Analyses et traitements de données microbiologiques issues de séquenceur à haut débit

Nori SADOUNI

Université Nice Sophia-Antipolis
Master Sciences de la vie & de la santé
Mention Génétique Immunité et Développement
Parcours Biologie Informatique et Mathématique BIM
2016-2017

Sous la direction du
Pr. Raymond RUIMY, chef du service bactériologie à l'Archet - Nice
&
du Dr. Olivier Croce, Responsable service Bioinformatique IRCAN

Institut de recherche sur le Cancer et le Vieillissement (IRCAN)
CNRS UMR 7284 - INSERM U 1081 - UNS Faculté de Médecine 28 avenue de Valombrose 06107 Nice Cedex 02 - France Tel. : +33 (0)493377782 / +33 (0)493377703

Table des matières

Table des figures

Remerciements

Je remercie tout d'abord l'Université de Nice Sophia-Antipolis, qui m'a permis de pouvoir réaliser ce stage. Ayant côtoyé d'autres stagiaires au sein de l'institut, j'ai constaté qu'il était rare pour un Master de pouvoir effectuer deux stages d'une durée de 6 mois. Ainsi, je remercie l'Université de nous offrir une telle immersion dans le monde du travail, où l'on peut vraiment se faire une idée du poste que l'on occupera. Je tenais également à remercier les différents enseignants que j'ai pu avoir, et qui m'auront permis d'avoir les outils pour correctement appréhender ce stage.

Je tenais également à remercier le Pr.Raymond RUIMY de m'avoir accueilli pour ce stage. Je remercie également grandement le Dr.Olivier CROCE, pour tout ses conseils, son aide sur l'aspect informatique, mais également microbiologique. Il m'aura été d'une aide précieuse. J'ai été au cours de ce stage très bien encadré. Nous avions des réunions hebdomadaires où l'on faisait le point. Je tenais également à le remercier pour sa patience, pour le temps qu'il a prit pour m'expliquer le fonctionnement de certains programmes, ses conseils dans certains scripts. Et pour tout le temps qu'il a du prendre sur son travail personnel pour m'épauler.

Je tenais à remercier le Dr.Eric RÖTTINGER de m'avoir fait une place dans ces bureaux pour m'accueillir durant toute la durée du stage à l'IRCAN. Ainsi que le Dr.Jacob WARNER avec qui j'aurais partagé le bureau pendant 6 mois. Un grand merci également au Dr. Julien CHERFILS pour les pauses passées entre père de famille, mais également pour les conseils professionnels qu'il a pu m'apporter. Le Dr.Alexandre OTTAVIANI, désolé de t'avoir battu sur Hearthstone.

Un grand merci aux autres stagiaires bio-informatique de l'IRCAN, Vincent GUERLAIS, Emilien LEDANT, Christopher CAVALLIER et Marine POULLET, pour leurs conseils, les bons moments que j'ai pu passé avec eux, et leur soutiens au cours de mon stage. Ainsi que le magnifique Florent TESSIER, ingénieur au pôle bioinformatique de l'IRCAN.

Et bien évidemment, ma famille pour leur soutien tout au long de ma scolarité, et également mon épouse, Lilia, qui a su être présente et m'épauler tout au long de la durée de mon stage.

À ma fille Hanna.

Abstract

High throughput sequencing is now fast and cheap enough to be considered a viable part of the toolbox for investigating bacteria. Bacterial genome analysis is increasingly performed by diverse groups in research, clinical and public health labs alike, who are interested in a wide range of topics related to bacterial genetics and evolution. The main subjects of the microbiology are the study of pathogenicity and antimicrobial resistance. Whole genome sequencing is generally performed when the bacteria in question show interesting characteristics. Practical deployment of the technology is hindered by the bioinformatic challenge of analyzing results accurately.

We propose here the study of two strains isolated from clinical samples at hospital Archet 2, Nice. The first strain is a new bacterium species names *Nissabacter archeti*, isolated from pustule scalp of a 29-year-old man. This strain is representative of a new genus within the *Enterobacteriaceae* family. The second one is a new bacterium species of Klebsiella genus isolated from a patient (69 year old man) with a polycystic kidney disease.

These two strains showed an increased virulence and antimicrobial resistance. Whole genome sequencing was performed on these two bacteria. This study covered trimming of raw data, assembly, finishing, annotation, genome comparison and extracting common typing information. To identify the most closely related strains, we performed phylogenetic analysis based on several approaches. We made the analysis on the 16S rRNA structure, housekeeping genes and on the whole genome.

Here we present the main characteristics of *Klebsiella sp* and *Nissabacter archeti*.

Keywords : Computational genomic, Bacterial, Next generation sequencing, De novo genome analysis, clinical research.

1 Introduction

1.1 Généralités sur la notion d'espèce bactérienne

1.1.1 La taxonomie chez les bacteries

La taxonomie est la science qui permet de classer les organismes en groupes d'affinité ou taxons. Elle a pour but de leur attribuer une identité. Chez une bactérie, l'assignation d'une bactérie à un taxon permet d'en déduire les caractéristiques écologiques, épidémiologiques voire thérapeutiques que possède ce taxon.

1.1.2 Notion d'espèce

La notion d'espèce telle qu'elle a été définie dans la taxonomie des eucaryotes n'est pas applicable chez les bactéries. En bactériologie, une espèce est définie comme un ensemble de souches possédant des caractéristiques biochimiques, morphologiques et génétique similaires. Il s'agit d'une notion non figée, elle varie en fonction des critères choisis, avec le temps et l'évolution des techniques. En 2002, l'International Committee on Systematic of Procaryotes (ICSP) a eu la tâche de redéfinir les critères et de proposer une définition de l'espèce bactérienne [28]. Ainsi deux bactéries sont considérées de la même espèce si l'hybridation ADN à ADN est supérieure à 70%.

1.1.3 Limitations de la détermination d'espèce bacterienne

Cependant, de par la lourdeur opératoire de la technique des hybridations ADN-ADN (radioactivité, grosses quantités d'ADN, souvent des problèmes de reproductibilité d'un laboratoire à un autre...), le comité de Stackebrandt encourage l'utilisation d'autres techniques dont les résultats sont comparables à ceux des hybridations ADN-ADN et qui sont permises par les avancées technologiques et de biologie moléculaire. De nos jours, les études phylogénétiques sont basées sur l'utilisation de marqueurs moléculaires et des séquences de gènes. Nous exposerons brièvement quelques techniques permettant de classifier les espèces bactériennes.

1.2 Identification d'espèces

1.2.1 Identification par l'ARN ribosomal 16S

L'ARN ribosomique 16S (ARNr 16S) est l'ARN ribosomique constituant la petite sous-unité des ribosomes des procaryotes. La première classification des bactéries basée sur la séquence du gène de l'ARN ribosomal 16S a été proposée par Woese en 1977 [29]. À l'heure actuelle le séquençage de l'ARNr 16S est devenu le principal outil de la taxonomie bactérienne et il est d'ailleurs considéré comme référence pour toute description d'une nouvelle espèce [28]. Le gène de l'ARNr 16S se compose à la fois de régions très bien conservées et de régions hypervariables. Les régions conservées peuvent servir de sites de liaison d'amorces universelles pour l'amplification du gène entier ou de fragments du gène, tandis que les régions hypervariables contiennent des séquences spécifiques à l'espèce qui peuvent faire la distinction entre différentes espèces de bactéries. Avec plus de 100 000 séquences du gène de l'ARNr 16S disponibles dans les bases de données publiques, l'amplification et le séquençage de l'ARNr 16S peuvent identifier de manière relativement

précise les bactéries. Un niveau d'identité de séquence de 98.7% est généralement accepté pour associer 2 bactéries comme étant de la même espèce[27]. Cependant, cette technique montre des limites. En effet, il est compliqué d'utiliser seulement l'ARNr 16S, certaines bactéries présentent un très fort taux d'identité (supérieur à 99%) alors qu'elles ont des phénotypes très différents et ne sont pas de la même espèce. Ceci est démontré par l'hybridation ADN-ADN. [25]. De plus, la présence de plusieurs copies de l'ARNr 16S au sein du génome handicape l'établissement de la phylogénie dans certains groupes bactériens [21].

1.2.2 Multi Locus Sequence Typing : MLST

Cette technique repose sur l'analyse de certains gènes très conservés chez les bactéries. Ces gènes sont nommés des gènes de ménage. Sept à dix gènes sont généralement nécessaires pour cette technique [19]. Chaque fragment présente des variations de séquence d'une souche à l'autre. Les différentes séquences possibles de chaque gène représente un allèle. La combinaison des allèles trouvés chez une espèce définit la séquence type : ST qui est caractéristique d'une espèce. Ainsi une souche sera représentée par une combinaison d'allèles. Une autre souche possédant la même combinaison d'allèles sera de la même espèce. Cette technique est plus précise que l'analyse de l'ARNr 16S. Elle possède une résolution plus élevée et nous permettra de descendre au niveau des sous-espèces. En revanche le MLST n'est pas recommandé pour étudier des espèces éloignées.

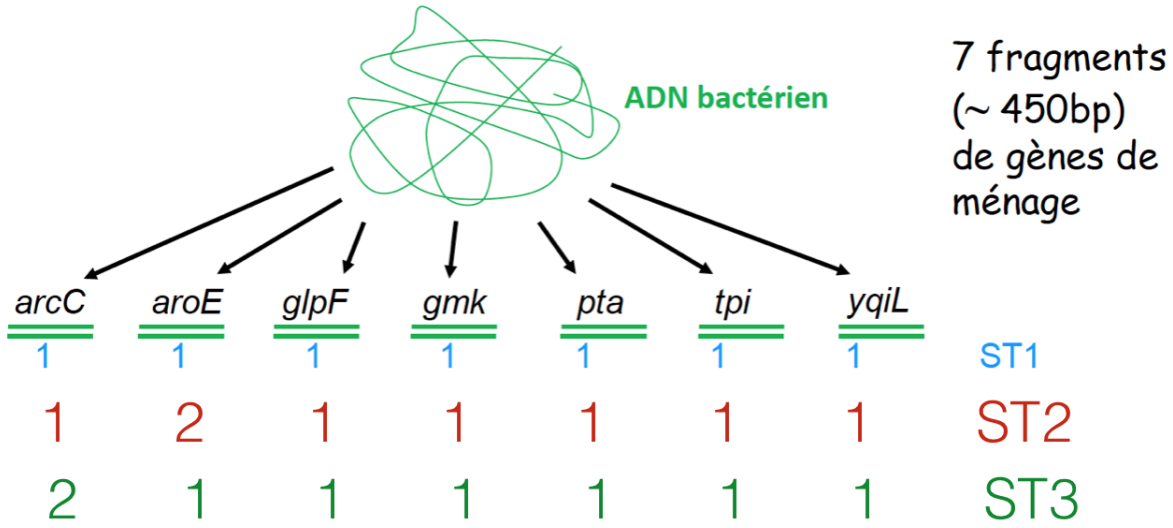


FIGURE 1 – La technique du MLST.
Elle se base sur l'analyse des gènes de ménages. Ici les gènes de ménages sont ceux utilisés pour *Staphylococcus aureus*. Ainsi chaque gène aura un allèle. Chaque souche aura une combinaison d'allèle pour chaque gène qui définira un Sequence Type : ST. Une souche possédant une variation sur un seul gène aura une combinaison d'allèles différents et sera considérée comme un ST différent. Par exemple ici, la deuxième souche présente une variation de séquence sur le gène aroE. Ainsi il s'agit d'un autre ST.

1.2.3 Spectrométrie de masse : MALDI-TOF

Cette technique est sûrement une des plus en vogue actuellement, surtout dans le domaine de la recherche clinique. La désorption-ionisation laser assistée par matrice (en anglais Matrix Assisted Laser Desorption Ionisation ou MALDI) est une technique d'ionisation douce utilisée en spectrométrie de masse. Elle permet l'ionisation d'échantillon. Une fois ionisés, ces composants vont se fragmenter. Dans le cas du MALDI-TOF, le spectromètre de masse est couplé à un analyseur de temps de vol où les molécules fragmentées en fonction de leurs poids et de leurs charges vont passer. Un profil de pic très précis des molécules présentes dans les échantillons est obtenu.

Dans le cas de la microbiologie, une souche est étalée sur une cible. L'échantillon est bombardé par un faisceau laser. Il en ressort un profil de pics, qui est comparé avec une base de données contenant les profils de pics de bactéries. Si le profil est reconnu dans la base de données, il s'agit d'une espèce connue. Sinon, d'autres méthodes sont à utiliser afin d'identifier l'espèce en question.

Cette technique est souvent utilisée en premier temps pour obtenir rapidement une identification de la souche, en effet elle ne nécessite pas d'extraction d'ADN ni de manipulation particulière. Elle est reproductible, efficace et peu coûteuse (hors coup de l'appareil).

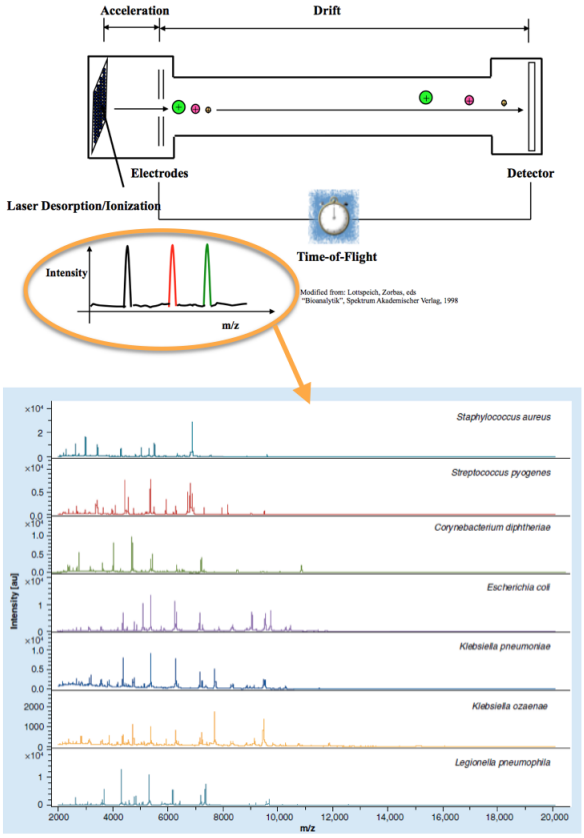


FIGURE 2 – La technique du MALDI-TOF. L'échantillon est étalé sur une matrice qui se fera bombarder par des faisceaux lasers. Les échantillons seront séparés en fonction de leur poids moléculaire m et leur charge z. Par la suite un profil de pics est obtenu. Cette technique est très discriminante, en effet chaque espèce de bactérie aura un profil de pics spécifique. Le profil de pics montre l'intensité en fonction de la masse de l'échantillon sur la charge.

1.3 Le séquençage à haut débit

1.3.1 Les avantages du séquençage à haut débit

La technologie du séquençage à haut débit a eu un énorme impact sur la science moderne et en médecine. De part son coût en baisse et sa rapidité, il est considéré à l'heure actuelle comme un outil à part entière en microbiologie pour étudier les bactéries [17]. Utilisé en recherche fondamentale, clinique ou encore en santé publique, il est utilisé pour un large panel d'études allant de la génétique microbienne à la génétique évolutive.

L'obtention d'un génome de bactérie peut être maintenant réalisée directement au laboratoire dans un délai de quelques heures à quelques jours grâce à des séquenceurs de type Illumina MiSeq, Ion Torrent PGM, ou Roche 545 FLX par exemple.

Cette technique nous donnera beaucoup plus d'informations car on aura à notre disposition le génome complet. Ainsi on pourra étudier les gènes que possède la bactérie et ses caractéristiques.

1.3.2 Technique de séquençage à haut débit

Cette technique est généralement utilisée lorsqu'il n'y a pas de génome connu similaire ou proche de cette bactérie dans les bases de données publiques. Ainsi il s'agira de faire du séquençage *De Novo*. C'est à dire obtenir la séquence d'un organisme pour lequel il n'existe pas une séquence de référence dans les bases de données. À l'issue du séquençage, il y aura un traitement bio-informatique des données. Ceci afin de les rendre interprétable et d'en tirer des informations. Par exemple les gènes de résistances aux antibiotiques ou des gènes de virulences. Ce traitement consistera à regrouper les fragments de lecture (reads) se chevauchant en séquence continue que nommée contigs (pour séquences contiguës). Le but étant d'avoir le moins de contigs possible et les plus longs possible, c'est-à-dire le génome le plus complet.

Ce processus est facilité par la production de lectures longues (exemple : PacBio qui produit de longues séquences de lecture) mais aussi de lecture "paired end" et "mate pair". Les banques dites "paired end" permettent le séquençage de deux extrémités de fragments courts, de tailles inférieures à 1kb alors que le séquençage de banques dites "mate pair" produit des lectures par paires des extrémités de fragments d'une taille de plusieurs kilobases.

1.3.3 Traitement bio-informatique des données

Les données nécessitent un traitement bio-informatique. En effet à la sortie du séquenceur, un fichier au format fastq est obtenu, il s'agit de fichiers bruts non interprétable. Ils contiennent les séquences de lectures et pour chaque séquence des scores de qualités.

Les principales étapes de traitements pour des données de génomique microbienne issues de séquenceur à haut débit sont :

— Tri des données : Filtrer les séquences de faibles qualités

— L'assemblage : assembler les séquences de lectures en séquences continues

— Scaffolding / Finishing : assembler les séquences continues en long segments voir obtenir le génome entier

— Annotation : annotation des séquences pour identifier les gènes présents ou ajouter d'autres informations en liens avec les bases publiques

En fonction de la nature des données, le traitement sera différent et on utilisera des outils différents. Le but étant d'optimiser le traitement afin d'avoir le jeu de données le plus propre possible, pour y extraire le maximum d'informations.

1.4 Le contexte clinique du stage

1.4.1 Identification d'une souche bactérienne en laboratoire clinique

Ce stage est orienté recherche clinique. Il est réalisé sous la tutelle du Pr.Ruimy, chef de service en bactériologie à l'hôpital de Nice l'Archet. Le processus d'identification d'une souche dans un laboratoire clinique sera brièvement expliqué.

Le point de départ est l'hospitalisation d'un patient pour une infection. Un bilan bactériologique ainsi qu'une hémoculture sont réalisés. Ces techniques vont révéler la présence ou non d'agent infectieux. Si le bilan ou l'hémoculture est positive, il y a infection bactérienne. La souche est isolée puis cultivée. Très souvent, la première technique d'identification est le MALDI-TOF, qui permettra en quelques minutes l'identification de la souche. Si on obtient aucune identification avec la technique du MALDI-TOF, l'ARNr du 16S est alors séquencée. Si d'après la séquence d'ARNr du 16S il n'y a aucune identification, ou qu'elle est ambiguë et si la souche présente un intérêt particulier (virulence accrue, signe de résistances aux antibiotiques) alors la souche est séquencée.

Le séquençage nous permettra d'avoir le génome complet de la bactérie. Ceci permettra de pouvoir l'étudier plus en profondeur, d'avoir plus d'informations sur les caractéristiques de la bactérie, d'avoir la liste de ses gènes notamment gènes de résistances aux antibiotiques et gènes de virulence. Une phylogénie peut être faite afin de connaitre et identifier les bactéries proches. De la génomique comparative peut être faites afin de voir quelles caractéristiques sont partagées avec d'autres espèces proches.

1.4.2 Échantillons à notre disposition

Lors de ce stage nous avons eu à notre disposition deux échantillons qui ont été isolés à l'hôpital de l'Archet à Nice.

La souche *Klebsiella sp* a été prélevée sur un patient ayant une polykystose rénale greffée. Il a récemment été hospitalisé pour une fièvre intense. Gardé sous observation à l'hôpital de l'Archet, il a été observé une persistance de la fièvre. Mis sous Tazocilline, qui est un antibiotique utilisé pour des souches multi-résistantes, il est observé une amélioration de l'état de santé du patient. La souche a été isolée et un MALDI-TOF a été effectué afin d'identifier l'espèce bactérienne. Le profil de pics n'a donné aucune identification dans la base de données. l'ARNr 16S a été séquencé, et encore une fois aucun résultat probant n'a été obtenu. Il a toutefois été déterminé qu'il s'agissait d'une entérobactérie, mais aucune identification précise au niveau du genre ou de l'espèce. Les éléments laissent facilement à penser qu'il s'agit d'une nouvelle espèce bactérienne. De plus, elle montre des signes de résistances aux antibiotiques et a provoqué une grosse fièvre chez ce patient signe du virulence accrue. Face à l'intérêt que présente cette souche, son génome a été séquencé.

Le second échantillon a été prélevé sur un patient à Nice à l'hôpital de l'Archet. Il provient d'un pustule sur le cuir chevelu d'un homme âgé de 29 ans. La souche a été isolée et un MALDI-TOF a été réalisé afin d'identifier la souche. Il n'y a eu aucun résultat d'identification. Le séquençage de l'ARNr 16S a été effectué. Là encore aucune identification. D'après les résultats du séquençage de l'ARNr 16S l'espèce en question est considérée comme un nouveau genre chez les entérobactéries. L'analyse du séquençage de l'ARNr 16S a ici déjà fait l'objet d'un article [23]. Ainsi cette souche a été nommée *Nissabacter archeti* (en référence à Nice et l'hôpital de l'Archet où la souche a été isolé). Afin d'en connaitre plus sur ce nouveau genre, le génome de la bactérie a été séquencé à l'URMITE (Marseille).

2 Objectif

La thématique principale de ce stage est d'analyser des génomes bactériens issus de séquenceurs à haut débit de seconde génération. Il s'agit de nouvelles espèces de bactéries, voir de nouveau genre de bactéries. Ainsi, il s'agit de traiter, analyser, étudier ces génomes avec une approche De Novo.

Deux génomes seront principalement étudiés lors de ce stage. Les objectifs seront donc d'obtenir le génome le plus propre possible et le plus complet possible. Il s'agira de traiter les séquences, annoter leurs génomes afin d'obtenir le plus d'informations possible ces bactéries. On pourra déterminer par la suite les caractéristiques de la souche, tel que les gènes de résistances aux antibiotiques, ou les gènes de virulence. Une étude phylogénétique sera menée sur ces deux souches. Il s'agira de déterminer les bactéries connues les plus proches. Ainsi on pourra faire de la génomique comparative. C'est-à-dire comparer les caractéristiques des souches entre elles et comparer leurs contenus génétiques.

De nombreux logiciels sont à notre disposition pour réaliser chaque étape. Ainsi il s'agira également de comparer quelques outils afin d'avoir la meilleure combinaison d'outils à utiliser en tenant compte de la spécificité de nos jeux de données. Ceci afin de choisir pour chaque jeu de donnée les logiciels les mieux adaptés.

3 Matériels et Méthodes

3.1 Séquençage des échantillons

Nous présenterons ici la façon dont ont été générées nos données. Elles ont toutes deux été séquencées à Marseille à l'URMITE.

3.1.1 *Nissabacter archeti*

Cette souche a été séquencée grâce à un séquenceur de type Illumina MiSeq en paired-end. La préparation de banques pour le séquençage Illumina en paired end consiste à fragmenter l'ADN génomique à des tailles inférieures à 0.8kB. Avec ce type de préparation de banques d'ADN on aura une orientation des séquences qui sera de type : Forward - Reverse (sens - antisens), le séquençage est convergeant. Les séquences de lectures (reads) seront séparées par une distance connue qui correspond à la taille des amorces (chez Illumina varie de 200 à 500 paires de bases).

3.1.2 *Klebsiella sp*

Cette souche a été séquencée grâce à un séquenceur de type Illumina MiSeq en Mate-pair. La préparation de banques pour le séquençage en mate pair est différente. L'ADN est fragmenté en morceaux compris entre 2kB et 15kB. Ses fragments sont circularisés. Les fragments circulaires sont à nouveau fragmentés puis séquencés. Ici le séquençage sera de type Reverse - Forward (antisens - sens), c'est-à-dire divergeant. Les séquences de lectures seront beaucoup plus grandes et seront séparées par une distance qui correspond à la taille de l'amorce (chez illumina : 2 à 5kB).

3.2 Reconsitution des génomes

Il s'agira ici de présenter les différentes étapes de traitements de données microbiologiques issues d'un séquenceur à haut débit de second génération. La méthodologie pour la souche *Nissabacter archeti* sera d'abord exposée puis celle de *Klebsiella sp.* Le traitement sera quasiment identique, une étape supplémentaire sera utilisée pour *Klebsiella sp.* Il s'agira ici d'évaluer la meilleure combinaison d'outils à utiliser en tenant compte de la spécificité de nos jeux de données.

3.2.1 Traitement des données de *Nissabacter archeti*

Qualité du séquençage et tri des données

La première étape de traitement de données que nous effectuons est le contrôle de la qualité du séquençage et le tri les séquences de faible qualité. Pour évaluer la qualité de notre séquençage nous sommes passés par FASTQc (v0.11.4)[2]. Ce logiciel prend les fichiers bruts obtenus à la sortie du séquenceur qui sont des fichiers au format .fastq. Il va nous renvoyer un fichier au format HTML que l'on peut ouvrir via un navigateur web.

Si la qualité du séquençage est acceptable, nous pouvons alors passer à l'étape suivante qui consiste à filtrer les éléments indésirables. Nous nous sommes basés sur une publication qui compare différents logiciels permettant le trimming (tri des séquences) [7]. Basé sur cet article nous avons sélectionné Trimmomatic (v0.36)[6]. Plus gourmand en puissance

de calcul et donc en temps, c'est celui qui sera le plus sensible à la détection de séquences de faible qualité (scores inférieurs indiqués comme étant inférieur à 20). L'analyse est lancée en mode PE (paired-end) pour tenir compte des deux brins du séquençage. Les reads possédant un score inférieur à 20 (mauvaise qualité) sont filtrés. Les reads trop courts (inférieur à 36 bases) sont rejetés ainsi que les séquences chimériques. Trimmomatic renvoie 4 fichiers, 2 concernent le brin sens et anti-sens pour lesquels les reads filtrés sont appariés (fichiers paired) et 2 autres fichiers, un pour le sens l'autre pour l'anti-sens qui contiennent les reads non appariés (fichiers unpaired). À l'issue de cette étape, la qualité des échantillons est à nouveau évaluée en utilisant FASTQc.

Assemblage des séquences de lecture en contig

Pour l'assemblage plusieurs logiciels ont été testés. Nous avons restreint le choix des assembleurs testés en tenant compte des travaux de Magoc [18]. Cette publication compare différents assembleurs. Nous avons choisi ceux qui montrent la meilleure efficacité en terme de nombre de contigs et de longueur d'assemblage des contigs. Nous avons sélectionné Velvet (v1.2.10)[30] et SPAdes (v3.9.1)[4]. Les deux utilisent la méthode des graphes de Bruijn pour réaliser l'assemblage. Cette méthode nécessite de choisir une ou plusieurs tailles de motifs (k-mers). Pour évaluer la qualité d'un assemblage nous utiliserons Quast (v4.4) [12], qui renvoie des statistiques permettant de juger de la qualité d'un assemblage.

Assemblage avec Velvet (v1.2.10)

Velvet traite les données séquencées en Paired-end. Il va utiliser deux modules pour faire de l'assemblage velveth et velvetg. Dans un premier temps Velveth est lancé, il faut saisir la taille du k-mers choisis pour réaliser l'assemblage. Des taille de k-mers allant de 21 à 111 ont été définies avec un pas de 6. Ce qui correspond à 15 k-mers différents. Velvet prend en entrée les fichiers paired du séquençage ainsi que les unpaired. Ensuite Velvetg va réaliser l'étape d'assemblage des k-mers en fragments continus. Nous avons réalisé cette étape pour chacun des k-mers produit par Velveth (soit avec les 15 k-mers différents). la couverture du séquençage est içi laissée en mode automatique, la taille des adaptateurs soit 400 pour cette souche et la longueur minimum des contigs que l'on souhaite à la fin (huit cent nucléotides).

Assemblage avec SPAdes (v3.9.1)

SPAdes a été utilisé en mode paired-end. Il faut saisir l'orientation du séquençage, forward reverse pour *Nissabacter archeti*. Il faut saisir quel fichier correspond au brin sens et anti-sens ainsi que la nature du séquençage (paired-end). On a pris une large gamme de k-mers : 21,33,55,77,99,101,127. Nous avons testé plusieurs combinaisons, celle ci a été retenue comme la plus efficace. Le seuil de la couverture du séquençage a été laissé en mode automatique, le mode careful a été activé ce qui permettra une étape de correction des mésappariements. Les fichiers en entrés sont renseignés dans un fichier de configuration. SPAdes fournis en sortie d'analyse un fichier "contigs" ainsi qu'un fichier "scaffolds". En effet il réalise une étape de scaffolding en interne.

Finition sur les données du génome

À partir du fichier scaffolds créé par SPAdes, nous avons utilisé Gapfiller (v1.10) [24]afin de reboucher les trous (gaps) présents dans le génome. Gapfiller prend en entré le fichier scaffolds, et les fichiers bruts du séquençage avec lesquels il va reboucher les gaps. Concernant les options, on saisit la taille de l'insert qui est de 400 ici, l'orientation du séquençage (forward-reverse) et le nombre de bases des reads devant se chevaucher avec le scaffold à 50. 30 itérations ont été fixées pour reboucher les gaps. À l'issue de cette étape, les scaffolds ont été découpés au niveau des gaps grâce à des scripts personnels écrit en Python (v2.7.6). Un filtre a également été mis en place afin de ne pas retenir les contigs de tailles inférieurs à 800 nucléotides. À cette étape nous avons à notre disposition un fichier contenant les contigs. Il représente le génome de *Nissabacter archeti*.

3.2.2 Traitement des données de *Klebsiella sp*

Les mêmes étapes ont été appliquées afin de réaliser le traitement des données de Klebsiella sp. Cependant de part la différence entre le séquençage réalisé en mate-pair et celui réalisé en paired-end, des changements d'options sont nécessaires et une étape supplémentaire sera réalisée. Ainsi on précisera ici les différences avec le précédent traitement sans revenir sur les points similaires.

Qualité et tri des données

Ainsi on a tout d'abord vérifié la qualité du séquençage via FASTQc. Il s'avère qu'il y avait encore présence d'adaptateur Illumina TruSeq (identifié par FASTQc). Les séquences adaptatrices ont été coupées (séquence : 'AGATCGGAAGAGC'). Les mêmes critères de tri ont ensuite été appliqués ici.

Assemblage des séquences de lecture en contigs

La même comparaison a été effectuée ici afin d'avoir le logiciel le mieux adapté à nos données.

Assemblage avec Velvet 1.2.10

La différence majeure ici est due au fait que Velvet prend en entrée des fichiers séquencés en Paired-end. L'orientation du séquençage étant différente il a fallu passer par la séquence inverse complémentaire pour avoir un assemblage cohérent. Ceci a été réalisé grâce à un script écrit en python version 2.7.6. Les options sont similaires ici sauf la taille de l'insert qui est de 4000 bases.

Assemblage avec SPAdes 3.9.1

SPAdes dispose d'un mode mate-pair. Ce qui évite de passer par la séquence complémentaire inverse. Il suffit de renseigner dans le fichier de configuration le type de séquençage. SPAdes tiendra compte du fait que en mate-pair les reads sont plus longs. Il faut également saisir l'orientation du séquençage qui est ici reverse-forward (orientation du séquençage pour un séquençage en mate-pair chez Illumina). Les mêmes k-mers ont été utilisés. Les autres options sont identiques.

Améliorations et finitions des données

À partir du fichier scaffold que génère SPAdes, les gaps ont été rebouchés de la même manière via Gapfiller. Concernant les options, nous avons choisi d'utiliser le nombre de bases des reads devant se chevaucher avec le scaffold à 130. Les options modifiées sont la taille de l'insert, ici à 4000 bases, l'orientation du séquençage (reverse-forward). Les autres options ont été laissées par défaut. À la fin de cette étape les scaffolds ont été éclatés au niveau des gaps pour former des contigs.

De part la longueur des fragments d'ADN obtenus par la technique du mate-pair, il est possible ici d'ajouter une étape supplémentaire qui est de regrouper les contigs en fragments plus longs : les scaffolds. Pour cette étape nous avons comparé : SSPACE (v3.0) [5], Opera (v2.0.6) [10] et scaffmatch (v0.9) [20] Nous nous sommes basés sur un article [13] comparant différent logiciel permettant le scaffolding pour les sélectionner. SSPACE est cité comme générant très souvent de meilleurs résultats que les autres [13]. Opera a récemment sorti une nouvelle version (2016). Scaffmatch est plus récent que le papier comparant les scaffolders ainsi nous avons choisi de le comparer aux deux autres.

Pour l'utilisation de SSPACE, on passe par la création d'un fichier de configuration qui contient quelques options telle que le choix du logiciel pour l'alignement des séquences (BWA a été choisi) et les fichiers bruts du séquençage au format fastq. SSPACE accepte les fichiers appariés (paired) ainsi que les non appariés (unpaired). On y entre la taille de l'insert (4000 nucléotides) et l'orientation du séquençage (reverse-forward). L'option relative à la taille des reads (-m) a été choisi à 130 pour *Klebsiella sp.*

Opera va passer par une étape de pré-traitement, où il construit une carte grâce aux contigs et aux reads bruts. Pour se faire nous avons utilisé BWA. On saisit comme options le type de séquenceur utilisé (Illumina), le fichier contenant les contigs et les deux fichiers bruts (paired uniquement). On a choisi la taille du k-mers à 100. La taille des reads à 250, la taille de l'insert à 4000. L'option permettant de le paralléliser a été utilisée.

Scaffmatch prend en entrée le fichier de contigs ainsi que les fichiers bruts du séquençage. La longueur de l'insert est saisie à 4000, l'orientation du séquençage en reverse-forward. Scaffmatch est peu personnalisable comparé à SSPACE, mais dispose de plus d'options qu'Opera.

À l'issue du scaffolding notre échantillon comporte des gaps. Gapfiller a été utilisé afin d'en reboucher le maximum. Les scaffolds ont ensuite été éclatés en contigs au niveau des gaps.

3.3 Analyses génomiques

Une fois les données triées, optimisées et assemblées, on va pouvoir tirer des informations du génome. Ici la méthodologie des deux génomes sera traitée simultanément. En effet les mêmes analyses ont été effectuées pour les deux souches.

Un megablast du génome complet sur NCBI a d'abord été fait sur la base de données "Genomes".

3.3.1 Extraction In Silico et analyse de l'ARNr 16S

La séquence d'ARNr 16S a été extraite. RNAmmer [16] permet d'extraire de façon *In silico* la séquence 16S. Pour cela il faut entrer le règne de l'espèce à analyser (bactérie) et entrer l'option pour obtenir les sous unités ribosomales.

Un blastn [14] a été effectué sur la séquence d'ARNr du 16S. La base de données choisie est "Genomes" sur NCBI. Les paramètres par défaut ont été conservés. Les cent premiers résultats sont affichés. Parmi les résultats, 16 bactéries d'espèces différentes ont été sélectionnées. Les séquences alignées ont été téléchargées. Une phylogénie a ensuite été réalisée sur ces séquences. Nous avons choisi cette base de données car par la suite d'autres études phylogénétiques vont être menées et nous souhaitons comparer les résultats des différentes méthodes. Un blastn a été réalisé sur la base de données "prokaryotic 16S ribosomal RNA" sur le NCBI. La phylogénie a été faite via MEGA (v7.0) [15]. Un alignement local a été réalisé dans un premier temps en utilisant l'algorithme MUSCLE[8]. Les paramètres sont laissés par défaut. Un arbre phylogénétique a été réalisé à partir de cet alignement. La méthode utilisée est celle du Maximum Likelihood couplée à un test de phylogénie de type bootstrap avec 300 itérations. Les autres paramètres ont été laissés par défaut.

3.3.2 Annotation du génome

L'annotation a été faite via PROKKA (v1.11) [26] qui permet une bonne personnalisation de cette étape. Il a été préféré a RAST [3], moins personnalisable. Ainsi il faut sélectionner le règne (bactérie), préciser qu'il s'agit d'une gram-, l'identification des ARN a également été réalisée. PROKKA va interroger plusieurs bases de données, UniprotKB dans un premier temps, et d'autres bases de données telles que HMM. Ceci permettra d'avoir des résultats plus fins. Il sortira par la suite plusieurs type de fichiers tel que le fichier Genbank au format gbk contenant le génome, les gènes codants, leurs positions. Un fichier au format ffn contenant seulement les séquences en ADN des gènes codants et un fichier au format faa qui contient les séquences en acides aminés des gènes codants.

3.3.3 Phylogénie basée sur les gènes de ménages

Nous avons extrait les séquences des gènes considérés comme des gènes de ménages chez les entérobactéries (AtpD, RpoB, InfB, GyrB) [11]. Un blastn a été réalisé sur chacun de ses gènes de ménages sur la base de données "Genomes" du NCBI. 16 espèces ont été sélectionnées. Afin d'avoir des résultats comparables entre cet méthode et le 16S, les mêmes espèces ont été sélectionnées. Les séquences alignées ont été téléchargées et une phylogénie a été réalisée à l'aide de MEGA. Les mêmes paramètres ont été utilisés, l'alignement a été fait par MUSCLE, l'arbre a été construit par Maximum Likelihood avec un test de bootstrap à 300 itérations.

3.3.4 Phylogénie basée sur la totalité du génome

D'après la phylogénie précédemment établie, 11 espèces ont été sélectionnées afin de réaliser une expérience d'hybridation ADN-ADN In Silico. Leurs génomes ont été téléchargés et l'hybridation ADN-ADN In Silico a été réalisé via GGDC (v2.1) [22] . GGDC prend en entrée deux fichiers et renvoie la distance relative des deux génomes.

Il a fallu faire toutes les combinaisons de génome à génome. Pour onze génomes cela correspond à 55 combinaisons. Une fois tout ses combinaisons effectuées une matrice de distance a été construite. MEGA a ensuite pu reconstruire un arbre phylogénétique à partir de cette matrice. La méthode utilisée est celle du Neighbor-joining, sans bootstrap. Les valeurs de la matrice étant figées aucune itération n'est possible ici.

3.3.5 Comparaison de la composition du génome à deux espèces proches

Deux espèces ont été choisies comme étant les plus proches d'après les phylogénies précédemment établies. Leurs génomes ont été annotés de façon locale grâce a PROKKA. Une étude des orthologues a été faite. Cette analyse a été effectuée via orthoMCL [9]. Pour cela le fichier faa que génère PROKKA a été utilisé. OrthoMCL estime comme orthologues des protéines dons l'e-value d'un blastp est inférieure à 10^5 et le pourcentage d'identité d'au moins 50%. Il en ressort un fichier contenant les orthologues connus chez nos bactéries. Un diagramme de Venn a ensuite été fait à partir de ces résultats afin de visualiser le nombre de protéines en commun et celles uniques à notre espèce.

Via rpsblast [1], on a également effectué une classification des protéines en catégories fonctionnelles. En se basant sur la base de données des COG (Cluster of Orthologous Genes) on a pu envoyer nos séquences codantes. Ceci nous a permis d'identifier les catégories fonctionnelles dans lesquelles sont impliqués nos gènes.

4 Résultats

4.1 Reconstitution des génomes

4.1.1 Qualité et tri des données

Cette étape est importante car elle permet de partir sur de bonnes bases pour l'analyse. Cependant il faut veiller à ne pas être trop stringent lors de cette étape afin de ne pas enlever trop de séquence. Ce qui rendrait l'étape d'assemblage plus compliquée.

Nissabacter archeti

Grâce à FASTQc on sait qu'à l'issue du séquençage que cet échantillon est composé de 702676 séquences par brin. La longueur des reads est de 250 nucléotides, le pourcentage GC est de 57% et il y a présence de séquences présentant un score inférieur à 20 (considérées comme des séquences de mauvaises qualités). À l'issue du tri 586364 reads pairs sont conservés (84.45%). Pour les reads unpaired : 87.618 reads sont conservés (12.47%) pour le brin forward et 6684 conservées (0.95%) pour le brin reverse. Au final 22.010 reads n'ont pas été retenus soit 3.13%. Toutes les séquences ont un score supérieur à 28 (considéré comme très bon). Les reads font 232 nucléotides en moyenne après trimming.

Klebsiella sp

Le séquençage a produit 969448 séquences par brin. La longueur des reads est de 250 nucléotides, le pourcentage GC est de 52% , il y a présence de séquence de faible qualité et de séquences sur-représentés.

Après l'étape de tri des données, on a 600684 reads pairs conservé (61.96%). Pour les reads unpaired, 217.107 (22.39%) séquences ont été conservé sur le brin forward et 118.049 (12.18%) pour le brin reverse. Au final 33.608 reads n'ont pas été retenu soit 3.47%. Les séquences ont toutes un très bon score (supérieur à 28), il n'y a aucune séquence sur-représentée et la longueur moyenne des reads est de 250 nucléotides.

4.1.2 Étape d'assemblage

Une fois les données triées, elles ont été assemblées en *De Novo*. L'assemblage De Novo consiste à regrouper les fragments de séquences d'ADN se chevauchant en séquences continues que l'on nomme : contig et ceci sans l'aide d'un génome de référence connu dans les bases de données publiques. Grâce à certains critères on pourra comparer plusieurs résultats d'assemblage entre eux. Un bon assemblage est caractérisé par un nombre faible de contigs (le moins de fragment de génome possible), la taille du plus long fragment et le N50. Le N50 est une mesure évaluant la continuité d'un assemblage. Il s'agit de la taille du segment (contig ou scaffold) telle que la moitié de la somme des bases de tous les segments (assemblage) soit comprise dans des segments de taille supérieure.

Velvet s'est révélé être moins performant sur l'ensemble des données. Nous avons testé 15 k-mers différents. Il nous renvoie un nombre supérieur de contigs et un N50 inférieur, signe d'un assemblage moins bon. Les détails sont disponibles dans la figure 3.

SPAdes nous donne de meilleurs résultats sur l'ensemble des données. Le nombre de contigs est environ cinq fois inférieur à celui obtenu avec Velvet. Le N50 est quant à lui quasi trois fois supérieur à celui obtenu avec Velvet.

	Contig le plus grand	Nb de contigs (≥800nt)	N50
SPAdes 3.9.1 *Klebsiella sp*	1.565.619	16 scaffold : 10	1.269.645
SPAdes 3.9.1 *Nissabacter archeti*	494.061	67 scaffold : 54	120.086
Velvet 1.2.10 *Klebsiella sp k-mers : 93*	649.300	35	265.232
Velvet 1.2.10 *Nissabacter archeti k-mers : 55*	129.183	375	43.032

FIGURE 3 – L'assemblage a été fait avec SPAdes 3.9.1 et Velvet 1.2.10. Nous comparons ici les résultats des assemblages. Les données exposées nous permettent d'évaluer un assemblage. Il faut avoir le plus petit nombre de contigs possible et le N50 le plus élevé. SPAdes montre un nombre de N50 bien plus élevé et un nombre de contigs inférieur.

4.1.3 Amélioration de l'assemblage

Nissabacter archeti

Le fichier de scaffold sorti par SPAdes est repris. Les gaps ont été bouchés avec Gapfiller et les scaffolds brisés au niveaux des gaps. Gapfiller a permis de reboucher un gap, 221 nucléotides inconnus (N) ont été remplacés par des bases grâce a Gapfiller, laissant un total de 12 gaps et 911 N. À cette étape le fichier est composé de 54 scaffolds. Les scaffolds ont été brisés au niveau de ces gaps.

À l'issue de ces étapes le génome est constitué de 66 contigs. Le plus large est de 494.061 bases, le N50 est de 120.148. La taille totale du génome est de 5.14MB.

Klebsiella sp

Le fichier scaffolds produit SPAdes est composé au total de 10 scaffolds. via Gapfiller on a tenté de rebouché les gaps du fichiers scaffolds. Il y a eu 2 gaps rebouchés sur 6 et 3098 nucléotides inconnus remplacés par des bases sur 5211. Les scaffolds ont ensuite été brisé au niveau des gaps en contigs.

Nous avons tester esnuite tester plusieurs assembleurs : SSPACE, Opera et Scaffmatch. SSPACE est le logiciel nous ayant fournis les meilleurs résultats (figure 4).

	Scaffold le plus grand	Nb de scaffolds (≥800nt)	N50
SSPACE 3.0 *Klebsiella sp*	4.042.670	5	4.042.670
Opera 2.0.6 *Klebsiella sp*	1.567.706	14	1.269.642
Scaffmatch 0.9 *Klebsiella sp*	1.567.706	14	1.269.642

FIGURE 4 – L'assemblage a été fait avec SSPACE 3.0, Opera 2.0.6 et Scaffmatch 0.9. Nous comparons ici les résultats de l'étape de scaffolding. Les données exposées nous permettent d'évaluer le scaffolding. Il faut avoir le plus petit nombre de contigs possible et le N50 le plus élevé. SSPACE montre le meilleur résultat pour l'étape de scaffolding.

SSPACE a été utilisé pour regrouper les contigs en scaffolds et Gapfiller utilisé pour reboucher les gaps. À l'issue de cette étape, nous avons cinq scaffolds dont le plus grand est de 4.042.670 bases. Le N50 de ce génome est de 4.042.670 il y a 4407 N (bases inconnus dans les gaps) et 9 gaps. Les scaffolds ont à nouveau été découpés en contigs, il y en a un total de 13. La longueur total du génome est de 5.22MB.

4.2 Analyses génomiques

4.2.1 *Nissabacter archeti*

Le megablast du génome contig par contig montre un penchant pour le genre *Serratia*. En effet, sur les 66 contigs, 42 montrent en top position une identification avec une bactérie du genre Serratia. Quelques contigs ont été identifié comme étant des plasmides.

Étude de la séquence d'ARNr 16S

La séquence d'ARNr 16S a été identifié. Le gène est complet, sa longueur est de 1530 bases. D'après le blastn de l'ARNr 16S réalisé sur la base de données "Genomes", on a un pourcentage d'homologie qui est de 96.99% avec la bactérie *Serratia rubidaea*. Sur la base de donnée 16S rRNA le top hit est *Ewingella americana* avec 98% d'identité, *Serratia rubidaea* est second avec 97% d'homologie. Une phylogénie a été effectuée d'après les résultats sur la base de données "Genomes" avec un test de bootstrap. Une valeur de "bootstrap" (pourcentage de 0 à 100%) est associée à chaque branche de l'arbre indiquant le nombre de fois où cette branche a été retrouvée au fil des répétitions et juger ainsi de leur crédibilité. En d'autres termes, la valeur de "bootstrap" indique une évaluation de la résistance d'un noeud à la perturbation des données. Le résultat est visualisable dans la figure 5. On peut voir que notre bactérie est proche de *Rahnella aquatilis* ainsi que des deux *Serratia*. Les valeurs de bootstraps sont relativement faibles (inférieures à 90%).

Annotation du génome

À l'issue de l'annotation on a pu savoir que notre génome est composé de 4755 gènes. On sait également que notre génome contient 76 ARNt et 94ARN non codant. L'annotation a mis en évidence les gènes de résistances aux antibiotiques et gènes de virulence. Un total de 48 gènes sont annotés comme étant des gènes de résistances. Certains sont annotés comme étant des gènes multi-résistants.

Analyse des gènes de ménages

Les résultats sont similaires à l'étude du 16S. Beaucoup de *Serratia* ressortent. Une phylogénie a été réalisée d'après les résultats du blastn. Les résultats seront exposés sous forme d'arbres sur la figure 5 avec à chaque fois l'espèce présentant le plus fort taux d'homologie basé sur les résultats du blastn. L'arbre le plus pertinent est présenté en figure 5. Les autres arbres seront consultables en annexe. Sur AtpD la bactérie présentant le plus fort taux d'homologie est *Serratia marcescens* avec 91.3%. Pour gyrB c'est *Serratia marcescens* qui a le plus fort taux d'homologie avec 86.3%. InfB : *Serratia rubidaea* avec 85.1%. RpoB : *Serratia marcescens* avec 89.2%. Une chose intéressante est que sur tous les arbres, *Nissabacter archeti* est toujours au sein d'un noeud regroupant *Serratia marcescens*, *Serratia rubidaea* et *Rahnella aquatilis*.

Phylogénie basée sur la totalité du génome

La distance de génome à génome est déterminée par la similarité des séquences sur la totalité du génome. GGDC utilise un alignement local (blast+) pour calculer cette distance. La distance est transformée en une valeur analogue à l'hybridation ADN-ADN

(présenté en introduction) par une régression linéaire généralisé (GLM). Les paramètres de cette régression ont été obtenus depuis des jeux de données de références. Pour ces jeux de données la corrélation entre hybridation ADN-ADN et la distance de génome à génome sont connues. Plus il y a d'espèces et plus le nombre de combinaisons d'hybridations ADN-ADN *In silico* augmentent.11 espèces de bactéries ont été retenues pour réaliser une hybridation ADN-ADN In Silico. Les espèces proches ont été conservées. N'ont pas été conservés *Shigella boydii, Pectobacterium atrosepticum, Pectobacterium wasabiae, Erwinia carotovora, Cedecea neteri. Serratia marcescens* et *Serratia rubidaea* ont été identifiées comme étant les plus proches. L'arbre est visualisable sur la figure 5.

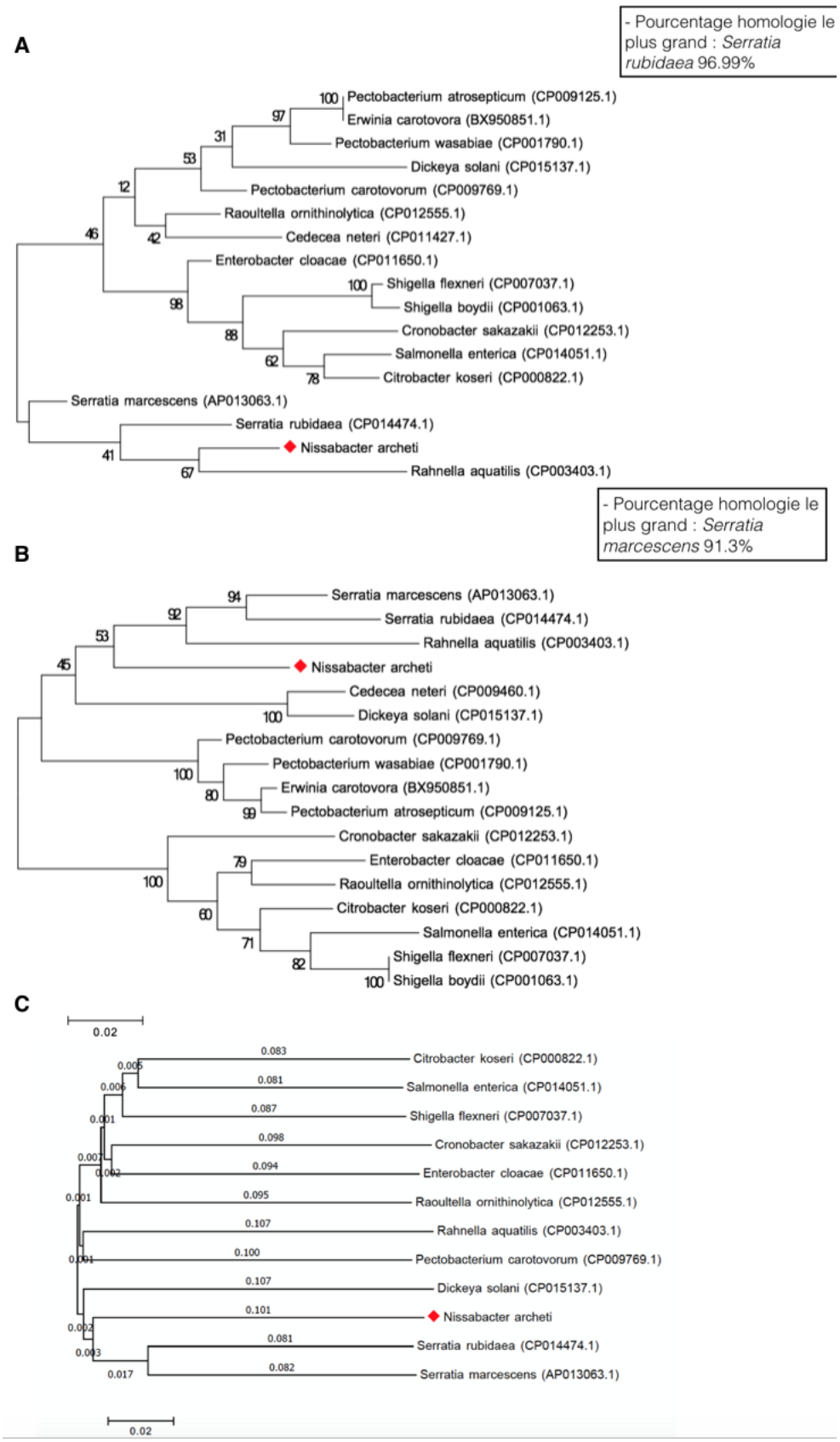


FIGURE 5 – Cette figure expose plusieurs arbres phylogénétiques. Les blastn qui ont servi de bases pour la construction des arbres ont tous été réalisés sur la base de données "Genomes" du NCBI. Les mêmes espèces ont été sélectionnées entre les différentes études pour pouvoir comparer la position de notre bactérie.**A**. Cet arbre est basé sur la séquence de l'ARNr 16S. L'espèce montrant le plus fort taux d'homologie est *Serratia rubidaea*. **B**. Cette arbre est basé sur le gène AtpD. *Serratia marcescens* est l'espèce montrant le plus fort taux d'homologie. **C**. Phylogénie basée sur la totalité du génome. *Serratia marcescens* et *Serratia rubidaea* sont proches de notre bactérie.

Comparaison du contenu génomique avec des espèces proches

Les orthologues sont des protéines similaires que l'on retrouve chez des espèces différentes. On entends par similaires des protéines dont l'identité (bases à bases) est supérieure à un certain seuil. OrthoMCL estime comme orthologues des protéines dont l'e-value d'un blastp est inférieure à 10^5 et le pourcentage d'identité d'au moins 50%. Il en ressort que notre souche possède 2339 protéines en commun avec ses deux espèces. 145 protéines avec Serratia marcescens seulement et 137 avec Serratia rubidaea seulement. Au final elle dispose de 541 protéines uniques à son espèce.

A.

	Nissabacter archeti	*Serratia rubidaea*	*Serratia marcescens*
Nb bases	$5.1.10^6$	$4.9.10^6$	$5.1.10^6$
Gènes	4755	4621	4876
Signal peptide	432	436	504
rRNA	9	22	22
tRNA	76	78	90
miscRNA	94	76	89
tmRNA	1	1	1

B.

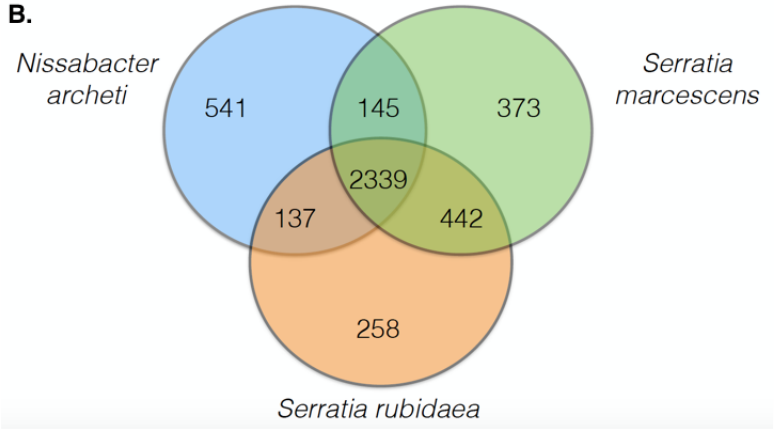


FIGURE 6 – **A**. Comparaison du contenu génomique de *Nissabacter archeti* avec *Serratia marcescens* et *Serratia rubidaea*. **B**. Comparaison des orthologues entre *Nissabacter archeti* avec *Serratia marcescens* et *Serratia rubidaea*. Notre bactérie a 541 protéines qui ne sont pas partagées avec ces deux souches.

Les COG sont une classification de protéines en fonction de catégories fonctionnelles. Il s'agit donc de classer les protéines dans une ou plusieurs catégories, en fonction de leur similarité avec d'autres protéines précédemment classées (base de données des COGs). Pour cela on utilise typiquement rpsblast qui ne recherche pas des similarités d'alignement comme pour un blast normal mais qui cherche des signatures.

RPSblast nous renvoie les protéines classées en catégories fonctionnelles de notre bactérie. Une comparaison avec les deux autres espèces est également faites (résultat figure 7). Chez *Nissabacter archeti* la catégorie la plus importante (hors fonction inconnue et fonction générale) est : Transport et métabolisme des carbohydrates. Chez les deux autres bactéries il s'agit de " Transport et métabolisme des acides aminé". Mais la différence pour chaque catégorie entre chaque bactérie est très proche.

Espèces		*Nissabacter archeti*	*Serratia rubidaea*	*Serratia marcescens*
Code	Description	% total		
J	Translation, ribosomal structure and biogenesis	3.72	3.85	3.59
A	RNA processing and modification	0.02	0.02	0.02
K	Transcription	7.39	7.84	8.64
L	Replication, recombination and repair	3.56	2.8	2.69
B	Chromatin structure and dynamics	0	0.02	0.02
D	Cell cycle control, cell division, chromosome partitioning	0.83	0.69	0.76
Y	Nuclear structure	0	0	0
V	Defense mechanisms	0.95	1.03	1.01
T	Signal transduction mechanisms	3.66	3.47	3.65
M	Cell wall/ membrane biogenesis	4.6	4.78	4.85
N	Cell motility	2.17	2.16	2.24
Z	Cytoskeleton	0.02	0	0
U	Intracellular trafficking and secretion, and vesicular transport	2.47	2.64	2.41
O	Posttranslational modification, protein turnover, chaperones	2.67	2.92	2.92
C	Energy production and conversion	4.74	5.14	4.97
G	Carbohydrate transport and metabolism	8.18	7.03	7.66
E	Amino acid transport and metabolism	7.96	8.77	9.17
F	Nucleotide transport and metabolism	1.88	2.12	1.85
H	Coenzyme transport and metabolism	3.32	3.61	3.34
I	Lipid transport and metabolism	2.17	2.12	2.66
P	Inorganic ion transport and metabolism	4.33	3.61	5.62
Q	Secondary metabolites biosynthesis, transport and catabolism	1.36	2.82	2.35
R	General function prediction only	10.02	10.62	10.28
S	Function unknown	23.97	20.29	19.28

FIGURE 7 – Tableau des COGs chez *Nissabacter archeti*.Ce tableau reprend les protéines classées en catégories fonctionnelles. On peut observer que pour chaque catégorie un code est attribué. Pour chaque catégorie, on a le pourcentage relatif des proteines impliqués dans cette catégorie. Les COGs sont comparés avec les deux autres espèces.

4.2.2 *Klebsiella sp*

Dans un premier temps nous avons fait un megablast du génome complet contig par contig. Ceci nous a permis de voir que notre bactérie était très proche du genre *Klebsiella*. En effet sur le plus gros contig les deux premiers résultats sont *Klebsiella michiganensis* et *Klebsiella oxytoca*. Une espèce de *Salmonella* semble également être proche, ainsi qu'une espèce de *Citrobacter*. Une chose intéressante est qu'un contig a été identifié avec 100% d'homologie comme étant un phage phiX174. Ce phage est utilisé dans les kits de séquençage Illumina. Ainsi nous avons pu mettre en évidence il s'agit d'une contamination. Le contig en question de longueur égale à la taille du plasmide a été enlevé du génome.

Étude de la séquence d'ARNr 16S

La séquence d'ARNr 16S a ensuite été extrait via RNAmmer. Nous avons extrait deux séquence d'ARNr 16S. Un des 16S n'est pas complet, le second est complet avec une longueur de 1528 nucléotides. La séquence d'ARNr 16S a été analysée par blastn et une phylogénie a été faite. D'après la séquence d'ARNr 16S le génome est proche de *Citrobacter Koseri*(figure 8). Mais le pourcentage d'homologie est de 97.9% loin des 98.7% nécessaire pour dire qu'ils sont de la même espèce. De plus les résultats sont assez serrés et de nombreuses *Klebsiella* sont présents à des taux légèrement inférieurs (97.85% pour Klebsiella michiganensis et 97.78% pour Klebsiella oxytoca). Ces résultats ne sont pas suffisants et disposant du génome nous allons approfondir les analyses afin de placer au mieux cette nouvelle bactérie. De plus sur la base de donnée 16S rRNA, *Klebsiella oxytoca* est en top position avec une homologie de 98.6%. La seconde espèce est *Citrobacter amalonaticus* avec une homologie à 98.23%.

Annotation du génome

À l'issue de l'annotation on a pu savoir que notre génome est composé de 4874 gènes. Nous savons également qu'il y a 77 ARNt et 98 ARN non codant. Ceci a mis en évidence les gènes de résistances aux antibiotiques et les gènes de virulence. Un total de 41 gènes sont annotés comme étant des gènes de résistances. Cette souche dispose de gènes de résistances multiple (annotés Multridrug résistance protein).

Analyse des gènes de ménages

Afin de continuer l'analyse phylogénétique, des gènes de ménages ont été récupérés. Les séquences génomiques de AtpD, GyrB, InfB et RpoB ont été récupérées et analysées par blastn. Les résultats sont similaires à l'étude du 16S. Beaucoup de *Klebsiella* ressortent, quelques *Citrobacter* et *Salmonella* également. Une phylogénie a été réalisé d'après les résultats du blastn. Les résultats sont exposés sous forme d'arbres. Le résultat pour AtpD est visualisable sur la figure 8. Les autres arbres seront mis en annexe.

Phylogénie basée sur la totalité du génome

Par la suite, 11 espèces de bactéries ont été retenues pour réaliser une hybridation ADN-ADN *In silico*. Elles ont été hybridées via GGDC, la matrice de distance a été

construite et un arbre phylogénétique créé à partir de la matrice. *Klebsiella oxytoca* et *Klebsiella michiganensis* ont été identifiés comme étant les plus proches (figure 8). N'ont pas été conservées *Shigella flexneri, Raoultella ornithinolytica, Dickeya solani, Serratia marcescens, Serratia rubidaea*. D'après la figure on voit que notre bactérie est située au beau milieu du genre des *Klebsiella*. Les deux autres espèces proches sont *Klebsiella pneumoniae* et *Klebsiella variicola*. Cette bactérie est considérée comme une nouvelle espèce du genre*Klebsiella*. Elle est actuellement nommée *Klebsiella sp* en attendant qu'une identité lui soit attribuée.

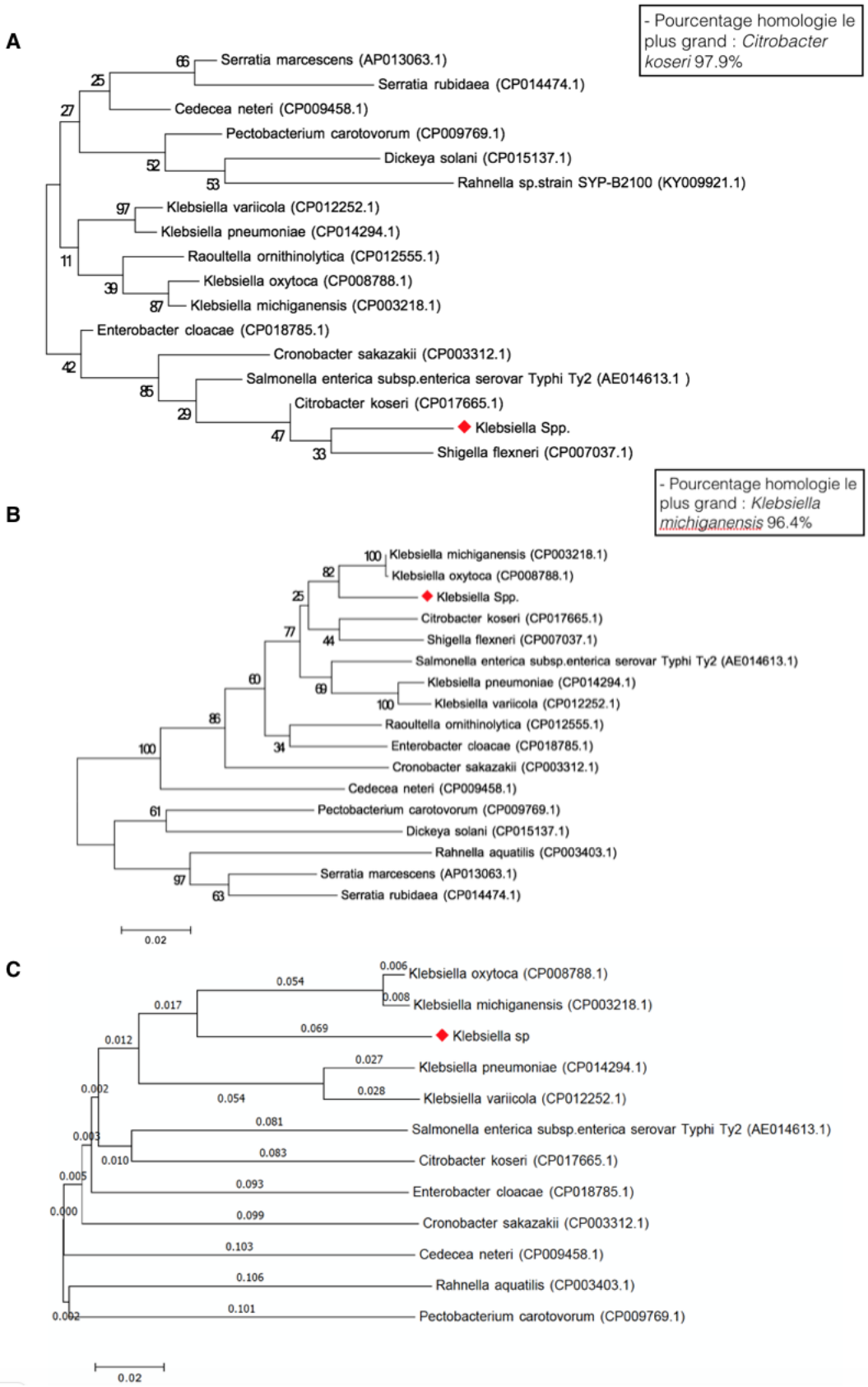


FIGURE 8 – Phylogénie de *Klebsiella sp.* Cette figure expose plusieurs arbres phylogénétiques. Les blastn qui ont servi de bases pour la construction des arbres ont tous été réalisés sur la base de données "Genomes" du NCBI. Les mêmes espèces ont été séléctionnées entre les différentes études pour pouvoir comparer la position de notre bactérie. **A**. Cette arbre est basé sur la séquence de l'ARNr 16S. L'espèce montrant le plus fort taux d'homologie est *Citrobacter koseri*. **B**. Cette arbre est basé sur le gène AtpD. *Klebsiella michiganensis* est l'espèce montrant le plus fort taux d'homologie. **C**. Phylogénie basé sur la totalité du génome. *Klebsiella oxytoca* et *Klebsiella michiganensis* sont proches de notre bactérie.

Comparaison du contenu génomique avec des espèces proches

Le génome des deux Klebsiella a été annoté localement et une étude des orthologues a été faite. Il en ressort que notre souche possède 2970 protéines en commun avec ses deux espèces. 29 protéines avec Klebsiella oxytoca et 36 avec Klebsiella michiganensis. Au final elle dispose de 220 protéines uniques à son espèce. Cette liste de protéines uniques à notre bactérie a été extraite et relayée au Pr.RUIMY.

A.

	Klebsiella sp	*Klebsiella oxytoca*	*Klebsiella michiganensis*
Nb bases	$5.2.10^6$	$6.6.10^6$	$7.2.10^6$
Gènes	4874	6436	7115
Signal peptide	437	614	686
rRNA	12	25	25
tRNA	77	86	86
miscRNA	98	134	134
tmRNA	1	1	1

B.

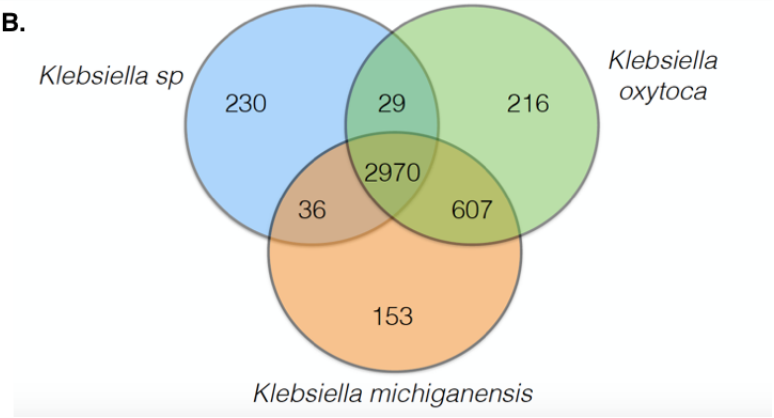


FIGURE 9 – **A**.Comparaison du contenu génomique de *Klebsiella sp* avec *Klebsiella oxytoca* et *Klebsiella michiganensis*. **B**. Comparaison des orthologues entre *Klebsiella sp* avec *Klebsiella oxytoca* et *Klebsiella michiganensis*. Notre bactérie a 230 protéines qui ne sont pas partagés avec ses deux souches.

Une comparaison des COGs avec les deux autres espèces a été faite. Le tableau est visualisable sur la figure 10. La catégorie la plus importante chez notre bactérie est "Transport et métabolisme des acides aminés" (hors "fonction inconnue" et "fonction générale"). Pour la totalité des catégories, nous pouvons observer que la différence entre bactérie est très maigre. Ce type de données est nécessaire pour la description de nouvelle espèce.

Espèces		*Klebsiella sp*	*Klebsiella oxytoca*	*Klebsiella michiganensis*
Code	Description	% total	% total	% total
J	Translation, ribosomal structure and biogenesis	3.67	2.7	3.05
A	RNA processing and modification	0.02	0.01	0.01
K	Transcription	7.66	7.68	7.73
L	Replication, recombination and repair	3.03	5.83	4.08
B	Chromatin structure and dynamics	0	0	0
D	Cell cycle control, cell division, chromosome partitioning	0.67	0.64	0.64
Y	Nuclear structure	0	0	0
V	Defense mechanisms	1.06	1.17	1.07
T	Signal transduction mechanisms	3.52	3.27	3.45
M	Cell wall/ membrane biogenesis	4.46	3.78	3.98
N	Cell motility	1.08	1.74	1.58
Z	Cytoskeleton	0	0	0
U	Intracellular trafficking and secretion, and vesicular transport	2.19	2.94	2.46
O	Posttranslational modification, protein turnover, chaperones	2.82	2.34	2.47
C	Energy production and conversion	6.07	4.49	4.93
G	Carbohydrate transport and metabolism	9.32	8.69	9.41
E	Amino acid transport and metabolism	10.41	7.19	8.49
F	Nucleotide transport and metabolism	2.09	1.47	1.64
H	Coenzyme transport and metabolism	3.78	2.86	3.32
I	Lipid transport and metabolism	2.29	1.84	2.09
P	Inorganic ion transport and metabolism	6.32	4.69	5.28
Q	Secondary metabolites biosynthesis, transport and catabolism	2.11	1.91	1.97
R	General function prediction only	9.18	8.21	8.77
S	Function unknown	18.25	26.55	23.59

FIGURE 10 – Tableau des COGs chez *Klebsiella sp*. Ce tableau reprend les protéines classées en catégories fonctionnelles. On peut observer que pour chaque catégorie un code est attribué. Pour chaque catégorie, on a le pourcentage relatif des protéines impliquées dans cette catégorie. Les COGs sont comparés deux autres espèces considérées comme étant les plus proches.

5 Discussion

5.1 Reconstitution des génomes

5.1.1 Qualité et tri des données

FASTQc renvoie un fichier contenant une multitude de critères relatives à la qualité du séquençage. Par exemple il va nous donner la longueur moyennes des phases de lecture obtenues (reads), leurs qualités (critère divisé en trois section : très bon $score \geq 28$, moyen $28 \leq score \leq 20$, mauvais $score \leq 20$), le pourcentage de séquence GC (nous donnera une idée d'une possible contamination), la présence de séquences sur-représentées (présence d'adaptateur, ou erreur de séquençage par exemple). Il s'agit là de quelques exemples de statistique que nous renvoie FASTQc, les principaux utilisés. FASTQc nous donne au total pas moins de 11 statistiques différentes qui nous permettent de juger la qualité du séquençage. L'étape de contrôle de qualité du séquençage n'est pas obligatoire, mais très recommandée. Elle va nous permettre d'optimiser le tri des séquences.

Nissabacter archeti

D'après les résultats on a pu voir qu'il y avait dans notre séquençage des séquences de faibles qualités qui devaient être triées. Ceci dans le but de partir sur la base la plus propre possible pour la suite du traitement et pour pouvoir tirer le plus d'informations possible sur notre souche. Au final en triant toutes les données possédant un score inférieur à 20, les séquences trop courtes, les séquences sur-représenté, 3,13% de nos séquences n'ont pas été retenu. Ce qui révèle une bonne qualité du séquençage. Ici la plupart des séquences sont contenus dans les fichier paired. Il y a 12,47% des séquences dans le fichier unpaired du brin forward et 0,95% des séquences dans le fichier unpaired du brin reverse. Les fichiers unpaired pourront apporter une information pour l'étape d'assemblage.

Klebsiella sp

Via FASTQc on a pu voir qu'il y avait dans ce jeu de données présence d'adaptateurs de type Illumina TruSeq. Ces adaptateurs sont utilisés dans les kits de séquençage Illumina. Ils sont normalement filtrés à la fin du séquençage mais il arrive parfois que malgré le filtre appliqué, il reste des séquences adaptatrices. FASTQc les a mis en évidence et elles ont été supprimées. Au final seulement 3,47% des reads ont été trié. Ce qui révèle une bonne qualité du séquençage. On peut noter ici que 22,39% des séquences sont contenus dans le fichier unpaired du brin forward et 12,18% dans le fichier unpaired du brin reverse. Il faudra tenir compte des fichiers unpaired pour l'assemblage et le scaffolding. Ne pas le faire représenterait une grosse perte d'informations.

5.1.2 Étape d'assemblage

Pour cette étape nous avons comparé deux logiciels. Le but était d'obtenir le meilleur assemblage possible.

Velvet s'est révélé être moins performant sur l'ensemble des données. Et ceci sur la totalité des k-mers. Velvet traite de façon individuelle chaque k-mers. Nous en avons

choisi plusieurs afin d'avoir un spectre large. Mais pour *Klebsiella sp* et *Nissabacter archeti* il nous renvoie un nombre supérieur de contigs et un N50 inférieur. De plus Velvet est plus compliquer à lancer. Il faut préalablement aller dans le fichier de configuration du programme changer la longueur maximal des k-mers autorisés qui est limité à 31 par défaut.

SPAdes lui peut traiter les données séquencés en mate-pair et paired-end. On peut le lancer de façon très simple en entrant les fichiers input et output. Il dispose d'un mode avancé plus personnalisable. Il passe par l'utilisation d'un fichier de configuration. On pourra y entrer l'orientation des fichiers séquencés (Reverse/Forward ou Forward/Reverse), le type de séquençage (mate-pair, paired-end), le brin sens et l'anti sens. Une fois les fichiers entrés SPAdes dispose d'un large panel d'option qui permettent d'optimiser l'assemblage. On va pouvoir choisir plusieurs k-mers. Ce logiciel va se servir de plusieurs k-mers pour l'assemblage et non pas d'une seule (graphe de bruijn à multi k-mers). SPAdes dispose d'une option permettant de réduire le nombre de mésappariement. Il va également détecter automatiquement la taille de l'amorce. SPAdes est plus complet, plus personnalisable et donc mieux adaptable à nos données. De plus, il réalise directement une étape de scaffolding avec de bons résultats. Ainsi c'est ce dernier qui a été retenu pour nos jeux de données.

Nissabacter archeti

À la fin de cette étape nous avons donc 67 contigs pour *Nissabacter archeti*. Ce nombre est relativement élevé pour un assemblage. Le génome est haché. Ceci est du au fait que cette échantillon a été séquencé en paired-end. Le paired-end génère des fragments de moins d'un kilobase. L'idéal aurait été de le séquencer en mate-pair et de coupler ce séquençage avec du paired-end. Le mate-pair générant de long fragment, il permet d'avoir une bonne couverture du génome. Les fragments générés en paired-end viendront combler les quelques gaps laissés par l'assemblage.

Klebsiella sp

Les statistiques obtenues avec les données du génome de *Klebsiella sp* sont plus flatteuses que celle obtenues avec *Nissabacter archeti*. En effet après l'assemblage nous avons 16 contigs. Ce qui est relativement correcte. En se basant sur l'article comparant les assembleurs [18], on peut voir que notre assemblage est très bien évalué et se situe bien par rapport aux autres. La plupart ont un N50 approchant les 120.000. Notre assemblage à un N50 de 1.269.645. Ceci s'explique par le séquençage en mate-pair comme expliqué plus haut qui génère des fragments long. Ceci nous permet d'avoir une bonne couverture du génome et peut de contig. Cependant, on peut avoir le même raisonnement que plus haut et se dire que si un séquençage en paired-end avait été couplé à ce séquençage en mate-pair. Nous aurions encore moins de contigs.

5.1.3 Amélioration de l'assemblage

Nissabacter archeti

Le scaffolding n'a pas été possible lors de cette étape. Les fragments étant trop petits, il n'y avait aucune amélioration des données après scaffolding. Gapfiller a totalement rebouché un gap et partiellement les autres. Ceci est du au fait que nous avons augmenté la stringence de Gapfiller grâce à l'option -m qui demande un nombre de bases minimum du read brut se chevauchant avec le contig ou scaffolds pour reboucher le gap. Ceci a permis de ne pas reboucher les gaps de façon aléatoire. De plus, il fallait veiller à ne pas avoir de gap à 1 N. Le fait d'avoir très peut de N dans son génome est flatteur, mais ça peut être révélateur d'un mauvais usage de Gapfiller. En effet ce dernier va reboucher les gaps avec les mauvais reads. Il va allonger le contig ou le scaffold jusqu'à arriver au contig suivant et va chevaucher le contig ou scaffold. Afin de ne pas le chevaucher il va couper 1 nucléotide avant le contig ou scaffold suivant. Ceci explique les gaps à 1 N qui sont témoins d'une mauvaise finition de données. Au final à l'issue de ces étapes, c'est ce fichier que nous conserverons comme génome.

Klebsiella sp

Les données de *Klebsiella sp* permettent le scaffolding de part la taille des fragments que génère le mate-pair. Nous avons tester SSPACE, Opera et Scaffmatch pour réaliser cette étape.

Opera est plutôt simple à installer, il nous renvoie un fichier récapitulatif sur les données ce qui est intéressant et il est rapide. Par contre il a peu d'options, il ne tient pas compte de la longueur de l'amorce (distance qui sépare les reads), il a de nombreuses dépendances à installer et il ne tient pas compte pas fichiers filtrés non appariés (Unpaired).

Scaffmatch lui dispose de plus d'options, on peut y préciser la longueur de l'amorce, l'orientation du séquençage (Reverse/Forward) et il est également rapide. Cependant son fichier de sortie est relativement brut et nécessite un traitement informatique. Il ne renvoie pas de fichier récapitulatif sur l'analyse et ne prend pas les fichiers filtrés non appariés (Unpaired).

SSPACE dispose de nombreuses options. On écrit préalablement un fichier de configuration. On saisit également le sens du séquençage si Reverse Forward ou Forward Reverse. Ensuite il faut lancer le programme avec le fichier de configuration et les options tel que le nombre de bases minimum se chevauchant entre contig pour être regrouper en scaffolds, la taille minimale des contigs utilisés pour faire des scaffolds, la couverture des reads brut pour assembler deux contigs entre eux et former un scaffolds. Ce logiciel est parallèlisable. Il est le plus performant, il renvoie un fichier récapitulatif contenant les statistiques de l'analyse. Il est rapide, très personnalisable.

Concernant Gapfiller, on a veillé aux mêmes détails que pour *Nissabacter archeti*. C'est-à-dire ne pas avoir dans notre génome de gap à 1 N, quitte à avoir un peu plus de gaps non rebouché.

À l'issue de cette étape, nous avons brisé les scaffolds en contigs au niveau des gaps. La conservation des contigs plutôt que des scaffolds est préférée dans ce cas car on cherche de la fiablité des données (structure du génome) au détriment de chiffres c'est à dire le minimum de morceau). Dans la mesure où on cherche à identifier les gènes, scaffolds ou contigs cela ne changera pas ce contenu génique. De plus si les scaffoldeurs (opera,

sspace, etc..) donnent des résultats de scaffolding différents. De même le fait de ne pas arriver à boucher des gaps qui sont supposés être relativement petits (inférieur à 100N) incite à la prudence et il se pourrait que ces scaffolds ne soit pas au bons endroits. Ainsi nous préférons nous baser sur les contigs. À noter que le scaffolding (formation de gaps) dont les gaps on été remplis, a eu son utilité pour réduire le nombre de contigs

Choisir les outils les mieux adaptés à ce jeu de donnée nous a permis d'obtenir un génome relativement complet. En effet, nous avons peu de contigs, les statistiques sont bonnes (valeur N50 élevé, gros contigs). Ceci nous permet de partir sur de bonnes bases afin d'extraire au mieux les informations du génome.

5.2 Analyses génomiques & phylogénétiques

Plusieurs phylogénies de *Nissabacter archeti* et de *Klebsiella sp* ont été réalisées et ceci par des approches différentes. Nous allons tenter d'interpréter les résultats que nous apportent chaque méthode et comparer les résultats entre eux.

5.2.1 Phylogénies à différentes résolutions de *Nissabacter archeti*

Étude de la séquence d'ARNr 16S

Basé sur la séquence d'ARNr du 16S la bactérie montre un fort taux d'homologie avec *Serratia marcescens*. Cependant les pourcentages d'identités sont beaucoup trop proches pour pouvoir se baser uniquement sur ce résultat. Les valeurs de bootstraps sont faibles, on ne peut totalement ce fier à cette arbre. Une information importante ressort en revanche, l'arbre est divisé en deux. Un amas de bactéries au dessus et 2 bactéries dans la branche inférieur avec notre bactérie. Le 16S nous permet de déduire que notre bactérie est proche de *Serratia marcescens*, *Serratia rubidaea* et *Rahnella aquatilis*

Analyse des gènes de ménages

L'étude des gènes de ménages montre que notre bactérie est proche du genre *Serratia*. Les résultats sont en accord avec le résultat de l'analyse du 16S et confortent cette idée. La difficulté ici est qu'il y a 4 arbres. Les mêmes espèces ont été sélectionnées d'un gène à l'autre ce qui permet de pouvoir comparer les arbres. Nous avons tenté de faire à partir de ces 4 arbres un arbre consensus. Ceci s'est révélé très compliqué de part les valeurs de bootstrap qui sont parfois faibles. Ce qui révèle une fragilité de la branche. *Nissabacter archeti* est toujours éloignée des autres groupes.

Phylogénie basée sur la totalité du génome

Cette étude est sans doute celle qui apporte le plus d'informations sur les espèces proches de notre bactérie. En effet contrairement aux autres techniques, on ne se base pas sur un fragment de séquences ou quelques gènes. Ici la totalité du génome est pris en compte. Les résultats confirment ici que *Nissabacter archeti* est proche du genre Serratia. Un doute subsistait quant à la proximité de *Rahnella aquatilis*. En effet elle est proche sur la totalité des gènes de ménage et sur le 16S. Cette analyse nous permet d'écarter cette hypothèse.

Serratia marcenscens et *Serratia rubidaea* ont été sélectionnées comme les espèces les plus proche de notre souche. Basé sur le 16S, les gènes de ménages et la totalité du génome c'est celle qui sont les plus proches. Le pourcentage d'identité du 16S est proche de *Serratia rubidaea* et la totalité des gènes de ménages sont proches de ces deux bactéries.

5.2.2 Analyses génomiques

Le megablast du génome a permis d'avoir une idée du genre proche de *Nissabacter archeti*. Il a également mis en évidence la présence de plasmide. 3 contigs ont été identifiés comme possédant potentiellement un plasmide. Les plasmides sont intéressants à étudier car ils peuvent être à l'origine d'une virulence accrue par exemple. Ainsi il serait intéressant d'en savoir plus sur l'origine de ses plasmides.

Annotation du génome

Malgré un génome relativement haché, de bonnes informations ont été mises en avant sur ce génome. La séquence de l'ARNr 16S est complète, les gènes de ménages aussi sont complet. D'après la littérature, il apparait qu'une bactérie à environ 3000 à 4000 gènes voir plus selon la taille de son génome. Notre annotation a permis d'en annotée 4755. Une liste de gènes considérés comme étant de gènes de résistances à des antibiotiques a été extraite. La résistance aux antibiotiques étant un des enjeux majeur de la microbiologie, il est intéressant de savoir quelle gènes sont présent chez cette souche. Cette souche possède par exemple des gènes de résistances à la Bicyclomycine et des gènes multi-résistants tel que MdtA, MexB, etc. La Bicyclomycine possède un large spectre d'action et est particulièrement efficace contre les bactéries gram-. Cet antibiotique agit également contre les gram+. Le gène MdtA va conférer une résistance à la novobiocine qui est active principalement sur les staphylocoques et les streptocoques. Peut-être que notre bactérie a hérité de ce gène par transfert horizontale ? Il serait intéressant de regarder au cas par cas les résistances dont dispose *Nissabacter archeti*.

Comparaison du contenu génomique avec des espèces proches

Serratia marcenscens et *Serratia rubidaea* ont été sélectionnées comme les espèces les plus proche de notre souche. Une comparaison de leurs contenue génétique a été faite. Ils semblent avoir le même nombre de gènes ($\pm$ 150 sur les deux espèces). Il sera intéressant ici d'étudier les gènes qui sont uniques à notre bactérie. On a déjà à notre disposition la liste de gène unique à *Nissabacter archeti*.

Les voies métaboliques sont représentées par un code. Il y en a un total de 23. On sait grâce a RPSblast dans quelle voies sont impliquées nos gènes. Cette analyse apporte des informations sur notre souche et une étude des COG est très souvent demandé pour une publication sur le génome de la bactérie.

5.2.3 Phylogénies à différentes résolutions de *Klebsiella sp*

Étude de la séquence d'ARNr 16S

Aucun pourcentage d'identité supérieur à 98,7% n'a été identifié ici. Ce qui confirme qu'il s'agit d'une nouvelle espèce. L'espèce présentant le plus fort taux d'homologie sur la base de donnée "genomes" est *Citrobacter koseri*. Cependant les pourcentages d'identités sont si proches les uns des autres qu'on ne peut se baser uniquement sur ce résultat. Parmi les 100 espèces présentant le meilleur taux d'homologie au niveau du 16S, de nombreuses Klebsiella ressortent. Sur la base de données "16s rRNA", le top hit est représenté par *Klebsiella oxytoca*. Ces éléments nous dirigent vers que l'idée cette bactérie est proche du genre *Klebsiella*. L'arbre phylogénétique est plus difficilement interprétables. Les valeurs de bootstraps sont ici très basses (33,47,29). L'arbre n'est pas totalement fiable. Le problème avec l'analyse de l'ARNr 16S ici est que les valeurs sont tellement proches entre elles qu'il est très compliqué de pouvoir se baser uniquement sur ce résultat.

Analyse des gènes de ménages

L'analyse des gènes de ménages est ici un plus compliquée que pour *Nissabacter archeti*. En effet, d'un gène à l'autre nous avons des résultats différents. Ressortent principalement *Klebsiella oxytoca, Klebsiella michiganensis, Citrobacter koseri*. La valeur des bootstraps nous font douter sur la certitude des branches. Elles sont parfois très basses (par exemple 12,23,21 sur l'arbre concernant RpoB). Ceci nous permet d'avoir une idée plus précise des espèces proches mais on préférera passer par la totalité du génome.

Phylogénie basée sur la totalité du génome

Comme dit précédemment, cette technique est surement la plus informative. Elle nous montre que notre bactérie est proche de *Klebsiella michiganensis* et *Klebsiella oxytoca*. On peut même voir que notre bactérie se situe au beau milieu du genre des Klebsiella avec à proximité également *Klebsiella pneumoniea* et *Klebsiella variicola*. Ceci nous confirme que cette espèce est une nouvelle espèce du genre des Klebsiella dont le nom reste à définir.

5.2.4 Analyses génomiques

Le fait d'avoir mis de coté le phage phiX174 est un bon signe quant à la qualité de notre assemblage. Ca séquence a totalement été retrouvée de façon *De Novo* et mis de coté dans un contig.

Annotation du génome

Nous partions sur de bonnes bases concernant ce génome. L'annotation n'a posé aucun problème et le nombre de gènes est cohérent avec le nombre de gènes moyen présent chez une bactérie. Les gènes étiquetés comme étant des gènes de résistances aux antibiotiques ont été extrait et transférer au Pr.RUIMY. Cette souche possède des résistances à la Fosmidomycine et la Tétracycline. La Tétracycline est utilisée depuis plusieurs dizaines d'années pour le traitement des infections respiratoires principalement pour limiter les

infections dues aux germes sensibles aux cyclines notamment dans leurs manifestations respiratoires, telles que les infections bronchopulmonaires et contre la brucellose, ainsi que pour le traitement de fond en dermatologie de certaines formes d'acné sévère et de psoriasis. La Fosmidomycine bloque la croissance des souches multi-résistantes. Il est donc inquiétant de voir que cette souche possède un gène de résistance à cet antibiotique.

Comparaison du contenu génomique avec des espèces proches

Sur la totalité des résultats *Klebsiella michiganensis* et *Klebsiella oxytoca* ont été choisies comme étant les plus proches. Pour deux des gènes de ménages elles présentent le plus fort taux d'homologie (InfB et AtpD), elles sont très nombreuses parmi les résultats du blastn sur la séquence du 16S et c'est incontestablement les plus proches si on se base sur la totalité du génome. Concernant la comparaison du contenu génétique, les deux autres espèces ont plus de gènes, mais leurs génomes est plus gros également. Proportionnellement ils ont le même nombre de gènes. Ils ont tous globalement 1000 gènes par mégabase. Une étude des gènes orthologues nous a permis également de mettre en évidence les caractéristiques partagés et les protéines uniques à notre bactérie. Il serait intéressant d'explorer les protéines présent dans cette liste.

6 Conclusion et Perspectives

Une des thématiques principale de la microbiologie est la résistance aux antibiotiques des bactéries. C'est un fléau en microbiologie car les souches deviennent de plus en plus résistantes, ce qui pause un problème pour le traitement des patients. Ces souches présentaient un intérêt (virulence accrue, résistances à des antibiotiques), c'est pourquoi le Pr.RUIMY a demandé le séquençage de ses souches. Ainsi l'objectif principale du stage était de traiter et analyser deux jeux de données microbiologiques issues de séquenceur à haut débit.

L'une est un nouveau genre nommé *Nissabacter archeti*, l'autre était une bactérie inconnu maintenant classifiée dans le genre des *Klebsiella* et considérée comme une nouvelle espèce : *Klebsiella sp.*

Le séquençages de ses souches on a mis en évidence de nombreuses informations. Les informations principales étant la liste des gènes des deux espèces avec notamment les gènes de résistances aux antibiotiques et les bactéries proches de nos bactéries. La comparaison avec les espèces proches à permis de connaitre les caractéristiques que ses souches partagent. Il en ressort les caractéristiques communes et les caractéristiques uniques à nos souches.

Il aurait intéréssant de compler l'étude par une analyse des voies métaboliques dans lesquels sont impliqués les protéines à nos espèces. Il serait également intéressant d'examiner les plasmides présents dans nos souches.

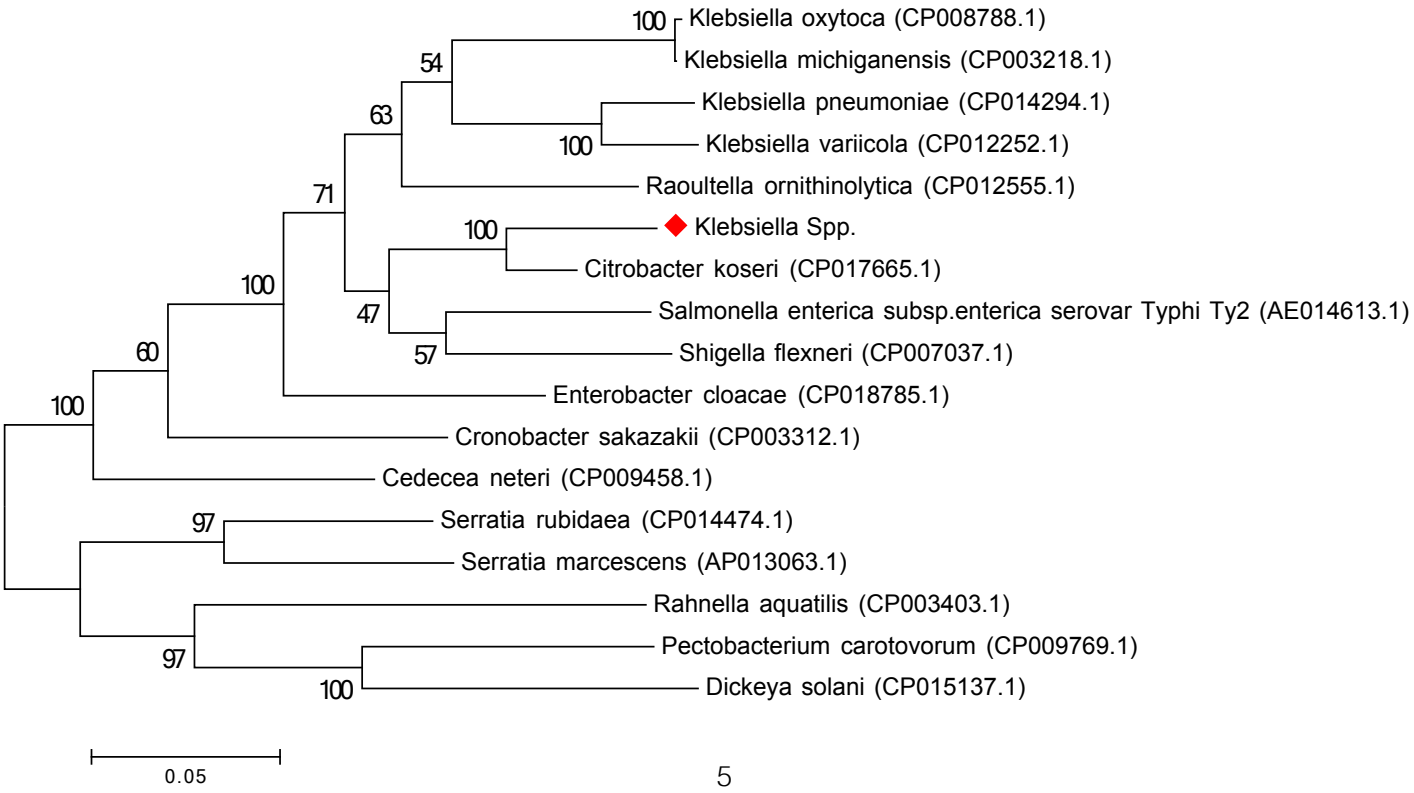
Les informations extraites sur ces deux bactéries seront publiées. En effet, nous avons tout les éléments à notre disposition pour pouvoir présenter les caractéristiques de nos nouvelles espèces bactériennes.

Tout ceci n'aurait pas été possible sans le séquençage à haut débit. À l'heure actuelle un séquençage de génome bactérien est de l'ordre de 300€. Les prix du séquençages ont baissé, ce qui le rend de plus en plus accessible. C'est pourquoi on hésite plus à séquencer un génome de bactérie lorsque celle-ci montre un intérêt particulier. D'autant plus quand elle n'a pas été identifié par les techniques d'analyses en laboratoire clinique. Avec les avancées technologiques, on peut facilement imaginer que le séquençage deviendra de plus en plus accessible. Notamment grâce au MinION par exemple. Il peut être directement branché à un ordinateur en USB 3.0 et génère des reads ultra-longs (de l'ordre de la centaine de kilobases). Ainsi on peut facilement imaginer que l'on aura pour un génome bactérien, un fragment continu, sans gaps. On aura une meilleure fiabilité des génomes obtenu. Ce qui facilitera énormément les étapes d'analyses bioinformatiques.

7 Annexes

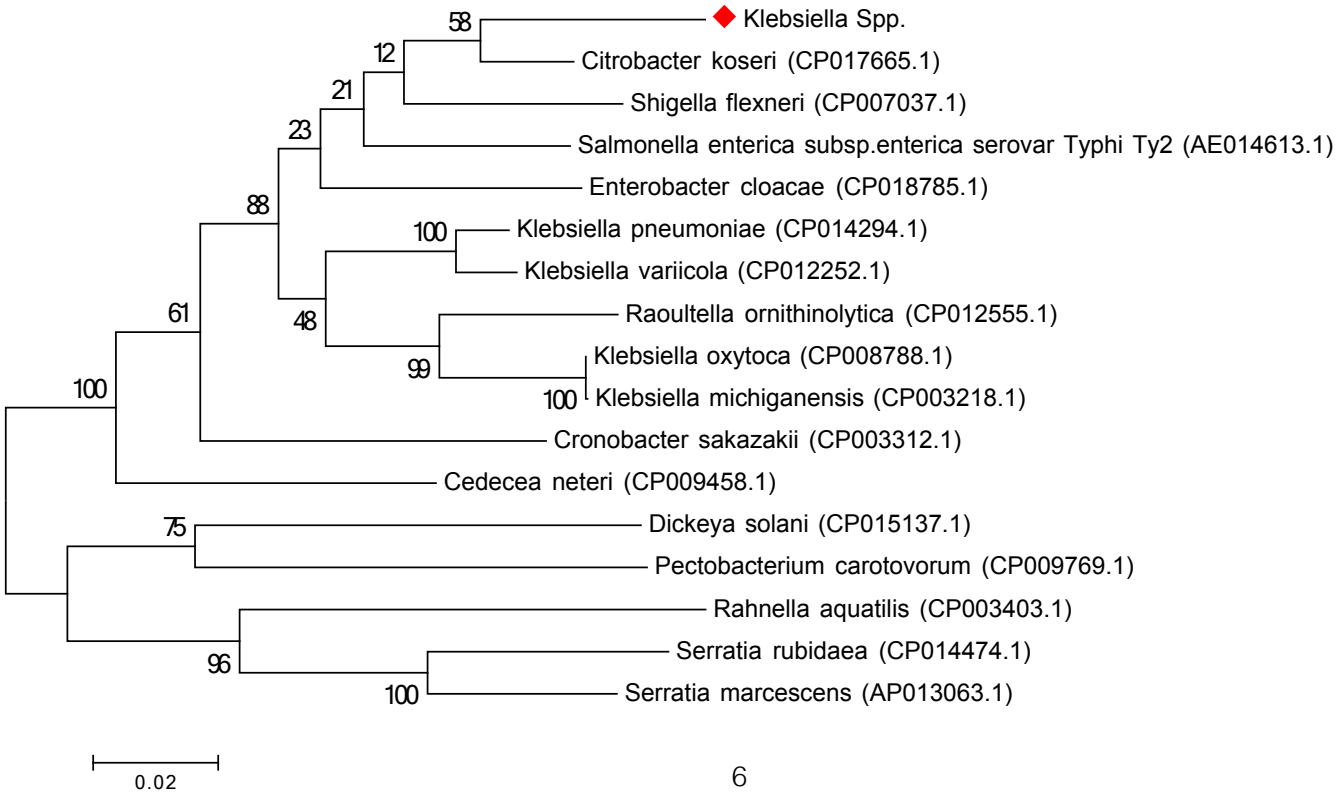
Klebsiella Sp -GyrB

- Pourcentage homologie le plus grand : *Citrobacter koseri* 94.5%



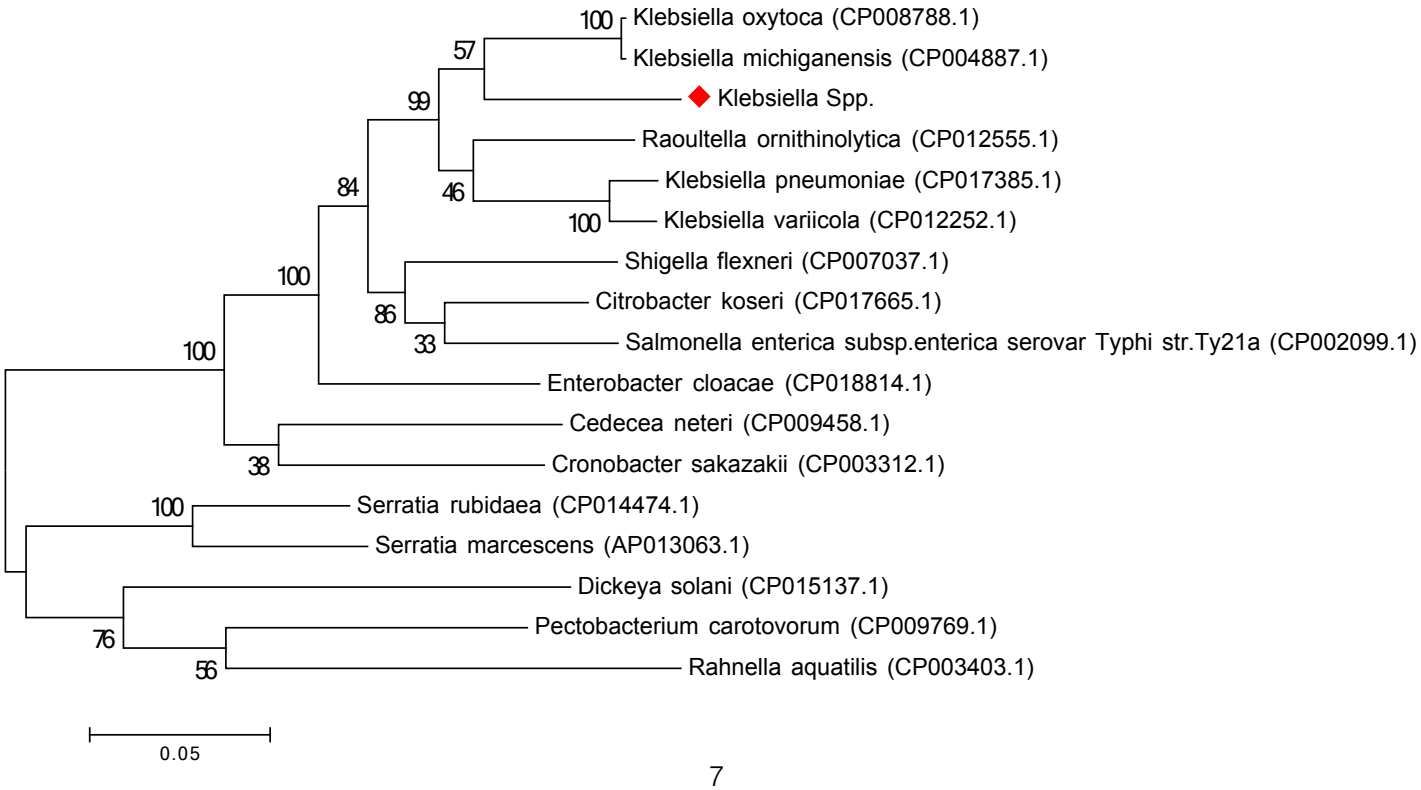
Klebsiella Sp -RpoB

- Pourcentage homologie le plus grand : *Citrobacter koseri* 95.1%



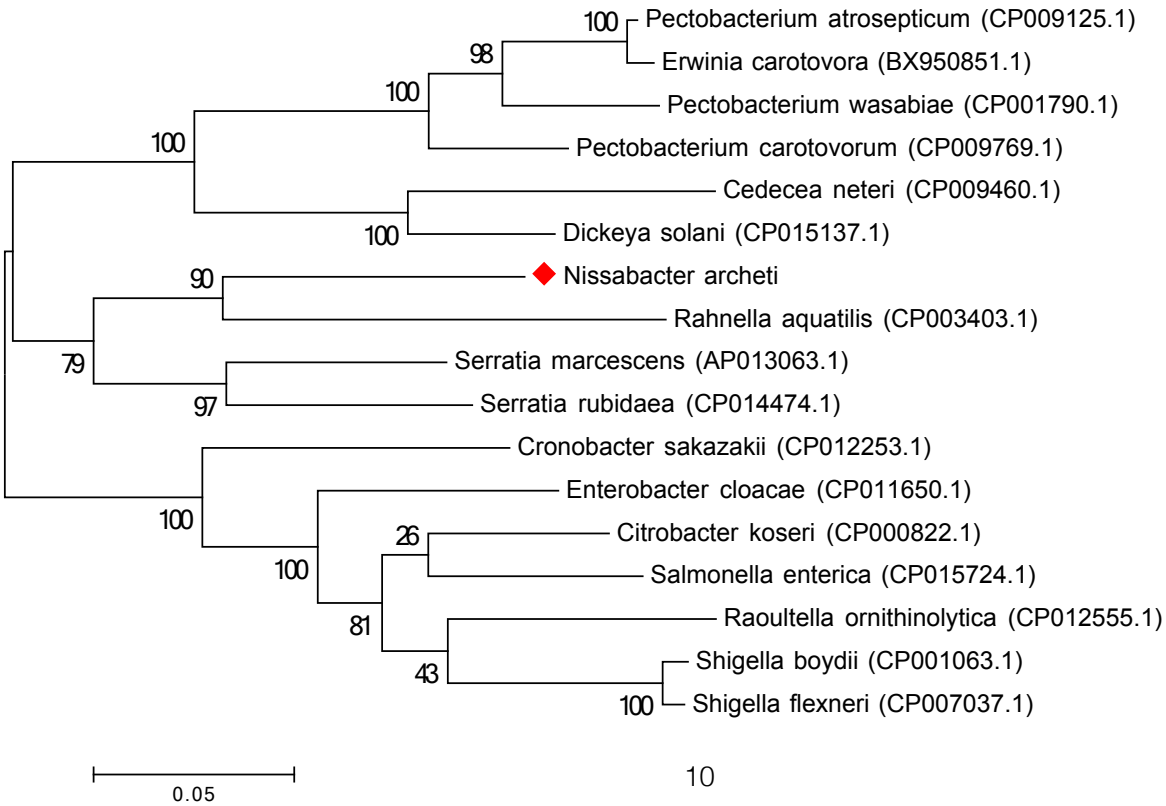
Klebsiella Sp -InfB

- Pourcentage homologie le plus grand : *Klebsiella oxytoca* 91.3%



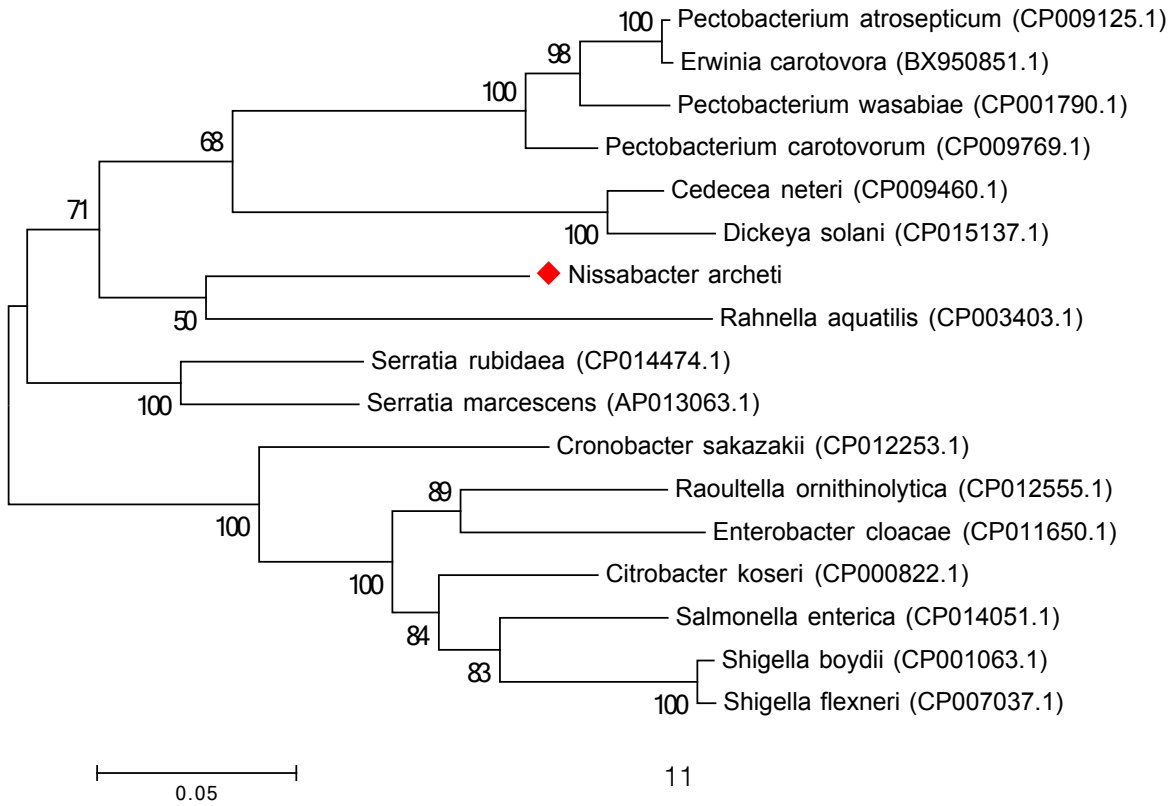
Nissabacter archeti - GyrB

- Pourcentage homologie le plus grand : *Serratia marcescens* 86.3%



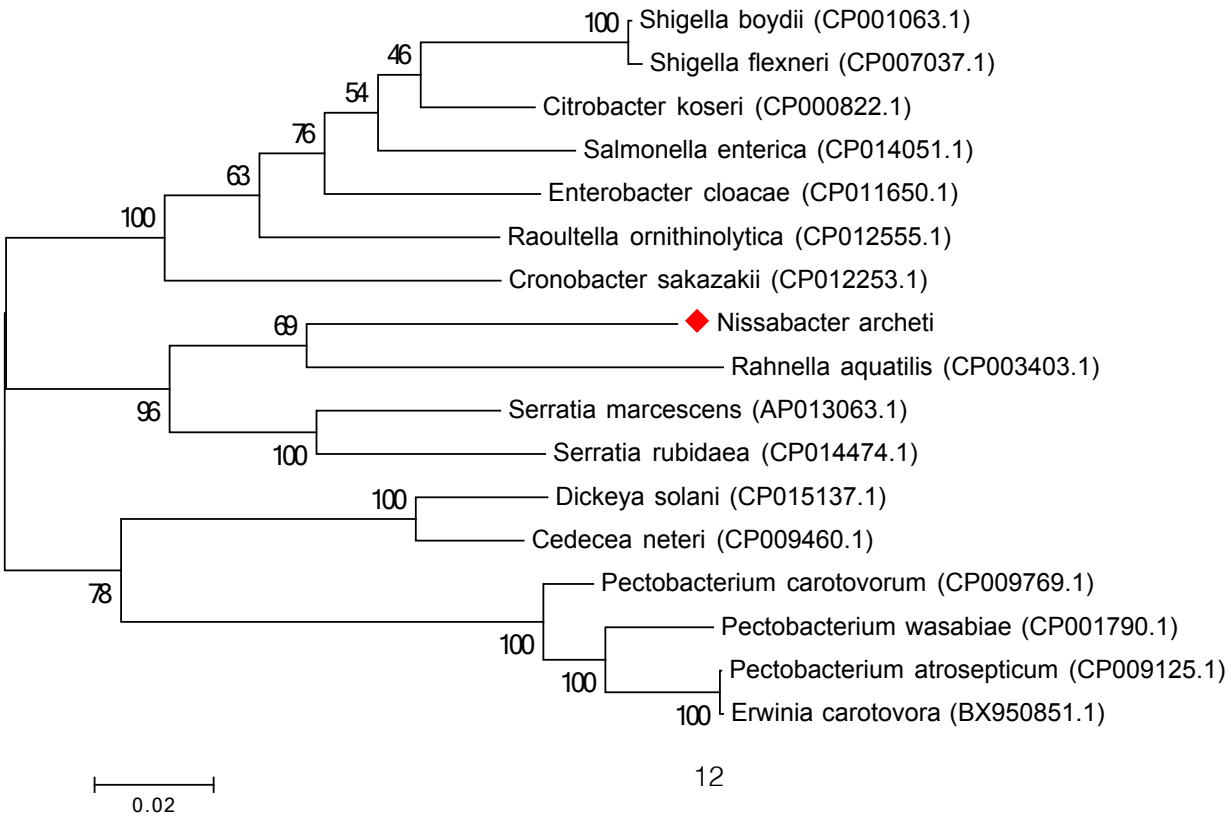
Nissabacter archeti - InfB

- Pourcentage homologie le plus grand : *Serratia rubidaea* 85.1%



Nissabacter archeti - RpoB

- Pourcentage homologie le plus grand : *Serratia marcescens* 89.2%



Références

[1] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. Nucleic acids research *25*, 3389–3402.

[2] Andrews, S. et al. (2010). FastQC a quality control tool for high throughput sequence data. - .

[3] Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M. et al. (2008). The RAST Server : rapid annotations using subsystems technology. BMC genomics *9*, 75.

[4] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin et al. (2012). SPAdes : a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology *19*, 455–477.

[5] Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.

[6] Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic : a flexible trimmer for Illumina sequence data. Bioinformatics *1*, btu170.

[7] Del Fabbro, C., Scalabrin, S., Morgante, M. and Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One *8*, e85024.

[8] Edgar, R. C. (2004). MUSCLE : multiple sequence alignment with high accuracy and high throughput. Nucleic acids research *32*, 1792–1797.

[9] Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., Shanmugam, D., Roos, D. S. and Stoeckert, C. J. (2011). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new Ortholog groups. Current protocols in bioinformatics , 6–12.

[10] Gao, S., Bertrand, D., Chia, B. K. and Nagarajan, N. (2016). OPERA-LG : Efficient and exact scaffolding of large, repeat-rich eukaryotic genomes with performance guarantees. Genome biology *17*, 102.

[11] Glaeser, S. P. and Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Systematic and applied microbiology *38*, 237–245.

[12] Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013). QUAST : quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

[13] Hunt, M., Newbold, C., Berriman, M. and Otto, T. D. (2014). A comprehensive evaluation of assembly scaffolding tools. Genome biology *15*, R42.

[14] Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. and Madden, T. L. (2008). NCBI BLAST : a better web interface. Nucleic acids research *36*, W5–W9.

[15] Kumar, S., Stecher, G. and Tamura, K. (2016). MEGA7 : Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Molecular biology and evolution , msw054.

[16] Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T. and Ussery, D. W. (2007). RNAmmer : consistent and rapid annotation of ribosomal RNA genes. Nucleic acids research *35*, 3100–3108.

[17] Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R. and Pallen, M. J. (2012). High-throughput bacterial genome sequencing : an embarrassment of choice, a world of opportunity. Nature Reviews Microbiology *10*, 599–606.

[18] Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L. J. and Salzberg, S. L. (2013). GAGE-B : an evaluation of genome assemblers for bacterial organisms. Bioinformatics *29*, 1718–1725.

[19] Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A. et al. (1998). Multilocus sequence typing : a portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences *95*, 3140–3145.

[20] Mandric, I. and Zelikovsky, A. (2015). ScaffMatch : scaffolding algorithm based on maximum weight matching. Bioinformatics , btv211.

[21] Marchandin, H., Teyssier, C., de Buochberg, M. S., Jean-Pierre, H., Carriere, C. and Jumas-Bilak, E. (2003). Intra-chromosomal heterogeneity between the four 16S rRNA gene copies in the genus Veillonella : implications for phylogeny and taxonomy. Microbiology *149*, 1493–1501.

[22] Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P. and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC bioinformatics *14*, 60.

[23] Mlaga, K., Lotte, R., Montaudié, H., Rolain, J.-M. and Ruimy, R. (2017). ?Nissabacter archeti ?gen. nov., sp. nov., a new member of Enterobacteriaceae family, isolated from human sample at Archet 2 Hospital, Nice, France. New Microbes and New Infections *17*, 81–83.

[24] Nadalin, F., Vezzi, F. and Policriti, A. (2012). GapFiller : a de novo assembly approach to fill the gap within paired reads. BMC bioinformatics *13*, S8.

[25] Rosselló-Mora, R. and Amann, R. (2001). The species concept for prokaryotes. FEMS microbiology reviews *25*, 39–67.

[26] Seemann, T. (2014). Prokka : rapid prokaryotic genome annotation. Bioinformatics , btu153.

[27] Stackebrandt, E. and Ebers, J. (2006). Taxonomic parameters revisited : tarnished gold standards. Microbiology today *33*, 152.

[28] Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A., Kämpfer, P., Maiden, M. C., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H. G. et al. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. International journal of systematic and evolutionary microbiology *52*, 1043–1047.

[29] Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain : the primary kingdoms. Proceedings of the National Academy of Sciences *74*, 5088–5090.

[30] Zerbino, D. R. and Birney, E. (2008). Velvet : algorithms for de novo short read assembly using de Bruijn graphs. Genome research *18*, 821–829.