prélèvements de la saison humide et sèche (Tableau 11). L'abondance relative des Proteobacteria varie le plus entre les deux saisons avec une diminution de 23,39 points. La part des bactéries non attribuables à un phylum est plus élevée dans l'échantillon S avec une abondance relative de 17,19 % contre 7,22 % dans l'échantillon H.

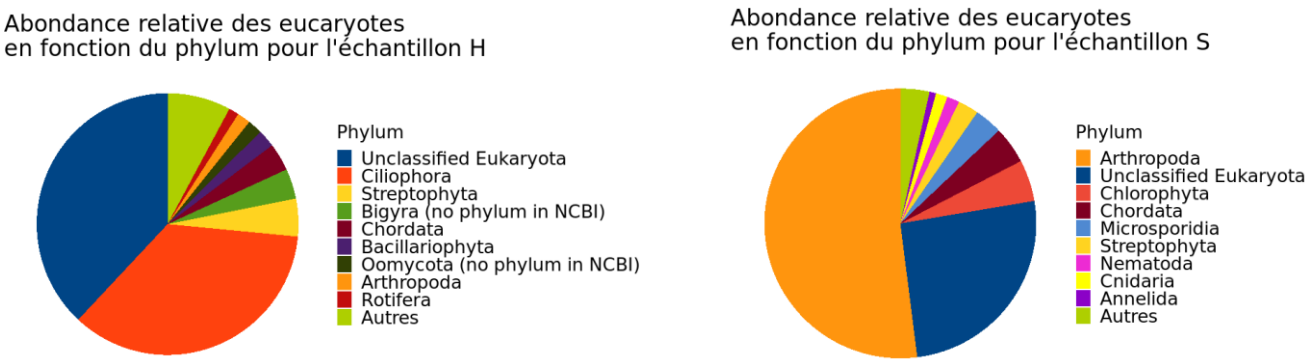

Figure 11 : Assignation taxonomique des contigs eucaryotes en fonction du phylum pour les échantillons H et S

Les phyla assignés aux Eucaryotes entre la saison humide et sèche varient grandement (Figure 11). Les arthropoda majoritairement présents dans l'échantillon S avec 51,99 % d'abondance (Tableau 11) diminue grandement dans l'échantillon H (1,65 %). La proportion de séquences eucaryotes non classifiées : 38,03 %; est la première en termes d'abondance dans ce même échantillon. Elles représentent, en proportion, le deuxième phylum dans l'échantillon S (25,69 %). La part des eucaryotes appartenant à un autre phylum que ceux majoritairement présents est 2,2 fois plus importante dans l'échantillon H (7,83 %) que dans l'échantillon S (3,45 %).

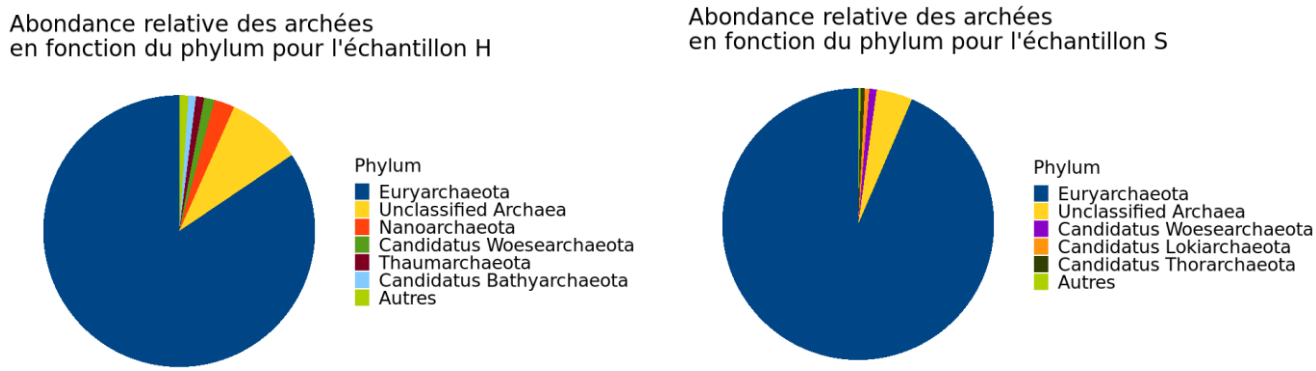

Figure 12 : Assignation taxonomique des contigs archéens en fonction du phylum pour les échantillons H et S

Le phylum archéen le plus abondant est Euryarchaeota pour les deux échantillons (Figure 12). Le deuxième phylum le plus abondant correspond aux archées non classifiées. Ces deux phyla représentent 93,36 % des séquences de l'échantillon H et 97,34 % des séquences de l'échantillon S (Tableau 11). Enfin, le phylum Candidatus woesarchaeota est commun aux deux échantillons, contrairement aux phyla Candidatus bathyarchaeota, nanoarcheota et thaumarchaeota spécifiques de l'échantillon H. Les phyla Candidatus lokiarchaeota et Candidatus thorarchaeota sont spécifiques de l'échantillon S.

Tableau 11 : Récapitulatif de l'abondance et des pourcentage d'abondance obtenus pour les différents phyla avec Squeezemeta pour les bactéries, archées et eucaryotes

Bactéries	Échantillon H	Phylum	Proteobacteria	Bacteroidetes	Unclassified Bacteria	Verrucomicrobia	Firmicutes	Planctomycetes	Ignavibacteriae	Tenericutes	Spirochaetes	Autres
		Nombre lectures	192 840 986	54 910 434	21 181 860	6 922 477	4 415 325	4 393 265	2 667 477	1 587 972	1 208 968	3 204 230
		Abondance	65,74	18,72	7,22	2,36	1,51	1,5	0,91	0,54	0,41	1,09
	Échantillon S	Phylum	Proteobacteria	Bacteroidetes	Unclassified Bacteria	Chloroflexi	Firmicutes	Planctomycetes	Balneolaeota	Deinococcus-Thermus	Spirochaetes	Autres
		Nombre lectures	59 523 771	37 928 974	24 156 159	5 009 289	3 663 855	3 494 405	2 509 188	2078447	1 117 996	1 074 878
		Abondance	42,35	26,98	17,19	3,56	2,61	2,49	1,79	1,48	0,8	0,76
Eucaryotes	Échantillon H	Phylum	Unclassified Eukaryota	Ciliophora	Streptophyta	Bigyra (no phylum in NCBI)	Chordata	Bacillariophyta	Oomycota (no phylum in NCBI)	Arthropoda	Rotifera	Autres
		Nombre lectures	4 625 391	4 310 613	567 036	452 865	417 501	266 128	217 355	200 333	152 122	952 626
		Abondance	38,03	35,44	4,66	3,72	3,43	2,19	1,79	1,65	1,25	7,83
	Échantillon S	Phylum	Arthropoda	Unclassified Eukaryota	Chlorophyta	Chordata	Microsporidia	Streptophyta	Nematoda	Cnidaria	Annelida	Autres
		Abondance	1 331 229	657 730	127 413	113 937	84 937	63 557	37 763	35 726	19 956	88 374
		Pourcentage	51,99	25,69	4,98	4,45	3,32	2,48	1,47	1,4	0,78	3,45

Archées	Échantillon H	Phylum	Euryarchaeota	Unclassified Archaea	Nanoarchaeota	Candidatus Woesearchaeota	Thaumarchaeota	Candidatus Bathyarchaeota	Autres
		Abondance	329 540	34 767	9495	4857	3829	3483	4249
		Pourcentage	84,45	8,91	2,43	1,24	0,98	0,89	1,09
	Échantillon S	Phylum	Euryarchaeota	Unclassified Archaea	Candidatus Woesearchaeota	Candidatus Lokiarchaeota	Candidatus Thorarchaeota	Candidatus Bathyarchaeota	Autres
		Abondance	11 955 921	548 175	110 172	68 622	63 887	31 295	5315
		Pourcentage	93,07	4,27	0,86	0,53	0,5	0,24	0,04

3.2.3 Kraken2 et Squeezemeta

La comparaison des phyla bactériens assignés par les logiciels Squeezemeta et Kraken2 montre des compositions similaires (Figure 13). Les Proteobacteria dominent les autres phyla avec les deux programmes. L'abondance plus importante des Proteobacteria pour les échantillons de la saison humide par rapport à la saison sèche est confirmée, +23 points avec Squeezemeta et +19,88 points avec Kraken2. Le deuxième phylum le plus abondant, celui des Bacteroidetes, confirme une diminution semblable entre saison sèche et humide, -6,63 % avec Kraken2 et -8,26 points avec Squeezemeta. Les Firmicutes, 4ème et 5ème en abondance, et les Planctomycetes, 5ème et 6ème en abondance pour Kraken2 et Squeezemeta respectivement, sont les deux autres phyla communs. Ils montrent une variation de leur abondance relative proche entre échantillons. Pour les Firmicutes, +1,66 point avec Kraken2 et +1,1 point avec Squeezemeta entre la saison humide et sèche. Pour les Planctomycetes +0,99 point avec Kraken2 et +1,08 point avec Squeezemeta. Le phylum des Actinobacteria est présent en forte abondance lorsque détecté par Kraken2 mais ne l'est pas avec Squeezemeta.

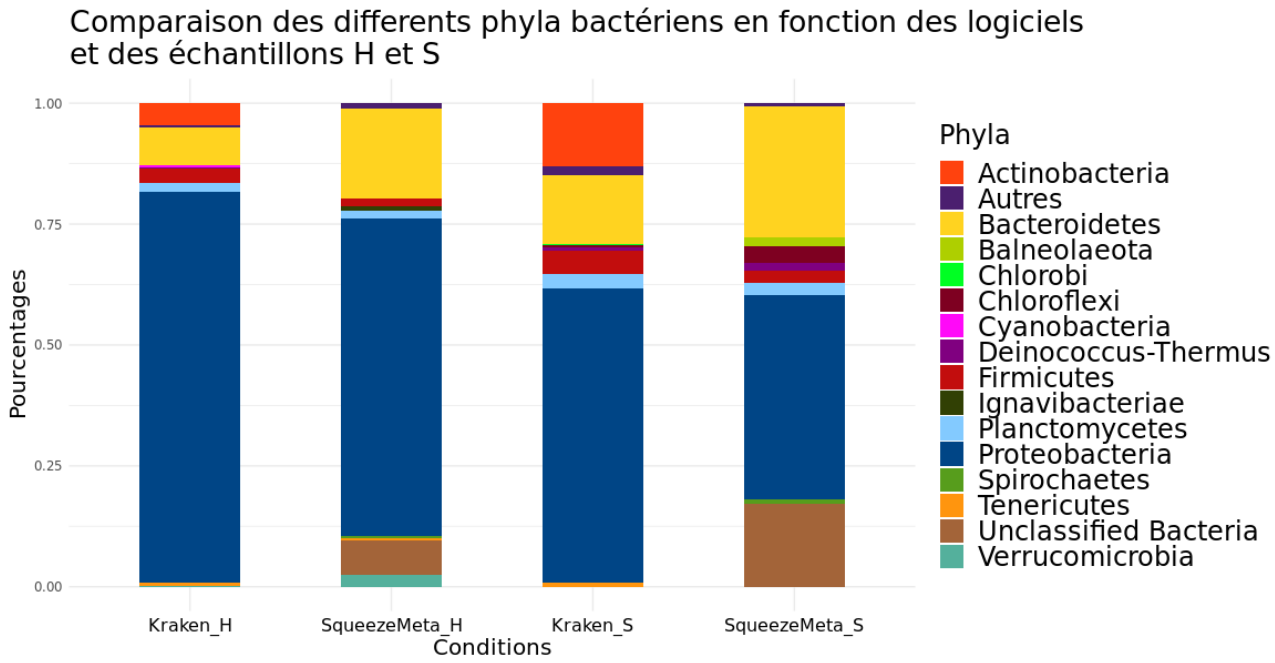

Figure 13 : Comparaison des phyla bactériens assignés par Kraken2 et Squeezemeta pour les deux échantillons

L'abondance et les phyla détectés par Squeezemeta et Kraken2 varient grandement (Figure 14). Les Chordata, retrouvés pour les deux logiciels et tous les échantillons, est le phyla principal avec Kraken2 pour les deux saisons. Toutefois, il n'est que le 5ème et 4ème phyla pour les échantillons H et S avec Squeezemeta. Le phylum Arthropoda est lui aussi retrouvé pour l'ensemble des échantillons. Il est le premier phylum de l'échantillon S pour Squeezemeta avec 51,99 % d'abondance et le deuxième phylum avec Kraken2avec une abondance de 27,24 %. Cependant pour l'échantillon H, Squeezemeta n'observe qu'une abondance de 1,67 % de ce phylum contre 25,38 % avec Kraken2.

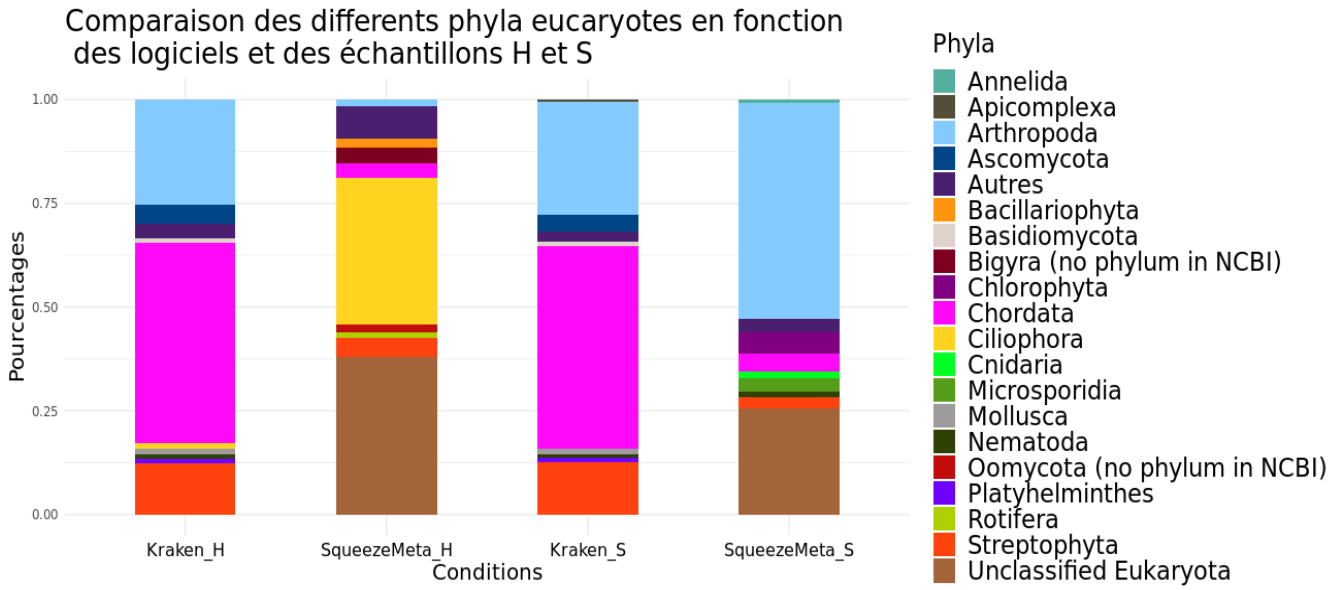

Figure 14 : Comparaison des phyla eucaryotes assignés par Kraken2 et Squeezemeta pour les deux échantillons

Kraken2 et Squeezemeta classifient tous les deux le phylum Euryarchaeota comme le phylum majoritaire (Figure 15). De plus, tous les deux observent l'augmentation de l'abondance de ce phylum entre la saison humide et sèche : +12,13 points pour Kraken2 et +8,62 points pour Squeezemeta. Les autres phyla observés sont différents entre Kraken2 et Squeezemeta, à l'exception des Thaumarchaeota détectés dans l'échantillon H avec les deux logiciels.

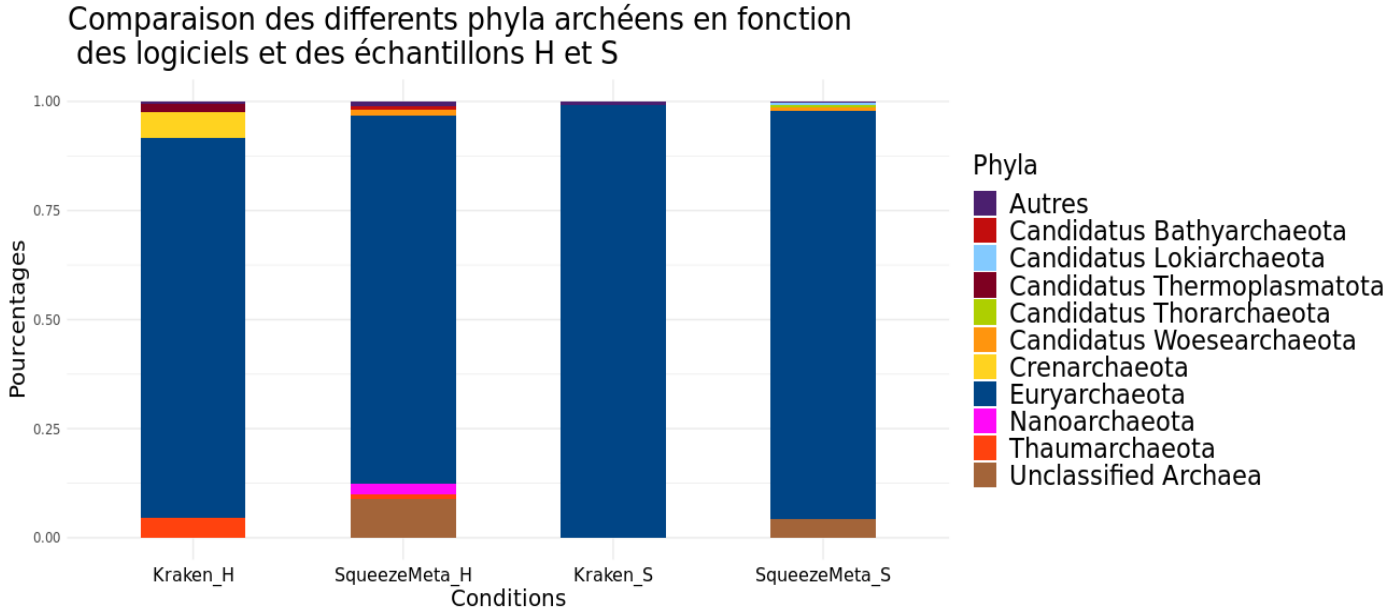

Figure 15 : Comparaison des phyla archéens assignés par Kraken2 et Squeezemeta pour les deux échantillons

3.3 Affiliation fonctionnelle

Squeezemeta permet en plus de l'affiliation taxonomique d'obtenir l'assignation fonctionnelle des gènes présents dans les échantillons. Pour l'échantillon H : 63 114 fonctions ont été assignées avec la base de données COG, 15 230 avec la base de données KEGG et 10 511 avec PFAM. Pour l'échantillon S : 62 102 fonctions ont été assignées avec la base de données COG, 13 208 avec la base de données KEGG et 8 771 avec PFAM. Pour pouvoir comparer la prévalence d'une fonction par rapport à une autre, Squeezemeta normalise l'abondance en prenant en compte la profondeur de séquençage et la longueur du gène. Cette abondance normalisée est appelée TPM par le logiciel.

Les fonctions appartiennent à des voies métaboliques. En regroupant les fonctions par voies métaboliques, il est possible d'obtenir une vision globale des différences d'expression des gènes entre saisons. Pour les fonctions déterminées par KEGG, les 20 premières voies métaboliques sont communes d'une saison à l'autre (Tableau 12). Exception faite des voies métaboliques de la phosphorylation oxydative et du métabolisme de la porphyrin, plus exprimées en saison sèche que humide. Par ailleurs, pour les voies métaboliques communes entre les deux conditions, le TPM est systématiquement plus faible en saison sèche.

Sur les 20 voies métaboliques les plus abondantes obtenues avec la base de données COG, 19 sont communes entre les deux saisons (Tableau 13). Tout comme les voies métaboliques de KEGG lors de la saison sèche, celles déterminées par COG sont systématiquement moins abondantes.

Enfin, pour les voies métaboliques obtenues avec PFAM, 16 des 20 premières voies les plus abondantes sont communes entre les saisons (Tableau 14). Cependant la variation de leur ordre d'apparition est plus importante entre les saisons qu'avec les autres bases de données.

Tableau 12 : 20 voies métaboliques principales obtenues avec la base de données KEGG; TPM correspond à l'abondance normalisée; Les couleurs permettent de retrouver facilement les fonctions entre échantillons

Voies métaboliques échantillon H	TPM	Voies métaboliques échantillon S	TPM
Unclassified	402 953,42	Unclassified	362 254,59
Signaling and cellular processes; Transporters	34 620,29	Signaling and cellular processes; Transporters	27 682,21
Unclassified: metabolism; Enzymes with EC numbers	24 440,61	Unclassified: metabolism; Enzymes with EC numbers	20 569,95
Poorly characterized; Function unknown	17 617,43	Poorly characterized; Function unknown	15 789,79
Membrane transport; ABC transporters	13 385,66	Membrane transport; ABC transporters	8 845,02
Genetic information processing; Transcription factors	9 424,87	Genetic Information Processing; Translation; Ribosome	7 405,34
Genetic Information Processing; Translation; Ribosome	8 312,70	Genetic information processing; Transcription factors	6 419,32
Genetic information processing; Ribosome biogenesis	7 965,48	Metabolism; Peptidases	6 327,63
Genetic information processing; Transfer RNA biogenesis	6 300,23	Genetic information processing; Ribosome biogenesis	5 534,52
Metabolism; Peptidases	6 271,03	Genetic information processing; Transfer RNA biogenesis	4 967,34
Unclassified: genetic information processing; Replication and repair	6 170,60	Unclassified: genetic information processing; Replication and repair	4 768,00
Poorly characterized; General function prediction only	5 296,70	Poorly characterized; General function prediction only	4 380,79
Metabolism of cofactors and vitamins; Porphyrin and chlorophyll metabolism	4 607,69	Genetic Information Processing; Translation; Aminoacyl-tRNA biosynthesis	4 056,04
Genetic Information Processing; Translation; Aminoacyl-tRNA biosynthesis	4 530,13	Unclassified: signaling and cellular processes; Transport	4 023,57
Unclassified: signaling and cellular processes; Transport	4 129,33	Metabolism; Energy metabolism; Oxidative phosphorylation	3 790,47
Signaling and cellular processes; Prokaryotic defense system	4 116,99	Signaling and cellular processes; Prokaryotic defense system	3 723,26
Metabolism; Nucleotide metabolism; Purine metabolism	4 100,44	Metabolism; Nucleotide metabolism; Purine metabolism	3 337,87
Genetic information processing; DNA repair and recombination proteins	4 034,83	Genetic information processing; DNA repair and recombination proteins	3 167,57
Genetic information processing; Chromosome and associated proteins	4 013,87	Metabolism of cofactors and vitamins; Porphyrin and chlorophyll metabolism	3 155,67
Genetic information processing; Chaperones and folding catalysts	3 654,97	Genetic information processing; Transcription machinery	2 838,70

Tableau 13 : 20 voies métaboliques principales obtenues avec la base de données COG; TPM correspond à l'abondance normalisée; Les couleurs permettent de retrouver facilement les fonctions entre échantillons

Voies métaboliques échantillon H	TPM	Voies métaboliques échantillon S	TPM
Unclassified	223 039,91	Unclassified	210 575,60
Function unknown	148 191,88	Function unknown	124 590,87
General function prediction only	52 999,67	General function prediction only	44 898,61
Amino acid transport and metabolism	44 419,33	Energy production and conversion	34 143,24
Energy production and conversion	38 952,30	Amino acid transport and metabolism	33 348,06
DNA replication, recombination and repair	31 946,26	DNA replication, recombination and repair	27 726,82
Inorganic ion transport and metabolism	31 665,19	Translation, ribosomal structure and biogenesis	25 438,09
Translation, ribosomal structure and biogenesis	30 719,71	Inorganic ion transport and metabolism	25 355,10
Signal transduction mechanisms	28 563,33	Cell envelope biogenesis, outer membrane	23 055,05
Cell envelope biogenesis, outer membrane	28 457,52	Carbohydrate transport and metabolism	21 361,05
Post Translational modification, protein turnover, chaperones	24 090,36	Signal transduction mechanisms	21 144,60
Carbohydrate transport and metabolism	23 708,22	Post Translational modification, protein turnover, chaperones	18 162,59
Transcription	23 034,18	Transcription	16 639,70
Lipid metabolism	19 284,99	Coenzyme metabolism	13 544,20
Coenzyme metabolism	17 739,99	Lipid metabolism	12 485,89
Nucleotide transport and metabolism	11 975,45	Nucleotide transport and metabolism	10 034,17
Secondary metabolites biosynthesis, transport and catabolism	9 166,93	Secondary metabolites biosynthesis, transport and catabolism	5 690,17
Cell motility and secretion	6 751,76	Cell motility and secretion	4 317,30
Cell division and chromosome partitioning	5 628,63	Cell division and chromosome partitioning	4 263,86
Replication, recombination and repair	3 654,20	Cell wall/membrane/envelope biogenesis	3 381,58

Tableau 14 : 20 voies métaboliques principales obtenues avec la base de données PFAM; TPM correspond à l'abondance normalisée; Les couleurs permettent de retrouver facilement les fonctions entre échantillons

Voies métaboliques échantillon H	TPM	Voies métaboliques échantillon S	TPM
Unclassified	438 957,11	Unclassified	402 973,38
Unknown fonction	8 968,46	Unknown fonction	7 787,66
ABC transporter	5 914,75	Tetratricopeptide repeat	5 212,89
Tetratricopeptide repeat	4 605,79	ABC transporter	3 980,00
Response regulator receiver domain	4 090,84	Response regulator receiver domain	3 078,00
Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase	3 182,98	Methyltransferase domain	2 340,28
Enoyl-(Acyl carrier protein) reductase	2 975,56	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase	2 249,97
AcrB/AcrD/AcrF family	2 753,74	AcrB/AcrD/AcrF family	2 208,44
Pyridine nucleotide-disulphide oxidoreductase	2 465,74	Enoyl-(Acyl carrier protein) reductase	2 007,39
Methyltransferase domain	2 389,22	short chain dehydrogenase	1 963,60
AAA domain	2 134,32	Pyridine nucleotide-disulphide oxidoreductase	1 852,45
Binding-protein-dependent transport system inner membrane component	2 117,87	Glycosyl transferases group 1	1 798,69
AMP-binding enzyme	2 025,39	AAA domain	1 668,27
Enoyl-CoA hydratase/isomerase	1 993,42	Reverse transcriptase (RNA-dependent DNA polymerase)	1 464,75
Glycosyl transferases group 1	1 859,40	Binding-protein-dependent transport system inner membrane component	1 447,56
Aldehyde dehydrogenase family	1 853,15	AMP-binding enzyme	1 380,90
Diguanylate cyclase, GGDEF domain	1 844,48	PAS fold	1 369,54
Helix-turn-helix domain	1 625,64	Sigma-54 interaction domain	1 209,66
Major Facilitator Superfamily	1 437,34	Major Facilitator Superfamily	1 115,33
Aminotransferase class I and II	1 388,73	NAD dependent epimerase/dehydratase family	1 091,53

4. Discussion :

L'assemblage des lectures en contigs peut produire des artéfacts et des chimères. Les premiers correspondent à des séquences créées par le logiciel mais n'ayant pas d'équivalent réel dans les données. Les deuxièmes correspondent à l'association au sein d'un même contig de deux séquences appartenant à deux organismes distincts. Un moyen de s'affranchir des artéfacts est la suppression des contigs pour lesquels le nombre de lectures alignées ou la profondeur est trop faible. Pour le deuxième, les chimères sont observables grâce à une variation importante de la profondeur au sein d'un contig. Le logiciel Valet (Hill, 2015) permet de mettre en évidence les contigs potentiellement chimériques ainsi que les coordonnées des régions potentiellement en défaut sans passer par une base de données. Afin d'être significatif, ce logiciel requiert en entrée des contigs d'une taille minimale de 1 000 pb. Par manque de ressources informatiques, cette étape n'a pas été réalisée. MetaQuast (Mikheenko et al., 2016), une extension de Quast (Gurevich et al., 2013), permet lui aussi la détection de contigs mal assemblés mais nécessite une référence pour fonctionner ce qui n'est pas optimal lorsque le contenu de l'échantillon est inconnu, comme dans le cas de cette analyse.

L'approche par k-mer de Kraken2 permet d'obtenir une classification taxonomique au niveau du genre voire de l'espèce. Cependant, Tamames et al. (2019) ont montré que pour la classification de lectures courtes Kraken2 obtient les meilleurs résultats avec des échantillons intestinaux. Pour d'autres écosystèmes comme ceux marins ou thermaux, le nombre d'espèces classifiées diminue et le taux d'erreur augmente fortement. Il atteint plus de 10 % de lectures incorrectement classifiées au rang taxonomique Famille pour un échantillon marin contre moins 1 % pour un échantillon intestinal. Toutefois, le fait d'utiliser Kraken2 sur des contigs plutôt que sur des lectures courtes permet d'augmenter la précision de l'assignation tout en gardant le même rappel comme montré par Tran et Phan dans leur article. En effet, l'assemblage des lectures courtes en contigs diminue le nombre d'espèces prédites mais améliore la qualité de la prédiction grâce à l'augmentation de la taille des séquences : le taux de faux positifs est donc diminué. Cependant, il est fort probable que Kraken2 ait assigné incorrectement un certain nombre de contigs notamment lorsque la taille de ces derniers diminue (en moyenne 89 % des contigs des assemblages sont inférieurs à 1 000 pb). Les résultats produits par Kraken2 peuvent donc être incorrects, et ce particulièrement pour les rangs taxonomiques bas comme le genre ou l'espèce.

Les bases de données utilisées peuvent être l'une des raisons des écueils observés avec Kraken2. Premièrement, ces dernières peuvent contenir des informations erronées. Dans les faits, une séquence rentrée dans GenBank par un chercheur n'est modifiable que par ce dernier. Si son assignation taxonomique n'est pas correcte : problème de nom, séquence mal assignée, etc. ; seul son auteur peut la modifier. Deuxièmement, il existe la présence de contaminants dans les bases de données. La majorité des séquences présentes dans GenBank sont des "draft" génomes, soit des génomes pour lesquels un assemblage a été généré mais fragmenté en de nombreux contigs (jusqu'à plusieurs milliers). L'ensemble des contigs d'un "draft" génome sont associés à la même espèce. Néanmoins, certains peuvent à tort avoir été assignés à une espèce alors qu'ils appartiennent à une autre, totalement différente. Le NCBI, gestionnaire de GenBank, vérifie la présence de ces contaminants et demandent aux auteurs de les supprimer. Ainsi, il est possible de retrouver de nombreux contaminants dans les séquences issues des "draft" génomes. L'utilisation de la base de données RefSeq peut être une bonne alternative. Cette base utilise les données de Genbank et les valide via plusieurs filtres afin de ne conserver que les séquences garantissant un minimum de curation (Tatusova et al., 2015). Enfin, les bases de données ne sont pas complètes et sont biaisées vers les organismes

majoritairement étudiés par les chercheurs ou dont le milieu est plus étudié. Si une séquence n'est pas présente dans la base de données, Kraken2 ne peut lui attribuer un taxon. Ce qui expliquerait la grande proportion de séquences non classifiées. De plus, Kraken2 repose sur la recherche de k-mers exacts afin d'attribuer une lecture à un taxon. Si cette dernière diffère de quelques nucléotides la classification ne peut être effectuée.

Pour pallier à l'utilisation de bases de données, des méthodes basées sur le binning peuvent être envisagées (Sedlar et al., 2017). Comme expliqué précédemment, le binning non supervisé permet, en se basant sur les paramètres intra-échantillon (composition en k-mer, teneur en GC, etc.) et inter-échantillon (variation de l'abondance), de regrouper les séquences susceptibles d'appartenir à un même organisme. Les méthodes basées sur la composition ont comme avantage d'obtenir une visualisation claire d'un microbiome. Mais elles nécessitent en contrepartie d'avoir des séquences relativement longues et des échantillons peu complexes, au risque de ne pas être fiables. Les méthodes basées sur l'abondance offrent des propriétés complémentaires. Elles peuvent réaliser la classification de petits fragments mais nécessitent un certain nombre d'échantillons pour fonctionner, et ne permettent pas d'obtenir une visualisation claire et précise du microbiote. Enfin les méthodes hybrides, qui combinent les deux méthodes précédentes, permettent d'obtenir les meilleurs résultats comme montré dans le CAMI challenge (Sczyrba et al., 2017).

Malheureusement ces techniques se prêtent très mal à cette étude. Le fait que les prélèvements aient dû être regroupés en seulement deux conditions a fait perdre une grande partie des possibilités offertes par l'approche par différence d'abondance. Par ailleurs, l'échantillon qui en résulte est beaucoup plus complexe que chaque échantillon pris indépendamment, ce qui découle en un assemblage plus fragmenté et donc plus difficile à utiliser pour des méthodes basées sur la composition. Pour toutes ces raisons, le binning n'a pas été retenu comme une option viable pour la classification taxonomique et fonctionnelle.

Squeezemeta, contrairement à Kraken2, réalise sa taxonomie sur des protéines par homologie (Tamames et Puente-Sánchez, 2019). Les séquences peuvent diverger d'un point de vue nucléotidique mais conserver les mêmes acides aminés. Cela pourrait expliquer la plus grande proportion de séquences classifiées avec ce logiciel par rapport à Kraken2. Pour obtenir des protéines, il faut trouver des gènes. Squeezemeta utilise Prodigal un logiciel spécialisé dans la détection *ab initio* de séquences codantes bactériennes. L'un des problèmes rencontrés est que ce logiciel n'a pas été conçu pour la détection de gènes eucaryotes, et peut donc passer à côté d'un certain nombre d'entre eux. Ceci expliquerait que le pourcentage de séquences eucaryotes est très faible avec Squeezemeta par rapport aux résultats de Kraken2. Mais également les variations importantes de l'abondance et de la présence des phyla classifiés par les deux méthodes. Au vu des résultats très similaires obtenus avec Kraken2 pour les deux échantillons, la composition en Eucaryotes ne semble cependant pas varier significativement d'une saison à l'autre.

Les eucaryotes jouent un rôle important dans l'ensemble des niches écologiques. L'étude de la diversité de ces organismes est principalement focalisée sur les branches animales, végétales et fongiques. Elle ne prend que peu en compte les protistes, qui regroupent la majorité des eucaryotes unicellulaires. Or cette branche constitue une part importante des lignées eucaryotes (Burki, 2014). Ces micro-organismes sont difficiles à assembler et à annoter dans des échantillons métagénomiques dû à leur faible profondeur de séquençage, à leur taille de génome plus grande que celle des bactéries et à leur sous-représentation dans les bases de données.

De plus, la recherche de gènes chez les eucaryotes est une tâche difficile, particulièrement à cause de l'alternance entre introns et exons. Des logiciels comme

Augustus ou GeneMark-ES (Lomsadze, 2005) permettent de prédire *ab initio* la présence de gènes eucaryotes dans des séquences. Néanmoins, l'utilisation de ces logiciels pose plusieurs contraintes comme la nécessité d'entraîner un modèle sur des séquences proches de celle de l'échantillon recherché. Or le pouvoir prédictif de ces modèles diminue rapidement s'il est appliqué à des individus distants phylogénétiquement de ceux ayant permis son entraînement (Korf, 2014). Un échantillon métagénomique est par nature complexe, rendant la découverte de gènes difficile.

D'autres approches ont été développées comme le logiciel MetaEuk (Levy Karin et al. 2020) qui permet grâce à une référence d'identifier les protéines eucaryotes, puis annoter taxonomiquement les contigs des échantillons métagénomiques via une approche par homologie. L'un des inconvénients de MetaEuk est la nécessité de ne fournir que des séquences eucaryotes en entrée, et donc de connaître au préalable quels contigs de l'édchantillon appartiennent à ce domaine. C'est pour réaliser cette tâche que le logiciel EukRep (West et al., 2017) a été spécialement créé. EukRep est basé sur une approche par k-mer et machine learning afin d'identifier les séquences eucaryotes en amont de la prédiction de gènes. Là encore, ce logiciel présente quelques inconvénients comme la taille des contigs : au minimum 1 000 pb. De plus, le nombre de faux positifs générés est corrélé positivement avec la taille des séquences en entrée.

Il est donc possible d'obtenir une classification taxonomique plus représentative des grandes séquences eucaryotes présentes dans les données métagénomiques que celle générée avec le pipeline Squeezemeta. Cette classification complémentaire permettrait aussi de confirmer les résultats concordants produits avec Kraken2.

Toutefois les résultats obtenus avec Squeezemeta et Kraken2 permettent d'obtenir une vision générale et fiable de la composition en micro-organismes, particulièrement pour les bactéries et les archées. Squeezemeta classifie un plus grand nombre de séquences que Kraken2 et ce avec un taux de faux positif faible, inférieur à 1 % lors des tests réalisés par Tamames et al. (2019), sur des échantillons marins, intestinaux et thermaux.

Les résultats obtenus pour les bactéries entre Squeezemeta et Kraken2 sont concordants à l'exception des Actinobacteria très peu détectées avec Squeezemeta. Cela peut venir du fait que Squeezemeta classifie plus de séquences que Kraken2, et qu'en proportion les Actinobacteria semblent moins présentes. Une autre hypothèse serait la plus grande abondance de ce phylum dans la partie fragmenté de l'assemblage pour laquelle il est plus difficile de retrouver des gènes. Il en résulterait une diminution des séquences assignées à ce phylum, préféré à la catégorie "non classifiée", là où Kraken2 ne serait pas handicapé par cette contrainte.

Enfin, des résultats similaires sont observables entre les deux logiciels pour l'abondance des phyla archéens. Soit une forte augmentation des séquences archéennes lors de la saison sèche et une nette domination du phylum Euryarchaeota pour l'ensemble des échantillons. L'augmentation massive des Archées lors de la saison sèche peut être due à une mauvaise manipulation lors du prélèvement. En effet, lors de la saison sèche, le niveau d'eau est bas, il est donc difficile de ne pas récupérer un peu de sédiments en même temps que l'eau. Les proportions de bactéries, eucaryotes et archéennes déterminées par Kraken2 pour l'échantillon S sont d'ailleurs très proches de celles déterminées par Wurzbacher entre 0 et 10 cm de la couche sédimentaire du lac Stechlin en Allemagne (Wurzbacher et al., 2017). Dans cette étude, cette proportion était de 70:20:10 (bactéries : eucaryotes : archées) contre 60:32:8 pour les résultats de Kraken2 sur l'échantillon prélevé en saison sèche. Le phylum majoritaire à cette strate est celui des Euryarchaeota, suivi de celui des Thaumarchaeota. La composition et la proportion des phylum archéens déterminées par Squeezemeta sont aussi proches de celles trouvées dans la couche sédimentaire la plus récente du lac Alinen Mustajärvi en Finlande (Rissanen

et al., 2019), avec une prédominance du phylum Euryarchaeota et la présence en plus faible proportion des phyla Woesearchaeota et Bacthyarcheota.

Il est à noter cependant que l'abondance relative issue de Kraken2 et Squeezemeta n'est pas complète car elle ne prend pas en compte les lectures n'ayant pas été incluses dans l'assemblage. Cela représente 10 % des lectures de l'échantillon H et 25 % des lectures pour l'échantillon S. Une possibilité pour obtenir une abondance relative complète serait de prendre en compte ces lectures dans l'analyse. Kraken2 étant déjà spécialisé dans l'analyse de lectures courtes, cela ne poserait aucune difficulté technique de lui rajouter les lectures non assemblées. Squeezemeta propose une option « singletons » permettant d'inclure ces lectures comme étant des contigs. Cependant, cette étape a un coût informatique important et va nécessairement augmenter le temps d'analyse.

L'approche de classification en se basant sur les gènes réalisée par Squeezemeta permet aussi d'obtenir les différentes fonctions et voies métaboliques présentes dans l'écosystème du lac. Pour les voies métaboliques les plus abondantes, les variations entre saisons ne sont pas importantes. Cependant, pour l'ensemble des résultats produits avec les trois bases de données, l'abondance des voies métaboliques en saison sèche est systématiquement plus faible. Une explication à ce phénomène serait la moins grande proportion de fonctions assignées par les logiciels durant cette saison.

L'analyse des fonctionnalités du lac est toujours à l'étude, certaines fonctions particulières pourraient être recherchées afin d'observer l'adaptation des micro-organismes aux variations des paramètres physico-chimiques entre saisons, comme par exemple des voies métaboliques impliquées dans le traitement des métaux lourds.

5. Conclusion

La métagénomique est une discipline jeune et complexe. De nombreuses approches ont été mises en place par les chercheurs dans le but de déterminer la composition en micro-organismes d'un écosystème. Les analyses, qui peuvent être produites, dépendent ainsi du type d'expérience réalisé mais aussi des méthodes bioinformatiques employées.

L'analyse présentée dans ce mémoire a permis d'extraire la taxonomie et les fonctions des organismes présents dans le lac Dayat Morsli, tout en prenant en compte sa particularité d'échantillonnage. L'assembleur Metaspades a été choisi pour réaliser un assemblage métagénomique *de novo* car il génère le plus grand nombre de grands contigs, et est aussi le plus représentatif des échantillons comparé à Megahit.

L'assignation taxonomique des séquences se base sur deux logiciels, Kraken2 et Squeezemeta. Ces derniers proposent deux approches de classification différentes et reposent sur deux bases de données. Cela permet d'obtenir des résultats fiables et concordants pour la taxonomie bactérienne et archéenne au rang phylum. Les variations de l'abondance des Proteobacteria, des Bacteroidetes, des Firmicutes et des Planctomycetes sont bien observées avec les deux logiciels pour les bactéries. Cette approche nécessite cependant confirmation pour les eucaryotes, qui sont sûrement sous-estimés par Squeezemeta avec une abondance moyenne de 2,78 % pour les deux saisons contre 29,57 % avec Kraken2. L'augmentation des archées lors de la saison sèche pourrait provenir d'une erreur lors du prélèvement. Notamment car leur composition en saison sèche et leur abondance est proche de celle déjà caractérisée dans d'autres lacs.

L'assignation fonctionnelle des gènes s'est faite sur trois bases de données différentes et a permis d'obtenir les voies métaboliques exprimées dans les deux échantillons. Celles-ci ne montrent pas de variations importantes d'une saison à l'autre mais ces résultats nécessitent des analyses complémentaires notamment avec des voies métaboliques plus spécifiques.

Plusieurs axes d'améliorations peuvent être apportés à ce travail. L'ajout d'une étape dédiée aux eucaryotes permettrait d'améliorer leur caractérisation dans l'échantillon. Réaliser une recherche de marqueur génique type permettrait aussi d'améliorer l'affiliation taxonomique de certaines séquences. L'inclusion de l'ensemble des séquences assemblées et non assemblées, permettrait d'obtenir une abondance relative complète et ainsi réaliser des analyses de diversité avec l'utilisation par exemple de l'indice de Shannon pour la diversité inter-échantillon ou l'indice de Bray-Curtis pour l'analyse de diversité intra-échantillon. Finalement, une analyse plus fine et approfondie de la taxonomie pourrait mettre en évidence des variations saisonnières invisibles au rang taxonomique phylum.

L'auteur rappelle que le présent travail fera l'objet d'une publication une fois les analyses complètement finalisées.

Bibliographie :

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, et Christopher Quince. 2014. « Binning Metagenomic Contigs by Coverage and Composition ». *Nature Methods* 11 (11): 1144-46. https://doi.org/10.1038/nmeth.3103.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, et David J. Lipman. 1990. « Basic Local Alignment Search Tool ». *Journal of Molecular Biology* 215 (3): 403-10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2020. « Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with BioBakery 3 ». Preprint. Microbiology. https://doi.org/10.1101/2020.11.19.388223.

Ben Bayer, Wided, Nathalie Casse, Mohamed Bey Baba Hamed, Françoise Denis, Vanina Pasqualini, Marie Vaugoyeau, et Aurore Caruso. 2019. « First Characterization of Physicochemical and Biological Variables of the Salt Wetland Dayat Morsli in Oran (Algeria) ». *Journal of African Earth Sciences* 160 (décembre): 103652. https://doi.org/10.1016/j.jafrearsci.2019.103652.

Benbayer, Wided. 2019. « Caractérisation bio-physicochimique des eaux du lac de 'Dayat Morsli' et approche métagénomique ».

Breitwieser, Florian P, Jennifer Lu, et Steven L Salzberg. 2019. « A Review of Methods and Databases for Metagenomic Classification and Assembly ». *Briefings in Bioinformatics* 20 (4): 1125-36. https://doi.org/10.1093/bib/bbx120.

Buchfink, Benjamin, Klaus Reuter, et Hajk-Georg Drost. 2021. « Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND ». *Nature Methods* 18 (4): 366-68. https://doi.org/10.1038/s41592-021-01101-x.

Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, et al. 1996. « Complete Genome Sequence of the Methanogenic Archaeon, Methanococcus Jannaschii ». *Science* 273 (5278): 1058-73. https://doi.org/10.1126/science.273.5278.1058.

Burki, F. 2014. « The Eukaryotic Tree of Life from a Global Phylogenomic Perspective ». *Cold Spring Harbor Perspectives in Biology* 6 (5): a016147-a016147. https://doi.org/10.1101/cshperspect.a016147.

Consortium, The C. elegans Sequencing. 1998. « Genome Sequence of the Nematode C. Elegans: A Platform for Investigating Biology ». *Science, New Series* 282 (5396): 2012-18.

Dauga, Catherine, Joël Doré, et Abdelghani Sghir. 2005. « La diversité insoupçonnée du monde microbien ». *médecine/sciences* 21 (3): 290-96. https://doi.org/10.1051/medsci/2005213290.

Dohm, Juliane C, Philipp Peters, Nancy Stralis-Pavese, et Heinz Himmelbauer. 2020. « Benchmarking of Long-Read Correction Methods ». *NAR Genomics and Bioinformatics* 2 (2): lqaa037. https://doi.org/10.1093/nargab/lqaa037.

Eddy, Sean R. 2009. « A New Generation of Homology Search Tools Based on Probabilistic Inference ». *Genome Informatics. International Conference on Genome Informatics* 23 (1): 205-11.

European Commission. Joint Research Centre. 2016. *LUCAS Soil Component: Proposal for Analysing New Physical, Chemical and Biological Soil Parameters.* LU: Publications Office. https://data.europa.eu/doi/10.2788/884940.

Finn, Robert D., Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, et al. 2014. « Pfam: The Protein Families Database ». *Nucleic Acids Research* 42 (D1): D222-30. https://doi.org/10.1093/nar/gkt1223.

Fleischmann, Robert D., Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, et al. 1995. « Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd ». *Science, New Series* 269 (5223): 496-498+507-512.

Ghurye, Jay S., Victoria Cepeda-Espinoza, et Mihai Pop. 2016. « Metagenomic Assembly: Overview, Challenges and Applications ». *The Yale Journal of Biology and Medicine* 89 (3): 353-62.

Gilbert, Jack A, Janet K Jansson, et Rob Knight. 2014. « The Earth Microbiome Project: Successes and Aspirations ». *BMC Biology* 12 (1): 69. https://doi.org/10.1186/s12915-014-0069-1.

Grossart, Hans-Peter, Ramon Massana, Katherine D. McMahon, et David A. Walsh. 2020. « Linking Metagenomics to Aquatic Microbial Ecology and Biogeochemical Cycles ». *Limnology and Oceanography* 65 (S1). https://doi.org/10.1002/lno.11382.

Gunde-Cimerman, Nina, Ana Plemenitaš, et Aharon Oren. 2018. « Strategies of Adaptation of Microorganisms of the Three Domains of Life to High Salt Concentrations ». *FEMS Microbiology Reviews* 42 (3): 353-75. https://doi.org/10.1093/femsre/fuy009.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, et Glenn Tesler. 2013. « QUAST: Quality Assessment Tool for Genome Assemblies ». *Bioinformatics* 29 (8): 1072-75. https://doi.org/10.1093/bioinformatics/btt086.

Handelsman, Jo, Michelle R. Rondon, Sean F. Brady, Jon Clardy, et Robert M. Goodman. 1998. « Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products ». *Chemistry & Biology* 5 (10): R245-49. https://doi.org/10.1016/S1074-5521(98)90108-9.

Heather, James M, et Benjamin Chain. 2016. « The Sequence of Sequencers: The History of Sequencing DNA », 8.

Hill, Christopher Michael. 2015. « Novel methods for comparing and evaluating single and metagenomic assemblies », 179.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, et al. 2019. « EggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses ». *Nucleic Acids Research* 47 (D1): D309-14. https://doi.org/10.1093/nar/gky1085.

Hyatt, Doug, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, et Loren J Hauser. 2010. « Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification ». *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.

Imelfort, Michael, Donovan Parks, Ben J. Woodcroft, Paul Dennis, Philip Hugenholtz, et Gene W. Tyson. 2014. « GroopM: An Automated Tool for the Recovery of Population Genomes from Related Metagenomes ». *PeerJ* 2 (septembre): e603. https://doi.org/10.7717/peerj.603.

Johnson, Sarah Stewart, Marc Gerard Chevrette, Bethany L. Ehlmann, et Kathleen Counter Benison. 2015. « Insights from the Metagenome of an Acid Salt Lake: The Role of Biology in an Extreme Depositional Environment ». Édité par Paul Jaak Janssen. *PLOS ONE* 10 (4): e0122869. https://doi.org/10.1371/journal.pone.0122869.

Kanehisa, M. 2000. « KEGG: Kyoto Encyclopedia of Genes and Genomes ». *Nucleic Acids Research* 28 (1): 27-30. https://doi.org/10.1093/nar/28.1.27.

Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, et Mao Tanabe. 2016. « KEGG as a Reference Resource for Gene and Protein Annotation ». *Nucleic Acids Research* 44 (D1): D457-62. https://doi.org/10.1093/nar/gkv1070.

Kang, Dongwan D., Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, et Zhong Wang. 2019. « MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies ». *PeerJ* 7 (juillet): e7359. https://doi.org/10.7717/peerj.7359.

Korf, Ian. 2004. « Gene finding in novel genomes ». *BMC Bioinformatics* 5 (1): 59. https://doi.org/10.1186/1471-2105-5-59.

Laczny, Cedric C, Tomasz Sternal, Valentin Plugaru, Piotr Gawron, Arash Atashpendar, Houry Margossian, Sergio Coronado, Laurens der Maaten, Nikos Vlassis, et Paul Wilmes. 2015. « VizBin - an Application for Reference-Independent Visualization and Human-Augmented Binning of Metagenomic Data ». *Microbiome* 3 (1): 1. https://doi.org/10.1186/s40168-014-0066-1.

Levy Karin, Eli, Milot Mirdita, et Johannes Söding. 2020. « MetaEuk—Sensitive, High-Throughput Gene Discovery, and Annotation for Large-Scale Eukaryotic Metagenomics ». *Microbiome* 8 (1): 48. https://doi.org/10.1186/s40168-020-00808-x.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, et Tak-Wah Lam. 2015. « MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph ». *Bioinformatics* 31 (10): 1674-76. https://doi.org/10.1093/bioinformatics/btv033.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et 1000 Genome Project Data Processing Subgroup. 2009. «

The Sequence Alignment/Map Format and SAMtools ». *Bioinformatics* 25 (16): 2078-79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Heng. 2013. « Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM ». *arXiv:1303.3997 [q-bio]*, mai. http://arxiv.org/abs/1303.3997.

Lomsadze, A. 2005. « Gene Identification in Novel Eukaryotic Genomes by Self-Training Algorithm ». *Nucleic Acids Research* 33 (20): 6494-6506. https://doi.org/10.1093/nar/gki937.

Mande, S. S., M. H. Mohammed, et T. S. Ghosh. 2012. « Classification of Metagenomic Sequences: Methods and Challenges ». *Briefings in Bioinformatics* 13 (6): 669-81. https://doi.org/10.1093/bib/bbs054.

Metzker, Michael L. 2010. « Sequencing Technologies — the next Generation ». *Nature Reviews Genetics* 11 (1): 31-46. https://doi.org/10.1038/nrg2626.

Mikheenko, Alla, Vladislav Saveliev, et Alexey Gurevich. 2016. « MetaQUAST: Evaluation of Metagenome Assemblies ». *Bioinformatics* 32 (7): 1088-90. https://doi.org/10.1093/bioinformatics/btv697.

Milanese, Alessio, Daniel R Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, et al. 2019. « Microbial Abundance, Activity and Population Genomic Profiling with MOTUs2 ». *Nature Communications* 10 (1): 1014. https://doi.org/10.1038/s41467-019-08844-4.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, et Pavel A. Pevzner. 2017. « Metaspades: A New Versatile Metagenomic Assembler ». *Genome Research* 27 (5): 824-34. https://doi.org/10.1101/gr.213959.116.

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. « Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation ». *Nucleic Acids Research* 44 (D1): D733-45. https://doi.org/10.1093/nar/gkv1189.

Ounit, Rachid, Steve Wanamaker, Timothy J Close, et Stefano Lonardi. 2015. « CLARK: Fast and Accurate Classification of Metagenomic and Genomic Sequences Using Discriminative k-Mers ». *BMC Genomics* 16 (1): 236. https://doi.org/10.1186/s12864-015-1419-2.

Pfeiffer, Franziska, Carsten Gröber, Michael Blank, Kristian Händler, Marc Beyer, Joachim L. Schultze, et Günter Mayer. 2018. « Systematic Evaluation of Error Rates and Causes in Short Samples in Next-Generation Sequencing ». *Scientific Reports* 8 (1): 10950. https://doi.org/10.1038/s41598-018-29325-6.

Rissanen, Antti J, Sari Peura, Promise A Mpamah, Sami Taipale, Marja Tiirola, Christina Biasi, Anita Mäki, et Hannu Nykänen. 2019. « Vertical Stratification of Bacteria and Archaea in Sediments of a Small Boreal Humic Lake ». *FEMS Microbiology Letters* 366 (5). https://doi.org/10.1093/femsle/fnz044.

Rusch, Douglas B, Aaron L Halpern, Granger Sutton, Karla B Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, et al. 2007. « The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific ».

Édité par Nancy A Moran. *PLoS Biology* 5 (3): e77. https://doi.org/10.1371/journal.pbio.0050077.

Sanger, F., S. Nicklen, et A. R. Coulson. 1977. « DNA Sequencing with Chain-Terminating Inhibitors ». *Proceedings of the National Academy of Sciences* 74 (12): 5463-67. https://doi.org/10.1073/pnas.74.12.5463.

Schadt, E. E., S. Turner, et A. Kasarskis. 2010. « A Window into Third-Generation Sequencing ». *Human Molecular Genetics* 19 (R2): R227-40. https://doi.org/10.1093/hmg/ddq416.

Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. « Critical Assessment of Metagenome Interpretation—a Benchmark of Metagenomics Software ». *Nature Methods* 14 (11): 1063-71. https://doi.org/10.1038/nmeth.4458.

Sedlar, Karel, Kristyna Kupkova, et Ivo Provaznik. 2017. « Bioinformatics Strategies for Taxonomy Independent Binning and Visualization of Sequences in Shotgun Metagenomics ». *Computational and Structural Biotechnology Journal* 15: 48-55. https://doi.org/10.1016/j.csbj.2016.11.005.

Sharpton, Thomas J. 2014. « An Introduction to the Analysis of Shotgun Metagenomic Data ». *Frontiers in Plant Science* 5 (juin). https://doi.org/10.3389/fpls.2014.00209.

Shen, Wei, Shuai Le, Yan Li, et Fuquan Hu. 2016. « SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation ». Édité par Quan Zou. *PLOS ONE* 11 (10): e0163962. https://doi.org/10.1371/journal.pone.0163962.

Stanke, Mario, Mark Diekhans, Robert Baertsch, et David Haussler. 2008. « Using Native and Syntenically Mapped CDNA Alignments to Improve de Novo Gene Finding ». *Bioinformatics* 24 (5): 637-44. https://doi.org/10.1093/bioinformatics/btn013.

Stein, J L, T L Marsh, K Y Wu, H Shizuya, et E F DeLong. 1996. « Characterization of Uncultivated Prokaryotes: Isolation and Analysis of a 40-Kilobase-Pair Genome Fragment from a Planktonic Marine Archaeon. » *Journal of Bacteriology* 178 (3): 591-99. https://doi.org/10.1128/JB.178.3.591-599.1996.

Strous, Marc, Beate Kraft, Regina Bisdorf, et Halina E. Tegetmeyer. 2012. « The Binning of Metagenomic Contigs for Microbial Physiology of Mixed Cultures ». *Frontiers in Microbiology* 3. https://doi.org/10.3389/fmicb.2012.00410.

Sunagawa, S., L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, et al. 2015. « Structure and Function of the Global Ocean Microbiome ». *Science* 348 (6237): 1261359-1261359. https://doi.org/10.1126/science.1261359.

Tamames, Javier, Marta Cobo-Simón, et Fernando Puente-Sánchez. 2019. « Assessing the Performance of Different Approaches for Functional and Taxonomic Annotation of Metagenomes ». *BMC Genomics* 20 (1): 960. https://doi.org/10.1186/s12864-019-6289-6.

Tamames, Javier, et Fernando Puente-Sánchez. 2019. « SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline ». *Frontiers in Microbiology* 9 (janvier): 3349. https://doi.org/10.3389/fmicb.2018.03349.

Tatusova, Tatiana, Stacy Ciufo, Scott Federhen, Boris Fedorov, Richard McVeigh, Kathleen O'Neill, Igor Tolstoy, et Leonid Zaslavsky. 2015. « Update on RefSeq Microbial Genomes Resources ». *Nucleic Acids Research* 43 (D1): D599-605. https://doi.org/10.1093/nar/gku1062.

Tran, Quang, et Vinhthuy Phan. 2020. « Assembling Reads Improves Taxonomic Classification of Species ». *Genes* 11 (8): 946. https://doi.org/10.3390/genes11080946.

Turnbaugh, Peter J., Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, et Jeffrey I. Gordon. 2007. « The Human Microbiome Project ». *Nature* 449 (7164): 804-10. https://doi.org/10.1038/nature06244.

Vavourakis, Charlotte D., Rohit Ghai, Francisco Rodriguez-Valera, Dimitry Y. Sorokin, Susannah G. Tringe, Philip Hugenholtz, et Gerard Muyzer. 2016. « Metagenomic Insights into the Uncultured Diversity and Physiology of Microbes in Four Hypersaline Soda Lake Brines ». *Frontiers in Microbiology* 7 (février). https://doi.org/10.3389/fmicb.2016.00211.

Wang, Ying, Haiyan Hu, et Xiaoman Li. 2015. « MBBC: An Efficient Approach for Metagenomic Binning Based on Clustering ». *BMC Bioinformatics* 16 (1): 36. https://doi.org/10.1186/s12859-015-0473-8.

West, Patrick T., Alexander J. Probst, Igor V. Grigoriev, Brian C. Thomas, et Jillian F. Banfield. 2017. « Genome-Reconstruction for Eukaryotes from Complex Natural Microbial Communities ». Preprint. Genomics. https://doi.org/10.1101/171355.

Wood, Derrick E., Jennifer Lu, et Ben Langmead. 2019. « Improved Metagenomic Analysis with Kraken2 ». *Genome Biology* 20 (1): 257. https://doi.org/10.1186/s13059-019-1891-0.

Wu, Yu-Wei, et Yuzhen Ye. 2011. « A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-Tuples ». *Journal of Computational Biology* 18 (3): 523-34. https://doi.org/10.1089/cmb.2010.0245.

Wurzbacher, Christian, Andrea Fuchs, Katrin Attermeyer, Katharina Frindte, Hans-Peter Grossart, Michael Hupfer, Peter Casper, et Michael T. Monaghan. 2017. « Shifts among Eukaryota, Bacteria, and Archaea Define the Vertical Organization of a Lake Sediment ». *Microbiome* 5 (1): 41. https://doi.org/10.1186/s40168-017-0255-9.

Ye, Simon H., Katherine J. Siddle, Daniel J. Park, et Pardis C. Sabeti. 2019. « Benchmarking Metagenomics Tools for Taxonomic Classification ». *Cell* 178 (4): 779-94. https://doi.org/10.1016/j.cell.2019.07.010.

Yue, Yi, Hao Huang, Zhao Qi, Hui-Min Dou, Xin-Yi Liu, Tian-Fei Han, Yue Chen, Xiang-Jun Song, You-Hua Zhang, et Jian Tu. 2020. « Evaluating Metagenomics Tools for Genome Binning with Real Metagenomic Datasets and CAMI Datasets ». *BMC Bioinformatics* 21 (1): 334. https://doi.org/10.1186/s12859-020-03667-3.