Mémoire

Auteur : Lequeux, Alina
Promoteur(s) : Wilmotte, Annick; Cornet, Luc
Faculté : Faculté des Sciences
Diplôme : Master en biochimie et biologie moléculaire et cellulaire, à finalité didactique
Année académique : 2022-2023
URI/URL : http://hdl.handle.net/2268.2/18582

Nouveaux génomes de cyanobactéries antarctiques pour la phylogénomique, la taxonomie et la recherche des adaptations génomiques à l'environnement

Alina Lequeux

En vue de l'obtention du diplôme de Master en Biochimie et Biologie Moléculaire et Cellulaire (BBMC), à finalité didactique

Année académique 2022 - 2023

Université de Liège – Département des sciences de la vie – Faculté des sciences

Laboratoire de Physiologie et génétique bactériennes (InBioS)

Promoteur : Dr Annick Wilmotte

Co-promoteur : Dr Luc Cornet

Résumé

Titre : *Nouveaux génomes de cyanobactéries antarctiques pour la phylogénomique, la taxonomie et la recherche des adaptations génomiques à l'environnement*

L'objectif de ce mémoire consiste à caractériser et à déterminer les particularités génomiques de souches de cyanobactéries antarctiques de la collection BCCM (Belgian Coordinated Collections of Microorganisms), à l'aide de l'outil bioinformatique « GEN-ERA » qui utilise différents logiciels pour les analyses génomiques et phylogénétiques. L'étude s'est fait en différentes étapes : (1) l'assemblage des génomes et leur comparaison par ANI (*Average Nucleotide Identity*) ; (2) l'analyse du taux de contamination et la sélection de souches ; (3) l'étude phylogénomique et la réalisation d'arbres phylogénétiques ; (4) la détermination taxonomique des souches ; (5) la détermination de la fonction de gènes spécifiques, de voies métaboliques et des métabolites secondaires.

Les résultats obtenus indiquent que les souches de la collection étudiées présentent des séquences non-cyanobactériens, tels que provenant de protéobactéries par exemple, ce qui confirme le caractère non-axénique des souches. Nous avons sélectionné six souches de bonnes qualités génomiques pour la suite des analyses. Celles-ci sont réparties au sein de trois familles de cyanobactéries : Oscillatoriaceae (3 souches), Leptolyngbyaceae (1 souche) et Nostocaceae (2 souches). Nous nous sommes ensuite focalisés sur les Oscillatoriaceae. Un arbre phylogénomique a été inféré et a servi de support pour la détermination taxonomique. Quatre banques de données taxonomiques ont été utilisées (NCBI, GTDB, SILVA, CyanoSeq) montrant des discordances entre elles, rendant la détermination des taxons difficile. Un embranchement comprenant trois de nos souches de la collection a été associé au genre *Laspinema*. Avec les résultats de ANI et GGDC, nous mettons en évidence l'existence de cinq nouvelles espèces au sein de ce genre, dont deux représentées par des souches de la collections BCCM/ULC et notamment une espèce ne présentant que des souches d'origine antarctique. Nous avons ensuite réalisé des analyses de gènes fonctionnels spécifiques et de métabolites secondaires. Le genre entier aurait perdu la voie de béta-oxydation et aucun gène spécifique à l'ensemble du genre n'a été identifié. Cependant, jusqu'à 14 métabolites secondaires ont été identifiées dans les génomes d'Oscillatoriaceae bien que les génomes de souches d'origine antarctique présentent un nombre nettement plus faible de métabolites secondaires. Avec nos analyses, nous n'avons pas trouvé de métabolites ou de voies spécifiques aux souches antarctiques qui justifieraient leurs résistances au froid et aux UVs.

Mots clés : cyanobactéries, Antarctique, métagénomique, phylogénomique, GEN-ERA

Mémoire réalisé par Lequeux Alina durant l'année académique 2022-2023 dans le laboratoire InBIOS de l'Université de Liège (Uliège) avec la supervision des Dr Annick Wilmotte (Promoteur) et Luc Cornet (Co-promoteur).

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au succès de mon mémoire et qui m'ont soutenu tout au long de ce parcours.

Je tiens à remercier dans un premier temps ma promotrice, Dr Annick Wilmotte, maitre de recherches du FRS-FNRS, directrice de la collection BCCM/ULC dans le Département des sciences de la vie au sein de l'Unité de recherche InBios de l'Université de Liège, pour sa disponibilité, ses judicieux conseils et son apport de connaissances avisées. Je la remercie de m'avoir permis de réaliser mon mémoire au sein de son laboratoire.

Je tiens également à remercier sincèrement mon co-promoteur Dr Luc Cornet, chercheur scientifique en génomique et bio-informaticien dans le Département des sciences de la vie de l'Université de Liège. Il m'a encadré et aidé tout au long de mon mémoire. Son expertise, sa disponibilité et son soutien constant ont été essentiels pour l'avancement de mon travail. Ses conseils éclairés, son investissement et sa confiance en mes capacités m'ont permis de réaliser et de terminer mon mémoire.

Je souhaite également adresser mes remerciements à tous les professeurs de la Faculté de sciences qui m'ont accompagnée tout au long de mon parcours académique. Leurs enseignements m'ont permis de développer mes compétences et de me préparer au mieux pour ma carrière professionnelle.

Enfin, Je tiens à témoigner toute ma reconnaissance aux personnes suivantes, pour leur aide dans la réalisation de ce mémoire :

Dr Anne-Catherine Ahn, Valentina Savaglia et Beatriz Roncero-Ramos d'avoir réalisé le séquençage des génomes et de m'avoir permis de travailler sur leurs souches.

Dr Marcelo Vaz. Je le remercie de m'avoir accordé de son temps pour répondre à mes questions et de m'avoir aidé dans mes recherches.

Professeur Denis Baurain, de m'avoir soutenu dans mon parcours et dans mon choix de mon mémoire.

Ma famille, pour leur soutien et leurs encouragements tout au long de mes études.

Table des matières

1 Liste des abréviations

ANI : Average Nucleotide Identity

BCCM : Belgian Coordinated Collections of Microorganims

BGC : gènes biosynthétiques de métabolites secondaires

CDPS : cyclodipeptide synthase

dDDH: hybridation ADN-ADN digitale

DDH : hybridation ADN-ADN expérimentale

DSMZ : Leibniz Institute DSMZ- German Collection of Microorganisms and Cell Cultures (Deutsche Sammlung von Mikroorganismen und Zellkulturen, en allemand)

VNTR : répétitions en tandem à nombre variable

GEBA : Genomic Encyclopedia of Bacteria and Archaea

GEN-ERA : Culture collections in the GENomic ERA

GGDC: Genome-to-Genome Distance Calculator

GOE : le grand évènement d'oxydation

HGT : transfert horizontal de gènes

ICBN : International Code of Botanical Nomenclature

ICNP : International Code of Nomenclature of Procaryotes

InBios : Integrative Biological Sciences

Itol : Interactive Tree Of Life

MAG : Metagenome-assembled genomes

MDV : Vallées sèches de Mc Murdo

NRPS : synthétase peptidique non ribosomique

OG : gènes orthologues de protéines

RiPP(-like) : peptides synthétisés par le ribosome et modifiés après la traduction

RRE(-containing) : RiPP recognition element

TYGS : Type (Strain) Genome Server

ULC : University of Liège Collection

2 Introduction

2.1 Contexte

Mon mémoire a été réalisé dans le laboratoire de physiologie et génétique bactérienne de Madame Annick Wilmotte, dans l'unité de recherche InBios de l'Université de Liège. L'étude utilise la boîte à outils bio-informatique « GEN-ERA » développée par Luc Cornet pour réaliser des analyses de génomique microbienne au sein des collections BCCM (Belgian Coordinated Collections of Microorganisms) (Cornet et al., 2023). Un des objectifs de ce projet est de décrypter la phylogénie et la taxonomie des cyanobactéries. Il consiste également à détecter des génomes spécifiques aux environnements extrêmes et des composés bioactifs expliquant leur adaptation (https://bccm.belspo.be/content/bccm-collections-genomic-era). Mon mémoire fait partie de cet objectif, qui est de déterminer les particularités génomiques de souches de cyanobactéries d'Antarctique par rapport à ceux venant d'autres biomes. Il consiste également à tester l'efficacité de l'outil « GEN-ERA », disponible pour tout débutant en bio-informatique qui souhaiterait réaliser des analyses (méta)génomiques.

La première tâche de mon mémoire a été de réaliser l'assemblage de 11 génomes de la collection BCCM/ULC. Ces génomes ont ensuite été comparés par ANI avec des génomes de cyanobactéries téléchargés depuis les banques de données GenBank et RefSeq. Une première sélection a été réalisée sur les génomes ayant le meilleur pourcentage d'ANI. La détermination du taux de contamination des génomes et la sélection des souches les moins contaminées ont ensuite été effectuées. À partir des souches sélectionnées, la phylogénie avec des génomes d'autres cyanobactéries venant de GenBank et RefSeq a été réalisée. Lorsque l'arbre phylogénétique a été obtenu, une seule famille de cyanobactéries a été choisie pour la suite des analyses : les Oscillatoriaceae. Nous nous sommes surtout focalisés sur le genre *Laspinema*, un nouveau genre découvert en 2018 (Heidari et al., 2018). Ensuite, la détermination de la taxonomie la plus correcte des souches de cette famille a été réalisée, ainsi que l'étude de la fonction des gènes et des métabolites secondaires.

Mon travail s'est donc déroulé en cinq phases majeures (**Figure 1**) : (1) l'assemblage des génomes et la comparaison par ANI ; (2) l'analyse du taux de contamination et la sélection de souches ; (3) l'étude phylogénomique et la réalisation d'arbres phylogénétiques ; (4) la

détermination taxonomique des souches sélectionnées ; (5) la détermination de la fonction des gènes et des métabolites secondaires.

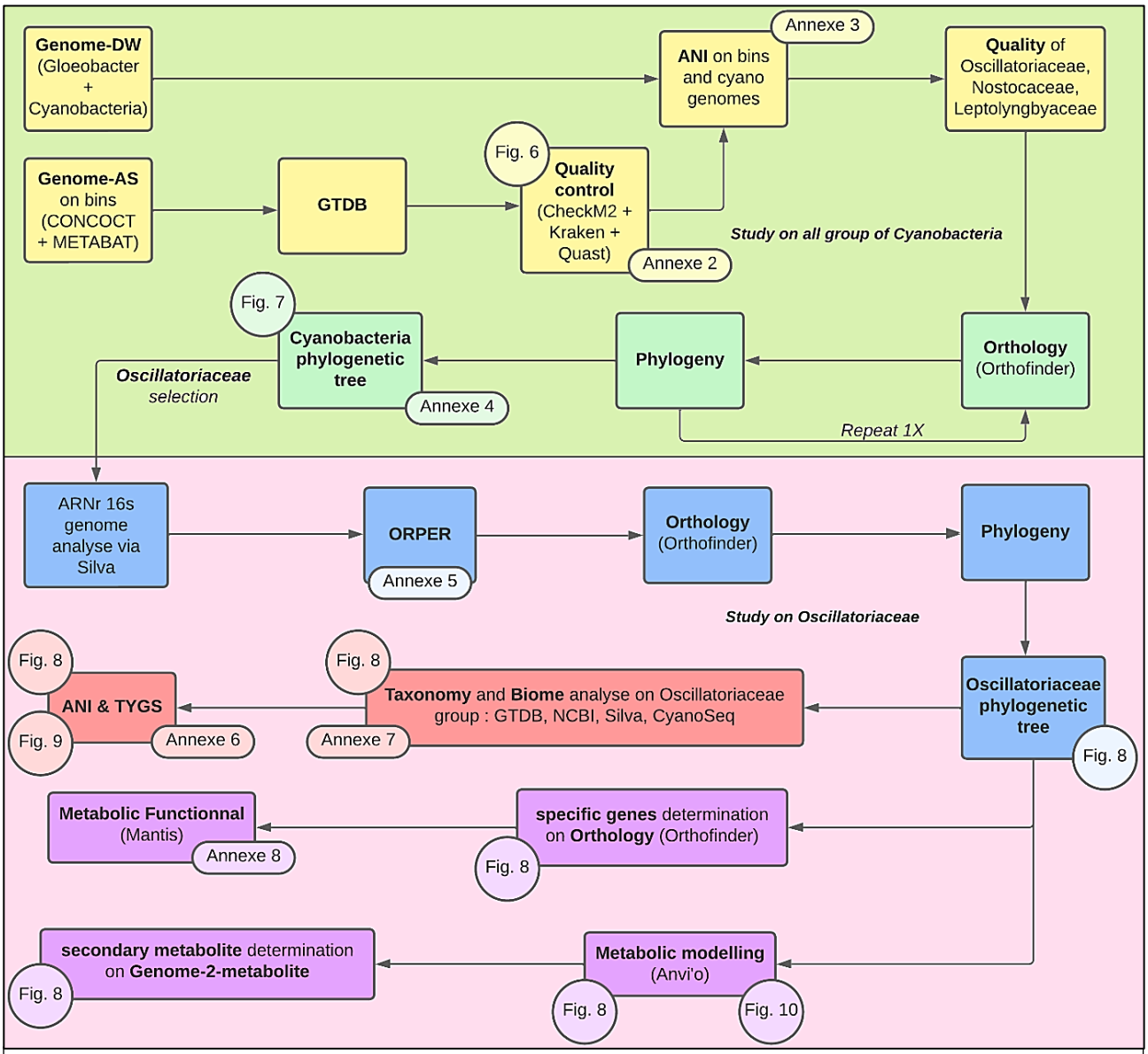


Figure 1 : Diagramme présentant les étapes générales réalisées pour cette étude. Celle-ci est divisée en 2 étapes majeures : l'étude de l'ensemble des cyanobactéries et des souches de BCCM/ULC (fond vert) et l'étude de la famille des Oscillatoriaceae (fond rose). Ces deux étapes sont subdivisées en sous étapes : l'assemblage des génomes, la comparaison par ANI, l'analyse du taux de contamination et la sélection des souches (jaune) ; l'étude phylogénomique et la réalisation d'arbre phylogénétique des cyanobactéries (vert) et des Oscillatoriaceae (bleu) ; la détermination taxonomique (rouge) ; l'étude des gènes : détermination de la fonction des gènes et des métabolites secondaires (mauve). Les flèches indiquent l'ordre et le sens des étapes. Les trois flèches partant de « Oscillatoriaceae phylogenetic tree » indiquent que ces étapes ont été réalisées en parallèle. Pour la majorité des étapes, des bulles sont jointes indiquant le numéro des figures (bulle ronde) ou des annexes (bulle ovale) qui présentent les résultats associés. Les figures en lien avec l'introduction ne sont pas présentées.

2.2 BCCM/ULC

BCCM/ULC est une collection publique, contenant l'une des plus grandes collections de cyanobactéries (sub)polaires documentées au monde. La collection BCCM/ULC vise aussi à rassembler une partie représentative des souches de cyanobactéries terrestres, d'eau douce et marines avec un focus sur la diversité polaire de différentes origines écologiques (tapis limnétiques, croûtes de sol, cryoconites, endolithes…). Actuellement, plus de 200 souches ont été caractérisées par des analyses phénotypiques (morphologie basée sur des observations microscopiques) et génotypiques (ARNr 16S). Cette identification a montré que les souches de BCCM/ULC appartiennent aux ordres Synechococcales, Oscillatoriales, Pleurocapsales, Chroococcidiopsidales et Nostocales (Cornet, Ahn, et al., 2021) (https://bccm.belspo.be/about-us/bccm-ulc).

La collection BCCM/ULC est hébergée par l'unité de recherche InBios de l'Université de Liège et est dirigée par Dr Annick Wilmotte. Divers projets sont réalisés sur les cyanobactéries par une approche de classification polyphasique, incluant l'isolement des souches et des méthodes indépendantes de la culture (séquençage des amplicons du gène ARNr 16S, métagénomique, génomique comparative). L'équipe du Dr Annick Wilmotte participe à des expéditions sur le terrain dans l'Antarctique afin de collecter des échantillons. Enfin, des recherches taxonomiques sont menées pour améliorer la classification du phylum des cyanobactéries, en comparant les caractères morphologiques et moléculaires des souches (https://bccm.belspo.be/about-us/bccm-ulc).

2.3 Pourquoi étudier des cyanobactéries d'Antarctique ?

Nous pouvons découper cette grande question en deux sous-questions pour mieux comprendre son intérêt : (1) Pourquoi étudier les cyanobactéries ? (2) Pourquoi étudier des souches antarctiques ?

Premièrement, les cyanobactéries jouent un rôle important au sein de la biosphère, en tant que producteur primaire dans les écosystèmes marins. Elles produisent également des molécules bioactives et des toxines puissantes (Demay et al., 2019 ; Dextro et al., 2021 ; Mazard et al., 2016). Les cyanobactéries sont également connues pour produire de nombreux métabolites secondaires dont l'activité la plus fréquemment détectée est la cytotoxicité (42% des familles de métabolites) (Demay et al., 2019). Les ordres Oscillatoriales (46,5 %) et Nostocales (29 %) sont ceux qui produisent le plus de métabolites bioactifs variés (Demay et al., 2019).

Les cyanobactéries peuvent produire des efflorescences nocives (= cyanoHAB) dans de nombreux plans d'eau à travers le monde. De nombreuses espèces produisent des toxines, ce qui les rend particulièrement préoccupantes pour l'approvisionnement en eau potable, les loisirs et la pêche dans divers plans d'eau (Burford et al., 2020).

Les cyanobactéries ont également un potentiel dans les domaines biomédical et alimentaire. Certaines cyanotoxines peuvent être bénéfiques et être utilisées comme médicament pour les thérapies cancéreuses (Zanchett & Oliveira-Filho, 2013). D'autres molécules bioactives sont intéressantes dans d'autres voies d'application, comme l'utilisation de leur voie de biosynthèse pour produire des métabolites (Mazard et al., 2016), mais aussi comme source alimentaire (la spiruline (Ciferri & Tiboni, 1985)), ou encore l'utilisation de leurs capacités de photosynthèse, de fixation de l'azote et d'autotrophie en biotechnologie (J. S. Singh et al., 2016 ; R. Singh et al., 2017). Les cyanobactéries peuvent être aussi utiles dans les domaines de la pharmacologie, de la cosmétique, de l'agriculture, de l'industrie alimentaire et comme source de biocarburant (R. Singh et al., 2017). Ces molécules peuvent avoir différentes bioactivités : antibactériennes, antifongiques, anticancéreuses, immunosuppressives, anti-inflammatoires et antituberculeuses (Demay et al., 2019 ; Gerwick & Fenner, 2013).

Deuxièmement, l'Antarctique est un continent assez hostile au vivant, aux conditions extrêmes, caractérisé par de faibles températures, majoritairement inférieures à 0°C (Jadhav et al., 2022 ; Zakhia et al., 2008), de faibles concentrations de nutriments (Wait et al., 2006), de l'irradiance solaire élevée et des radiations UVs importantes (Velichko et al., 2021). Cependant, des communautés microbiennes complexes et diverses sont présentes dans certains lacs recouverts de glace de manière permanente ou dont la couverture fond en été (Jungblut et al., 2016). Dans cet environnement, la présence de compétiteurs (les plantes, …) et de consommateurs (zooplancton, …) sont manquants ou très réduits. La couverture de glace dans ces lacs protège ces microorganismes du vent et favorise les gradients de densité de salinité. Dans ces tapis microbiens, peuplés par une communauté dense de microorganismes stratifiés verticalement, nous retrouvons des cyanobactéries, des eucaryotes microbiens, des microalgues phototrophes et des bactéries hétérotrophes (Jungblut et al., 2016 ; Peeters et al., 2012). Les cyanobactéries sont les principaux producteurs primaires dans ces habitats, remplissant plusieurs rôles écologiques vitaux tels que la fixation du C et du N_2 atmosphériques (Velichko et al., 2021). Elles ont développé différentes stratégies pour résister au froid et aux UV solaires. Elles possèdent différentes molécules telles que des molécules cryoprotectrices, des caroténoïdes

ou encore des molécules photoprotectrices (les acides aminés de type mycosporine (MAA)) (Quesada & Vincent, 1997). Pour faire face aux effets nocifs de l'exposition importante aux UV solaires, surtout en été, où il peut y avoir 24 heures de lumière, elles ont développé des stratégies comme la dissipation de la chaleur et la motilité glissante (Castenholz & Garcia-Pichel, 2013 ; Karsten, 2008 ; Quesada & Vincent, 1997).

Finalement, ces microorganismes d'Antarctique représentent une source potentielle de nouveaux biocatalyseurs exploitables dans des conditions extrêmes (Simon & Daniel, 2011). Les Vallées sèches de McMurdo (MDV) sont des régions avec un certain nombre de lacs recouverts de glace en permanence, abritant des communautés microbiennes complexes et diverses (Priscu et al., 1999). Ces propriétés font de cet environnement l'un des habitats lacustres les plus physiquement stables sur Terre, facilitant les recherches dans ces domaines (Jungblut et al., 2016). Les études des microorganismes du lac de Fryxell, faisant partie des vallées sèches de Mc Murdo (Jungblut et al., 2016 ; Lumian et al., 2021) indiquent l'importance de l'existence d'une communauté microbienne capable de s'adapter au froid. Ces exemples indiquent l'intérêt d'étudier ces microorganismes afin de comprendre comment ils arrivent à vivre et à évoluer en milieu polaire.

2.4 Cyanobactéries

2.4.1 Origine

Les cyanobactéries sont un groupe ancien de procaryotes photosynthétiques (Demoulin et al., 2019).

Il y a 4 milliards d'années, la Terre primitive avait une atmosphère sans oxygène mais riche en gaz comme H_2S, CH_4, NH_3, H_2O, CO_2 (Kasting et al., 2001). Des micro-organismes anaérobies phototrophes anoxygéniques capables de consommer le H_2 comme donneur d'électrons existaient à cette époque. Différentes preuves paléontologiques et géochimiques indiquent l'existence de cyanobactéries dès l'Archéen. Des microfossiles ont été retrouvés dans des schistes, des sédiments et des stromatolithes (Shestakov & Karbysheva, 2017). Les premières cyanobactéries seraient apparues il y a environ 2,6 milliards d'années dans des plans d'eaux chaudes peu profonds (Blank & SÁnchez-Baracaldo, 2010). Au cours de la lente évolution des cyanobactéries qui a duré des centaines de millions d'années, leur appareil photosynthétique a évolué et leur a permis d'utiliser la lumière comme source d'énergie et l'eau comme source de carburant pour

produire de l'oxygène dans ces eaux (Garcia-Pichel, 1998 ; Shestakov & Karbysheva, 2017). À cette époque, les cyanobactéries étaient les seuls procaryotes capables de réaliser la photosynthèse oxygénique puisqu'elles possèdent les deux photosystèmes alors que les autres microorganismes phototrophes anoxygéniques ont des photosystèmes de type PSI ou PSII (Cardona et al., 2019 ; Sánchez-Baracaldo & Cardona, 2020). Petit à petit, cette production d'oxygène va se retrouver dans l'atmosphère et provoquer le grand évènement d'oxydation (GOE), il y a 2,4 milliards d'années (Buick, 2008 ; Canfield, 2004 ; Holland, 2006). Le développement massif des cyanobactéries sur toute la planète, au début du Protérozoïque, a entraîné l'augmentation progressive de la concentration en oxygène détectée dans les roches à partir de 2,32 Ga (Javaux, 2007). Cette oxygénation a modifié la chimie des premiers océans, passant de conditions anoxiques à un état intermédiaire anoxique et sulfidique pour finalement atteindre une teneur en oxygène atmosphérique proche du niveau actuel (~10 %) à la fin du Néoprotérozoïque (**Figure 2**) (Javaux, 2007 ; Peretó, 2011). Ce changement atmosphérique a fourni les conditions favorables à l'apparition des nouveaux micro-organismes aérobies et a obligé les procaryotes anaérobies sensibles à l'oxygène à se déplacer dans d'autres habitats (Shestakov & Karbysheva, 2017).

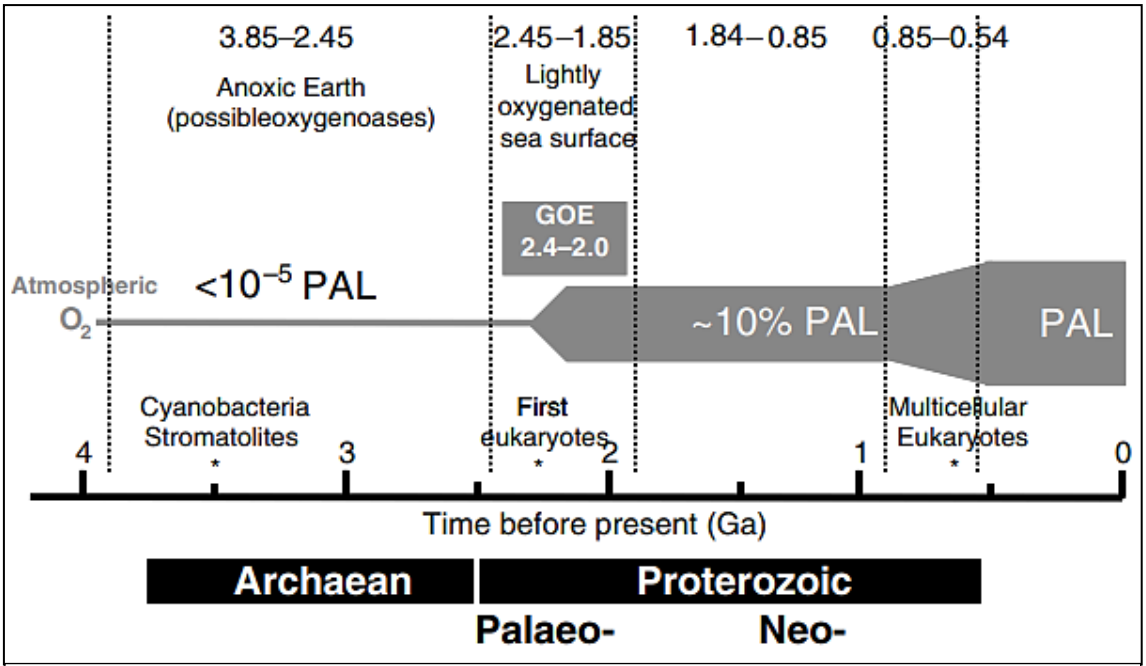


Figure 2 : Ligne du temps hypothétique représentant l'évolution du taux d'oxygène sur Terre. L'accumulation d'oxygène atmosphérique au cours des temps géologiques (en Ga = 10^9 ans) est exprimé en termes de fraction du niveau atmosphérique actuel (PAL) (Peretó, 2011).

À partir du dernier ancêtre commun eucaryote (LECA), les cyanobactéries ont permis la diversification de la vie complexe (Roger et al., 2017), ainsi que le développement de leurs compétiteurs et consommateurs, réduisant ainsi la prolifération des cyanobactéries à des écosystèmes restreints (notamment les milieux extrêmes). Il y a environ 1,5 milliard d'années, une endosymbiose entre une cyanobactérie ancestrale et un eucaryote unicellulaire hétérotrophe protiste aurait donné naissance au plaste (de Vries & Archibald, 2017 ; Sánchez-Baracaldo et al., 2022). Les premiers eucaryotes photosynthétiques auraient émergé et se seraient diversifiés en glaucophytes, algues rouges, algues vertes et plantes terrestres. Ce sont des organismes qui portent toujours actuellement un dérivé de cyanobactérie sous forme de plaste avec une majorité des gènes cyanobactériens transférés dans leur noyau (de Vries & Archibald, 2017 ; Sánchez-Baracaldo et al., 2022). En tant que producteurs primaires, les eucaryotes photosynthétiques dominent désormais la plupart des environnements terrestres et marins (Sánchez-Baracaldo et al., 2017).

Les cyanobactéries ont colonisé différents habitats dans le monde dont les milieux marins, les eaux douces, les milieux terrestres, les sources thermales, et d'autres milieux extrêmes (Whitton, 2012). Les cyanobactéries jouent toujours aujourd'hui un rôle clé dans les cycles mondiaux du carbone et de l'azote, en plus d'être un élément important dans de nombreux réseaux trophiques aquatiques (Burford et al., 2020 ; Sánchez-Baracaldo et al., 2022).

2.4.2 La photosynthèse et les plastes

En tant que phototrophes, toutes les cyanobactéries possèdent des membranes internes appelées thylakoïdes hébergeant l'appareil photosynthétique. Celui-ci est constitué de deux photosystèmes et de leurs pigments, sauf chez *Gloeobacter* où la photosynthèse se déroule dans la membrane plasmique (Grettenberger, 2021 ; Nakamura et al., 2003). Les thylakoïdes sont disposés directement dans les cellules de cyanobactéries, et non au sein d'organites comme les chloroplastes, présents chez les eucaryotes (Komárek & Kaštovský, 2003 ; Rippka et al., 1979).

La photosynthèse est apparue chez les eucaryotes par l'endosymbiose d'une cyanobactérie ancestrale, apparentée à *Gloeomargarita lithophora*, au sein d'un hôte eucaryotique hétérotrophe qui a donné le chloroplaste (Ponce-Toledo et al., 2017). Cette endosymbiose primaire s'est produite chez l'ancêtre d'Archaeplastida (Cardona et al., 2019 ; de Vries & Archibald, 2017). Ensuite, une seconde endosymbiose au sein des algues a permis la propagation de la photosynthèse à d'autres eucaryotes. La plupart des études

phylogénétiques placent la divergence de la lignée des chloroplastes près de la racine des cyanobactéries (ramification précoce) (Couradeau et al., 2012; Criscuolo & Gribaldo, 2011; B. Li et al., 2014; Ponce-Toledo et al., 2017; Shih et al., 2013), bien que quelques études insèrent des chloroplastes plus haut dans l'arbre (ramification tardive) (Dagan et al., 2013; Deusch et al., 2008) ou les nichent dans des clades dérivés, par exemple, les Nostocales (Ochoa de Alda et al., 2014) ou apparentés au sous-groupe V (*Cyanothece)* (Deschamps et al., 2008). Cependant, la position précise du plaste dans la diversité cyanobactérienne reste un phénomène difficile à dater, car les génomes de chloroplastes ont subi une réduction de leur taille par rapport à leurs parents cyanobactériens (Sánchez-Baracaldo et al., 2017). De plus, les microfossiles cyanobactériens exploitables sont peu nombreux et les algues eucaryotes plus complexes et récentes utilisées pour ses études rendent les analyses plus complexes (Demoulin et al., 2019).

2.4.3 Classification taxonomique

Autrefois, la taxonomie était limitée à l'analyse de la morphologie. Les cyanobactéries étaient traditionnellement décrites comme des algues et appelées « cyanophytes » ou « algues bleu-vert », puisqu'elles ont un appareil photosynthétique de type algal qui comprend la chlorophylle et les deux photosystèmes (Dextro et al., 2021 ; Shestakov & Karbysheva, 2017). Jusqu'à la fin du 20e siècle, le système de nomenclature des cyanobactéries suivait L'International Code of Botanical Nomenclature (ICBN). À la fin des années 1970, Stanier et ses collègues ont reconnu la nature procaryotique des cyanobactéries (ressemblant à des bactéries Gram-négatives photosynthétiques) et ont proposé de suivre l'International Code of Nomenclature of Prokaryotes (ICNP) (Bergey's Manual of Systematic Bacteriology (R. Castenholz & Waterbury, 1989 ; Stanier & van Niel, 1962). Les deux codes ont été utilisés en parallèle depuis des décennies pour nommer les cyanobactéries, provoquant ainsi une certaine confusion. La grande majorité des taxons a été décrit en utilisant le code de nomenclature botanique (ICBN) car plus facile d'utilisation. De plus, celui-ci accepta les taxons validement décrits par le code bactériologique, mais non inversement (Demoulin et al., 2019 ; Dextro et al., 2021). La confusion a été récemment résolue par une révision de l'International Code of Nomenclature of Prokaryotes (ICNP) (Oren et al., 2021) qui réciproque la reconnaissance des taxons validement décrits selon l'autre code de nomenclature.

Les cyanobactéries ont été divisées en deux groupes : les cyanobactéries unicellulaires et multicellulaires. Elles ont ensuite été classées selon des critères morphologiques en cinq

sous-sections correspondant aux cinq anciens ordres de cyanobactéries (Komárek et al., 2014 ; Rippka et al., 1979 ; Schirrmeister et al., 2011 ; Shih et al., 2013) :

- Section I : Chroococcales. Ce sont des microorganismes unicellulaires qui ne se divisent que par fission binaire (Komárek et al., 2014 ; Rippka et al., 1979 ; Schirrmeister et al., 2011 ; Shih et al., 2013).
- Section II : Pleurocapsales. Ce sont aussi des microorganismes unicellulaires qui se divisent par de multiples fissions en un ou plusieurs plans et sont solitaires ou disposées en colonies. Les Pleurocapsales peuvent également produire de petites cellules facilement dispersées (baéocytes) après division par de multiples fissions (Komárek et al., 2014 ; Rippka et al., 1979 ; Schirrmeister et al., 2011 ; Shih et al., 2013).
- Section III : Oscillatoriales. Au sein des cyanobactéries filamenteuses multicellulaires, les Oscillatoriales ont des cellules végétatives disposées en filaments perpendiculairement à l'axe de croissance (Komárek et al., 2014 ; Rippka et al., 1979 ; Schirrmeister et al., 2011 ; Shih et al., 2013).
- Section IV : Nostocales. Ce sont des cyanobactéries filamenteuses multicellulaires, qui possèdent des cellules spécifiques, les hétérocystes pour fixer l'azote. Elles peuvent également présenter des cellules résistantes au stress environnemental appelées akinètes. Le genre *Nostoc* est le plus représenté et le plus polyphylétique des cyanobactéries (Komárek et al., 2014 ; Rippka et al., 1979 ; Schirrmeister et al., 2011 ; Shih et al., 2013).
- Section V : Stigonematales. Elles ont les mêmes caractéristiques que les Nostocales mais peuvent aussi se diviser selon plusieurs plans et former de véritables trichomes ramifiés (Komárek et al., 2014 ; Rippka et al., 1979 ; Schirrmeister et al., 2011 ; Shih et al., 2013).

La classification en sous-sections est pratique mais ne reflète pas la phylogénie et engendre une distribution déséquilibrée des phyla (Shih et al., 2013). Des espèces présentant des caractéristiques physiologiques, environnementales et génétiques différentes ont été assignées aux mêmes sections (Demoulin et al., 2019 ; Shestakov & Karbysheva, 2017). Depuis, la classification taxonomique des cyanobactéries a été continuellement réévaluée grâce au développement de la microscopie électronique et des méthodes de biologie moléculaire. L'utilisation du gène de l'ARNr 16S dans la reconstruction phylogénétique du phylum des cyanobactéries au cours des dernières décennies a provoqué des changements drastiques des rangs taxonomiques supérieurs et

a abouti à la description de nouveaux taxons(Komárek et al., 2014 ; Lefler et al., 2023 ; Schirrmeister et al., 2011 ; Shestakov & Karbysheva, 2017). Le gène de l'ARNr 16S fournit une vision plus robuste et précise des relations évolutives des cyanobactéries par rapport à la morphologie seule, permettant la création de clades monophylétiques (Komárek et al., 2014 ; Lefleret al., 2023). Dans la taxonomie actuelle des cyanobactéries, l'approche « polyphasique » est utilisée pour avoir une classification plus « juste », basée sur des critères morphologiques, physiologiques, écologiques et moléculaires (Komárek, 2016 ; Wilmotte et al., 2017). Grâce aux avancées technologiques, Komárek et son équipe ont proposé, en 2014, une nouvelle classification taxonomique des cyanobactéries permettant une révision des positions taxonomiques de certaines espèces ou genres « anciens » de cyanobactéries. Cette classification comprend huit ordres : Nostocales, Chroococcidiopsicales, Spirulinales, Pleurocapsales, Chroococcales, Oscillatoriales, Synechococcales, Gloeobacterales (Hentschke & Junior, 2022 ; Komárek et al., 2014 ; Shestakov & Karbysheva, 2017). Récemment, Strunecký et al ont révisé les ordres et les familles de cyanobactéries sur la base d'analyses phylogénomiques d'ARNr 16S, améliorant notre compréhension et résolvant la polyphylie de ces rangs. Dix nouveaux ordres et quinze nouvelles familles sont proposés pour satisfaire autant que possible l'exigence de monophylétisme à tous les rangs taxonomiques (**Figure 3**) (Strunecký et al., 2023). En 2012, le nombre d'espèces des cyanobactéries était estimé à 8000 sur base des taxons déjà présents dans la base de données Algaebase (Guiry, 2012) mais ne prenait pas en compte les espèces cryptiques. C'est probablement une sous-estimation.

Les études de la diversité microbienne reposent sur des bases de données bien organisées pour une attribution taxonomique précise. Les bases de données qui existent actuellement ne traitent ni correctement ni avec précision le phylum des cyanobactéries (Lefler et al., 2023). Les taxonomies des bases de données telles que NCBI, CyanoSeq, GTDB ou SILVA (Lefler et al., 2023 ; Ramos et al., 2017) sont souvent incompatibles entre elles pour la taxonomie et la nomenclature des cyanobactéries (Lefler et al., 2023). Ces erreurs taxonomiques et nomenclaturales conduisent à des conclusions inexactes et trompeuses pour les chercheurs menant des études sur la diversité (Lefler et al., 2023).

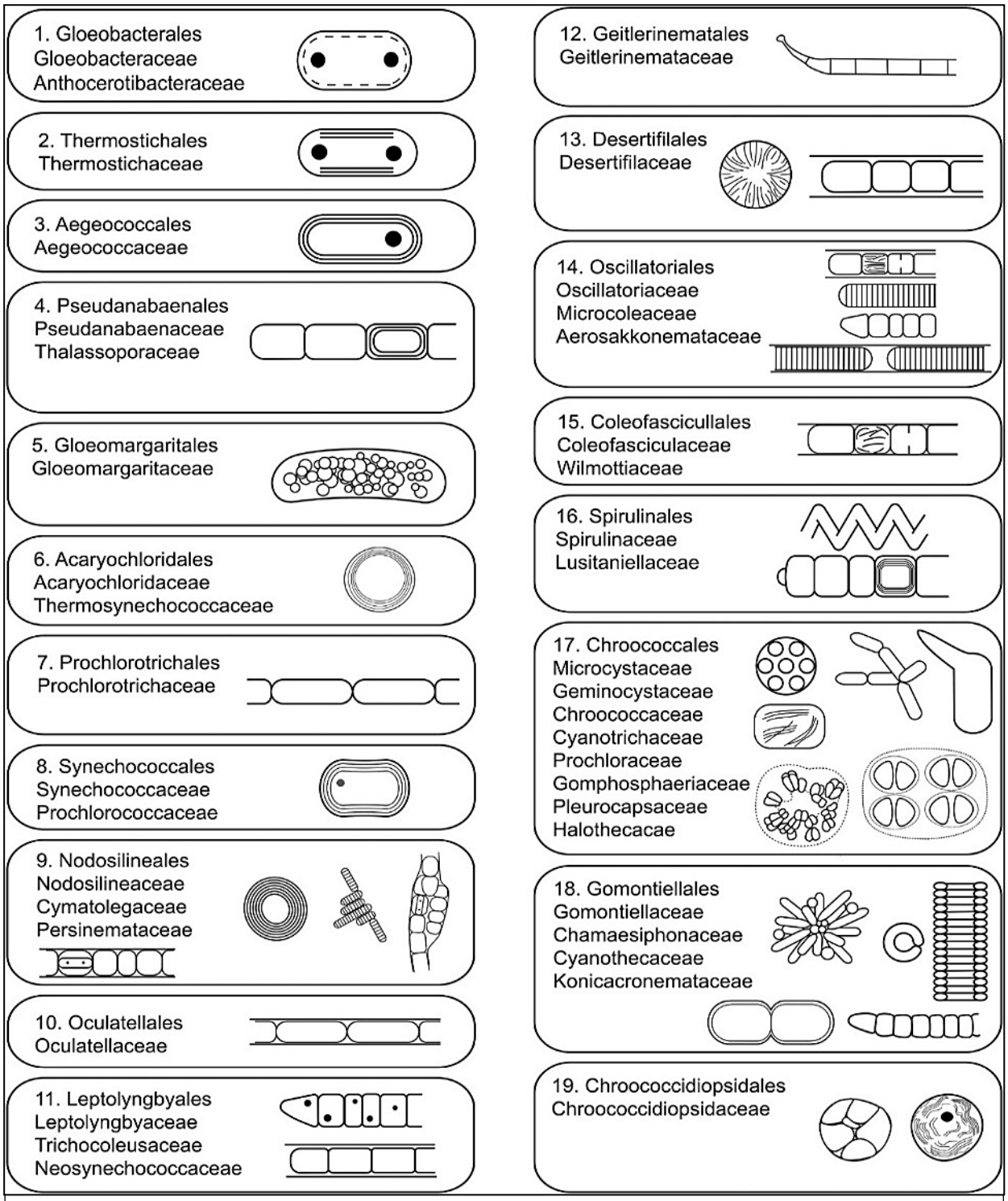


Figure 3 : Vue schématique des ordres et des familles de cyanobactéries (à l'exception des Nostocales), et certains des caractères taxonomiques morphologiques importants utilisés pour les distinguer (Strunecký et al., 2023).

2.4.4 ANI pour la détermination taxonomique

Historiquement, l'hybridation ADN-ADN expérimentale (DDH) a été la « norme de référence » pour la délimitation des espèces. Une valeur DDH ≥ 70 % était reconnue comme la limite d'espèce entre deux souches de bactéries (Wayne et al., 1987). Suite à la croissance du nombre de génomes complets séquencés, la technique expérimentale complexe a été remplacé par une mesure bioinformatique de la similarité entre les génomes (Konstantinidis & Tiedje, 2005). C'est ainsi que l'« Average Nucleotide Identity » (ANI) remplace maintenant le DDH car il est plus rapide et fournit des données reproductibles. Deux souches sont considérées comme appartenant à la même espèce, si elles ont plus de 95 % d'ANI entre elles (Konstantinidis & Tiedje, 2005). Entre 73% et 95% d'ANI (au niveau interspécifique) deux souches sont considérées comme appartenant au même genre. Pour être plus précis, la limite de démarcation du genre a été fixée à 73,11%, selon Barco et al. (2020).

L'ANI indique le pourcentage moyen de similarité des séquences partagées entre une paire de génomes (Goris et al., 2007 ; Konstantinidis & Tiedje, 2005). ANI aligne et compare des ensembles de séquences d'un génome de départ (query) à la séquence d'un autre génome de référence. Les gènes sont considérés comme semblables lorsqu'une séquence de gène correspond à au moins 60 % d'identité sur 70 % de la longueur du gène dans le génome de référence, en utilisant l'outil de recherche d'alignement local de base (BLAST) (Palmer et al., 2020).

Depuis sa description initiale, ANI a subi un certain nombre d'ajustements afin d'accélérer son utilisation, notamment par rapport à la fragmentation du génome. À la suite de cela, l'outil de référence pour les procaryotes est devenu fastANI, une nouvelle méthode pour estimer l'ANI à l'aide d'une cartographie de séquences sans alignement. FastANI est précis pour les génomes complets et incomplets, et calcule les valeurs ANI par paires parmi tous les génomes procaryotes disponibles dans la base de données NCBI. À plus de de 95 % d'ANI, les génomes sont considérés comme faisant partie de la même espèce. Entre 95% et 83% d'ANI, ils sont considérés faisant partie du même genre (Jain et al., 2018). La limite du genre pour FastANI n'est pas abordée. Barco et al précise uniquement qu'il existe une différence de 10%, correspondant à une zone d'incertitude (pour FastANI par rapport à ANI). FastANI est l'outil de référence utilisé pour déterminer la taxonomie de génome dans GTDB (Parks et al., 2019), bien que d'autres outils soient en phases de test, tels que

GGDC (=Genome-to-Genome Distance Calculator) de DSMZ qui s'inspire de l'hybridation ADN-ADN et utilise la même valeur seuil pour l'espèce à 70% (Meier-Kolthoff et al., 2013).

2.4.5 *Laspinema sp.*

Laspinema sp., un genre récemment décrit, a été trouvé pour la première fois dans les sources thermales (Heidari et al., 2018) et dans la vase recouvrant le bassin de maturation des Thermes de Balaruc-Les-Bains (Duval et al., 2020). L'étymologie du nom vient du grec : « *Láspi »* pour boue, saleté et « *-nema* » pour fil, et fait référence à l'habitat du genre et à son apparence (Heidari et al., 2018). Seule la taxonomie basée sur l'ARNr 16S a pu être utilisée sur ce genre jusqu'à maintenant (Heidari et al., 2018 ; Stanojković et al., 2022). Actuellement, le genre *Laspinema* englobe plusieurs espèces (*L. thermale*, *L. etoshii et L. lumbricale*) (Duval et al., 2020 ; Heidari et al., 2018) ainsi que des espèces *de Laspinema* précédemment mal identifiées comme *Oscillatoria acuminata* PCC 6304 (Zimba et al., 2021). Malheureusement, en raison du sous-échantillonnage de ce genre, son histoire évolutive n'est toujours pas entièrement résolue (Stanojković et al., 2022). Nous avons déterminé que certaines souches de la collection BCCM/ULC sont des espèces du genre *Laspinema* et pourraient agrémenter ce nouveau genre de nouvelles informations génomique et biochimiques.

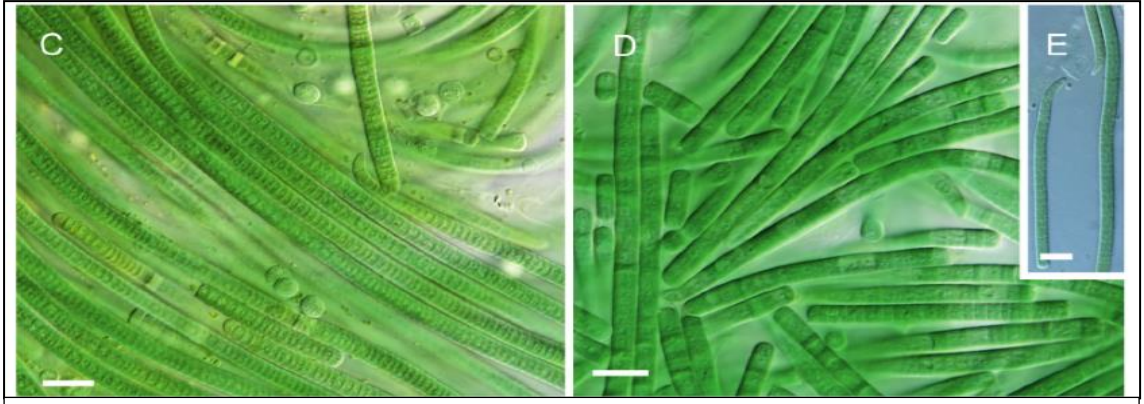


Figure 4 : Photos de la souche *Laspinema thermale* HKS5 (Heidari etal., 2018)

Les deux études décrivent ce nouveau genre de la même manière. *Laspinema sp.* PMC 878.14, isolée de boues thermales et *Laspinema thermale* HKS5 (**Figure 4**), isolée de tapis benthique de sources thermales, sont des filaments droits ou ondulés bleu-vert ou vert olive souvent progressivement atténués aux extrémités (Duval et al., 2020 ; Heidari et al., 2018). Les gaines sont minces et peu visibles, voire absentes. Les cellules sont plus courtes que larges, de 3 à 5 µm de large et de 2 à 3 µm de long. Les trichomes cylindriques sont

non ramifiés, toujours légèrement rétrécis aux parois transversales, finement granulées, courbés, crochus et intensément mobiles. Les cellules apicales sont allongées, coniques, courbées ou droites, à extrémité arrondie sans calyptre, avec des parois cellulaires externes épaissies (Duval et al., 2020 ; Heidari et al., 2018). Les thylakoïdes ont une disposition radiale avec plusieurs espaces inter-thylakoïdaux. D'autres composants ont été observés dans les cellules tels que de nombreuses petites vésicules gazeuses ou aérotopes (qui favorisent la flottaison des filaments), des carboxysomes, des granules de cyanophycine (réserves d'azote), du glycogène, des lipides, ou encore du poly-ß-hydroxybutyrate agissant comme sources de carbone et d'énergie, et des granules de polyphosphate (réserves de phosphore) (Duval et al., 2020 ; Heidari et al., 2018).

2.5 Données génomiques

2.5.1 Séquençage

Depuis quelques années, les technologies de « séquençage de nouvelle génération (NGS) », telles qu'Illumina, sont utilisées pour le séquençage de génomes cyanobactériens, engendrant une augmentation importante des données de séquençage. Ces technologies de séquençage « à lecture courte » permettent le séquençage massif (Hu et al., 2021 ; Mardis, 2013). Cependant, du fait qu'il s'agit de technologies à lecture courte, le séquençage de génome entier ne peut être terminé et contient des lacunes. Cela réduit la qualité moyenne des génomes dans les bases de données publiques, entrainant des artéfacts. Le principal défi pour ces technologies de séquençage est la présence de répétitions. Il en existe deux types : global, comme dans l'opéron d'ARNr de procaryote où une séquence est répétée dans tout le génome, ou local, comme les répétitions en tandem à nombre variable (VNTR) (van Belkum et al., 1998) dans le génome de procaryotes où quelques paires de bases sont répétées en tandem (Koren & Phillippy, 2015). Ces problèmes sont résolus par le séquençage à lecture longue, le séquençage de troisième génération, tel que Pacific Biosciences ou Oxford Nanopore (Koren & Phillippy, 2015). Ces technologies de « lecture longue » peuvent surmonter les problèmes rencontrés avec les lectures courtes, telles que les répétitions à l'échelle du génome et la détection de variantes structurelles. Le séquençage à longue lecture améliore la qualité des génomes en permettant l'assemblage complet des génomes microbiens (Koren & Phillippy, 2015).

2.5.2 Métagénomique

Le séquençage « Shotgun » des eaux de la mer des Sargasses, par Venter et ses collègues en 2004, a montré que la métagénomique est un moyen réalisable d'accumuler des connaissances génomiques au sein de l'environnement. Dans cette étude, ils ont récupéré près de 1,5 Gbp de séquences d'ADN microbien à partir de populations microbiennes de trois sites marins en utilisant en masse le séquençage complet des génomes à partir d'eau de mer filtrée. Cela conduit à la découverte de près de 70 000 nouveaux gènes (Alves et al., 2018 ; Simon & Daniel, 2011 ; Venter et al., 2004). Une étape cruciale dans l'analyse taxonomique de grands ensembles de données métagénomiques est le « binning » qui consiste à associer des séquences issues d'un mélange de différents organismes à des groupes phylogénétiques selon leurs origines taxonomiques (Simon & Daniel, 2011).

L'étude métagénomique des cyanobactéries permet de découvrir de nouvelles souches et nouvelles biomolécules intéressantes provenant de milieux extrêmes (Simon & Daniel, 2011). En 2018, Luc Cornet et son équipe ont réalisé l'analyse métagénomique de nouvelles souches cyanobactériennes (sub)polaires. Ils ont étudié des souches de cyanobactéries non stériles et ont créé un pipeline métagénomique qui facilite l'assemblage des génomes cyanobactériens, avec une quantité suffisante d'ADN génomique. Ils ont réalisé 14 nouveaux assemblages de cyanobactéries, dont 11 souches (sub)polaires, et 13 assemblages d'organismes appartenant à leur microbiome, principalement des Proteobacteria er Bacteroidota (Cornet, Bertrand, et al., 2018).

2.5.3 Risque de contamination

Selon Cornet & Baurain (2022), trois sources de contaminations majeures sont possibles : informatiques, biologiques et expérimentales (**Figure 5**). Les sources de contaminations variées peuvent être regroupé en deux groupes principales :

1) Redondant où le segment génomique est présent plusieurs fois à cause d'un remplacement d'une séquence étrangère qui contient le gène attendu au sein du génome entier, (Cornet & Baurain, 2022)

2) Non redondant où un segment génomique a été ajouté ou remplacé dans l'assemblage. Il en existe deux sous-catégories :

    a. Une contamination véritablement non redondante : où une région génomique supplémentaire est ajoutée et n'a aucune région homologue avec

l'organisme cible. Elle est présente en raison de l'inclusion d'un organisme taxonomiquement différent (Cornet & Baurain, 2022).

b. Une contamination par remplacement de génome : où un segment génomique fait défaut dans l'organisme cible et est remplacé par une région génomique étrangère comprenant des gènes attendus (Cornet & Baurain, 2022).

Une source de contamination possible est l'existence de relations trophiques complexes entre les cyanobactéries et des bactéries de type Proteobacteria ou Bacteroidota. L'obtention de cyanobactéries axéniques (= dépourvues de germes contaminants, donc "stériles") est difficile à cause des communautés bactériennes vivant en étroite relation avec les cyanobactéries dans la nature (Cornet, Bertrand, et al., 2018). Cornet et son équipe ont ainsi montré qu'une grande partie (52 %) des génomes de cyanobactéries accessibles au public sont contaminés par de séquences étrangères. L'analyse des génomes indiquent jusqu'à 41,5 % des séquences génomiques contaminées par des protéobactéries et des bactéroïdes (Cornet, Meunier, et al., 2018).

Chaque année, l'augmentation du nombre de génomes accroit le risque de contamination des génomes (Cornet & Baurain, 2022). L'introduction de séquences étrangères peut se produire à différents moments du processus de séquençage (**Figure 5**) : contamination lors de la culture par des organismes indésirables (un champignon, par exemple), l'inclusion d'ADN indésirable (l'ADN humain, par exemple) lors de l'extraction d'ADN, le séquençage sur des plateformes partagées, le séquençage d'organisme chimérique (un bacille-cyanobactérie ou un OGM par exemple), la présence d'erreurs taxonomiques dans les bases de données, la fusion de régions génomiques similaires lors de l'assemblage métagénomique(Cornet & Baurain, 2022). Cornet & Baurain soulignent l'importance de développer des protocoles bioinformatiques minutieux pour le traitement des données génomiques et diminuer la présence de contamination lors des études génomiques.

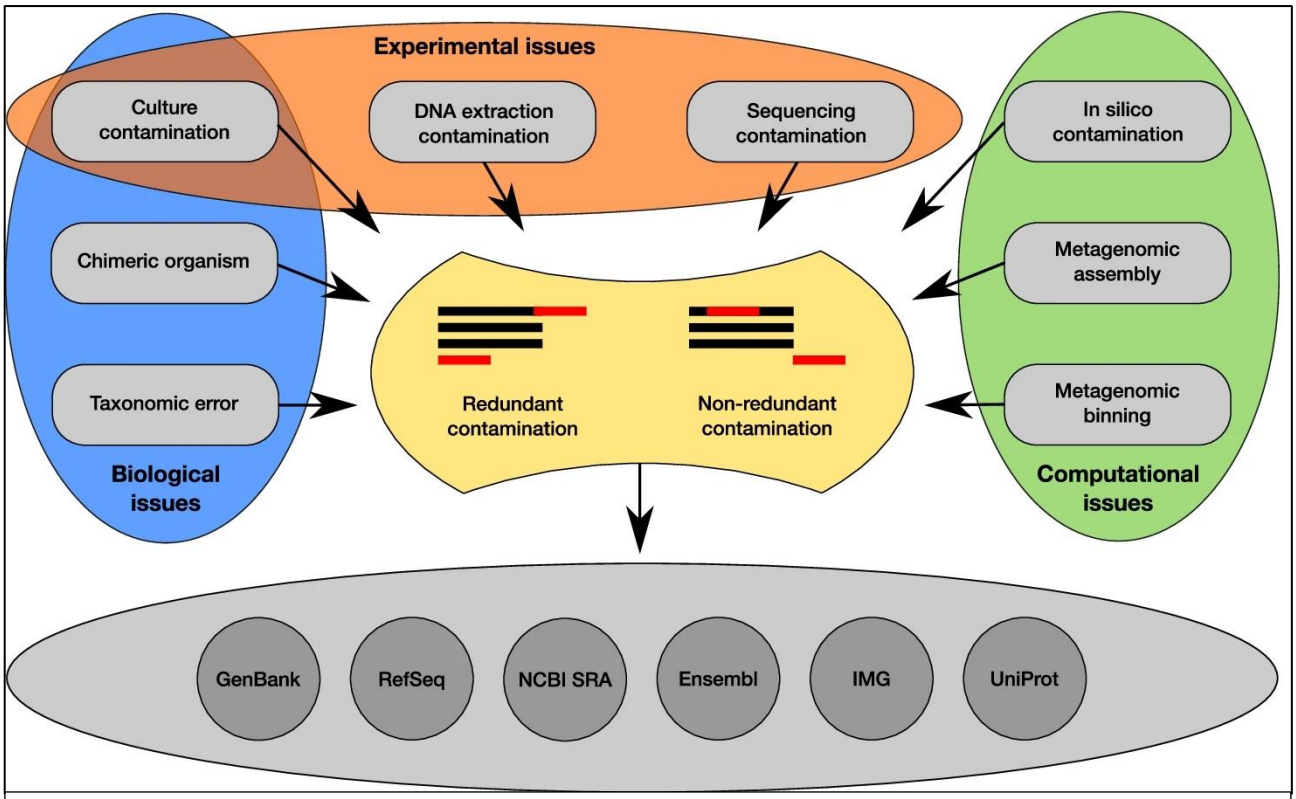


Figure 5 : Sources de contamination génomique. Trois types de problèmes conduisent à la contamination des données de séquences génomiques : biologiques, expérimentales et informatiques. La contamination de cultures "pures" peut être due à des causes expérimentales (par exemple, l'introduction accidentelle de micro-organismes contaminants) et biologiques (par exemple, la présence d'un endosymbionte). Une contamination redondante est lorsqu'un segment génomique est présent plusieurs fois dans un génome (par exemple, plusieurs ARNr SSU provenant d'organismes différents). Une contamination non redondante se produit lorsqu'une région génomique de l'organisme principal, celle qui est attendue, est remplacée par la région correspondante d'un organisme étranger (par exemple, l'ARNr SSU de l'organisme principal est remplacé par l'ARNr SSU d'un organisme étranger). Un segment d'ADN supplémentaire, ne faisant pas partie de l'organisme principal mais appartenant à un contaminant, serait également considéré comme une contamination non redondante (par exemple, de l'ADN eucaryote dans un génome bactérien) (Cornet & Baurain, 2022)

2.6 Phylogénie

Depuis la fin des années 90, de nombreuses études phylogénétiques basées sur des loci uniques (comme l'ARNr 16S ou certaines protéines) ont été publiées : (Cornet, Wilmotte, et al., 2018 ; Shih et al., 2013 ; Taton et al., 2006). Malheureusement, les phylogénies basées sur ces marqueurs uniques montrent leurs limites face à l'ajout de nouveaux taxons aux arbres phylogénétiques de plus en plus grands (Cornet, Ahn, et al., 2021; Cornet, Magain, et al., 2021). En 2013, le projet de séquençage CyanoGEBA (Genomic Encyclopedia of Bacteria and Archaea) a permis d'améliorer la couverture génomique des taxons cyanobactériens (Shih et al., 2013). Depuis lors, le nombre de génomes de cyanobactéries accessibles au public a considérablement augmenté. En décembre 2022, il y avait 3870 assemblages de génomes de cyanobactéries dans les banques de données

GenBank et RefSeq. L'explosion des données génomiques a permis de nouvelles recherches en phylogénie sur des sujets variés, tel que la taxonomie et l'étude de génomes entiers (Ramos et al., 2017 ; Strunecký et al., 2021), l'adaptation à l'environnement grâce au transfert horizontal de gènes (Chen et al., 2020), l'origine et l'évolution des cyanobactéries (Cornet, Ahn, et al., 2021 ; Sánchez-Baracaldo & Cardona, 2020).

Enfin, aucune étude phylogénétique n'a été réalisée, jusqu'à présent, sur le genre *Laspinema*, puisque c'est un nouveau genre découvert en 2018 et qu'il est encore confondu avec d'autres genres cyanobactériens (Heidari et al., 2018 ; Stanojković et al., 2022).

2.7 GEN-ERA

GEN-ERA est une boite à outils bioinformatique conçue pour les études génomiques sur les bactéries et les levures. GENERA est l'acronyme utilisé comme titre du projet BELSPO « Culture collections in the Genomic Era » (2019-2022). La boite à outils créée lors de ce projet contient 14 workflows Nextflow (DI Tommaso et al., 2017), soutenu par 11 conteneurs Singularity (Kurtzer et al., 2017). Chaque workflow est accompagné d'un script Python, inclus dans le conteneur, pour l'analyse et le formatage des résultats. Chaque workflow peut être exécuté par une seule ligne de commande, augmentant la reproductibilité des analyses (Cornet et al., 2023). GEN-ERA est disponible sur GitHub (https://github.com/Lcornet/GENERA). Les outils utilisés dans le cadre de ce mémoire seront décrits dans la partie Matériels et Méthodes. Toutes les lignes de commandes utilisées dans le cadre de ce mémoire sont disponibles dans l'**annexe 1**, rendant ces analyses complètement reproductibles.

Objectifs

L'objectif principal de cette étude est d'étudier les génomes des souches provenant de la collection BCCM/ULC afin de les classer phylogénétiquement, de les identifier au niveau taxonomique et de déterminer leur métabolisme en utilisant GEN-ERA comme outil bioinformatique (https://bccm.belspo.be/content/bccm-collections-genomic-era). Pour se faire, voici sur quoi notre attention s'est focalisée tout au long de l'étude :

1. Est-ce que les souches de la collection sont contaminées et, si oui, par quels types d'organismes ?
2. À quelle famille et quel genre cyanobactérien peuvent-elles être associées ?
3. Où se placent-elles dans un arbre phylogénétique de cyanobactéries ?
4. Quel nom taxonomique pouvons-nous leur attribuer ?
5. Est-ce que ce sont de nouvelles espèces ?
6. Est-ce que leurs gènes ont des fonctions spécifiques qui leur sont propres ?
7. Possèdent-elles des métabolites secondaires qui justifient leur résistance au froid et aux UVs ?
8. Quelles sont les particularités des génomes des souches antarctiques par rapport aux autres cyanobactéries ?

Voici les grandes questions majeures que nous allons essayer de répondre dans ce mémoire.

Afin de faciliter la compréhension et la lecture du mémoire, des groupes de génomes sont désignés et nommés comme suit :

- L'embranchement *Laspinema* correspond au groupe 1 (**G1**)
- Les 3 souches d'Antarctique se nomment groupe 2 (**G2**)
- Les 3 souches de Stanojkovic et al sont le groupe 3 (**G3**)

3 Matériels et Méthodes

3.1 Origine des souches de la collection BCCM/ULC

Toutes les souches viennent de la collection BCCM/ULC. Elles viennent toutes d'Antarctique sauf ULC722 qui vient d'un lac brésilien (**Table 1**). ULC722, ULC307 et ACSII 155, sont les seules souches où il n'y a pas d'informations sur le site (https://bccm.belspo.be/catalogues/catalogue-search?collection=ULC) car les données n'ont pas encore été publiées publiquement.

La souche ULC307 a été isolée d'échantillons d'un nid de Labbe avec beaucoup de restes de pétrels de neiges dans les Montagnes Sør Rondane (Cornet, Ahn, et al., 2021). Les souches ULC008, ULC002 et ULC029 sont originaires du même projet « EC project MICROMAT » en 1997-1998. Ils ont été isolés d'échantillons prélevés dans le « Progress Lake » (pour ULC008, https://bccm.belspo.be/catalogues/bm-details?accession_number=ULC%200008) et d'échantillons de tapis microbiens benthiques de lacs des collines du Larsemann (pour ULC002 et ULC029, https://bccm.belspo.be/catalogues/bm-details?accession_number=ULC%200002; https://bccm.belspo.be/catalogues/bm-details?accession_number=ULC%200029). Les souches ULC096 et ULC102 ont été isolées d'échantillons venant des étangs de plate-forme de glace de Mc Murdo. La souche ULC096 a été isolée en 1998 par Prof. Castenholz (U.Oregon, USA) (https://bccm.belspo.be/catalogues/bm-details?accession_number=ULC%200096). La souche ULC102 a été isolée en 1996 par TL Nadeau. Cette souche a été envoyée par le Prof. Castenholz (U.Oregon, USA) en 2007 (https://bccm.belspo.be/catalogues/bm-details?accession_number=ULC%200102). La souche ULC128 a été déposée par le Prof. J. Elster (U. South Bohemia, Czech Republic) en 2007 sous le nom de JR29 et trouvé dans le littoral du lac Rouge, dans des tapis brun-rouge sur les roches du fond de lac dans l'île de James Ross (https://bccm.belspo.be/catalogues/bm-details?accession_number=ULC%200128). La souche ULC180 a été isolée en 2011 par Zorigto Namsaraev lors de la campagne BELDIVA en Antarctique, sur des biofilms noirs et des graviers en granite (https://bccm.belspo.be/catalogues/bm-details?accession_number=ULC%200180).

ID	BCCM Taxonomic Name	Habitat	Sequencing
ACSII155	*Nostoc minutum*	Terrestre, sol de steppe en Sibérie	Illumina
ULC002	*Nostoc sp.* ANT.LH52B.1	Microbial mat, Lake 52, Larsemann Hills, East Antarctica	Illumina
ULC008	*Nostoc sp.* ANT.PROGRESS.2.1	Microbial mat, Lake Progress, Larsemann Hills, Eastern Antarctica	Illumina & Nanopore
ULC029	*Stenomitos sp.* ANT.LH52B.3	Microbial mat, Lake 52, Larsemann Hills, East Antarctica	Illumina
ULC096	*Phormidium terebriforme* ANT-PANCREAS	Pancreas pond, Mc Murdo Ice Shelf, Antarctica	Illumina & Nanopore
ULC102	*Phormidium pseudopriestleyi* ANT-BRACK -2	Brack Pond, Mc Murdo Ice Shelf, Antarctica	Illumina & Nanopore
ULC128	*Microcoleus favosus* JR29	Periphyton in littoral of Red Lake, brown-red mats on bottom rocks, James Ross Island, Antarctic Peninsula	Illumina & Nanopore
ULC180	*Nostoc sp.* OTC7	Granite outcrop on the northern side, black biofilms on granitic gravel, Tanngarden, Dronning Maud Land, Sør Rondane Mountains, Antarctica	Illumina & Nanopore
ULC307	*Phormidium autumnale* P4	Sample of *Prasiola* in a locality with a skua nest and a lot of remains of Snow Petrels, First nunatak of the Pingvinane system, westerly from Utsteinen, Dronning Maud Land, Sør Rondane Mountains, Antarctica	Illumina & Nanopore
ULC722	*Phormidium cf. chalybeum* CCIBt3309	Phytoplankton, Alkaline Lake, Pantanal da Nhecolândia, Mato Grosso do Sul, Brazil	Illumina & Nanopore

Tableau 1 : *Présentation des données de taxonomie, d'habitat et de séquençage à propos des souches de la collection BCCM et ACSII155 (https://bccm.belspo.be/catalogues/catalogue-search?collection=ULC)*

3.2 Séquençage des génomes de la collection BCCM/ULC

Après extraction de l'ADN et préparation de la librairie avec le Nextera XT DNA Library preparation Kit, le séquençage de toutes les souches s'est effectué avec Illumina (MiSeq) (GIGA Genomics, Université de Liège) (Cornet, Bertrand, et al., 2018) (**Tableau 1**). Ensuite, certaines librairies (ULC008, ULC046, ULC096, ULC102, ULC128, ULC108, ULC307, ULC722) ont également été utilisées pour le séquençage de longues lectures sur

le séquenceur MinION (**Tableau 1**) après préparation du séquençage par barcoding et ligation (Oxford Nanopore Technologies, UK) (Cornet, Bertrand, et al., 2018).

Le séquençage des souches ACSII155, ULC008, ULC046, ULC128, ULC180 et ULC307 a été réalisé par Valentina Savaglia. Beatriz Roncero-Ramos a réalisé le séquençage des souches ULC002 et ULC029. Les souches ULC096, ULC102 et ULC722 ont été séquencées par Dr. Anne-Catherine Ahn.

3.3 Les outils de GEN-ERA

3.3.1 Téléchargement de génomes

Le premier outil, *Genome-downloader.nf*, permet de télécharger les génomes selon la taxonomie NCBI (GenBank et RefSeq). Un miroir local de la taxonomie NCBI est chargé avec le script setup-taxdir.pl V0.212670 de la suite Bio-MUST-Core (Denis Baurain, 2021). L'utilisateur doit spécifier le nom du groupe et le rang taxonomique (par exemple, « Gloeobacterales » et « ordre »). L'outil favorise le téléchargement des assemblages GCF de RefSeq par rapport aux assemblages GCA de GenBank. L'outil peut également télécharger les protéines des génomes si elles existent sur les serveurs NCBI (Cornet et al., 2023).

L'outil a été utilisé pour le téléchargement des génomes des *Gloeobacter*, comme groupe externe. J'ai précisé le niveau taxonomique (genre) et le groupe (*Gloeobacter*), activé l'option « refseq » et désactivé l'option « genbank ».

Ensuite, il a servi au téléchargement des génomes de toutes les cyanobactéries présentes dans GenBank et RefSeq afin de permettre la comparaison de génomes avec les bins pour l'analyse d'ANI. Les options « refseq » et « genbank » étaient donc activées. Le niveau taxonomique était précisé (phylum), ainsi que le groupe (Cyanobacteria).

Les instructions des nextflow sont en **annexe 1**.

3.3.2 Assemblage des génomes

Le deuxième outil, *Assembly.nf*, est dédié à la production de génomes. Ce flux de travail peut assembler des génomes et des métagénomes non seulement à partir de lectures courtes Illumina, mais également de données à lecture longue PacBio ou Nanopore, grâce à l'utilisation de SPAdes V3.15.3 sur les lectures courtes (Bankevich et al., 2012), metaSPAdes V3.15.3 pour les métagénomes (Nurk et al., 2017) et metaFlye V2.19.b1774 sur les lectures longues avec l'option métagénome (Kolmogorov et al., 2020). Une taille de

génome attendue doit être fournie par l'utilisateur pour tous les assemblages à lecture longue. Une option pour le « binning » métagénomique, regroupant les contigs en génomes individuels assemblés par métagénome (MAG) est possible grâce à MetaBAT2 V2.15.6 qui est utilisé pour les séquences procaryotes (Kang et al., 2019) et CONCOCT V1.1 (Alneberg et al., 2014) qui est plus efficace pour les données eucaryotes (Saary et al., 2020).

Cet outil a été utilisé sur les 11 souches de la collection BCCM afin de réaliser l'assemblage sur des lectures courtes et longues (Nanopore) et d'obtenir des bins via MetaBAT2 et CONCOCT. Pour chacune des souches, le nom des fichiers (nom.fastq) d'illumina (pour les options shortreadsR1 et shortreadsR2) et de Nanopore (pour les options ontreads) a été donné. L'option « metagenome » a été activée et la taille du génome a été définie à 5mb. J'ai indiqué à l'option « binner » d'utiliser tous les outils de binning (MetaBAT2 et CONCOCT).

Les lignes de commandes du nextflow sont visibles en **annexe 1**.

3.3.3 GTDB

GTDB.nf utilise GTDBTk V2.2.0-r207 (Chaumeil et al., 2022) pour la classification et l'identification taxonomique des génomes procaryotes selon GTDB (Cornet et al., 2023).

Cet outil a été utilisé sur les 11 souches de la collection BCCM et sur les 52 génomes sélectionnés pour la réalisation de l'arbre phylogénétique des cyanobactéries. Dans la commande nextflow, uniquement le nom du fichier d'entrée a été précisé : « genome ».

Les lignes de commandes du nextflow sont visibles en **annexe 1**.

3.3.4 Analyse des contaminations

L'outil *GENcontams.nf* est utilisé pour l'estimation de la contamination génomique, l'exhaustivité et la production de statistiques sur le génome avec l'utilisation de 7 algorithmes différents (Cornet et al., 2023).

L'estimation de la contamination (c'est-à-dire l'inclusion d'ADN étranger dans un assemblage de génome) est possible via plusieurs outils notamment CheckM V1.1.3 (Parks et al., 2015), CheckM2 V2.1.3 (Chklovski et al., 2022), GUNC V1.0.5 (Orakov et al., 2021) et Kraken 2 V2.1.2 (Wood et al., 2019) pour les génomes bactériens. Kraken permet la détection de contaminations eucaryotiques chez les bactéries (Cornet et al., 2023). La complétude est assurée par CheckM V1.1.3 (Parks et al., 2015), CheckM2 V2.1.3

(Chklovski et al., 2022). L'obtention de métriques classiques sur la contiguïté et la taille du génome est possible avec QUAST V5.1.orc1 (Gurevich et al., 2013).

GENcontams.nf a été utilisé à plusieurs reprises, à chaque fois avec les algorithmes CheckM2 V2.1.3, Kraken 2 V2.1.2 et QUAST V5.1.orc1. Il a d'abord été utilisé pour les 11 souches de la collection BCCM afin de réaliser une première sélection. Il y avait 21 fichiers fna en entrée pour l'analyse, venant des bins d'origine cyanobactérienne sélectionnés grâce aux résultats de GTDB. Le taux de contamination, de complétude et la qualité d'assemblage ont été analysés. Le choix de sélection s'est déroulé comme suit : (1) sélection des bins ayant un taux de complétude supérieur à 70% (CheckM2 V2.1.3) ; (2) sélection des bins ayant un taux de contamination inférieure à 10% pour Kraken 2 V2.1.2 ; (3) sélection des bins ayant un taux de contamination de 10% pour CheckM2 V2.1.3 (Bowers et al., 2017). L'outil a également été utilisé sur un plus grand nombre de génomes (455 génomes), dont les génomes de cyanobactéries téléchargés sur GenBank et RefSeq (422 génomes), des analyses ANI (27 génomes) et les 6 bins des souches ULC, afin de sélectionner les génomes de meilleures qualités pour la réalisation du premier arbre phylogénétique. À chaque fois, l'analyse a été lancée trois fois en modifiant le mode (Kraken, CheckM2 et QUAST). J'ai précisé le nom du répertoire d'où se trouvait les fichiers d'entrée (GENERA-input), ainsi que la terminaison des fichiers (fna), et le niveau taxonomique (species).

Les lignes de commandes des nextflow sont visibles en **annexe 1**.

3.3.5 ANI

L'outil *ANI.nf* calcule les distances moyennes des nucléotides par paires entre les génomes en utilisant fastANI V1.33 (Jain et al., 2018). Le mode « many to many » est installé par défaut, mais il est possible de le modifier en « one to many » (Cornet et al., 2023).

L'outil *ANI.nf* a été utilisé, la première fois, sur les 3870 génomes cyanobactériens (GenBank et Refseq) téléchargés avec les 6 bins sélectionnés de la collection BCCM/ULC (ULC002, ULC008, ULC029, ULC096, ULC102, ULC722), en mode « one to many ». Il y avait 3876 génomes en fichiers d'entrée (.fna). Deux listes ont été créées : une pour les bins ULC (« shortlist ») et une autre pour tous les génomes cyanobactérien (« list »).

La seconde fois, l'outil a été utilisé sur les 27 génomes de l'arbre des Oscillatoriaceae, en mode « many to many ». Une liste a été créée avec le nom des 27 génomes.

Les lignes de commandes des nextflow sont visibles en **annexe 1**.

3.3.6 Orthology

L'outil *Orthology.nf* a deux utilités : l'inférence orthologique et la détermination des gènes spécifiques des génomes.

L'inférence orthologique peut être effectuée avec OrthoFinder V2.5.4 (Emms & Kelly, 2019), disponible pour les deux domaines, ou avec OrthoMCL (L. Li et al., 2003), disponible pour les procaryotes, via le pipeline pangénomique d'Anvi'o V7. 1 (Eren et al., 2015). Les deux logiciels peuvent calculer les groupes orthologues (OGs) de protéines. L'inférence orthologique commence généralement à partir de protéomes complets, la prédiction de protéines étant réalisée par Prodigal V2.6.3 (Hyatt et al., 2010) comme prédiction des protéines des procaryotes. Après l'inférence orthologique, *Orthology.nf* fournit automatiquement les gènes core, partagés par tous les génomes fournis par l'utilisateur en unicopie (Cornet, 2023).

L'autre option de *Orthology.nf* permet à l'utilisateur de déterminer les gènes spécifiques, via une sous-liste d'organismes, sans intrus. La principale différence avec les gènes de base est que des OG candidats spécifiques subiront un enrichissement orthologue en exploitant les génomes de tous les organismes de l'inférence orthologue. L'enrichissement orthologue est réalisé avec Forty-Two V0.212670 (Irisarri et al., 2017 ; Simion et al., 2017).

L'outil *Orthology.nf* a été utilisé trois fois : (1) pour la réalisation du premier arbre phylogénétique des cyanobactéries avec 250 génomes, (2) du second arbre phylogénétique des cyanobactéries avec 52 génomes et (3) de l'arbre phylogénétique des Oscillatoriaceae avec 27 génomes. L'inférence orthologique a été réalisée à chaque fois en mode inférence à 60 % de corepresence et puis en mode OG avec 58%, 98% et 100% de corepresence respectivement. La « corelist » contenait le nom des génomes. L'option « core » était activée et les options « specific » et « anvio » étaient désactivées à chaque fois. En mode inférence, les fichiers dans « infiles » ont été précisés (séquence d'ADN complète des génomes, en nucléotides). En mode OG, les infiles sont des séquences orthologiques, au format « protéines » et l'option « coreunwanted » a été indiqué à 3, pour les 3 génomes de *Gloeobacter*, considérées comme groupe externe. Lorsque cette option est utilisée, les noms des 3 génomes *Gloeobacter* ont été supprimés de la « corelist ». En phylogénétique, un groupe externe (ou extra-groupe ; en anglais : outgroup) est un ensemble d'organismes qui sert de groupe de référence pour déterminer les relations évolutives d'un groupe interne.

L'outil a également été utilisé pour la recherche de gènes spécifiques des 27 génomes de l'arbre des Oscillatoriaceae. Trois groupes ont été analysés sur l'ensemble des génomes de l'arbre (les souches du genre *Laspinema* (**G1**), les souches d'Antarctique (**G2**) et ULC722). L'analyse a été lancée en mode inférence à 60% de corepresence et en activant les options « core » et « specific » avec une « specificlist » (les noms des trois groupes d'analyses). L'option « anvio » était désactivée. Le fichier « corelist » contenait le nom des 24 génomes de la famille (les *Gloeobacter* ne sont pas mis dans la liste). Le type de génome a été indiqué (nucléotide) ainsi que le nombre de « coreunwanted » (les 3 génomes de *Gloeobacter*).

Les lignes de commandes pour ces analyses sont visibles en **annexe 1**.

3.3.7 Phylogeny

Les OG protéiques et nucléotidiques sont utilisés pour l'analyse phylogénomique avec *Phylogeny.nf*. Les OG d'acides aminés peuvent être alignés avec MUSCLE V3.8.31 (Edgar, 2004). Ce workflow implémente l'inférence phylogénomique à l'aide de BMGE V1.12 (Criscuolo & Gribaldo, 2010) pour la sélection de sites alignés sans ambiguïté, SCaFoS V1.25 (Roure et al., 2007) pour la concaténation de séquences OGs et RAxML V8.2.12 (Stamatakis, 2014) pour la construction d'arbres (Cornet et al., 2023).

L'outil *Phylogeny.nf* a été utilisé à trois reprises pour la réalisation d'arbres phylogénétiques à partir des OGs obtenus dans *Orthology.nf*. Il a servi pour la réalisation du premier arbre phylogénétique des cyanobactéries (avec 206 OGs), du second arbre phylogénétique des cyanobactéries avec 52 génomes (avec 674 OGs et en mode Jackniffe) et de l'arbre phylogénétique des Oscillatoriaceae (avec 801 OGs). À chaque fois, un répertoire contenait les fichiers OGs ; un fichier idm comprenait les ID des génomes avec leurs noms taxonomiques ; la terminaison des fichiers a été précisée (ext=fa), ainsi que le mode (« prot » pour les protéines). L'option « align » a été activée et l'option « jackk » a été activée une seule fois pour la phylogénie des 52 génomes cyanobactériens.

Les lignes de commandes pour la réalisation de ces arbres sont détaillées dans **l'annexe 1**.

3.3.8 ORPER

ORPER.nf est conçu pour contraindre une phylogénie d'ARN ribosomique à petite sous-unité (ARNr SSU) avec un squelette phylogénomique (Cornet, Ahn, et al., 2021). Cet outil produit d'abord un arbre phylogénomique basé sur des protéines ribosomiques concaténées, extraites de génomes publics, puis contraint la phylogénie plus large de

l'ARNr SSU à l'aide de cet arbre phylogénomique de référence. Cette contrainte multilocus est utilisée pour réduire l'imprécision des analyses monogéniques (Cornet, Ahn, et al., 2021). ORPER permet, sur base de la diversité des ARNr SSU, de localiser de nouvelles lignées sans génome séquencé ou d'identifier des génomes proches de souches pour lesquelles seules des séquences d'ARNr SSU sont disponibles (Cornet et al., 2023).

L'outil « ORPER » a été utilisé afin de réaliser un arbre phylogénétique des gènes ARNr 16S des Oscillatoriaceae en désignant les *Gloeobacter* comme groupe externe de l'arbre. Un fichier « sequencesbins.fasta » a été créé contenant les gènes ARNr 16S des 3 bins ULC d'Oscillatoriaceae (ULC096, ULC102 et ULC722). Ce fichier a été indiqué pour l'option « SSU ». Trois gènes d'ARNr 16S pour chaque bin d'origine cyanobactérienne ont été déterminés à l'aide du logiciel barnapp. Le niveau taxonomique a été précisé, ainsi que le groupe (family, Oscillatoriaceae). L'outgroup a été précisé avec le niveau taxonomique (Gloeobacteraceae, family). Les options « refgenbank » et « outgenbank » ont été activées. Les options « cdhit » et « drep » ont été désactivées.

La commande nextflow est décrite en **annexe 1**.

3.3.9 Métabolique fonctionnelle et Modélisation métabolique

Metabolic.nf est dédié à l'annotation des fonctions protéiques à l'aide de Mantis (Queirós et al., 2021) et à la modélisation métabolique des procaryotes à l'aide d'Anvi'o (Eren et al., 2015) avec la base de données KEGG comme référence (Kanehisa & Goto, 2000). Le mode fonctionnel effectue une caractérisation fonctionnelle des séquences protéiques obtenues à partir des OGs de OrthoFinder, tandis que le mode modélisation fournit une modélisation des voies KEGG, basée sur la présence d'au moins 60 % des gènes impliqués dans une voie, pour les génomes procaryotes. Les parcelles de présence/absence des voies KEGG sont ensuite représentées graphiquement avec ggplot2, selon une liste de génomes spécifiée par l'utilisateur (Cornet et al., 2023).

Les deux modes ont été utilisés dans mon cas d'étude. L'étude fonctionnelle a été réalisée sur les gènes spécifiques obtenus par OrthoFinder pour les trois groupes d'analyse de l'arbre des Oscillatoriaceae, en mode « fonctionnal ». Les fichiers faa ont été concaténés en un seul fichier pour chaque groupe d'étude et déplacés dans le répertoire « infile » L'analyse « modelling » a été réalisée sur 13 souches de l'arbre des Oscillatoriaceae (partie inférieure de l'arbre) en mode « modelling ». Dans le répertoire « infile » se trouvait les fichiers fna, et un fichier « list » a été créé avec les IDs des génomes.

Les commandes des nextflow sont écrits en **annexe 1**.

3.4 Genome2metabolite

Cet outil se présente comme tous les autres outils de GEN-ERA, mais n'est pas public. Il utilise le pipeline antiSMASH V3.0. (Weber et al., 2015) et un outil autonome pour l'identification et l'analyse génomique de gènes biosynthétiques. Il utilise l'algorithme ClusterFinder pour avoir une meilleure détection des groupes des gènes biosynthétiques de métabolites secondaires (BGC) appartenant à un grand nombre de classes connues (saccharide, acide gras et putatif) (Weber et al., 2015).

Ensuite, le logiciel Palantir, un outil d'analyse post-traitement de rapports antiSMASH, est utilisé pour gérer et affiner les données du cluster de BGC, en améliorant l'annotation BGC, en délimitant les modules et en favorisant un accès facile aux sous-séquences à différents niveaux (cluster, gène, module et domaine). De plus, il peut analyser les rapports antiSMASH fournis par l'utilisateur et les reformater pour une utilisation directe ou un stockage dans une base de données relationnelle (Meunier et al., 2020).

L'outil *genome2metabolite.nf* a été utilisé sur 13 souches de l'arbre des Oscillatoriaceae (partie inférieure de l'arbre), en utilisant les nucléotides des souches (fichiers fasta, précisé pour l'option « fastadir ») et en précisant le taxon (Bacteria).

La description du nextflow est visible en **annexe 1**.

4 Résultats

4.1 Analyse de l'assemblage et du taux de contamination des génomes

J'ai utilisé 10 souches de la collection BCCM/ULC (ULC002, ULC008, ULC029, ULC046, ULC096, ULC102, ULC128, ULC180, ULC307, ULC722) et une autre souche, ACSII155. Les souches ont été séquencées par Illumina et Nanopore sauf pour ACSII155, ULC002 et ULC029 qui ont été séquencées uniquement par Illumina (**Tableau 1**). D'après les informations retrouvées dans le site web officiel de BCCM (https://bccm.belspo.be/catalogues/catalogue-search?collection=ULC), les souches ont une taxonomie et des origines variées (**Tableau 1**). ULC002, ULC008, ULC180 sont du genre *Nostoc*. ULC096, ULC102 et ULC722 sont décrits comme faisant partie du genre *Phormidium.* ULC029 et ULC128 sont identifiés comme faisant partie des genres *Stenomitos* et *Microcoleus* respectivement. Les souches ACSII155, ULC307 et ULC722 ne sont pas disponibles publiquement et ACSII155 n'a pas encore reçu de numéro ULC. L'assemblage de ces génomes a été réalisé via le workflow *Assemby.nf.* Des bins (= résultat de l'assemblage de contigs ou de séquences métagénomiques en un génome individuel (Kyrgyzov et al., 2020) ont été obtenu via deux outils métagénomique : CONCOCT (Alneberg et al., 2014) et MetaBAT2 (Kang et al., 2019). Le **tableau 2** présente le nombre de bins obtenus pour chaque souche et pour chaque outil de binning.

Souches	Nombre de bins de CONCOCT	Nombre de bins MetaBAT2
ACSII155	21	9
ULC002	13	5
ULC008	13	6
ULC029	8	1
ULC046	9	6
ULC096	5	3
ULC102	8	3
ULC128	25	6
ULC180	5	2
ULC307	9	3
ULC722	11	3

Tableau 2 : *nombre des bins en fonction de CONCOCT et MetaBAT2 obtenu par Assembly.nf, pour chaque souche.*

Les bins obtenus sont ensuite analysés avec GTDB, une base de données de taxonomie des procaryotes (https://gtdb.ecogenomic.org/), pour déterminer leurs taxonomies. Ensuite, les bins cyanobactériens ont été analysés avec le workflow *GENContams.nf*, en utilisant les

outils Kraken (Wood et al., 2019), ChekM2 (Chklovski et al., 2022) et QUAST (Gurevich et al., 2013), afin de vérifier leur qualité.

Le pourcentage de séquences d'origine cyanobactérienne est assez faible pour l'ensemble des échantillons analysés (**Figure 6**). Sur les 11 souches analysées, le pourcentage de séquences associées aux cyanobactéries est de 19,25 % de moyenne (7,5% pour ACSII155, 5% et 13% pour ULC002, 17,9% pour ULC008, 31,1% pour ULC029, 15,8 % et 16,5% pour ULC045, 39,1% pour ULC096, 47,6% pour ULC10, 2, 0% pour ULC128, 10,8% et 21,4% pour ULC180, 14% et 3,4% pour ULC307, 22,4% pour ULC722). De plus, la présence d'autres microorganismes au sein des bins est assez importante et montre que nos bins ne contiennent pas uniquement des génomes de cyanobactéries. ULC128 ne contient pas de bin cyanobactérien mais uniquement ceux de protéobactéries.

En observant la **figure 6**, nous pouvons remarquer que dans 5 cas sur 11, les taux de complétudes et de contaminations sont identiques entre MetaBAT2 et CONCOCT. En effet, ULC008 a eu 99,9% de complétude et 0,39% de contamination ; ULC029 a eu 100% de complétude et 2,67% de contamination ; ULC096 a eu 100% de complétude et 0,32% de contamination ; ULC102 a eu 99,99% de complétude et 0,68% de contamination ; ULC722 a eu 99,77% de complétude et 0,54% de contamination. Dans 4 cas sur 11, MetaBAT2 est meilleur que CONCOCT avec un meilleur taux de complétude et un plus faible taux de contamination (99,99% de complétude et 0,34% de contamination pour ULC002, 77,69% de complétude et 1,51 % de contamination pour ULC046, 7,82% de complétude et 0,05% de contamination pour ULC128 et 71,92% de complétude et 2,5% de contamination pour ULC180). Pour ACSII155 et ULC307, CONCOCT a donné de meilleurs résultats avec 55,19% de complétude et 2,21% de contamination ; et 68,99% de complétude et 3,41% de contamination respectivement. Nous pouvons également observer que la souche ULC722 est la seule souche qui contient plusieurs génomes d'origine variée dont des Proteobacteria (30%), Verrucomicrobia (13,5%), Bacteroidota (18,6% selon CONCOCT et 10,7% selon MetaBAT2) et des Cyanobacteria (22,4%).

Les résultats obtenus par GTDB et les analyses *GENContams.nf* ont permis de réaliser un premier tri des bins. Six bins MetaBAT2 d'origine cyanobactérienne avec un taux de complétude supérieur à 70%, un taux de contamination inférieur à 10% pour Kraken et un taux de contamination inférieur à 10 % pour CheckM2 (Bowers et al., 2017) ont été sélectionnés pour la suite des analyses : ULC002, ULC008, ULC029, ULC096, ULC102, ULC722 (**Annexe 2**). La sélection des bins s'est faite comme décrit au pt 3.3.4 de Matériels

et méthodes. Si pour une même souche, il restait des bins de CONCOCT et MetaBAT2, je choisissais celui de MetaBAT2 car celui-ci est plus précis pour les génomes de procaryotes (Kang et al., 2019).

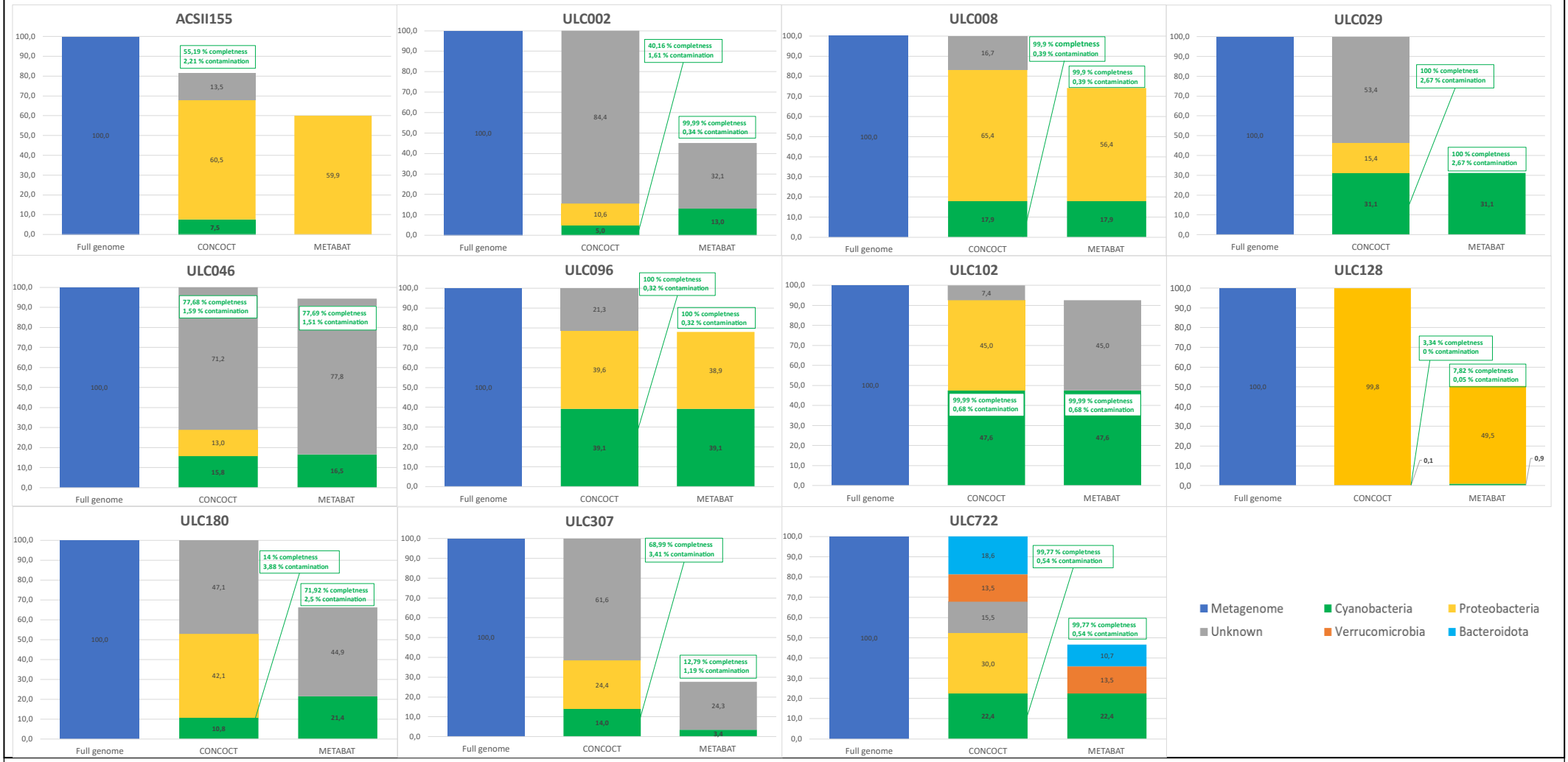


Figure 6 : Présentation des résultats des bins obtenus après l'analyse du Nextflow *GENContams.nf*. Pour chaque souche, les résultats de CONCOCT et MetaBAT2 sont présentés en pourcentage par rapport au génome entier (metagenome, en bleu foncé). Les proportions du type génomique des souches sont présentés en pourcentage pour les deux méthodes d'analyses : Cyanobacteria (vert), Proteobacteria (jaune), Verrucomicrobia (orange), Bacteroidota (bleu clair), Unknown (gris). Les pourcentages de complétude et de contamination sont aussi inscrits pour les bins cyanobactériens (ou verrucomicrobiens pour ULC128).

Afin des réaliser des phylogénies, des génomes cyanobactériens proches des bins devaient être sélectionnés. Cela a été réalisé en comparant mes bins a l'ensemble des génomes cyanobactériens disponibles publiquement, par Average Nucleotide Identity (ANI).

Pour réaliser l'analyse ANI sur mes bins sélectionnés, 3870 génomes de cyanobactéries présents dans les banques de données (GenBank et RefSeq) ont été téléchargés (via le Nextflow *Genome-downloader.nf*). À partir des génomes téléchargés et des 6 bins sélectionnés, l'analyse ANI a été lancée via le workflow *ANI.nf*. Après analyse, 27 génomes des banques de données, qui étaient les plus proches de mes bins : supérieur ou égal à 85 % de ANI (**Annexe 3**) ont été sélectionnés.

Ensuite, grâce aux résultats obtenus par GTDB sur mes bins et la recherche taxonomique sur NCBI des 27 génomes sélectionnés, les familles principales de mes 6 bins ont être déterminées : Oscillatoriaceae, Leptolyngbyaceae et Nostocaceae.

Ensuite, l'analyse CONTAMS a été relancée sur un plus grand nombre de génome, en tout 455 génomes. 422 génomes venaient des trois familles des cyanobactéries déterminées téléchargées dans les banques de données (RefSeq et GenBank), 27 génomes venaient des souches sélectionnées à l'étape d'ANI et 6 bins des souches ULC sélectionnés.

Lors de la sélection de ces génomes, des critères plus stricts que pour la sélection des bins ont été appliqués, 90 % de complétude et 5 % de contaminations, correspondant aux critères associés à des MAGs de haute qualité selon Bowers et al (2017). 246 génomes ont été conservés venant des analyses de ANI, de mes bins et des génomes des 3 familles de cyanobactéries. Le premier arbre phylogénétique a été réalisé avec ces 250 génomes (**Annexe 4**).

4.2 Le premier arbre phylogénétique

À partir des 250 génomes, des gènes core ont été générés via le workflow *Orthology.nf*, utilisant OrthoFinder (Emms & Kelly, 2019). Cet outil va comparer les gènes entre eux et déterminer lesquels sont orthologues. Le workflow a d'abord été utilisé en mode inférence à 60% de corepresence pour obtenir 63 OGs (= gènes orthologues de protéines). À partir de ces 63 OGs, le workflow a été relancé avec ces gènes en mode OG à 58 % de corepresence mais en excluant les 3 *Gloeobacter* de l'analyse. 206 core genes ont ainsi été obtenus et ont permis de déterminer le premier arbre phylogénétique **(Annexe 4).** Le fichier protML-final.tre obtenu a été ouvert sur Itol (Letunic & Bork, 2021) pour observer l'arbre phylogénétique. Ce premier arbre obtenu a permis d'avoir une vue d'ensemble d'où se situent mes 6 bins **(Annexe 4)**. Pour chacun des arbres réalisés, les génomes des 3 *Gloeobacter* ont été choisis comme groupe externe. Un premier tri des génomes a été réalisé et 52 génomes ont été sélectionnés. Pour la sélection des génomes, les génomes/souches qui étaient proches de mes bins ont été gardés. Les souches présentes sur une branche unique et ayant un bootstrap faible ont été supprimées.

À partir de ces 52 génomes, un nouvel arbre phylogénétique plus précis a été créé. 1200 coregenes ont été généré par l'analyse d'Orthology en mode inférence avec 60% de corepresence ; et puis 674 coregenes ont été obtenus en mode OG avec 98% de corepresence. À partir de ces 674 gènes, l'analyse Phylogeny a été lancée en ajoutant le mode Jackniffe **(Figure 7)**, afin d'avoir une corroboration de l'arbre avec celui produit en bootstrap. L'arbre phylogénomique obtenu a été complété en ajoutant des informations de taxonomies et d'habitats recherchés sur NCBI (https://www.ncbi.nlm.nih.gov/, Assembly) et GTDB (https://gtdb.ecogenomic.org/).

En observant la **figure 7** il a 6 types de biome présents pour 4 groupes de famille. Les habitats terrestres, d'eau douce, les milieux extrêmes et les symbioses sont présents dans les 3 familles des bins : Oscillatoriaceae, Leptolyngbyaceae et Nostocaceae. Les Oscillatoriaceae et Nostocaceae peuvent aussi vivre dans des milieux d'eaux salées (mer, lac salin, zone benthique, etc). Les *Gloeobacter* et une souche de *Nostoc* sont les seuls à avoir été décrits comme vivant dans les grottes et les rochers. Nous pouvons donc conclure que le type d'habitat des cyanobactéries étudiées ici est très diversifié allant des milieux aquatiques aux terrestres en passant par les milieux extrêmes (**Figure 7**).

La famille contenant le plus de souches est celle des Nostocaceae. Celle-ci contient deux bins : ULC008 et ULC002. Un bin est présent dans les Leptolyngbyaceae : ULC029. Les

Oscillatoriaceae contiennent 3 de 6 bins sélectionnés : ULC096, ULC102 et ULC722. Pour la suite des analyses, je me suis focalisée sur cette famille qui est un peu plus petite et contient le plus de mes bins (**Figure 7**).

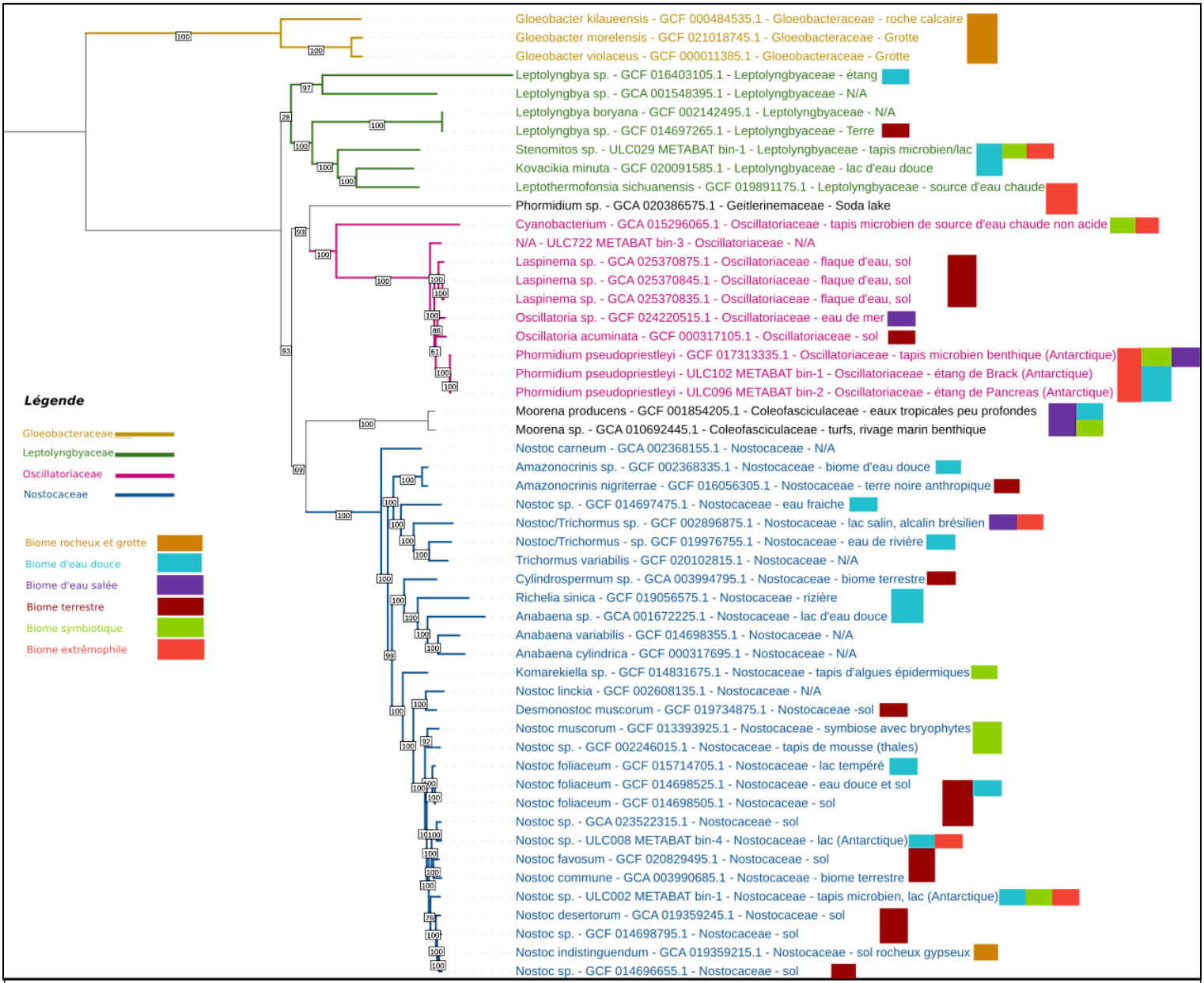
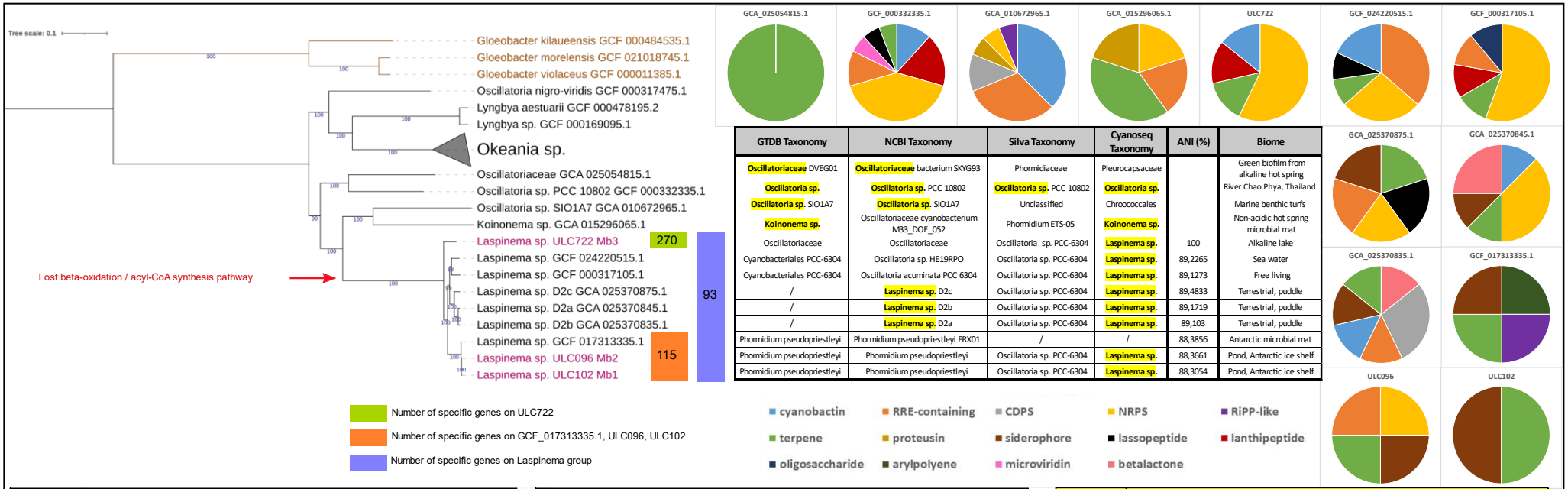


Figure 7 : Arbre phylogénétique obtenu en mode Jackniffe des 52 souches cyanobactériens comprenant mes 6 bins ULC. Le pourcentage de bootstrap est inscrit sur les branches de l'arbre. Les 4 familles majoritaires de l'arbre sont présentés : Gloeobacteraceae (brun-moutarde), Leptolyngbyaceae (vert foncé), Oscillatoriaceae (rose), Nostocaceae (bleu foncé). Ceux qui ne pouvaient être regroupés dans une de ces familles sont colorés en noir. Le nom taxonomique des souches a été déterminé via les banques de données NCBI et/ou GTDB. Le biome des souches, trouvés sur NCBI, est inscrit à côté des noms taxonomiques et a été regroupé en 6 types majeurs : le biome présent dans les roches et les grottes (brun clair) le biome d'eau douce (bleu clair), d'eau salée (mauve), le biome terrestre (brun foncé), le biome symbiotique (vert clair) et le biome extrémophile (rouge). Ces codes couleurs sont présentés à côté des dénominations de chaque souche, indiquant la diversité des habitats des souches.

4.3 L'arbre phylogénétique des Oscillatoriaceae

Nous avons remarqué que nos bins de souches ULC722, ULC102 et ULC96 étaient proches des 3 souches *Laspinema,* nouvellement décrites par Stanojkovic et al (2022) (**G3**). Nous avons donc décidé de comparer nos bins avec la souche de référence *Laspinema thermale* HK S5 (Heidairi et al., 2018), dont seul le gène d'ARNr 16S est décrit. ORPER (Cornet, Ahn, et al., 2021) a été utilisé sur les gènes ARNr 16S des trois bins ULC ainsi qu'avec la séquence de référence de *Laspinema thermale* HK S5 (**Annexe 5**), afin de comparer nos bins à la séquence de référence et afin de vérifier s'il ne manquait pas des génomes suite à la sélection réalisée par ANI. L'arbre obtenu par ORPER confirme que nos trois bins (ULC722, ULC102 et ULC96) sont proches du genre *Laspinema* (**Annexe 5**).

Avec cet arbre (**Annexe 5**) et le premier arbre phylogénétique (**Figure 7**), un croisement d'informations a été réalisé afin d'être plus représentatif sur le nombre de génomes associés aux souches de cette famille. 14 génomes déterminés par ORPER (**Annexe 5**) ont été ajoutés aux 7 génomes de la famille Oscillatoriaceae de l'arbre des cyanobactéries (**Figure 7**). À partir de cette sélection, les outils *Orthology.nf* et *Phylogeny.nf* ont été réutilisés sur les souches pour obtenir un arbre centré sur la famille des Oscillatoriaceae (**Figure 8**). Orthology a été utilisé en mode inférence avec 60% de corepresence sur 21 souches d'Oscillatoriaceae, les 3 bins et les 3 génomes de *Gloeobacter* (= outgroup). 1285 coregenes ont été générés et *Orthology.nf* a ensuite été utilisé en mode OG avec 100% de corepresence, afin de diminuer le nombre de gènes core. Ensuite, *Phylogeny.nf* a été lancé sur les 801 coregenes obtenus en mode OG. L'arbre d'Oscillatoriaceae obtenu via Phylogeny a été ouvert et analysé sur Itol où tous les Bootstrap sont à 100 sauf pour deux cas à 99% (entre deux groupes Oscillatoriaceae et deux groupes *Laspinema*) et un cas à 88% (entre deux espèces *Laspinema*, GCF_024220515.1 et GCF_000317105.1). En observant la **figure 7**, les 3 bins forment un groupe distinct par rapport aux autres souches du genre *Laspinema*. ULC722 semble être une espèce isolée aux autres souches du genre. ULC102, ULC096 et l'autre souche d'Antarctique (GCF_017313335.1) forment un groupe séparé des autres souches du genre *Laspinema*. Ce groupe est nommé groupe 2 (**G2**) pour la suite des explications.



GTDB Taxonomy	NCBI Taxonomy	Silva Taxonomy	Cyanoseq Taxonomy	ANI (%)	Biome
Oscillatoriaceae DVEG01	Oscillatoriaceae bacterium SKYG93	Phormidiaceae	Pleurocapsaceae		Green biofilm from alkaline hot spring
Oscillatoria sp.	Oscillatoria sp. PCC 10802	Oscillatoria sp. PCC 10802	Oscillatoria sp.		River Chao Phya, Thailand
Oscillatoria sp. SIO1A7	Oscillatoria sp. SIO1A7	Unclassified	Chroococcales		Marine benthic turfs
Koinonema sp.	Oscillatoriaceae cyanobacterium M33_DOE_052	Phormidium ETS-05	Koinonema sp.		Non-acidic hot spring microbial mat
Oscillatoriaceae	Oscillatoriaceae	Oscillatoria sp. PCC-6304	Laspinema sp.	100	Alkaline lake
Cyanobacteriales PCC-6304	Oscillatoria sp. HE19RPO	Oscillatoria sp. PCC-6304	Laspinema sp.	89,2265	Sea water
Cyanobacteriales PCC-6304	Oscillatoria acuminata PCC 6304	Oscillatoria sp. PCC-6304	Laspinema sp.	89,1273	Free living
/	Laspinema sp. D2c	Oscillatoria sp. PCC-6304	Laspinema sp.	89,4833	Terrestrial, puddle
/	Laspinema sp. D2b	Oscillatoria sp. PCC-6304	Laspinema sp.	89,1719	Terrestrial, puddle
/	Laspinema sp. D2a	Oscillatoria sp. PCC-6304	Laspinema sp.	89,103	Terrestrial, puddle
Phormidium pseudopriestleyi	Phormidium pseudopriestleyi FRX01	/	/	88,3856	Antarctic microbial mat
Phormidium pseudopriestleyi	Phormidium pseudopriestleyi	Oscillatoria sp. PCC-6304	Laspinema sp.	88,3661	Pond, Antarctic ice shelf
Phormidium pseudopriestleyi	Phormidium pseudopriestleyi	Oscillatoria sp. PCC-6304	Laspinema sp.	88,3054	Pond, Antarctic ice shelf

Name	Fonction/action
Cyanobactin	Ce sont des petits peptides cycliques produits par les cyanobactéries. Ces protéases participent aux modifications post-traductionnelles et ont des activités antipaludiques, antitumorales et d'inversion multimédicamenteuses.
Siderophore	Les sidérophores sont des chélateurs de fer synthétisés et sécrétés notamment par les micro-organismes pour leur permettre de puiser le fer essentiel à leur développement. Ces molécules ont une très haute affinité pour le fer ferrique Fe $^{3+}$.
Lassopeptide	Les lassopeptides sont un grand groupe de peptides macrocycliques. Ils appartiennentt à une nouvelle classe de produits naturels à structure très compacte et stable. Ils ont des activités antimicrobiennes, et sont des inhibitrices d'enzymes et des antagonistes de récepteurs.
RiPP-like	Les peptides synthétisés par le ribosome et modifiés après la traduction (RiPP) sont une classe de produits naturels en croissance rapide. Les peptides précurseurs de RiPP peuvent subir une adaptation enzymatique approfondie, produisant des produits structurellement et fonctionnellement divers, et leur logique de biosynthèse en fait des cibles de bio-ingénierie attrayantes. Les peptides synthétisés par voie ribosomique constituent une classe diversifiée de produits naturels d' origine ribosomique. Composés de plus de 20 sous-classes, les RiPP sont produits par une variété d' organismes, y compris les procaryotes, les eucaryotes et les archées, et ils possèdent un large éventail de fonctions biologiques.

Name	Fonction/action
Arylpolyene	Ce sont des pigments jaunes arylpolyène.
RRE-containing	C'est un domaine conservant un élément de reconnaissance du peptide précurseur RiPP(RRE). Le RRE se lie spécifiquement à un peptide précurseur et dirige les enzymes de modification post-traductionnelle vers leurs substrats.
Lanthipeptide	Ce sont des peptides synthétisés par les ribosomes et modifiés après la traduction (RiPP) qui présentent une grande variété d'activités biologiques, allant de l'antimicrobien à l'antiallodynique. Les lanthipeptides qui présentent une activité antimicrobienne sont appelés lantibiotiques.
CDPS	Les cyclodipeptides (CDP) représentent une famille diversifiée de petits peptides cycliques hautement stables. Ils sont répandus dans la nature et présentent une grande variété d'activités biologiques et pharmacologiques. Les synthases CDP (CDPS) et les synthétases peptidiques non ribosomales(NRPS) catalysent la biosynthèse de la structure centrale du CDP. Les CDP et leurs dérivés présentent un large éventail d'activités biologiques, telles des propriétés antibactériennes, antimicrobiennes, anticancéreuses et antiradicalaires. Ils ont également été développés pour transporter des molécules biologiquement actives à travers la barrière hémato-encéphalique.
Oligosaccharide	Les oligosaccharides sont des oligomères formés d'un nombre n d'oses (monosaccharides) par liaison glycosidique alpha ou bêta. Dans les cellules animales, les oligosaccharides des glycolipides et glycoprotéines membranaires sont situés du côté extracellulaire. Ils peuvent aussi avoir un rôle de réserve de sucre chez les végétaux (l'amidon).

Name	Fonction/action
Betalactone	Les β-lactones sont des hétérocycles et possèdent de puissantes propriétés médicinales. Les lactones, en général) contribuent à l'arôme de produits alimentaires naturels ou transformés, très divers (fruits, produits laitiers, etc.).
Microviridin	Les microviridines sont une classe d' inhibiteurs de sérine protéase produits par divers genres de cyanobactéries. Les microviridines font partie de la famille de produits naturels RiPP. Ce sont aussi des antifongiques.
Proteusin	Les proteusines sont une famille de peptides ribosomiques bactériens. Les seuls membres connus sont les cytotoxines de type polythéonamide. Les cyanobactéries contiennent un grand nombre de loci de proteusine putatifs. Cette diversité génétique offre un potentiel de découverte chimique et pharmacologique.
Terpene	Les terpènes sont une classe d'hydrocarbures, produits par de nombreuses plantes, des métabolites secondaires qui se rencontrent également chez les métazoaires. Certains terpénoïdes sont essentiels au métabolisme primaire, notamment la photosynthèse (chlorophylle, caroténoïdes et plastoquinone), la respiration (ubiquinone), la régulation du développement des plantes (gibbérellines et acide abscissique), etc. Les terpénoïdes ont une variété d'applications commerciales, y compris les produits pharmaceutiques, les nutraceutiques, les arômes/parfums et les produits chimiques industriels.
NRPS	Les synthétases de peptides non ribosomiques (NRPS) sont de grandes machineries multienzymatiques qui assemblent de nombreux peptides avec une grande diversité structurelle et fonctionnelle. Ces peptides comprennent plus de 20 médicaments commercialisés, tels que des antibactériens (pénicilline, vancomycine), des composés antitumoraux (bléomycine) et des immunosuppresseurs (cyclosporine).

Figure 8 : Etude phylogénétique, fonctionnelle et métabolique des génomes des souches sélectionnées de la famille Oscillatoriaceae. L'arbre phylogénétique de la famille, en haut à gauche, 27 souches, dont 8 ont été regroupés en *Okeania sp.* Les bins ULC sont présentés en rose et le groupe externe, les *Gloeobacter,* sont colorés en brun clair dans l'arbre. Le pourcentage de bootstrap est inscrit sur chaque branche. Pour chaque souche, le nom taxonomique et l'ID est inscrit afin de les identifier. À côté de l'arbre, un tableau taxonomique est présenté selon 4 banques de données : GTDB, NCBI, SILVA, CyanoSeq. Le choix du nom taxonomique est surlignés en jaune fluo pour chaque souche. Dans ce tableau, le pourcentage d'ANI du groupe *Laspinema* est présenté par rapport à ULC722, ainsi que le biome de chaque souche, trouvé sur NCBI. Le nombre de gène spécifique, déterminé par Orthology – OrthoFinder, pour 3 groupes est présentés : le groupe ULC722 (vert clair), le groupe des souches d'Antarctique (**G2**) (orange) et tout le groupe *Laspinema* (**G1**) (bleu). Ces gènes spécifiques ont servi à l'étude fonctionnelle présentée en **Annexe 8**. La flèche rouge indique que tout le groupe *Laspinema* (**G1**) a perdu une voie métabolique, celle de la béta-oxydation et de l'acyl-CoA synthèse. Les métabolites secondaires pour les 13 souches sélectionnées, en-dessous du groupe de *Okeania sp.,* obtenus via Genome2Metabolite, sont présentés sont formes de piechart et listés avec une description de fonction sous forme de tableau (en bas de la figure). Un code couleur a été respecté pour les deux formes de présentation : cyanobactine en bleu clair, sidérophore en brun, lassopeptide en noir, RiPP-like en mauve, arylpolyène en vert foncé, RRE-containing en orange, lanthipeptide en rouge, CDPS en gris, Oligosaccharide en bleu foncé, bétalactone en « pêche », microviridine en rose, proteusine en « moutarde », terpène en vert clair, NRPS en jaune.

L'analyse ANI a été réalisée sur les souches de l'arbre des Oscillatoriaceae. Sur la **figure 7**, nous avons reporté les valeurs d'ANI où ULC722 est face aux autres souches de *Laspinema*. En observant les résultats obtenus (**Figure 9**), nous pouvons remarquer qu'il y a un groupe (de ULC722 à ULC102) faisant partie du même genre, *Laspinema*, (à plus de 83 % d'ANI) qui est présenté. Ce groupe est désigné comme le groupe 1 (**G1**). Au sein de ce groupe (**G1**), ULC722 semble être une espèce isolée des autres souches, puisque le pourcentage d'ANI avec les autres souches du groupe est inférieur à 95 % et qu'il se trouve sur une branche isolée de l'arbre. L'analyse d'ANI indique aussi que les bins ULC096 et ULC102 sont très proches de la souche GCF_017313335.1, à plus de 95% d'ANI, formant un groupe à part entière (**G2**). GCF_024440515.1 et GCF_000317105.1 semblent également être deux espèces différentes du genre *Laspinema*. Le groupe de *Laspinema* de Stanojkovic et al (GCA_025370875.1, GCA_025370845.1 et GCA_025370835.1), dénommé groupe 3 (**G3**), semble être de la même espèce, à plus de 95% d'ANI, mais bien séparé des autres souches du groupe **G1**. Les quatre autres souches d'Oscillatoriaceae (GCA_025054815.1, GCF_000332335.1, GCA_010672965.1, GCA_015296065.1) sont chacune des espèces uniques différentes, car elles n'ont pas de valeurs significatives d'ANI

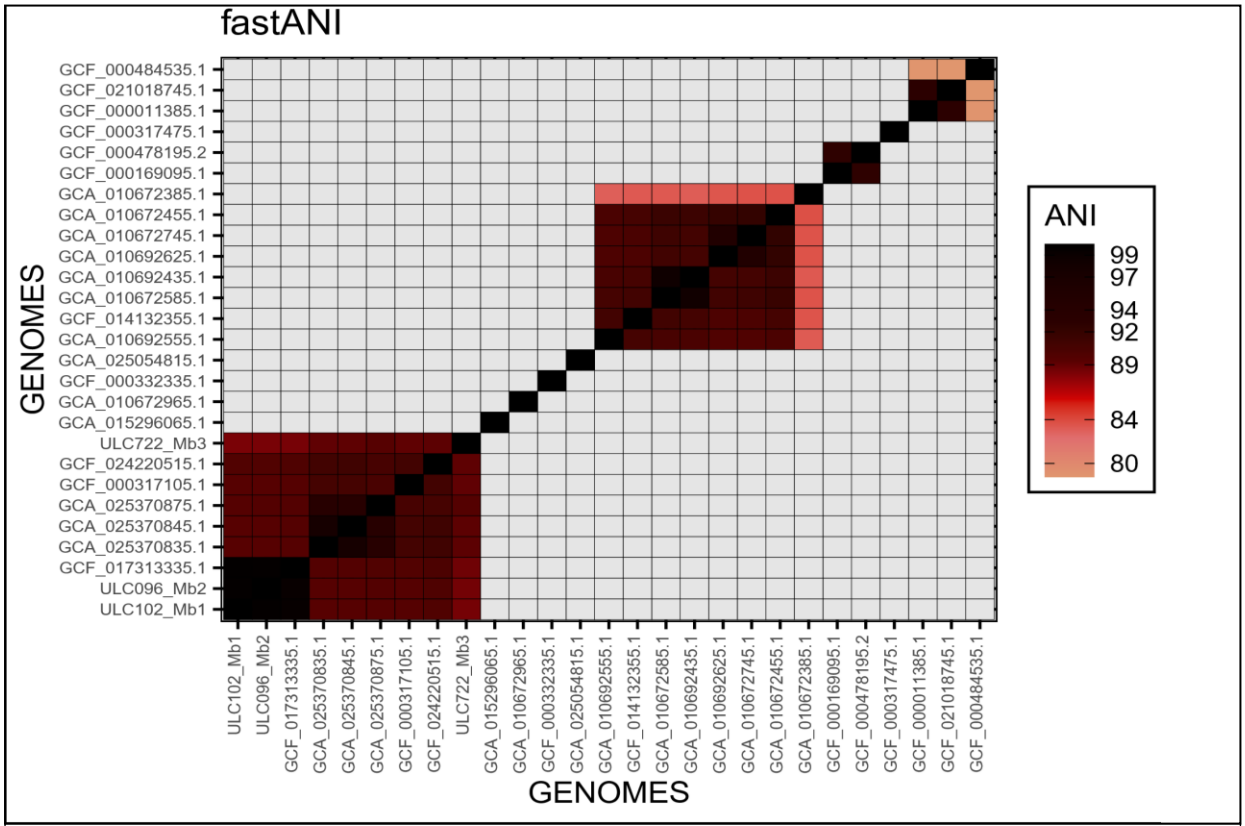


Figure 9 : Résultats de l'analyse ANI sur le groupe des Oscillatoriaceae, sous forme de matrice. Le pourcentage d'ANI d'une souche par rapport aux autres est présenté selon un code couleur qui s'assombrit. Si la case est claire, le pourcentage d'ANI est faible, proche de 80% ; si la case est foncée, noire, la pourcentage d'ANI est proche de 100%. Les cases en gris indiquent un résultat d'ANI inférieure à l'échelle présentée.

(cases grises de la **figure 9**) avec les autres souches de l'arbre. Ces résultats sont également visibles dans l'arbre phylogénétique (**Figure 8**).

Les résultats obtenus par GGDC sur l'outil de DSMZ (TYGS), basé sur l'hybridation ADN-ADN sont présentés en **Annexe 6**. L'outil n'a pu nommer les 9 souches du genre *Laspinema* (**G1**) et les considère toutes comme de nouvelles espèces. Les seules souches qui partagent le taux le plus élevé de DDH et semblent faire partie de la même espèce sont : ULC096, ULC102 et GCF_ 017313335.1 (97,3%, 97,1% et 96, 1%), ainsi que GCA_025370835.1 et GCA_025370845.1 (77,1%) (**Annexe 6**). Les autres souches ont un résultat de DDH inférieur à 70% et ne font pas partie de la même espèce, selon Meier-Kolthoff et al. (2013).

4.4 Taxonomie et biomes

La détermination taxonomique n'est pas chose aisée. Le choix taxonomique de l'ensemble des souches a été réalisé avec l'aide de Marcelo Vaz et la littérature associée et est décrit dans le tableau en **Annexe 7**.

Plusieurs taxonomies des cyanobactéries cohabitent actuellement, notamment grâce à 4 banques de données : GTDB (https://gtdb.ecogenomic.org/, (Waite et al., 2017) , NCBI (https://www.ncbi.nlm.nih.gov/, (Sayers et al., 2022), SILVA (https://www.arb-silva.de/aligner/, (Pruesse et al., 2012) et CyanoSeq (https://zenodo.org/record/7110927, (Lefler et al., 2023)). Une comparaison de ces 4 banques de données a été réalisée sur le genre *Laspinema* récemment défini. Les résultats montrent une discordance. GTDB donne des résultats pour 10 souches sur 13. NCBI a pu associer un nom taxonomique à toutes les souches ; de même pour SILVA et CyanoSeq, sauf pour GCF_017313335.1 qui ne possède pas de gènes codant pour l'ARNr 16S. SILVA et CyanoSeq sont des banques de données basées sur le gène ARNr 16S. Nous pouvons également observer sur la **figure 7** que pour GCA_025054815.1, les 4 banques de données n'ont pu que déterminer le nom de famille (Oscillatoriaceae pour le génome, Phormidiaceae et Pleurocapsaceae pour la séquence ARNr 16S) et pas le nom de genre ou d'espèce. Pour une seule souche (GCF_000332335.1), les 4 banques de données se sont accordées pour la nommer *Oscillatoria sp.* Pour la souche GCA_010672965.1, GTDB et NCBI se sont accordés pour le nom taxonomique *(Oscillatoria sp.* SIO1A7), là où SILVA n'a pu la classifier et CyanoSeq n'a trouvé que le nom d'ordre (Chroococcales). La souche GCA_015296065.1 est considérée comme faisant partie du genre *Koinonema* par GTDB et CyanoSeq. Nous avons remarqué qu'un clade (**G1**, de ULC722 à ULC096) correspondrait au nouveau genre *Laspinema* décrit par Heidari et al., en 2018. Le logiciel CyanoSeq a confirmé cette hypothèse. Les 3

bins ULC722, ULC096, ULC102 font partie de ce clade et sont considérés comme des *Laspinema*. Cette hypothèse peut être également observable dans l'arbre phylogénétique réalisé par ORPER (**Annexe 5**) où les 3 bins sont proches de la souche de référence *Laspinema thermale* HK S5 (Heidari et al., 2018). Pour les souches de l'embranchement *Laspinema* (**G1**), GTDB et NCBI ont donné des résultats variés, allant parfois uniquement de l'ordre au genre : Cyanobacteriales, Oscillatoriaceae, *Oscillatoria sp.*, *Oscillatoria acuminata*, *Phormidium pseudopriestleyi*. NCBI a également identifié les 3 souches *Laspinema* de Stanojkovic et al (**G3**). SILVA s'est référé à la souche *Oscillatoria sp.* PCC-6304 et a nommé toutes les souches du groupe ainsi.

Le biome des souches a été recherché dans le NCBI, plus précisément dans les métadonnées associées à la partie « bioproject ». Les types de milieux sont assez diversifiés : de milieu marin ou terrestre, de source chaude, de biofilm, etc. Les 3 souches *Laspinema sp.* de Stanojkovic et al (**G3**) viennent de flaques d'eau (milieu terrestre) en République tchèque. ULC722 est une souche isolée d'un lac salin de la région du Pantanal au Brésil. Les 2 autres bins de la collection BCCM/ULC viennent d'Antarctique : ULC096 et ULC102. Ils sont proches d'une autre souche originaire d'Antarctique GCF_017313335. Les 3 souches viennent de la plateforme de glace et des vallées sèches de Mc Murdo ; elles sont très proches géographiquement et écologiquement (Lumian et al., 2021). Nous pouvons voir dans l'arbre phylogénétique que ces trois souches forment un groupe isolé des autres (**G2**) (**Figure 8**).

4.5 Analyse des gènes fonctionnels

Afin de déterminer la fonction des gènes de mes souches, il faut d'abord leur rechercher des gènes spécifiques, des gènes qui leur sont propres et codant pour une protéine particulière.

Pour la recherche de gène fonctionnel sur mes souches, l'outil « Orthology » a été utilisé avec le logiciel OrthoFinder. L'outil réalise une comparaison des gènes afin de déterminer des gènes uniques sur un ensemble de génomes. L'analyse a été lancée trois fois sur trois groupes de génomes, par rapport à l'ensemble des 27 génomes de l'arbre des Oscillatoriaceae :

- Uniquement sur ULC722 ;
- Sur les 3 souches d'Antarctique (**G2**) : ULC096, ULC102, GCF_017313335 ;
- Sur les 9 souches de *Laspinema* (**G1**).

ULC722 a donné 270 gènes spécifiques, les 3 souches d'Antarctique (**G2**) en ont 115 et tout le groupe de *Laspinema* (**G1**) a 93 gènes spécifiques en commun (**Figure 8**).

Ces gènes spécifiques obtenus par OrthoFinder ont ensuite été analysés avec l'outil « Metabolic » en mode fonctionnel (Mantis) afin d'identifier leurs fonctions.

Aucune fonction spécifique des gènes spécifiques n'a été trouvée pour les souches d'Antarctique (**G2**). Les gènes étaient soit non classifiés (89), soit inconnus (26). Le groupe *Laspinema*, **G1**, a également 41 gènes non classifiés et 52 gènes inconnus. Cependant, un de gène spécifique a été retrouvé pour tout le groupe *Laspinema* **(G1).** C'est la protéine de liaison au substrat de Msme (= protéine de liaison et de transport du sucre (Sutcliffe et al., 1993)). Seul ULC722 a des gènes qui lui sont spécifiques avec des fonctions variées (**Annexe 8**). Il possède 61 gènes spécifiques de fonction connue, 26 gènes inconnus et 183 gènes non classifiés.

4.6 Analyse des gènes métaboliques

4.6.1 Metabolic Modelling

Ensuite, il était intéressant de rechercher les voies métaboliques et les métabolites secondaires spécifiques aux souches d'Antarctique du genre *Laspinema*, afin de comprendre leur spécificité et leur résistance au froid ou aux UVs.

La recherche de voies métabolique s'est faite avec l'outil « Metabolic » en mode modelling. Différentes voies métaboliques ont été analysées : le métabolisme des acides aminés, la biosynthèse des terpénoïdes et des polykétides, le métabolisme des glucides, le métabolisme énergétique, le métabolisme des lipides, le métabolisme des cofacteurs et des vitamines, le métabolisme des nucléotides et diverses voies des cyanobactéries tels que la photosynthèse oxygénique, l'assimilation des nitrates, l'assimilation de souffre et sulfate. Les résultats de « Metabolic » n'ont pas montré de données interpellantes sauf pour le métabolisme des lipides. Tout le groupe de *Laspinema* (**G1**) a perdu la voie de béta-oxydation (**Figure 8 et Figure 10**).

4.6.2 Métabolite secondaire

La recherche de métabolites secondaires a été réalisée via le logiciel *Genome2Metabolite.nf*. Les résultats obtenus donnent des informations intéressantes sur les métabolites secondaires et sont assez variés (**Figure 8**). Pour les 13 souches d'Oscillatoriaceae, 14 métabolites secondaires ont été recensés. Ils sont tous décrits dans

le tableau de la **figure 8**. Il y a des oligosaccharides, des arômes (béta-lactone), des proteusines, des terpènes, des pigments (arylpolyène), des chélateurs de fer (sidérophores), des peptides d'origines ribosomiques (= RiPP-like ; cyanobactine, microviridine, lanthipeptide, lassopeptide), des domaines RRE-containing, des synthases cyclodipeptides (CDPS) et des synthétases peptidiques non ribosomiques (NRPS).

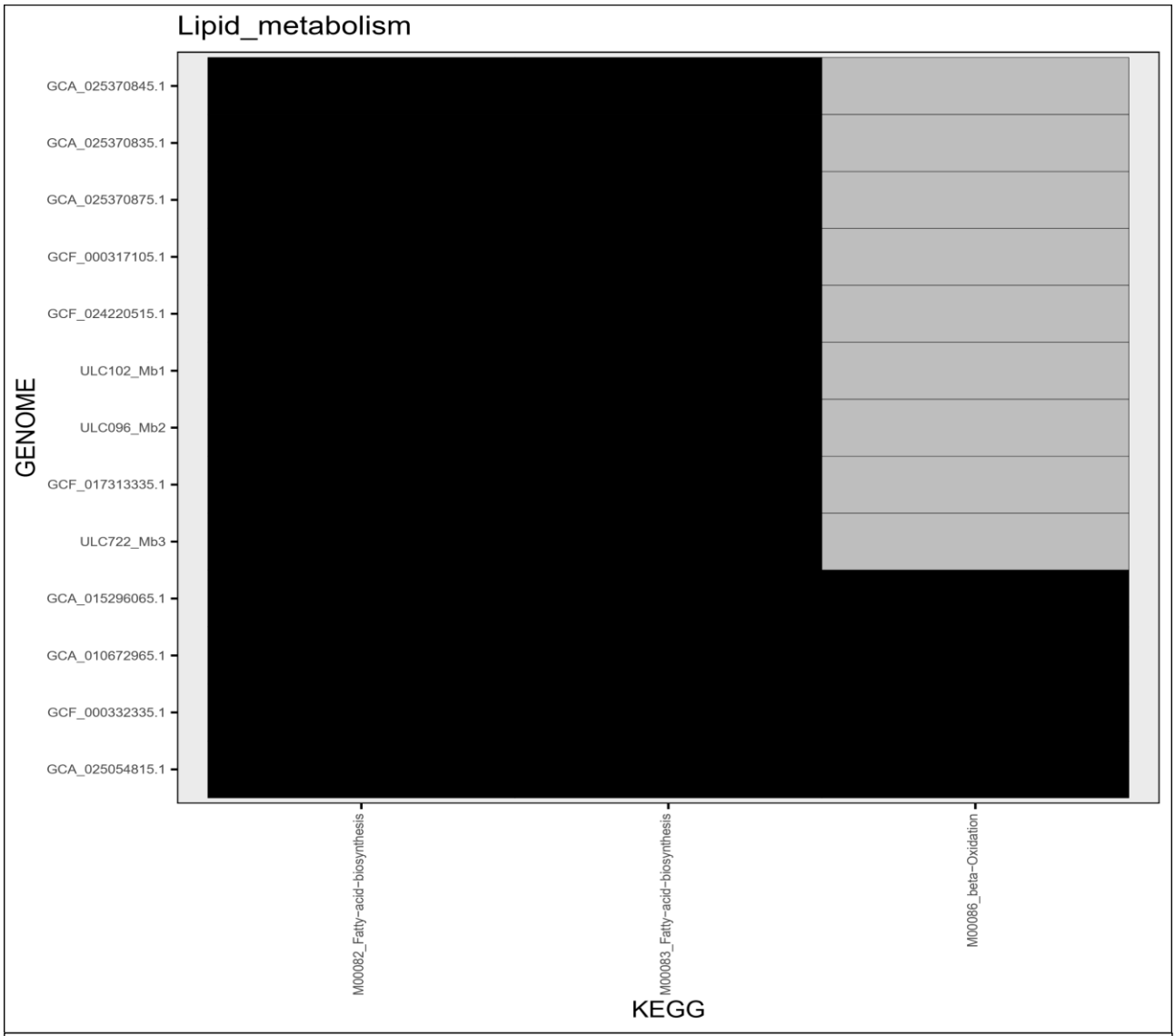


Figure 10 : résultat de l'analyse Metabolic Modelling sur les souches d'Oscillatoriaceae, sous forme de matrice. Les résultats présentés concernent ceux pour le métabolisme des lipides, pour trois voies métaboliques présentes dans KEGG. Une case en noire indique que la souche possède cette voie métabolique. Une case en gris indique que la souche ne possède pas cette voie métabolique.

Nous pouvons remarquer que le génome de la souche GCA_025054815.1 ne possèderait que des terpènes comme métabolites secondaires. Une seule souche, GCF_000332335.1 (*Oscillatoria sp.* PCC 10802 = *Oscillatoria princeps* NIVA CYA-150), possède 7 métabolites secondaires différents : dont principalement des NRPS, des lanthipeptdes, des cyanobactines, des RRE-containing, des microviridines, des lassopeptides et des terpènes.

Deux souches (GCA_010672965.1 et GCA_025370835.1) ont 6 métabolites secondaires différents, quatre souches (GCF_024220515.1, GCF_000317105.1, GCA_025370875.1 et GCA_025370845.1) en ont 5 et quatre autres (GCA_015296065.1, ULC722, GCF_017313335.1 et ULC096) en ont 4. La souche ULC722 possède différents peptides ribosomiques et non ribosomiques en plus des terpènes : cyanobactine, lanthipeptide, et NRPS. ULC102 est la seule souche à n'avoir que deux métabolites secondaires. Les 3 souches d'Antarctique (GCF_017313335, ULC096 et ULC102) (**G2**) ont peu de métabolites secondaires, avec une moyenne de 3,33 pour le groupe. Les trois souches (**G2**) ont en commun la présence de terpènes et de sidérophores. Les terpènes sont le groupe de métabolite secondaire présent dans tout l'embranchement sélectionné, sauf pour *Oscillatoria sp.* SIO1A7.

5 Discussion

Les premiers résultats obtenus concernent la pureté des souches de BCCM/ULC (**Figure 6**). Les résultats indiquent que celles-ci ne sont pas pures et qu'elles sont en présence de séquences autre que d'origine cyanobactérienne (Proteobacteria, Bacteroidota, Verrucomicrobia, ou d'origines inconnues). Ces résultats confirment que les souches étudiées n'étaient pas axéniques.

Le second résultat concerne la qualité des bins produits. Le taux de contamination varie de 0% à maximum 3,88% (dans le cas de ULC180) (**Figure 6**). Les bins ayant un taux de contamination inférieur à 1% sont ULC008, ULC096, ULC102, ULC128, ULC 722 pour CONCOCT et MetaBAT2 et ULC002 pour MetaBAT2 uniquement. ULC128 a un taux de contamination de 0% pour CONCOCT et 0,05 % pour MetaBAT2. Cependant, il ne contient aucune séquence de cyanobactéries, mais majoritairement de Proteobacteria. Cette souche a directement été éliminée pour la suite des analyses. En ce qui concerne le taux de complétude, il varie de 3,34 % (ULC128) à 100% (**Figure 6**). Les souches ayant un taux supérieur à 99% sont ULC008, ULC029, ULC096, ULC102, ULC722 pour CONCOCT et MetaBAT2, et ULC002 uniquement pour MetaBAT2. Finalement, les souches ayant la meilleure qualité de génomes cyanobactériens sont ULC002, ULC008, ULC096, ULC102 et ULC722 qui ont été sélectionné pour la suite des analyses. ULC029 a également été sélectionné parce qu'il a 100% de complétude et c'est la seule souche où MetaBAT2 a détecté uniquement du génome cyanobactérien.

La recherche de contamination redondante est possible via CheckM2 (Cornet & Baurain, 2022). CheckM2 est un outil basé sur l'apprentissage automatique pour prédire la qualité des isolats métagénomiques. CheckM2 construit des modèles adaptés pour prédire l'exhaustivité et la contamination des génomes bactériens et archéens sans tenir compte explicitement des informations taxonomiques (Chklovski et al., 2022). CheckM2 utilise le placement phylogénétique pour sélectionner des ensembles de marqueurs génétiques spécifiques. La contamination et la complétude sont estimées à l'aide de marqueurs spécifiques basés sur ce placement. Il utilise la « machine learning » pour prédire les contaminations et rend compte de la contamination intra-espèce lorsque l'identité en acides aminés entre deux marqueurs redondants est présente (Cornet & Baurain, 2022). CheckM2 fonctionne à la fois pour les contaminations redondantes et non-redondantes. Cependant, il a été montré CheckM2 sous-estime le taux de contamination dans plus de 97 % des cas pour les événements de remplacement uniques, alors qu'il n'a jamais sous-

détecté le type redondant (Cornet et al., 2022). CheckM2 ne fonctionne que sur les génomes procaryotes (Cornet & Baurain, 2022).

La recherche de contamination non-redondante est possible via Kraken (Cornet & Baurain, 2022). Kraken est un système de classification taxonomique rapide des données de séquence métagénomique (Wood et al., 2019). L'objectif premier de Kraken est de classer les reads dans les études métagénomiques. Néanmoins, il est capable de détecter les contaminants. Il construit sa base de données à partir des génomes et les divise en k-mers. Dans GEN-ERA, la base de données est issues de NCBI. Ces k-mers sont ensuite cartographiés sur les nœuds d'un arbre phylogénétique : plus ils sont largement partagés par plusieurs organismes, plus ils sont cartographiés profondément dans l'arbre. Cette méthode permet de couvrir l'ensemble du génome entier (Wood & Salzberg, 2014).

Dans mon cas d'étude, les contaminations génomiques m'obligent à supprimer des génomes qui ne sont pas assez « pures » et donc engendrent la possibilité de perdre des données utiles. Après analyse, 5 souches ont été supprimées : ACSII155, ULC046, ULC128, ULC180 et ULC307. L'analyse de qualité des génomes a permis la sélection de 6 souches de meilleure qualité (**Annexe 2**). Les résultats de QUAST sont présentés pour les 6 bins sélectionnés en **Annexe 2,** confirmant la sélection des meilleurs génomes. Le nombre de contigs est inférieur à 5 et ils ont un pourcentage de GC proche de 50% sauf pour ULC002 et ULC008 (41,46% et 41,78% respectivement).

Les résultats d'analyse ANI (**Annexe 3**) des 6 bins avec les génomes cyanobactériens ont permis de sélectionner 250 génomes pour la construction du premier arbre phylogénétique des cyanobactéries (**Annexe 4**). Ce premier arbre a permis de placer les souches et de déterminer à quelles familles elles appartiennent. La plus grande famille est celle des Nostocaceae. Les *Gloeobacter* ont été désignés comme groupe externe pour chaque arbre réalisé. Le rôle du groupe externe est de servir de point de comparaison pour le groupe interne et permet ainsi de créer un arbre phylogénétique enraciné (Farris, 1982). Les *Gloeobacter* sont un genre reconnu à la base de l'arbre phylogénétique des cyanobactéries (Nakamura et al., 2003 ; Ochoa de Alda et al., 2014 ; Ponce-Toledo et al., 2017 ; Sánchez-Baracaldo et al., 2017 ; Shih et al., 2013). De plus, certaines souches n'ont pas de nom d'espèce, mais uniquement le genre : *Nostoc sp., Anabaena sp., Trichormus sp., Leptolyngbya sp.* Ces informations indiquent qu'il y a encore un grand nombre de souche qui n'ont pas été identifiée avec précision dans les bases de données, pouvant poser problème lors de la recherche taxonomique pour de nouvelles souches.

Un second arbre des cyanobactéries a été réalisé pour être plus précis (**Figure 7).** Les 6 bins sont réparties au sein de 3 familles : Oscillatoriaceae (ULC096, ULC102 et ULC722) Leptolyngbyaceae (ULC029) et Nostocaceae (ULC002 et ULC008). D'après les recherches de Pessi et al (2019), les premiers organismes colonisateurs des milieux polaires seraient les cyanobactéries filamenteuses, telles que les *Leptolyngbya* et *Phormidium*. Celles-ci domineraient largement ces milieux actuellement. Cette explication justifierait pourquoi sur nos 6 bins de la collection, quatre font partie des Leptolyngbyaceae et Oscillatoriaceae (**Figure 7**). Ces microorganismes pionniers modifieraient le milieu en augmentant les niveaux de nutriments, la stabilité du sol et la protection contre les dommages physiques, permettant ensuite l'établissement de souches cosmopolites (Pessi et al., 2019).

Sur cet arbre, en plus d'avoir une meilleure disposition de nos souches, des informations du biome ont été indiquées. Nous pouvons observer que le type d'habitat des cyanobactéries est généralement très diversifié allant des milieux aquatiques aux terrestres ainsi que des milieux extrêmes (**Figure 7**) comme décrit dans différents articles (Chen et al., 2020 ; Dextro et al., 2023 ; Dvorák et al., 2017 ; Hentschke & Junior, 2022). Cependant, certaines souches n'ont pas d'information sur leur habitat. Il reste encore beaucoup de génomes pour lesquels il y a très peu d'informations décrites sur NCBI. Ce genre de situation peut poser problème pour des études tels que celle-ci où l'habitat peut donner des informations significatives sur nos souches cyanobactériennes, comme par exemple la résistance au froid ou à la salinité, etc. De plus, les informations concernant le milieu d'origine et le biome ne sont pas standardisées sur NCBI pour faciliter les comparaisons. Une classification de ce genre d'informations sous forme d'entonnoir pourrait grandement aider les chercheurs dans le traitement de données importantes et croissantes.

Ensuite, nous nous sommes focalisés sur l'étude d'une seule famille, celle qui contient le plus de souches de la collection BCCM/ULC, les Oscillatoriaceae. L'arbre phylogénétique des Oscillatoriaceae sur 27 génomes a permis de se rendre compte que les 3 bins ULC font partie d'un même clade, celui du genre *Laspinema* (dénommé groupe 1 (**G1**)). L'organisation de notre arbre phylogénétique (**Figure 8**) concorde avec ceux de Nadeau et al. (2001) et de Stanojkovic et al. (2022). L'arbre phylogénétique de Nadeau et al. (2001) indique la proximité des bins ULC096 et ULC102 (dénommé respectivement *Oscillatoria priestleyi* CMEE5020 Ant-Pancreas et *Phormidium autumnale* CMEE5034 Ant-Brack 2 dans son papier) avec *Oscillatoria accuminata* PCC6304. Stanojkovic et al (2022) confirment la proximité de la souche *Oscillatoria acuminata* PCC 6304 avec les 3 souches

Laspinema sp. D2a, D2b, D2c (**G3**). Selon lui, *Oscillatoria acuminata* PCC 6304 doit être considéré comme une espèce du genre *Laspinema*.

L'analyse ANI a été réalisée sur 13 souches des Oscillatoriaceae dont les 3 bins ULC. L'analyse des résultats d'ANI (**Figure 9**), nous a permis de nous rendre compte qu'il a 5 groupes d'espèces distincts au sein de l'embranchement des *Laspinema*. Les trois souches de l'Antarctique (**G2**) semblent être de la même espèce avec plus 95 % d'ANI, la limite intra-espèce (Jain et al., 2018). La souche ULC722, d'après les résultats d'ANI et de l'arbre phylogénétique, semble être une espèce bien isolée de l'ensemble de l'embranchement. Nous avons donc supposé que ces 5 groupes du genre *Laspinema* sont des nouvelles espèces. Pour confirmer nos hypothèses, nous avons envoyés les génomes de cet embranchement à TYGS, un serveur qui permet la classification taxonomique basée sur le dDDH (Meier-Kolthoff et al., 2022). Le rapport TYGS (**Annexe 6**) obtenu confirme que ces 9 souches sont bien des nouvelles espèces. La limite inter-espèce selon Meier-Kolthoff et al. (2013) est de 70%. GCF_017313335, ULC096 et ULC102 (**G2**) sont très proches d'après les résultats GGDC : 97,3 %, 97,1% et 96,1% et semblerait être de la même espèce comme nous l'avons supposé d'après les résultats d'ANI.

Depuis les années 1970, la délimitation des espèces chez les procaryotes était axée sur l'hybridation ADN-ADN (DDH) (Meier-Kolthoff et al., 2013). Cette méthode était la norme universelle offrant une comparaison à l'échelle du génome pendant plusieurs dizaines d'années, mais a été remplacée par l'analyse ANI parce que le DDH est une méthode trop grossière (c'est-à-dire que l'erreur expérimentale est trop élevée) et peut être influencé par des différences de tailles de génomes trop importantes entre les souches (Goris et al., 2007). L'analyse ANI est utilisée actuellement comme méthode de référence pour la comparaison de génome et la délimitation des espèces. Cependant, l'arrivée du DDH numérique (dDDH) via TYGS (DSMZ) semble surpasser ANI et peut être utilisée pour la délimitation des taxons au niveau sous-spécifique (Meier-Kolthoff et al., 2013 ; Meier-Kolthoff & Göker, 2019). Contrairement aux modèles linéaires utilisés pour ANI, dDDH est basé sur un modèle non linéaire appliqué à un ensemble de données empiriques plus large (Meier-Kolthoff et al., 2013). Ce modèle donne des estimations ponctuelles sur la même échelle que les valeurs DDH conventionnelles avec des intervalles de confiance (Meier-Kolthoff et al., 2013). Selon Meier-Kolthoff et al. (2019), il est donc raisonnable de conclure que dDDH dans TYGS est la méthode de choix pour la délimitation des espèces procaryotes, et ait surpassé l'ANI.

La détermination taxonomique des 13 souches de l'arbre des Oscillatoriaceae, réalisée avec l'aide de Marcelo Vaz, s'est faite sur base des résultats obtenus des 4 banques de données : GTDB, NCBI, SILVA et CyanoSeq (**Figure 8 et Annexe 7**). Les résultats taxonomiques de ces banques de données étaient assez variés. GTDB et NCBI s'accordent généralement dans leurs résultats. SILVA a déterminé tout le groupe *Laspinema* (**G1**) comme *Oscillatoria sp.* PCC-6304 (une souche de l'institut Pasteur). Ce genre de discordance dans les données taxonomiques peut poser problème dans la détermination d'une nouvelle souche. Dans les bases de données SILVA et GTDB, plusieurs taxons sont mal étiquetés au niveau du genre. Ces problèmes surviennent en raison d'une taxonomie inappropriée ou obsolète du NCBI. Par exemple, *Oscillatoria* PCC-6304 est incorrectement étiqueté comme le type d'*Oscillatoria* dans GTDB et dans SILVA, alors que cette souche correspond au genre *Laspinema* (Heidari et al., 2018 ; Stanojković et al., 2022). De plus, *Oscillatoria sp.* PCC-10802 est correctement étiqueté comme *Oscillatoria* tel que défini par (Mühlsteinova et al., 2018) dans SILVA. Cela crée une confusion au sein de SILVA car *Oscillatoria* englobe des genres disparates. De nombreux taxons microbiens sont constitués de lignées morphologiquement presque indiscernables mais génétiquement différenciées. Ces lignées cryptiques sont répandues parmi les procaryotes, ce qui pose des problèmes importants pour leur résolution taxonomique (Dvořák et al., 2014).

Malgré ces discordances, nous avons décidé de nommer tout cet embranchement (**G1**) comme faisant partie du genre *Laspinema*, comme désigné par les résultats de CyanoSeq. Nos résultats (arbre phylogénétique (**Figure 8**) et ANI (**Figure 9**)) soutiennent que nos bins ULC sont proches des souches *Laspinema* de Stanojkovic et al. (2022) (**G3**), ainsi que de *Laspinema thermale* HK S5 de Heidari et al. (2018) (**Annexe 5, ORPER**), justifiant notre choix de nommer toutes ces souches comme *Laspinema sp.*

Dans l'arbre des Oscillatoriaceae, la souche GCA_025054815.1 est la seule où le nom taxonomique est la famille (Oscillatoriaceae pour GTDB et NCBI, Phormidiaceae pour SILVA et Pleurocapsaceae pour CyanoSeq). La qualité du génome (**Annexe 9**) expliquerait pourquoi les banques de données ne pouvaient qu'identifier la famille et pas le genre pour cette souche. De plus, elle vient d'un biofilm, elle pourrait être contaminée par d'autres microorganismes. Les résultats de l'analyse « CONTAMS » sont les suivant : 98,82 % de complétude, 0,29 % de contamination pour CheckM2 et 12,69 % de contamination pour Kraken. Les résultats de QUAST indiquent qu'il a 34 contigs et 43,92 % de GC. De base, cette souche n'était pas sélectionnée après les analyses « CONTAMS », elle a été

sélectionnée après avoir obtenu l'arbre de la famille de Oscillatoriaceae sur ORPER dans le but d'avoir un arbre plus représentatif de cette famille. Cette souche est un exemple qui confirme que les banque de données comme GenBank possèdent des génomes contaminés auquel il faut prêter attention (Cornet, Meunier, et al., 2018).

Ensuite, des études métaboliques ont été réalisées sur les 13 souches de l'arbre des Oscillatoriaceae. La recherche de fonction des gènes a été réalisée avec les 3 groupes de gènes spécifiques déterminés par OrthoFinder. : ULC722, les souches d'Antarctique (**G2**) et le groupe des *Laspinema* (**G1).** D'après les résultats (**Annexe 8**), il semblerait qu'au sein des trois souches d'Antarctiques (**G2**), aucune fonction des gènes spécifiques n'a été identifié (89 gènes non classifiés et 26 gènes inconnus). Aucun gène spécifique justifiant leurs résistances au froid et aux UVs n'a été trouvé d'après les résultats de « Metabolic Fonctionnal ». De plus, toutes les souches du genre *Laspinema* (**G1**) a également des gènes non classifiés et inconnus (41 et 52 respectivement). Cependant, ce genre semble partager le gène d'une protéine de transport au sucre (Msme, spécifique des procaryotes selon KEGG), permettant d'avoir une réserve énergétique suffisante. Pour ULC722, 61 gènes spécifiques ont été identifiés, mais la souche possède aussi des gènes non classifiés (183) et inconnus (26). Le « groupe » où il y a le plus de gènes non classifiés est celui de ULC722 et où il y a le plus de gènes inconnus est celui de *Laspinema* **(G1).** Tous ces gènes spécifiques non identifiés correspondent à la matière noire microbienne en métagénomique. Ce sont des données que nous n'arrivons pas encore à traiter et à identifier, car ils ne sont pas présents dans les bases de données d'étude des gènes fonctionnels (Thomas & Segata, 2019).

À partir de ces gènes spécifiques, l'étude des voies métaboliques a été réalisée avec « Metabolic Modelling ». Les résultats (**Figure 8 et 10**) indiquent que l'embranchement des *Laspinema* (**G1**) a perdu la voie métabolique de béta-oxydation, qui permet la dégradation des acides gras. Pourtant, les longues chaines d'acide gras sont une source importante de carbone et d'énergie pour la croissance des bactéries (Black et al., 1992). Pour faire traverser les longues chaines d'acides gras à travers la membrane plasmique, deux protéines membranaires sont nécessaires. La première, fadL, a une forte affinité pour ces molécules et les fait d'abord traverser la partie supérieure de la membrane. La seconde, fadD, est une protéine de membrane interne associée à l'acyl CoA synthétase (Black et al., 1992). Cette enzyme catalyse l'estérification d'acide gras en thioester CoA métaboliquement actif, en utilisant une molécule d'ATP (Groot et al., 1976). Les acides gras doivent être activés afin de participer dans un grand nombre de réactions

cataboliques et métaboliques (Watkins, 1997). Parmi ces voies figurent la synthèse de triacylglycérol, de phospholipides, de plasmalogènes, de sphingolipides et d'esters de cholestérol, l'oxydation α et ß des acides gras, l'allongement des acides gras, la conversion des acides gras en alcools gras, l'insertion et l'élimination des doubles liaisons et l'acylation des protéines (**Figure 11**) (Watkins & Ellis, 2012). Les acides gras activés seront transportés vers la membrane mitochondriale par une navette, la carnitine, afin de subir la suite des réactions de la béta-oxydation (Watkins & Ellis, 2012).

Alors que la voie de béta-oxydation permet de former un composé acyl-CoA activé, intervenant dans différentes autres voies métabolique, pourquoi les souches du groupe du

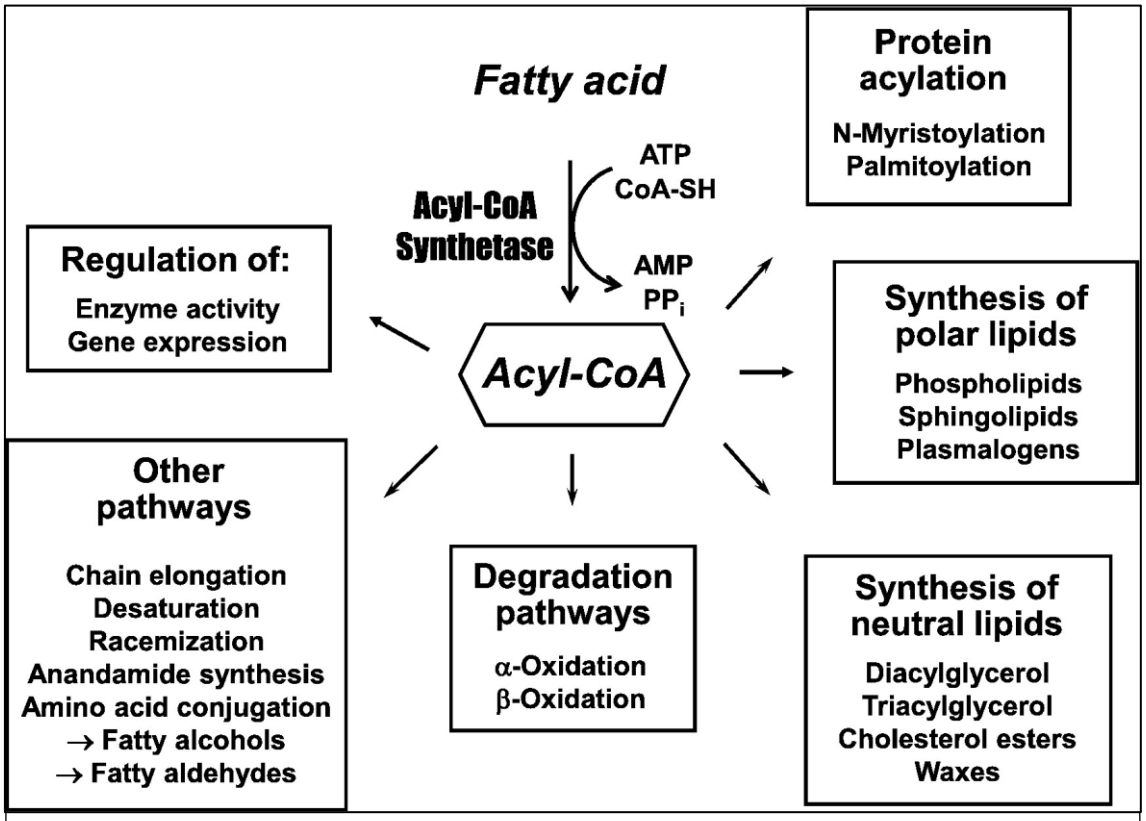


Figure 11 : Voies métaboliques possibles des acides gras activés, par thioestérification en CoA (Watkins & Ellis, 2012)

genre *Laspinema* (**G1**) auraient perdus cette voie ? Est-ce que l'apport en sucre (via le transporteur Msme) leur est suffisant pour survivre ? Ces résultats ne concordent pas avec ceux de Lumian et al (2021). Ils soutiennent que la souche du lac de Fryxell, GCF_017313335.1 est capable de réaliser la ß-oxydation des acides gras. Pourtant, la perte de cette voie entrainerait l'accumulation d'acides gras et l'inhibition de la consommation d'oxygène des mitochondries, induisant un stress oxydatif (Lim et al., 2018). Une explication possible serait qu'il existerait une autre voie de β-oxydation, adaptée au froid, et que les gènes n'ont pas été trouvés car ils seraient présents dans les gènes inconnus.

L'autre étude métabolique consiste à étudier les métabolites secondaires via *genome2metabolite.nf*. Selon Stanojkovic et al, le genre *Laspinema* pouvait avoir des stratégies d'adaptation environnementale distinctes associées aux réponses au stress et aux intensités lumineuses dans les microhabitats du sol. Les résultats (**Figure 8**) obtenus nous a donné une liste de métabolites secondaires présentes dans certaines souches des Oscillatoriceae, mais ne confirment pas les hypothèses de Stanojkovic et al.

Les métabolites secondaires présentes dans tout l'embranchement sélectionné, sauf pour *Oscillatoria sp.* SIO1A7*,* sont les terpènes, des composés intervenant dans le métabolisme primaire et la photosynthèse. Cela parait évident puisque les cyanobactéries sont des microorganismes photosynthétiques. Ils ont sûrement besoin des terpènes pour réaliser la photosynthèse. Dans ce cas, pourquoi est-il absent pour la souche *Oscillatoria sp.* SIO1A7 (GCA_010672965.1) ? D'après les résultats de qualité de génome (**Annexe 9**), le génome serait contaminé à 10,21% pour les résultats de Kraken, alors qu'il avait une complétude à 100%. La souche a beaucoup de contigs (182) et un pourcentage de GC a 47,90. Une contamination ne peut empêcher l'identification d'un composé. Cependant, si le gène est incomplet, cela peut exercer des influences. Bien que la complétude soit à 100%, il est possible que la séquence de gène soit incomplète et que les gènes liés aux terpènes ne soient pas présents. De plus, les résultats de QUAST indiquent qu'au sein du génome 19,68 N sont présents tous 100 kpb (**Annexe 9**). Cela pourrait expliquer pourquoi le gène lié aux terpènes n'a pas été identifié.

La souche GCA_025054815.1, identifiée comme Oscillatoriaceae n'a que les terpènes comme métabolites secondaires. Pourquoi n'a-t-elle pas d'autres métabolites secondaires comme les autres souches ? Il se peut que son génome soit incomplet également, malgré une complétude à 98,82 %, comme pour la discussion sur la souche précédente. Il manquerait des séquences de gènes liés à certains métabolites secondaires.

Aucunes des souches ne possèdent de cyanotoxines. Dans le désert froid du Pamir (Tadjikistan), Khomutovska et son équipe (2020) ont été surpris d'obtenir si peu de preuves de la production de toxines et émettent l'hypothèse que les conditions pourraient être trop difficiles pour sa production. Ils supposent que la production coûteuse en énergie de cyanotoxines n'est pas une caractéristique utile dans cet environnement. La période de végétation très courte, l'insolation élevée par les UV et la lumière visible, la température variable selon les saisons et la salinité souvent élevée peuvent exercer un stress important sur les cyanobactéries, les empêchant d'engager des coûts énergétiques supplémentaires

dans la production de toxines. Il est possible aussi que si les taxons toxigènes sont absents de ces communautés, c'est qu'ils auraient été éliminés par des forces évolutives. (Khomutovska et al., 2020). Ces hypothèses peuvent également s'appliquer sur nos souches antarctiques et expliquer l'absence de toxines.

Enfin, les trois souches d'Antarctique (ULC096, ULC102, GCF_017313335.1) (**G2**) ont en commun les terpènes et les sidérophores comme métabolites secondaires. Le sidérophore est un chélateur de fer et permettrait de capter le fer ferrique surtout si le microorganisme est en carence. ULC096 possède en plus des NRPS et des RRE-containing, des peptides ayant des actions variées tels qu'antibactérien, antitumoraux, immunosuppresseurs, etc. GCF_017313335.1 contient des gènes codant pour de l'arylpolyène (un pigment jaune) et des RiPP-like (des peptides naturels d'origine ribosomique). Ces métabolites ne semblent pas être spécifique aux milieux extrêmes polaires, car ils sont présents aux seins d'autres biomes terrestres, benthiques, lacs alcalins, etc. Il y a uniquement le sidérophore qui semble spécifique aux 6 souches de la partie inférieure de l'arbre (de GCA_025370875.1 à ULC102), alors que celles-ci viennent d'endroits différents : Antarctique (**G2**) et République Tchèque (**G3**). Les autres métabolites secondaires sont absents au sein des trois souches antarctiques (**G2**). Peut-être que ces souches n'ont pas besoin de métabolites supplémentaires pour survire dans un environnement hostile à la majorité des êtres-vivants, car les cyanobactéries n'ont pas de compétiteurs ou de consommateurs qui pourraient empêcher leur développement. Elles ont peut-être besoin des fonctions minimales pour survivre afin de ne pas dépenser trop d'énergie.

Le lac de Fryxell contient un tapis benthique microbien responsable du gradient d'oxygène et de sulfure dans cet environnement. La flore cyanobactérienne sous la glace des parties oxiques des lacs de la vallées sèches de Mc Murdo (MDV) est largement dominée par des genres d'Oscillatoriales, dont *P. pseudopriestleyi* (que nous considérons comme *Laspinema* dans notre cas d'étude) (Jungblut et al., 2016). Ce taxon a déjà été trouvé en grande abondance uniquement dans des étangs d'eau de fonte hypersalins peu profonds sur la plate-forme de glace McMurdo, en Antarctique (Jungblut et al., 2005). Ces étangs ont des niveaux élevés de sulfure, en raison de la présence de sédiments organiques riches en sels de sulfate de sodium, ce qui suggère que *P. pseudopriestleyi* a une tolérance élevée au sulfure. *Phormidium pseudopriestleyi* peut avoir la capacité de photosynthétiser à la fois de manière oxygénique et anoxygénique en passant à H_2S en tant que donneur d'électrons dans les puisards anoxiques du lac Huron, au Canada (Voorhies et al., 2012). Cette flexibilité métabolique pour utiliser H_2S comme donneur d'électrons serait avantageuse

dans le lac de Fryxell car cela permettrait une croissance photosynthétique pendant les périodes où il n'y a pas suffisamment de lumière pour la photosynthèse oxygénique. Des cyanobactéries et des diatomées tolérantes aux sulfures ont été trouvées capables de créer des microhabitats riches en oxygène, probablement saisonniers, dans les tapis qui maintiennent les assemblages d'aérobies et donc la richesse taxonomique (Jungblut et al., 2016). De plus, le sulfure inhibe la photosynthèse oxygénique en bloquant le transfert d'électrons entre H_2O et le complexe dégageant de l'oxygène dans la protéine D1 du photosystème II. Dans le lac de Fryxell, *Phormidium pseudopriestleyi* crée une couche de 1 à 2 mm d'épaisseur d'O_2 dans de l'eau sulfurée, démontrant qu'elle soutient la photosynthèse oxygénée en présence de sulfure (Lumian et al., 2021).

Les bins ULC096 et ULC102 viennent d'étangs de la vallées sèches de Mc Murdo et ont été dénommés respectivement comme *Oscillatoria priestleyi* CMEE5020 Ant-Pancreas et *Phormidium autumnale* CMEE5034 Ant-Brack 2 par Nadeau et al (2001). Les deux souches ont une bonne conductivité dans l'eau et vivent dans un habitat riche en sulfate de sodium (Nadeau et al., 2001). Il se pourrait que ces deux souches soient aussi tolérantes aux sulfides et puissent réaliser la photosynthèse anoxygénique avec le sulfide comme donneur d'électrons au lieu de l'eau. D'après les résultats de « Metabolic Modelling », toutes les 13 souches sélectionnées pour l'analyse ont la capacité d'assimiler du sulfure sauf *Oscillatoria sp.* SIO1A7 (GCA_010672965.1). Cependant, les résultats obtenus ne précisent pas la possibilité de réaliser la photosynthèse anoxygénique. Nous savons uniquement qu'elles peuvent assimiler des sulfures et les réduire.

6 Conclusion et perspectives

Tout d'abord, en utilisant les différents outils bio-informatiques de GEN-ERA, un grand nombre de résultats a pu être généré : analyse de qualité de génomes, réalisation de plusieurs arbres phylogénétiques, analyse des gènes fonctionnelles et de métabolites secondaires. Les résultats obtenus confirment que GEN-ERA est un outil bio-informatique très utile et permettant le traitement rapide des données métagénomiques et phylogénétiques.

Les objectifs de ce mémoire étaient d'étudier les génomes des souches de la collection BCCM/ULC, de les classer phylogénétiquement, de les identifier au niveau taxonomique et de déterminer leurs métabolismes. Les résultats suivants répondent aux questions décrites dans la section « 2.8 Objectifs » :

1. Les souches de la collection étaient, en effet, contaminées par des génomes de Proteobacteria, ou parfois de Verrumicrobia et Bacteroidota. Il y avait toujours un pourcentage de séquences dont *GENContams.nf* ne pouvait identifier l'origine (Unknown). À partir des résultats de qualité des séquences de génomes, 6 bins ont été sélectionnés : ULC002, ULC008, ULC029, ULC096, ULC102 et ULC722.

2. Ces bins appartenaient aux trois familles de Cyanobactéries : Oscillatoriaceae (ULC096, ULC102, ULC722), Leptolyngbyaceae (ULC029) et Nostocaceae (ULC008 et ULC002). ULC002 et ULC008 étaient tous les deux des *Nostoc sp.* ULC029 est associé à *Stenomitos sp.* ULC096, ULC102 et ULC722 font partie du genre *Laspinema.*

3. Nous nous sommes focalisés sur l'étude d'une seule famille : Oscillatoriaceae. Nos 3 bins se situent dans un embranchement (**G1**) regroupant d'autres souches du genre *Laspinema*, dont les souches de Stanojkovic et al (2022) (**G3**). ULC722 est isolé des autres souches de cet embranchement. ULC096 et ULC102 forment un sous-embranchement avec la souche GCF_017313335.1 du lac de Fryxell (Antarctique, **G2**).

4. L'ensemble de cet embranchement (**G1**) où nos 3 bins ULC096, ULC102 et ULC722 sont présents semble appartenir au genre *Laspinema sp.*

5. Grâce aux résultats d'ANI et GGDC (TYGZ), nous sommes convaincus que cet embranchement (**G1**) possède 5 nouvelles espèces appartenant à ce genre : les 3 souches d'Antarctique (**G2**), les 3 souches de Stanojkovic et al (2022) (**G3**),

GCF_00317105.1 (*Oscillatoria acuminata* PCC 6304), GCF_024220515.1 (*Oscillatoria sp.* HE19RPO) et ULC722.

6. Pour l'ensemble des souches du genre *Laspinema* **(G1),** un seul type de gènes spécifiques a été identifié, celui lié à la protéine du sucre Msme. Les autres gènes n'ont pu être identifiés et ont été considérés par « Metabolic Fonctionnal » soit comme non classifiés soit comme inconnus. Les 3 souches d'Antarctique **(G2)** n'ont aucun gène fonctionnel spécifique identifié. De plus, les souches du genre *Laspinema* **(G1)** auraient toutes perdu la voie de béta-oxydation. Une justification possible de cette disparition serait que cette voie demanderait trop d'énergie, mais reste à être confirmée.

7. Toutes les souches du genre *Laspinema* (sauf GCA_010672965.1) se partagent les terpènes comme métabolites secondaires. Les trois souches antarctiques **(G2)** (ULC096, ULC102 et GCF_017313335.1) ainsi que les 3 souches de Stanojkovic et al (2022) **(G3)** ont aussi des sidérophores comme métabolites secondaires. Il n'y pas de métabolites secondaires spécifiques aux cyanobactéries d'Antarctique. Elles possèdent des métabolites secondaires que les autres souches de biomes différents ont aussi.

8. Pour le moment, nous n'avons pas trouvé de métabolites ou de voies spécifiques aux souches d'Antarctique qui justifieraient leurs résistances au froid polaire et aux UVs. Cependant, une possibilité est qu'elles aient des copies supplémentaires de gènes impliqués dans ces résistances, et nous n'avons pas investigué cela. Il est possible également que ces gènes soient présents dans les gènes inconnus ou non classifiés des gènes spécifiques de nos souches d'Antarctique (la matière noire).

Nos résultats concordent pour dire que nos 3 bins ULC appartiennent au genre *Laspinema* et seraient des nouvelles espèces. Il faudrait étudier plus en détail ces souches pour décrire des nouvelles espèces, leur donner un nom et déposer les données dans les banques des données afin de rendre les recherches taxonomiques plus fructueuses dans le futur. Il reste encore beaucoup de recherches à faire à ce niveau. Un grand nombre de discordances taxonomiques entre les différentes banques de données sont encore présentes, surtout au sein des cyanobactéries. Nous pensons qu'une révision des noms taxonomiques des souches s'impose, surtout pour la souche *Oscillatoria sp.* PCC-6304, ainsi que pour *Phormidium pseudopriestleyi* FRX01, qui semblent être du genre *Laspinema.*

Enfin, dans cette étude, nous nous sommes focalisé sur la famille des Oscillatoriaceae et avons travaillé sur 3 bins de la collection BCCM/ULC. Il reste deux autres familles à

étudier (Leptolyngbyaceae et Nostocaceae) selon la même démarche : réalisation de l'arbre phylogénétique de la famille via ANI et ORPER ; l'analyse des données ANI et TYGC sur les souches de ces familles ; la détermination taxonomique via les 4 banques de données (NCBI, GTDB, SILVA, CyanoSeq), l'étude des gènes spécifiques, des voies métaboliques et des métabolites secondaires. Toutes ces données récoltées sur les bins ULC communiqueraient des informations pertinentes concernant les cyanobactéries d'Antarctique et permettrait de comprendre pourquoi elles sont résistantes au froid et aux UVs.

D'un point de vue plus général, il serait également utile que des progrès soient réalisés au niveau de l'identification des gènes de la « matière noire » car cela nous empêche, sans doute, de détecter des informations intéressantes sur les gènes des cyanobactéries des milieux extrêmes.

7 Annexes

7.1 Annexe 1 : Les lignes de commandes supplémentaires

Lignes de commandes pour l'étude de mon mémoire

1. Genome Downloader

PATH: /scratch/ulg/GENERA/alequeux/GENOME-DW

1.1. Téléchargement des génomes de Gloeobacter via RefSeq au lieu de Genbank, comme groupe externe.

$ nano -w Genome-downloader.job

$ nextflow run Genome-downloader.nf --taxolevel=genus --group=Gloeobacter --genbank=no --dRep=no --ignoreGenomeQuality=no --cpu=20 --refseq=yes

$ mv Genome-downloader_output Genome-downloarder_Gloeobacter

1.2. Téléchargement des génomes de cyanobactéries dans RefSeq et GenBank, dans le but de réaliser des comparaisons dans ANI :

$ nano -w Genome-downloader.job

$ nextflow run Genome-downloader.nf --taxolevel=phylum --group=Cyanobacteria --genbank=yes --dRep=no --ignoreGenomeQuality=no --cpu=20 --refseq=yes; mv Genome-downloader_output Genome-downloader_cyano_refseq_Genbank_output

2. Genome Assembly

PATH : /scratch/ulg/GENERA/alequeux/GENOME-AS

Assemblage de différents gènes de cyanobactéries, séquencées par illumina ou nanopore.

Les génomes de mes souches ont été recherchés dans un lien partagé :

$ cd /scratch/ulg/GENERA/SHARING/Alina

$ cp * /scratch/ulg/GENERA/alequeux/GENOME-AS

$ cd /scratch/ulg/GENERA/alequeux/GENOME-AS

Voici les instructions suivies pour l'analyse d'assemblage:

$ nano -w Assembly.job

$ nextflow run Assembly.nf --shortreadsR1=ULC102_R1.fastq --shortreadsR2=ULC102_R2.fastq --ontreads=ULC102_T.fastq --genomeSIZE=5mb --cpu=20 --metagenome=yes --binner=all; mv GENERA-assembly GENERA-assembly_ULC102

Voici un exemple pour les données de l'échantillon ULC102. Les mêmes "consignes" ont été réalisées pour chaque échantillon.

3. GTDB

PATH : /scratch/ulg/GENERA/alequeux/GTDB

3.1. GTDB sur l'ensemble des souches de la collection BCCM/ULC

L'analyse GTDB a été réalisée pour vérifier que les souches sont des cyanobactéries.

Les bins obtenus par l'assemblage ont été importé dans le répertoire GTDB. Une boucle a été créée pour faciliter les manipulations :

$ for nom in 002 008 029 046 128 180 307;

do cd /scratch/ulg/GENERA/alequeux/GTDB;

mkdir Genome_ULC$nom;

cd /scratch/ulg/GENERA/alequeux/GENOME-AS/GENERA-assembly_ULC$nom;

cp *.fasta /scratch/ulg/GENERA/alequeux/GTDB/Genome_ULC$nom;

cd /scratch/ulg/GENERA/alequeux/GTDB/Genome_ULC$nom;

rm Genome.fasta;

done

Ensuite, chaque fichier fasta a été transformé en fichier fna:

$ for chiffre in $(seq 1 25);

do mv CONCOCT_bin-$chiffre.fasta CONCOCT_bin-$chiffre.fna;

mv METABAT_bin-$chiffre.fasta METABAT_bin$chiffre.fna;

done

Ensuite pour chaque fichier de données, l'analyse GTDB a été réalisé:

$for fichier in ULC046;

do cd /scratch/ulg/GENERA/alequeux/GTDB/genome;

rm *;

cp /scratch/ulg/GENERA/alequeux/GTDB/Genome_$fichier/*.fna /scratch/ulg/GENERA/alequeux/GTDB/genome;

done

$ nano -w GTDB.job

$ nextflow run GTDB.nf --genome=genome --cpu=20; mv GENERA_GTDB GENERA_GTDB_ULC029

Ceci est un exemple d'échantillon (ULC029). La commande a été réalisée pour chaque souche en suivant les mêmes "consignes".

3.2. GTDB sur les 52 génomes :

$ cd ortho/

$ cd infiles_52genomes/

$ cp *.fna /scratch/ulg/GENERA/alequeux/GTDB/genome

$ for f in *.fna; do mv -- "$f" "GEN_$f"; done (pour renommer les fichiers)

$ cd GTDB/

$ nano -w GTDB.job

$ nextflow run GTDB.nf --genome=genome --cpu=20

$ mv GENERA_GTDB GENERA_GTDB_phylo

4. CONTAMS

PATH : /scratch/ulg/GENERA/alequeux/CONTAMS

4.1. Sur les souches de la collection BCCM/ULC

Pour déterminer le taux de contamination au sein de mes génomes, il faut d'abord déterminer quels sont les bins qui contiennent des gènes de cyanobactéries dans mes données :

$ cd /scratch/ulg/GENERA/alequeux/GTDB/GENERA_GTDB_ULC029/GTDB-classify

$ grep "yanobacteria" gtdbtk.bac120.summary.tsv | less

voici un exemple pour ULC029. La vérification a été réalisée pour toutes les souches de la même manière.

Les fichiers des bins ont été copiés dans un répertoire "GENERA-input" dans CONTAMS. Pour cela, Une boucle a été créée pour gagner du temps, voici un exemple:

$ for nom in 002 307;

do cd /scratch/ulg/GENERA/alequeux/GTDB/Genome_ULC$nom;

cp CONCOCT_bin-1.fna ULC$nom._CONCOCT_bin-1_Stenomitos.fna;

cp ULC$nom._CONCOCT_bin-1_Stenomitos.fna /scratch/ulg/GENERA/alequeux/CONTAMS/GENERA-input;

cp METABAT_bin1.fna ULC$nom._METABAT_bin-1_Stenomitos.fna;

cp ULC$nom._METABAT_bin-1_Stenomitos.fna /scratch/ulg/GENERA/alequeux/CONTAMS/GENERA-input;

done

Ensuite, l'analyse "CONTAMS" a été lancé :

$ nano -w GENcontams.job

$ nextflow run GENcontams.nf --genomes=GENERA-input --mode=kraken --ext=fna --cpu=20 --dbdir=/data/GENERA/CONTAMS --taxdump=/data/GENERA/CONTAMS/taxdump/ --taxlevel=species; mv GENERA-contams GENERA-contams_KRAKEN

Chaque mode a été testé via les mêmes consignes, en modifiant uniquement le mode : Kraken, CheckM2 et QUAST.

Les résulstats de CheckM2 indiquent le taux de contamination et de complétude. Les génomes qui ont une complétude supérieur à 70 % et qui ont un faible taux de contamination ont été sélectionnés.

Pour faciliter l'analyse de Kraken, une analyse a été relancée au niveau du phylum :

$ nano -w GENcontams.job

$ nextflow run GENcontams.nf --genomes=GENERA-input --mode=kraken --ext=fna --cpu=20 --dbdir=/data/GENERA/CONTAMS --taxdump=/data/GENERA/CONTAMS/taxdump/ --taxlevel=phylum; mv mv GENERA-contams GENERA-contams_KRAKEN_phylum

Une fois que le téléchargement est terminé, les resultats de Kraken ont été analysé afin de choisir les génomes qui ont maximum 10% de contamination.

4.2. Sur un plus grand nombre de génomes.

À l'aide des données de GTDB sur mes 6 bins, 3 familles de cyanoabactéries ont été déterminées pour l'étude : Leptolyngbyaceae, Oscillatoriaceae, Nostocaceae. L'analyse CONTAMS a été relancé sur les génomes téléchargés de ces familles dans Genome-downlaoder :

$ cd /scratch/ulg/GENERA/alequeux/ANI/GENERA_ANI/ANI-close-match/

$ for f in `cat gen2keep`; do cp /scratch/ulg/GENERA/alequeux/ANI/genome/$f.fna /scratch/ulg/GENERA/alequeux/CONTAMS/GENERA-input; done

$ cd /scratch/ulg/GENERA/alequeux/GENOME-DW/Genome-downloader_cyano_refseq_Genbank_output/

$ grep 'scillatoriaceae' Genomes.taxomonomy > liste1 (pour Oscillatoriaceae)

$ grep 'ostocaceae' Genomes.taxomonomy > liste2 (pour nostocacaea)

$ grep 'eptolyngbyaceae' Genomes.taxomonomy > liste3 (pour leptolyngbyaceae)

$ cat liste1 liste2 liste3 > lepto_nostoc_oscilla

$ cut -f1 lepto_nostoc_oscilla > liste_l_n_o

$ cd ../

$ for f in `cat liste_l_n_o`; do cp /scratch/ulg/GENERA/alequeux/GENOME-DW/Genome-downloader_cyano_refseq_Genbank_output/GENOMES/$f.fna /scratch/ulg/GENERA/alequeux/CONTAMS/GENERA-input; done

Ensuite j'ai lancé l'analyse CONTAMS, comme précédemment en mode CheckM2, Kraken et QUAST.

Les résultats obtenus pour CheckM2 et Kraken ont été téléchargés et analysés en réalisant un tri dans excel.

Les génomes devaient avoir une complétude supéieure à 90 % et un taux de contamination < ou = à 5 % pour CheckM2.

Ensuite, la première sélection a été comparée avec les résultats de Kraken et les génomes qui avaient un taux de contamination égale ou inférieure à 5 % ont été sélectionnés.

Au final, nous avons conservé 246 génomes venant des analyses d'ANI et du téléchargement des génomes de cyanobactéries.

Avec ces 246 génomes et les 3 génomes de *Gloeobacter*, nous allons réaliser notre premier arbre via OrthoFinder (cfr orthology).

5. ANI

PATH : /scratch/ulg/GENERA/alequeux/ANI

5.1 sur un grand nombre de génomes

Les génomes de mes 6 bins ont été copiés dans le répertoire « genome » de ANI :

$ cd /scratch/ulg/GENERA/alequeux/CONTAMS/INPUT/

$ cp ULC002._METABAT_bin-1_Stenomitos.fna ULC008_METABAT_bin-4_Nostoc.fna ULC029._METABAT_bin-1_Stenomitos.fna ULC096_METABAT_bin-2_Phormidium_pseudopriestleyi.fna ULC102_METABAT_bin-1_Phormidium_pseudopriestleyi.fna ULC722_METABAT_bin-3_Oscillatoriaceae.fna /scratch/ulg/GENERA/alequeux/ANI/genome

Les fichiers des bins ont été renomés pour que le nom soit plus court (voici un exemple, les mêmes démarches ont été réalisées pour les autres bins)

$ cd /scratch/ulg/GENERA/alequeux/ANI/genome

$ mv ULC002._METABAT_bin-1_Stenomitos.fna ULC002_Mb1.fna

$ echo -e "ULC002_Mb1\nULC102_Mb1\nULC096_Mb2\nULC029_Mb1\nULC008_Mb4\nULC722_Mb3\n" > shortlist

Cette commande permet de créer une liste de mes bins

$ cp shortlist /scratch/ulg/GENERA/alequeux/ANI

Tous les génomes de cyanobactéries téléchargés sont importés dans un des répertoires de ANI :

$ cd /scratch/ulg/GENERA/alequeux/GENOME-DW/Genome-downloader_cyano_refseq_Genbank_output/GENOMES

$ cp * /scratch/ulg/GENERA/alequeux/ANI/genome

$ find *.fna > list

création d'une liste de tous mes génomes qui serviront de fichier input

$ sed -i -e 's/.fna//g' list

$ cp list /scratch/ulg/GENERA/alequeux/ANI

$ cd /scratch/ulg/GENERA/alequeux/ANI

$ nano shortlist

Les noms de shortlist sont copiés dans le fichier list via : CTRL+C et CTRL+V

$ nano list

L'analyse ANI a été lancée :

$ nano -w ANI.job

$ nextflow run ANI.nf --genome=genome --list=list --mode=onetomany --shortlist=shortlist --cpu=20

$ mv GENERA_ANI GENERA_ANI_premier

Ensuite, un seul fichier de mes résultats obtenus a été créé :

$ cd /scratch/ulg/GENERA/alequeux/ANI/GENERA_ANI/ANI-close-match/

$ cat *.list | grep -v identical | sort -rV -k2 > close-list.sorted

Cette commande permet de mettre tous les fichiers ensemble, d'éliminer les lignes identiques, et de trier la seconde colonne numérique de manière décroissante.

$ less -N close-list.sorted

Cette commande permet de voir les premières lignes où la valeur de ANI est >= 95 %, je prends le nombre des premières lignes pour la commande suivante.

$ head -n10 close-list.sorted | cut -f1 | sort -f | uniq > gen2keep

cette liste correspond aux génomes à conserver pour la réalisation du premier arbre phylogénétique.

5.2 sur les génomes Oscillatoriaceae

Voici la commande nextflow :

$ nextflow run ANI.nf --genome=genome --list=list --mode=manytomany --idm=file.idm --cpu=20

Dans le répertoire, il y avait les 27 fichiers de génome (.fna) et le fichier « list » contenait le nom des fichiers.

6. Obtention du premier arbre phylogénétique des cyanobactéries

6.1. Orthology

PATH : /scratch/ulg/GENERA/alequeux/ortho

Après l'analyse des résultats de CONTAMS sur un grand nombre de génomes, une liste a été créée et chargée sur NIC 5 dans mon répertoire « ORTHO ». Mon fichier chargé a changé de nom en « corelist » et la commande suivante permet de supprimer "symbole" à la ligne:

$ perl -i.bak -ne 's/\r\n/\n/g; print;' corelist

$ for f in `cat corelist`; do cp /scratch/ulg/GENERA/alequeux/CONTAMS/GENERA-input/$f.fna /scratch/ulg/GENERA/alequeux/ortho/infiles; done

Dans mon répertoire « infiles » et mon fichier « corelist », les noms et les fichiers fna des 3 *Gloeobacter* ont été ajoutés.

$ nano -w Orthology.job

$ nextflow run Orthology.nf --infiles=infiles --mode=inference --core=yes --corelist=corelist --corepresence=60 --specific=no --anvio=no --type=nucleotide --cpu=20

Lorsque l'analyse orthology a été terminé, 63 coregenes ont été obtenus. J'ai donc relancé l'analyse en modifiant certaines choses.

- J'ai supprimé le nom des *Gloeobacter* dans la « corelist ».
- J'ai changé le nom de mon répertoire « infiles » pour en créer un autre :

$ mv infiles premier_ortho_tout

Ensuite, dans mon nouveau répertoire « infiles », les fichiers .fa du répertoire REDO (des résultats obtenus de la première analyse orthology) ont été importé.

$ cd GENERA-Orthology

$ mv REDO/* /scratch/ulg/GENERA/alequeux/ortho/infiles

$ nano -w Orthology.job

$ nextflow run Orthology.nf --infiles=infiles --mode=OG --core=yes --corelist=corelist --corepresence=58 --coreunwanted=3 --specific=no --anvio=no --type=protein --cpu=20 ; mv GENERA-Orthology GENERA-Orthology-5

206 coregenes ont été obtenus avec cette commande (j'avais essayé à 60, 50, 55 et 56 % de corepresence). Le but était de se rapprocher le plus possible de 100 coregenes. Ces coregenes ont été sélectionné pour réaliser la phylogénie et le fichier a été renommé :

$ mv GENERA-Orthology-58 GENERA-Orthology-58-206core

6.2. Phylogeny

PATH : /scratch/ulg/GENERA/alequeux/Phylo

Les fichiers des coregenes ont été transféré dans le répertoire "OGs".

$ cd /scratch/ulg/GENERA/alequeux/ortho/GENERA-Orthology-58-206core/

$ cd coreGenes/

$ cp *.fa /scratch/ulg/GENERA/alequeux/Phylo/OGs

Une liste a été créé pour les fichiers dans OGs.

$ cd ortho

$ cp corelist /scratch/ulg/GENERA/alequeux/GENOME-DW/Genome-downloader_cyano_refseq_Genbank_output/

$ cd /scratch/ulg/GENERA/alequeux/GENOME-DW/Genome-downloader_cyano_refseq_Genbank_output

$ for f in `cat corelist` ; do cat Genomes.taxomonomy | grep -w $f >> listephylo ; done (pour créer une liste des noms présents dans les fichiers OGs face à leurs noms taxonomiques)

$ less listephylo

$ cut -f 2-3 listephylo (couper les colonnes qui ne nous intéressent pas)

$ wc -l listephylo

$ cp listephylo /scratch/ulg/GENERA/alequeux/Phylo/

$ cd /scratch/ulg/GENERA/GENOME-DW/Genome-downloader_Gloeobacter/

$ cut -f 2-3 Genomes.taxomonomy > gloeo

$ cp gloeo /scratch/ulg/GENERA/alequeux/Phylo/

$ cd /scratch/ulg/GENERA/alequeux/Phylo/

$ cat listephylo gloeo > file.idm

$ nanon file.idm (pour ajouter le nom taxonomique manuellement de mes bins et ## dans le haut du fichier)

$ wc -l file.idm (pour vérifier si j'ai bien 250 "noms" + 2 lignes pour les caractère : #)

$ rm listephylo gloeo

Ensuite, l'analyse Phylogeny a été lancé :

$ nano -w Phylogeny.job

$ nextflow run Phylogeny.nf --OG=OGs --IDM=file.idm --cpu=20 --jackk=no --ext=fa --mode=prot --align=yes

$ mv GENERA-phylogeny GENERA-phylogeny_premier

7. Le second arbre des cyanobactéries

À partir de ces 52 génomes sélectionnés après l'analyse du premier arbre, l'analyse Orthology, Phylogeny en mode Jacknife ont été relancé.

7.1. Orthology

PATH : /scratch/ulg/GENERA/alequeux/ortho

$ cd ortho/

$ nano -w corelist

$ for f in `cat corelist`; do cp /scratch/ulg/GENERA/alequeux/ortho/premier_ortho_tout/$f.fna /scratch/ulg/GENERA/alequeux/ortho/infiles ; done

$ nano -w Orthology.job

$ nextflow run Orthology.nf --infiles=infiles --mode=inference --core=yes --corelist=corelist --corepresence=60 --specific=no --anvio=no --type=nucleotide --cpu=20

L'analyse a été relancée en mode OG, à partir des fichier REDO mis dans « infiles », car j'avais plus de 1200 coregenes et je dois en obtenir environ 600. J'ai également enlevé les 3 *Gloeobacter* de la « corelist ».

$ nano -w Orthology.job

$ nextflow run Orthology.nf --infiles=infiles --mode=OG --core=yes --corelist=corelist --corepresence=98 –coreunwanted=3 --specific=no --anvio=no --type=protein --cpu=20

Les 52 fichers OG ont été copié dans le répertoire « Phylo ».

7.2. Phylogeny

PATH : /scratch/ulg/GENERA/alequeux/Phylo

L'analyse Phylogeny a été relancée sur ces 52 génomes:

$ cd GENERA-Orthology_98_arbreOG/coreGenes/

$ cp *.fa /scratch/ulg/GENERA/alequeux/Phylo/OGs

$ cd ../

$ cp corelist /scratch/ulg/GENERA/alequeux/Phylo/

$ cd Phylo/

$ mv file.idm file.idm_premier

$ for f in `cat corelist` ; do cat file.idm_premier | grep -w $f >> listephylo ; done

$ nano listephylo (ajouter *Gloeobacter*)

$ mv listephylo file.idm

$ mv GENERA-phylogeny/ GENERA-phylogeny_premier

$ nano Phylogeny.job

$ nextflow run Phylogeny.nf --OG=OGs --IDM=file.idm --cpu=20 --jackk=yes --ext=fa --mode=prot --align=yes

8. **L'arbre des Oscillatoriaceae**

J'ai relancé Orthology et Phylogogeny pour obtenir un arbre centré sur les Oscillatoriaceae (en ajoutant ceux qui était présent dans ORPER -> *Oekeania, Lygnbya*...)

8.1. Orthology

PATH : /scratch/ulg/GENERA/alequeux/ortho

Orthology.job :

$ nextflow run Orthology.nf --infiles=infiles --mode=inference --core=yes --corelist=corelist --specific=no --anvio=no --type=nucleotide --corepresence=60 --coreunwanted=3 --cpu=20; mv GENERA-Orthology GENERA-Orthology-premier-Oscillatoriaceae

le répertoire « infiles » contenait : 21 souches Oscillatoriaceae, 3 ULC (722,102,096) et 3 *Gloeobacter* (outgroup) -> fichiers fna

la « corelist » contenait : le "nom" des 21 souches et des 3 ULC.

J'ai relancé en : mode=OG; Type = protein; corepresence=100; coreunwanted=3

le répertoire « infiles » contenait les fichier OGs.

8.2. Phylogeny

PATH : /scratch/ulg/GENERA/alequeux/Phylo

Phylogeny.job:

$ nextflow run Phylogeny.nf --OG=OGs --IDM=file.idm --cpu=20 --jackk=no --ext=fa --mode=prot --align=yes; mv GENERA-Orthology GENERA-Orthology-OG-100-Oscillatoriaceae

le file.idm contenait le nom des 21 souches Oscillatoriaceae, 3 ULC et 3 *Gloeobaster*.

9. ORPER

PATH : /scratch/ulg/GENERA/alequeux/ORPER

ORPER est une analyse qui se réalise sur le gène ARNr 16S.

Voici la commande pour obtenir les gènes 16S :

$ for f in `cat fna.list`; do barrnap \$f.fna --outseq \$f-barnap.fna --threads 1; done

Il faut donc chercher les fichiers 16S de mes génomes :

$ cd /scratch/ulg/GENERA/SHARING/Alina/16s

$ cp *.fna /scratch/ulg/GENERA/alequeux/ORPER

Dans mes fichiers des bins, les génomes de cyanobactéries déterminés via le site SILVA ont été sélectionnés. Les autres gènes ont été supprimés du fichier :

$ nano -w ULC722-16s.fna

$ nano -w ULC102-16s.fna

$ nano -w ULC096-16s.fna

$ cat ULC722-16s.fna ULC102-16s.fna ULC096-16s.fna > sequencesbins.fasta (attention pas de nom trop long)

$ nano -w ORPER.job

$ nextflow run ORPER.nf --reftaxolevel=family --refgroup=Oscillatoriaceae --refgenbank=yes --outgroup=Gloeobacteraceae --outtaxolevel=family --outgenbank=yes - cpu=20 --SSU=sequencesbins.fasta --cdhit=no --drep=no

Ensuite, ORPER a été relancé avec une séquence 16s de référence de *Laspinema sp HK S5* (SSU=sequencesbins*Laspinema*.fasta)

10. **Gene specific orthoFinder**

PATH : /scratch/ulg/GENERA/alequeux/ortho

L'analyse a été sur tous les génomes de l'arbre des Oscillatoriaceae :

$ nano -w Orthology.job

$ nextflow run Orthology.nf --infiles=infiles --mode=inference --core=yes --corelist=corelist --specific=yes --specificlist=specificlist --anvio=no --type=nucleotide --corepresence=60 --coreunwanted=3 --cpu=20

J'ai lancé l'analyse 3x avec 3 "specifictlist" différentes : les 3 *Phormidium*, ULC722 et de 722 aux *Laspinema* compris :

$ mv GENERA-Orthology GENERA-Orthology-gene-specific-Phormidium/

$ mv GENERA-orthology GENERA-Orthology-gene-specific-ULC722

$ mv GENERA-orthology GENERA-Orthology-gene-specific-ULC722àLaspinema

Pour déterminer le nombre de gène specific :

$ cd GENERA-Orthology-gene-specific-ULC722àLaspinema/

$ less final-specific.list

$ wc -l final-specific.list

11. Metabolic fonctionnel (MANTIS)

PATH : /scratch/ulg/GENERA/alequeux/metabo

À partir des mêmes fichiers faa utilisés pour orthology, l'analyse de l'étude métabolique fonctionnelle des gènes a été lancée :

$ cd ortho/GENERA-Orthology-gene-specific-Phormidium/specificGenes/

$ cat *.faa >> 3Phormidiums-specificgenes.faa # je dois concaténer en un seul fichier

$ cp 3Phormidiums-specificgenes.faa /scratch/ulg/GENERA/alequeux/metabo/infile/

$ cd /scratch/ulg/GENERA/alequeux/metabo/

$ nano -w Metabolic.job

$ nextflow run Metabolic.nf --infile=infile --mode=functional

Je recommence de même pour les 2 autres groupes et après j'observe les résulats :

$ cd GENERA_metabolic_3Phormidiums_specificgenes/MANTIS/

$ less consensus_annotation.tsv

J'ai téléchargé le fichier tsv et analysé les résultats dans excel.

12. Metabolic modelling (ANVIO)

Cet outil sert à l'étude des voies métaboliques. Il a été lancé sur les génomes de la partie inférieure de l'arbre des Oscillatoriaceae.

```
$ nano list  # nom des génomes de la partie inférieur de l'arbre

$ nano -w Metabolic.job

$ nextflow run Metabolic.nf --infile=infile --mode=modelling --list=list --cpu=20 --
kegg=/scratch/ulg/GENERA/Databases/ANVIO/kegg/; mv GENERA_metabolic/ GENERA_metabolic_modelling

$ mv GENERA_metabolic_modelling/ GENERA_metabolic_modelling_bas_arbre

$ cd GENERA_metabolic_modelling_bas_arbre/PDF/

$ less merge.pdf  # j'ai téléchargé ce fichier pour obsever les résulats.
```

12) Genome2metabolite

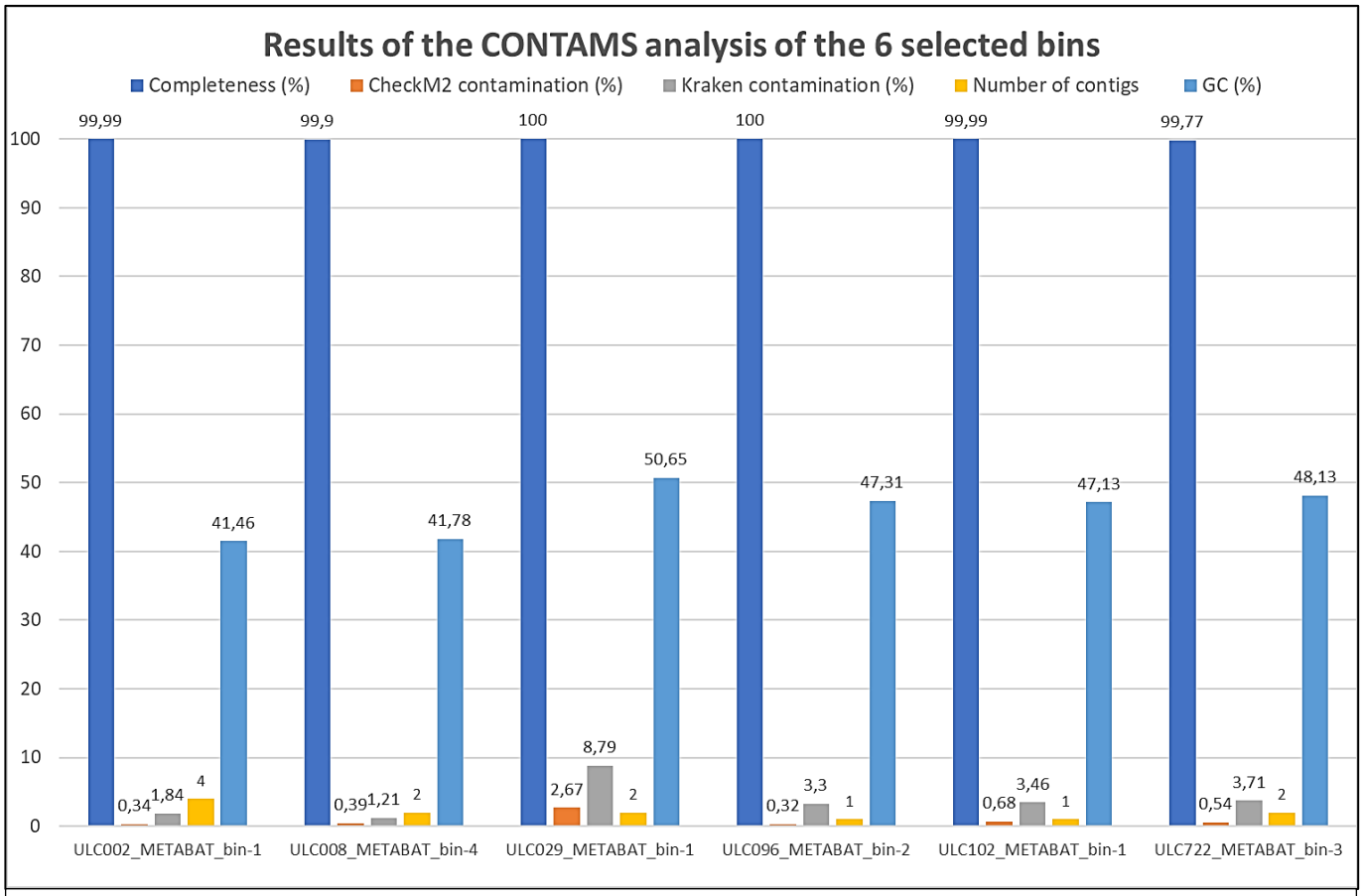
PATH : /scratch/ulg/GENERA/alequeux/meta2

Cet outil permet de détermienr les metabolites secondaires des génomes.

L'outil n'est pas présent dans GEN-ERA et a dû être importé comme suit :

```
$ mkdir meta2

$ cd /scratch/ulg/GENERA/SHARING/Alina/NRPS

$ cp * /scratch/ulg/GENERA/alequeux/meta2

$ cd /scratch/ulg/GENERA/alequeux/meta2

$ nano -w nextflow.config   # changer le pwd

$ nano -w genome2metabolite.job # changer le pwd

$ nextflow genome2metabolite.nf --fastadir=fasta --taxon=bacteria --cpu=20

$ mkdir fasta

$ cd metabo/infile/

$ cp *.fna /scratch/ulg/GENERA/alequeux/meta2/fasta  #partie basse de l'arbre

$ cd /scratch/ulg/GENERA/alequeux/meta2/fasta

$ find *.fna | wc -l

# Pour observer les résulats :

$ cd meta2/palantir-results/bgc_db_tables/

$ less clusters # j'ai téléchargé ce fichier et analysé sur excel
```

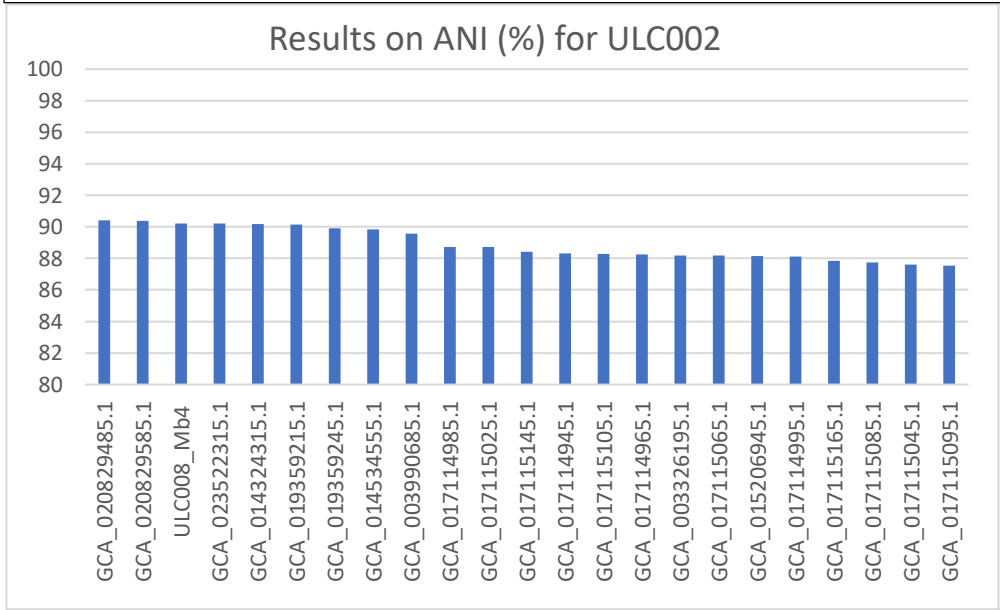
7.2 Annexe 2 : Graphique de synthèse des résultats obtenus par « CONTAMS » sur les 6 bins sélectionnés

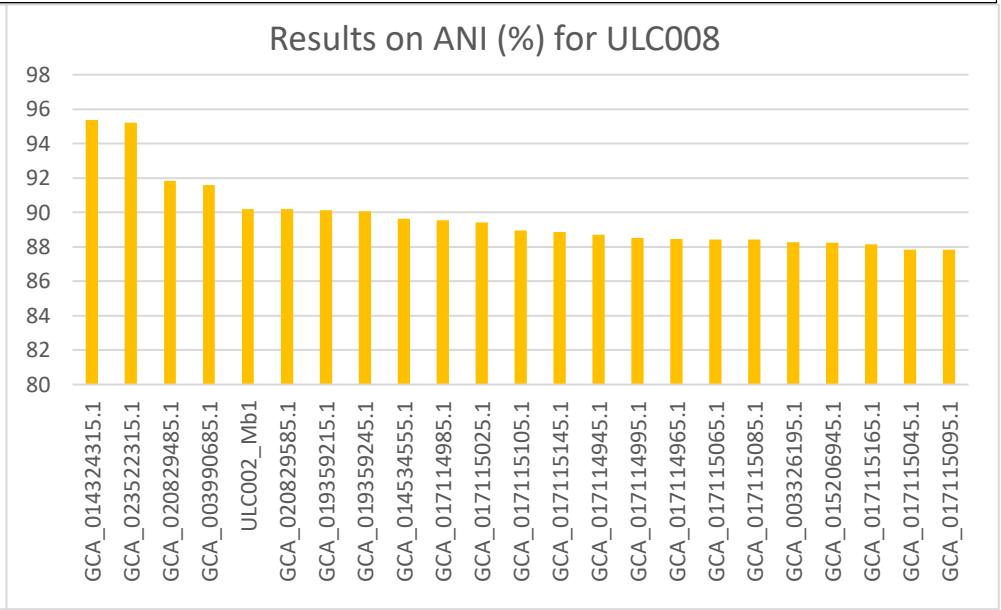


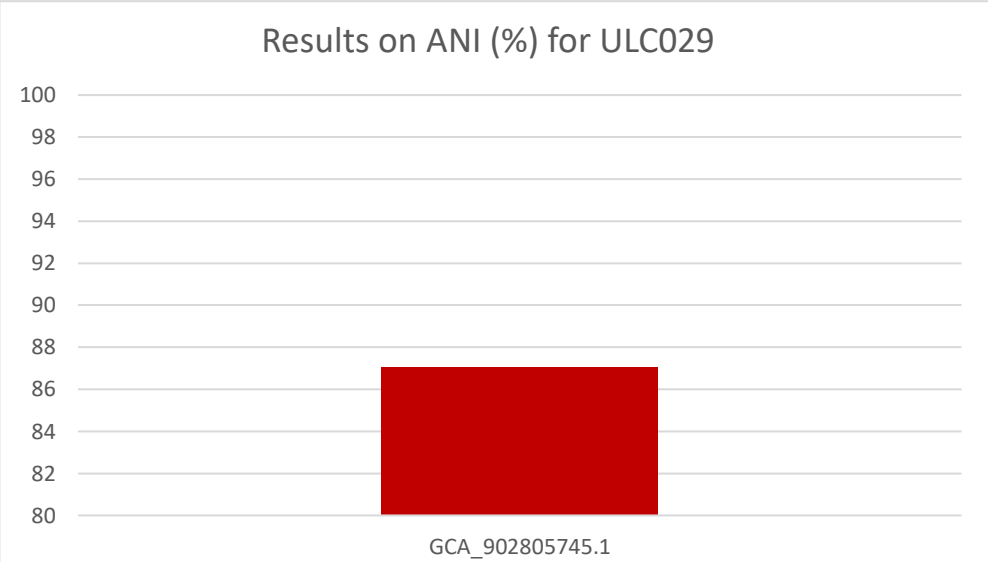
Annexe 1 : Graphique de synthèse des résultats obtenus par les analyses de « CONTAMS » sur les 6 bins sélectionnés. Les résultats sont présentés en pourcentage pour chaque souche sur chaque « bâtonnet ». Le taux de complétude, obtenu par CheckM2, est indiqué en bleu foncé ; le taux de contamination selon CheckM2 en orange ; le taux de contamination selon Kraken en gris, le nombre de contigs en jaune et le pourcentage de GC en bleu clair.

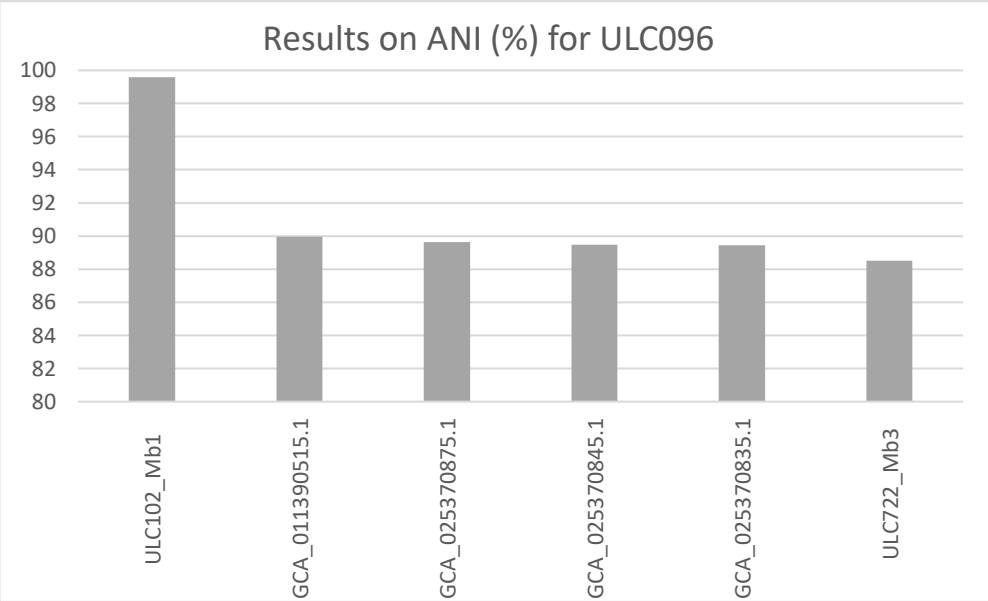
7.3 Annexe 3 : Graphiques des 6 bins indiquant le pourcentage d'ANI des génomes proches des bins.

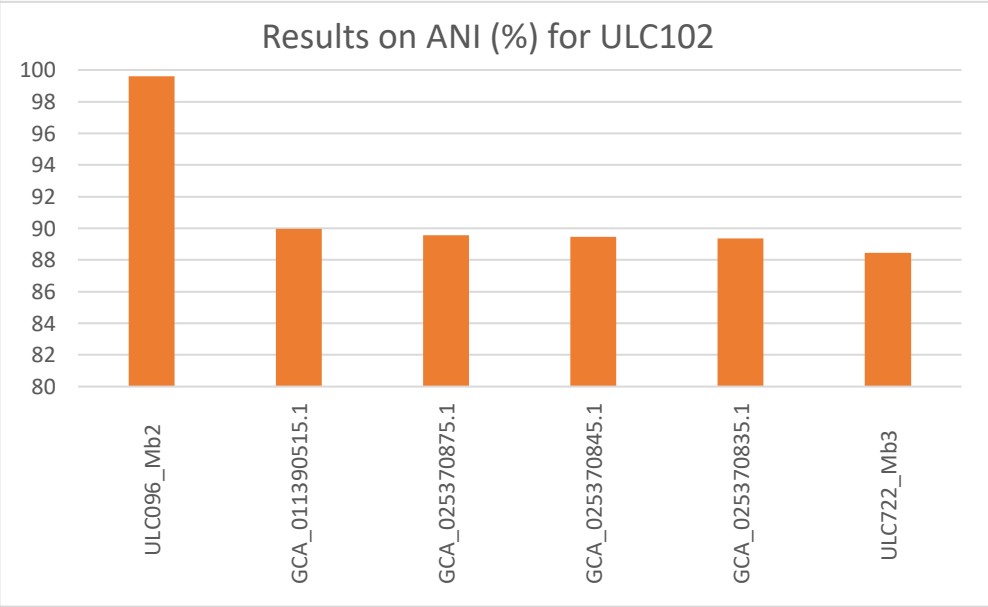
Annexe 2 : Graphiques des 6 bins (ULC002, ULC008, ULC029, ULC096, ULC102, ULC722) indiquant le pourcentage d'ANI des génomes proches des bins. Les génomes sélectionnés avaient un pourcentage supérieur ou égal à 85 % d'ANI.

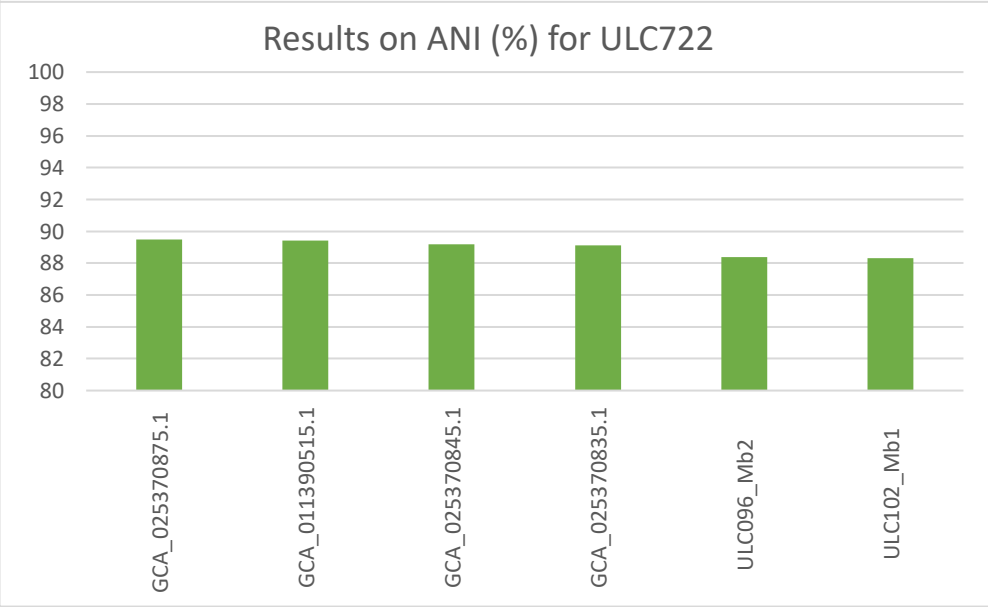




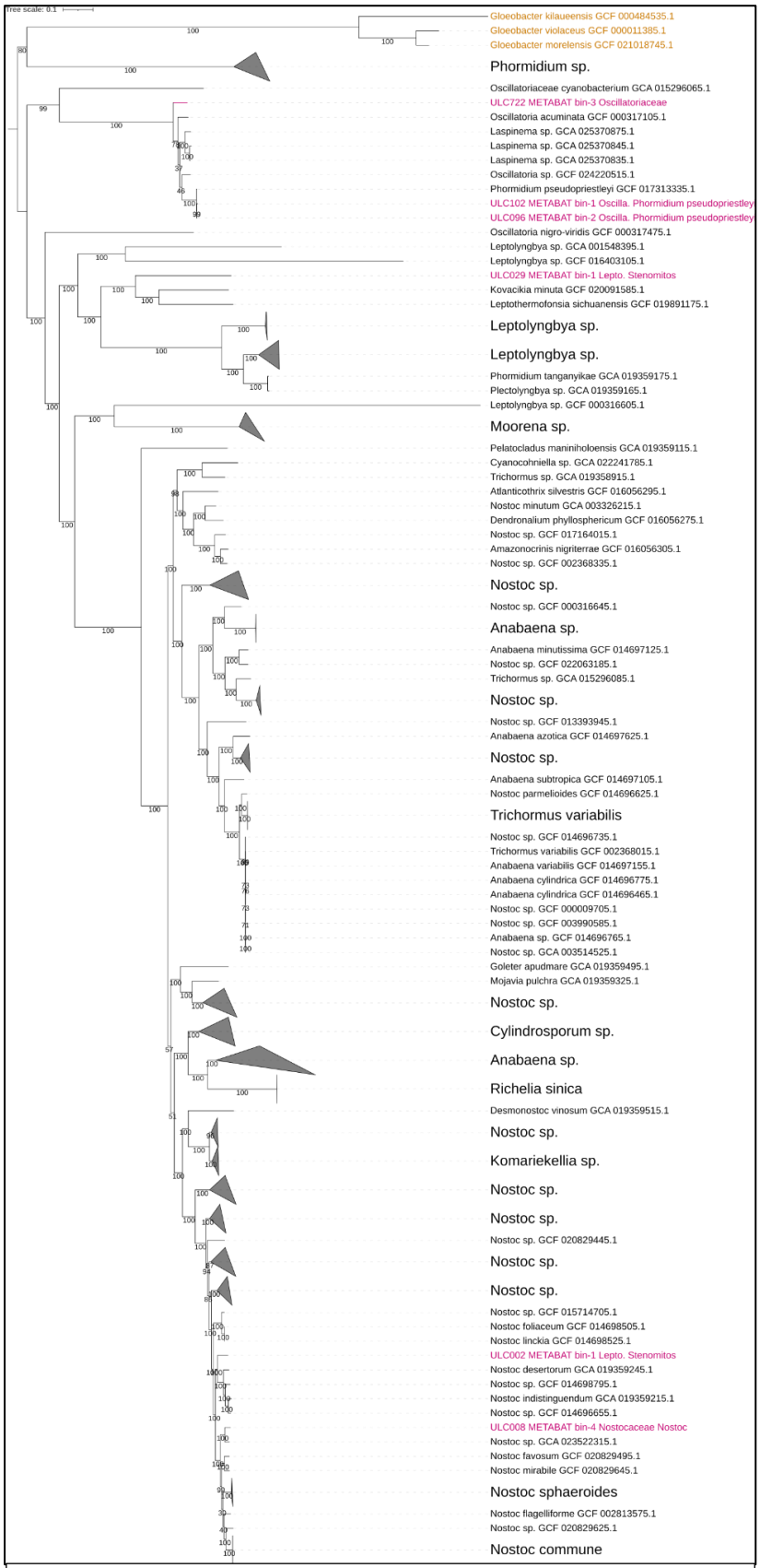

Results on ANI (%) for ULC029
100
98
96
94
92
90
88
86
84
82
80
GCA_902805745.1


Results on ANI (%) for ULC096
100
98
96
94
92
90
88
86
84
82
80
ULC102_Mb1
GCA_011390515.1
GCA_025370875.1
GCA_025370845.1
GCA_025370835.1
ULC722_Mb3


Results on ANI (%) for ULC102
100
98
96
94
92
90
88
86
84
82
80
ULC096_Mb2
GCA_011390515.1
GCA_025370875.1
GCA_025370845.1
GCA_025370835.1
ULC722_Mb3

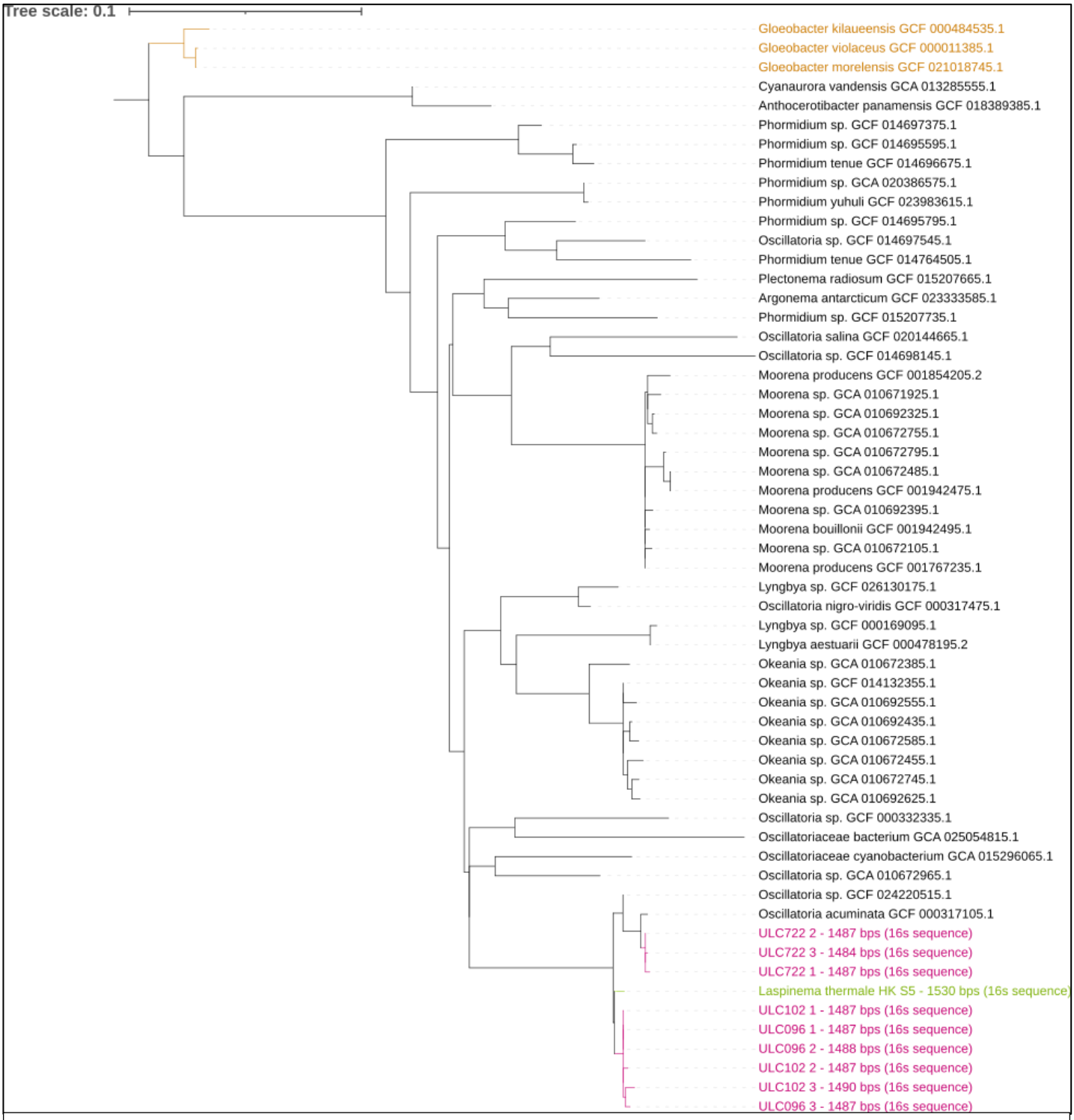

Results on ANI (%) for ULC722
100
98
96
94
92
90
88
86
84
82
80
GCA_025370875.1
GCA_011390515.1
GCA_025370845.1
GCA_025370835.1
ULC096_Mb2
ULC102_Mb1

7.4 Annexe 4 : Le premier arbre phylogénétique des cyanobactéries



Annexe 3 : Le premier arbre phylogénétique des cyanobactéries. Le pourcentage de bootstrap est inscrit sur les branches de l'arbre. Les 6 bins de la collection BCCM/ULC sont colorés en rose. Le groupe externe, les *Gloeobacter*, sont indiqués en jaune-moutarde. Le nom taxonomique des souches est celui déterminés par GTDB. Certaines souches faisant partie du même genre ont été regroupées selon le nom du genre (les triangles en gris).

7.5 Annexe 5 : Le premier arbre phylogénétique des Oscillatoriaceae obtenu par « ORPER »



Annexe 4 : le premier arbre phylogénétique des Oscillatoriaceae obtenu par ORPER, basé sur le gène d'ARNr 16S. Les 3 bins Oscillatoriaceae de la collection BCCM/ULC sont colorés en rose. Chaque bin possède trois séquences de ARNr 16S. La souche de référence, *Laspinema thermale HK S5* de Heidari et al, est présentée en vert. Le groupe externe, les *Gloeobacter*, sont indiqués en jaune-moutarde.

7.6 Annexe 6 : Présentation des 75 premiers résultats obtenus par GGDC, TYGS

Tableau 3 : *présentations des 75 premiers résultats du GGDC via TYGS (Meier-Kolthoff, 2013). Les résultats du DDH numérique est présenté en % avec le coefficient d'incertitude, pour chaque comparaison de génomes.*

Query	Subject	dDDH (%)	CI (%)
'GCF 017313335.1'	'ULC096 Mb2'	97,3	[96.3 - 98.1]
'GCF 017313335.1'	'ULC102 Mb1'	97,1	[96.0 - 97.9]
'ULC096 Mb2'	'ULC102 Mb1'	96,1	[94.7 - 97.1]
'GCA 025370835.1'	'GCA 025370845.1'	77,1	[74.2 - 79.9]
'GCF 024220515.1'	Megalodesulfovibrio paquesii DSM 16681	66	[63.0 - 68.8]
'GCA 025370845.1'	'GCA 025370875.1'	52,1	[49.4 - 54.7]
'GCA 025370835.1'	'GCA 025370875.1'	51,9	[49.2 - 54.5]
'GCF 000317105.1'	'GCF 024220515.1'	41	[38.5 - 43.6]
'GCA 025370845.1'	'GCF 024220515.1'	40,9	[38.4 - 43.4]
'GCA 025370835.1'	'GCF 024220515.1'	40,8	[38.3 - 43.3]
'GCF 000317105.1'	Tildeniella torsiva Uher 1998/13d	40,4	[37.9 - 42.9]
'GCA 025370845.1'	'GCF 000317105.1'	40	[37.5 - 42.5]
'GCA 025370835.1'	'GCF 000317105.1'	39,9	[37.4 - 42.4]
'GCF 017313335.1'	'GCF 024220515.1'	39,8	[37.3 - 42.4]
'GCA 025370875.1'	'GCF 024220515.1'	39,5	[37.0 - 42.0]
'GCA 025370875.1'	'GCF 000317105.1'	39,2	[36.7 - 41.7]
'GCF 024220515.1'	'ULC102 Mb1'	39,1	[36.6 - 41.6]
'GCF 024220515.1'	'ULC096 Mb2'	39,1	[36.6 - 41.6]
'GCA 025370875.1'	'GCF 017313335.1'	38,4	[35.9 - 40.9]
'GCF 000317105.1'	'GCF 017313335.1'	38,1	[35.6 - 40.6]
'GCA 025370835.1'	'GCF 017313335.1'	38	[35.5 - 40.5]
'GCA 025370845.1'	'GCF 017313335.1'	38	[35.6 - 40.5]
'GCA 025370875.1'	'ULC096 Mb2'	38	[35.5 - 40.5]
'GCA 025370875.1'	'ULC102 Mb1'	37,8	[35.4 - 40.4]
'GCF 000317105.1'	'ULC102 Mb1'	37,7	[35.2 - 40.2]
'GCF 000317105.1'	'ULC096 Mb2'	37,6	[35.1 - 40.1]
'GCA 025370845.1'	'ULC096 Mb2'	37,4	[35.0 - 40.0]
'GCA 025370835.1'	'ULC096 Mb2'	37,4	[34.9 - 39.9]
'GCA 025370835.1'	'ULC102 Mb1'	37,4	[34.9 - 39.9]
'GCA 025370845.1'	'ULC102 Mb1'	37,3	[34.8 - 39.8]
'GCA 025370875.1'	'ULC722 Mb3'	36	[33.6 - 38.5]
'GCF 024220515.1'	'ULC722 Mb3'	35,9	[33.5 - 38.4]
'GCF 000317105.1'	'ULC722 Mb3'	35,8	[33.4 - 38.3]
'GCA 025370845.1'	'ULC722 Mb3'	35,5	[33.0 - 38.0]
'GCF 017313335.1'	Aeromonas popoffii CIP 105493	35,5	[33.1 - 38.1]
'GCA 025370835.1'	'ULC722 Mb3'	35,5	[33.0 - 38.0]

'GCF 017313335.1'	'ULC722 Mb3'	35,1	[32.6 - 37.6]
'ULC102 Mb1'	'ULC722 Mb3'	34,8	[32.4 - 37.3]
'ULC096 Mb2'	'ULC722 Mb3'	34,8	[32.3 - 37.3]
'ULC096 Mb2'	Tildeniella torsiva Uher 1998/13d	30,9	[28.5 - 33.4]
'GCA 025370835.1'	Limnospira fusiformis SAG 85.79	30,9	[28.5 - 33.4]
'GCA 025370845.1'	Tildeniella torsiva Uher 1998/13d	30,3	[27.9 - 32.8]
'GCA 025370845.1'	Limnospira fusiformis SAG 85.79	29,6	[27.2 - 32.1]
'GCA 025370845.1'	Brasilonema octagenarum UFV-E1	28,6	[26.2 - 31.1]
'GCF 000317105.1'	Limnospira fusiformis SAG 85.79	28,6	[26.2 - 31.1]
'ULC722 Mb3'	Brevicoccus berkleyi PCC 7336	28,5	[26.1 - 31.0]
'GCF 024220515.1'	Limnospira fusiformis SAG 85.79	28,1	[25.7 - 30.6]
'ULC096 Mb2'	Dulcicalothrix desertica PCC7102	27,8	[25.4 - 30.3]
'ULC102 Mb1'	Dulcicalothrix desertica PCC7102	27,8	[25.5 - 30.3]
'ULC722 Mb3'	Dulcicalothrix desertica PCC7102	27,4	[25.0 - 29.9]
'GCF 000317105.1'	Rippkaea orientalis PCC 8801	26,9	[24.6 - 29.4]
'GCA 025370875.1'	Limnospira fusiformis SAG 85.79	26,8	[24.4 - 29.3]
'ULC722 Mb3'	Limnospira fusiformis SAG 85.79	26,7	[24.3 - 29.2]
'GCA 025370875.1'	Brevicoccus berkleyi PCC 7336	26,4	[24.1 - 28.9]
'GCF 000317105.1'	Dulcicalothrix desertica PCC7102	26,4	[24.0 - 28.9]
'ULC722 Mb3'	Planktothrix serta PCC 8927T	26,3	[23.9 - 28.8]
'GCA 025370845.1'	Brevicoccus berkleyi PCC 7336	26,3	[23.9 - 28.7]
'GCF 017313335.1'	Limnospira fusiformis SAG 85.79	26,3	[23.9 - 28.8]
'GCF 000317105.1'	Brevicoccus berkleyi PCC 7336	26	[23.7 - 28.5]
'ULC096 Mb2'	Xenococcus lajollai PCC 7305	25,9	[23.5 - 28.4]
'ULC102 Mb1'	Xenococcus lajollai PCC 7305	25,9	[23.5 - 28.4]
'GCF 000317105.1'	Neosynechococcus sphagnicola CAUP A 1101	25,9	[23.5 - 28.4]
'GCF 000317105.1'	Spirulina subsalsa PCC 9445	25,8	[23.4 - 28.3]
'ULC102 Mb1'	Rippkaea orientalis PCC 8801	25,8	[23.5 - 28.3]
'GCA 025370835.1'	Brevicoccus berkleyi PCC 7336	25,7	[23.4 - 28.2]
'GCA 025370835.1'	Dulcicalothrix desertica PCC7102	25,7	[23.4 - 28.2]
'ULC722 Mb3'	Xenococcus lajollai PCC 7305	25,7	[23.3 - 28.1]
'GCF 000317105.1'	Brasilonema octagenarum UFV-E1	25,6	[23.3 - 28.1]
'GCA 025370875.1'	Brasilonema octagenarum UFV-E1	25,5	[23.2 - 28.0]
'GCA 025370845.1'	Spirulina subsalsa PCC 9445	25,5	[23.1 - 28.0]
'GCF 000317105.1'	Xenococcus lajollai PCC 7305	25,4	[23.1 - 27.9]
'ULC096 Mb2'	Brevicoccus berkleyi PCC 7336	25,4	[23.1 - 27.9]
'ULC102 Mb1'	Brevicoccus berkleyi PCC 7336	25,4	[23.1 - 27.9]
'ULC102 Mb1'	Limnospira fusiformis SAG 85.79	25,1	[22.8 - 27.6]
'ULC722 Mb3'	Kovacikia minuta CCNU0001	25	[22.7 - 27.5]

7.7 Annexe 7 : Détermination et justification des noms taxonomiques des 13 souches Oscillatoriaceae

Tableau 4 : *présentation des résultats taxonomique obtenus par les 4 banques de données (GTDB, NCBI, SILVA, CyanoSeq). Le choix du nom taxonomique pour chaque souche est écrit en rouge et justifié à chaque fois dans la dernière colonne.*

ID	GTDB Taxonomie	NCBI Taxonomie	SILVA Taxonomie	Cyanoseq Taxonomie	Justification du choix taxonomique
GCA_025054815.1	d__Bacteria; p__Cyanobacteriota; c__Cyanobacteriia; o__Cyanobacteriales; f__Oscillatoriaceae; g__DVEG01; s__DVEG01 sp015295835	Oscillatoriaceae bacterium SKYG93 (uniquement famille)	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Phormidiaceae; uncultured;	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Chroococcales" "Pleurocapsaceae" NA	Le nom de la souche est "Oscillatoriaceae SKYG93". Elle n'a été identifiée qu'au niveau de la "famille". SILVA et CyanoSeq qui utilise la séquence d'ARN 16S ont donné de mauvais résultat. Phormidiaceae n'existe plus mais pourrait être accepté puisque dans le passé, cette famille "hébergeait" la plupart des morphotypes homocytés-filamenteux. En revanche, Pleurocapsaceae est complètement erroné, puisque cette famille comprend uniquement des types unicellulaires, en particulier ceux qui ont une sorte de thalle complexe.
GCF_000332335.1	d__Bacteria; p__Cyanobacteriota; c__Cyanobacteriia; o__Cyanobacteriales; f__Oscillatoriaceae; g__Oscillatoria; s__Oscillatoria sp000332335	*Oscillatoria sp.* PCC 10802	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Oscillatoriaceae; Oscillatoria PCC-10802;	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Oscillatoriales" "Oscillatoriaceae" "Oscillatoria"	Dans ce cas, il y a convergence des données ainsi qu'avec l'article de "Shih et al., 2013 PCC Cyanobacteria genomes". Le nom taxonomique *Oscillatoria* doit être utilisé. A l'origine, c'était *Oscillatoria princeps* NIVA CYA-150, selon O. Skulberg, un taxonomiste

GCA_010672965.1	d__Bacteria; p__Cyanobacteriota; c__Cyanobacteriia; o__Cyanobacteriales; f__PCC-6304; g__SIO1A7; s__SIO1A7 sp010672965	***Oscillatoria sp.*** SIO1A7	Unclassified	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Chroococcales" NA NA	Dans ce cas, toutes les bases de données sont erronées, sauf NCBI. La publication décrivant cette souche la considère comme une *Oscillatoria* (Leão et al., 2021).
GCA_015296065.1	d__Bacteria; p__Cyanobacteriota; c__Cyanobacteriia; o__Cyanobacteriales; f__PCC-6304; g__Koinonema; s__Koinonema sp015296065	Oscillatoriaceae cyanobacterium M33_DOE_052	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Oscillatoriaceae; Phormidium ETS-05;	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Oscillatoriales" "Oscillatoriaceae" **"Koinonema" sp.**	Dans ce cas, les bases de données sont correctes. Cependant, elles ne sont pas mises à jour. Cette souche correspond à la souche ET-05 comme indiqué par SILVA. Cependant, plus récemment, cette souche a été transférée dans le genre *Koinonema* (Buch et al., 2017). Selon Strunecky et al., 2023, ce genre appartient également aux Oscillatoriaceae/Oscillatoriales. Pour être le plus juste et correct possible, le nom Koinonema est choisis.
ULC722 METABAT bin-3	d__Bacteria; p__Cyanobacteria; c__Cyanobacteriia; o__Cyanobacteriales; f__Oscillatoriaceae;g__;s__	Oscillatoriaceae	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Oscillatoriaceae; Oscillatoria PCC-6304	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Oscillatoriales" "Oscillatoriaceae" **"*Laspinema*"**	Le genre *Laspinema* a été récemment décrit sur la base de morphotypes apparentés à *Phormidium* (Heidari et al., 2018). Selon Heidari, la souche PCC6304 est également un *Laspinema*. Une certaine confusion quant à l'identification finale de ce genre peut exister. Cependant, SILVA confirme les suppositions de Heidari et nous pouvons considérer que cette souche et les suivantes doivent être classées comme *Laspinema*.
GCF_024220515.1	d__Bacteria; p__Cyanobacteriota; c__Cyanobacteriia; o__Cyanobacteriales; f__PCC-6304; g__PCC-	*Oscillatoria sp.* HE19RPO	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Oscillatoriaceae;	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Oscillatoriales"	

	6304; s__PCC-6304 sp024220515		Oscillatoria PCC-6304;	"Oscillatoriaceae" "*Laspinema*"	
GCF_000317105.1	d__Bacteria; p__Cyanobacteriota; c__Cyanobacteriia; o__Cyanobacteriales; f__PCC-6304; g__PCC-6304; s__PCC-6304 sp000317105	*Oscillatoria acuminata* PCC 6304	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Oscillatoriaceae; Oscillatoria PCC-6304;	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Oscillatoriales" "Oscillatoriaceae" "*Laspinema*"	
GCA_025370875.1	/	*Laspinema sp.* D2c	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Oscillatoriaceae; Oscillatoria PCC-6304;	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Oscillatoriales" "Oscillatoriaceae" "*Laspinema*"	
GCA_025370845.1	/	*Laspinema sp.* D2c	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Oscillatoriaceae; Oscillatoria PCC-6304;	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Oscillatoriales" "Oscillatoriaceae" "*Laspinema*"	
GCA_025370835.1	/	*Laspinema sp.* D2c	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Oscillatoriaceae;	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Oscillatoriales"	

			Oscillatoria PCC-6304;	"Oscillatoriaceae" "*Laspinema*"	
GCF_017313335.1	d__Bacteria; p__Cyanobacteriota; c__Cyanobacteriia; o__Cyanobacteriales; f__PCC-6304; g__Phormidium; s__Phormidium pseudopriestleyi	*Phormidium pseudopriestleyi* FRX01	/	/	Cette souche est considéré comme faisant partie du groupe de *Laspinema* (visible sur l'arbre phylogénétique). Malheureusement pour cette souche, nous n'avions pas la séquence du gène d'ARNr 16S pour confirmer cette idée avec SILVA.
ULC096 METABAT bin-2	d__Bacteria; p__Cyanobacteria; c__Cyanobacteriia; o__Cyanobacteriales; f__Oscillatoriaceae; g__Phormidium; s__Phormidium pseudopriestleyi	*Phormidium pseudopriestleyi*	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Oscillatoriaceae; Oscillatoria PCC-6304;	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Oscillatoriales" "Oscillatoriaceae" "*Laspinema*"	Les deux souches ULC096 et ULC102 sont dans l'embranchement de l'ensemble des souches *Laspinema* (visible sur l'arbre phylogénétique). SILVA confirme cette hypothèse.
ULC102 METABAT bin-1	d__Bacteria; p__Cyanobacteria; c__Cyanobacteriia; o__Cyanobacteriales; f__Oscillatoriaceae; g__Phormidium; s__Phormidium pseudopriestleyi	*Phormidium pseudopriestleyi*	Bacteria; Cyanobacteria; Cyanobacteriia; Cyanobacteriales; Oscillatoriaceae; Oscillatoria PCC-6304;	"Bacteria" "Cyanobacteriota" "Cyanophyceae" "Oscillatoriales" "Oscillatoriaceae" "*Laspinema*"	

7.8 Annexe 8 : Présentation des gènes spécifiques fonctionnels déterminés par « Metabolic Fonctionnal »

Tableau 5 : *présentation des résultats de « Metabolic fonctionnal ». Chaque groupe d'analyse est présenté (ULC722, Laspinema, Les souches d'Antarctique) avec les différents OGs des 13 souches Oscillatoriaceae. Le « query » représente le numéro d'identifiant de l'analyse réalisée sur les contigs de génome. La colonne « description » reprend l'identification des gènes spécifiques ayant des fonctions. Dans certains cas, une fonction n'a pas été identifiée. « Unclassified » correspond aux gènes qui n'ont pas pu être classés avec précision. « Unknown » correspond aux gènes où aucune fonction connue n'a été identifiée.*

Genome Group	OG	Query	Description
ULC722	OG0022474	ULC722_Mb3@contig_137_pilon_120	Methyltransferase domain
	OG0022605	ULC722_Mb3@contig_137_pilon_2246	7-cyano-7-deazaguanine tRNA-ribosyltransferase
	OG0022854	ULC722_Mb3@contig_137_pilon_6181	Acetyl-coenzyme A synthetase N-terminus
	OG0022879	ULC722_Mb3@contig_137_pilon_6541	Anti-sigma-D factor RsdA to sigma factor binding region
	OG0022513	ULC722_Mb3@contig_137_pilon_700	Bacterial transferase hexapeptide / serine O-acetyltransferase
	OG0022696	ULC722_Mb3@contig_137_pilon_3482	CAAX prenyl protease-related protein
	OG0022507	ULC722_Mb3@contig_137_pilon_628	CAAX protease family protein
	OG0022683	ULC722_Mb3@contig_137_pilon_3358	CBS domain / Divalent cation (Mg) transmembrane transporter
	OG0022774	ULC722_Mb3@contig_137_pilon_4947	Chorion protein S16 / multicellular organism development
	OG0022557	ULC722_Mb3@contig_137_pilon_1455	C-myb, C-terminal
	OG0022902	ULC722_Mb3@contig_140_pilon_82	coagulation factor XIII A1 polypeptide
	OG0022834	ULC722_Mb3@contig_137_pilon_5873	Cobalt transport protein CbiN
	OG0022476	ULC722_Mb3@contig_137_pilon_122	coenzyme F420 hydrogenase subunit beta
	OG0022520	ULC722_Mb3@contig_137_pilon_872	coiled-coil domain-containing protein 151
	OG0022870	ULC722_Mb3@contig_137_pilon_6363	collagen type XXV alpha
	OG0022739	ULC722_Mb3@contig_137_pilon_4076	crotonobetainyl-CoA dehydrogenase
	OG0022803	ULC722_Mb3@contig_137_pilon_5448	Cytochrome P450 / iron heme binding / monooxygenase activity
	OG0022804	ULC722_Mb3@contig_137_pilon_5449	Cytochrome P450 / iron heme binding / monooxygenase activity
	OG0022810	ULC722_Mb3@contig_137_pilon_5489	daunorubicin C-13 ketoreductase
	OG0022478	ULC722_Mb3@contig_137_pilon_140	DegT/DnrJ/EryC1/StrS aminotransferase family
	OG0022556	ULC722_Mb3@contig_137_pilon_1429	Disaggregatase related repeat
	OG0022686	ULC722_Mb3@contig_137_pilon_3387	DNA-sulfur modification-associated
	OG0022504	ULC722_Mb3@contig_137_pilon_570	ER membrane protein complex subunit 1 / U3 small nucleolar RNA-associated protein 4
	OG0022579	ULC722_Mb3@contig_137_pilon_1856	Escherichia phage exonuclease subunit 2
	OG0022550	ULC722_Mb3@contig_137_pilon_1281	exonuclease 1
	OG0022585	ULC722_Mb3@contig_137_pilon_1922	G protein-coupled receptor 55

	OG0022840	ULC722_Mb3@contig_137_pilon_5959	Glycosyl transferase 4-like domain / Glycosyl transferases group 1
	OG0022536	ULC722_Mb3@contig_137_pilon_1091	hemolysin A
	OG0022671	ULC722_Mb3@contig_137_pilon_3218	Keratinocyte-associated protein 2
	OG0022841	ULC722_Mb3@contig_137_pilon_5962	L-malate glycosyltransferase
	OG0022630	ULC722_Mb3@contig_137_pilon_2594	Lymphocryptovirus nuclear antigen 1
	OG0022765	ULC722_Mb3@contig_137_pilon_4751	MAGE-like protein 2
	OG0022691	ULC722_Mb3@contig_137_pilon_3393	MazG nucleotide pyrophosphohydrolase domain
	OG0022678	ULC722_Mb3@contig_137_pilon_3277	MFS transporter, DHA2 family, methylenomycin A resistance protein
	OG0022780	ULC722_Mb3@contig_137_pilon_4984	Natural resistance-associated macrophage protein / manganese transmembrane transport protein
	OG0022604	ULC722_Mb3@contig_137_pilon_2244	NurA domain
	OG0022470	ULC722_Mb3@contig_137_pilon_115	O-Antigen ligase
	OG0022473	ULC722_Mb3@contig_137_pilon_119	peptidoglycan-N-acetylglucosamine deacetylase
	OG0022689	ULC722_Mb3@contig_137_pilon_3391	phospholipase D-like domain-containing protein DpdK
	OG0022795	ULC722_Mb3@contig_137_pilon_5296	Phospholipid-translocating P-type ATPase C-terminal
	OG0022662	ULC722_Mb3@contig_137_pilon_3124	polycystin 1
	OG0022477	ULC722_Mb3@contig_137_pilon_123	Polysaccharide biosynthesis protein / teichuronic acid exporter
	OG0022838	ULC722_Mb3@contig_137_pilon_5938	preprotein translocase subunit SecE
	OG0022687	ULC722_Mb3@contig_137_pilon_3388	protein DpdJ
	OG0022688	ULC722_Mb3@contig_137_pilon_3389	protein DpdJ
	OG0022489	ULC722_Mb3@contig_137_pilon_317	protein kinase MCK1
	OG0022503	ULC722_Mb3@contig_137_pilon_569	protein O-GlcNAc transferase
	OG0022620	ULC722_Mb3@contig_137_pilon_2432	protein O-GlcNAc transferase
	OG0022643	ULC722_Mb3@contig_137_pilon_2780	protein PBMUCL2
	OG0022773	ULC722_Mb3@contig_137_pilon_4925	putative membrane protein
	OG0022502	ULC722_Mb3@contig_137_pilon_521	selectin P ligand
	OG0022592	ULC722_Mb3@contig_137_pilon_2092	Selenoprotein, putative
	OG0022690	ULC722_Mb3@contig_137_pilon_3392	signal recognition particle subunit SRP72
	OG0022719	ULC722_Mb3@contig_137_pilon_3826	Spt5 C-terminal nonapeptide repeat binding Spt4
	OG0022522	ULC722_Mb3@contig_137_pilon_874	Transposase zinc-ribbon domain
	OG0022521	ULC722_Mb3@contig_137_pilon_873	trimerelysin I
	OG0022480	ULC722_Mb3@contig_137_pilon_169	Type IV pilin-like G and H, putative
	OG0022667	ULC722_Mb3@contig_137_pilon_3129	type IV secretion system protein VirB2
	OG0022673	ULC722_Mb3@contig_137_pilon_3235	Unconventional myosin-X coiled coil domain
	OG0022896	ULC722_Mb3@contig_140_pilon_42	Villin headpiece domain / actin binding / cytoskeleton organization
	OG0022669	ULC722_Mb3@contig_137_pilon_3131	voltage-gated sodium channel type X alpha
	OG0022468	ULC722_Mb3@contig_137_pilon_20	Unclassified

	OG0022481	ULC722_Mb3@contig_137_pilon_193	Unclassified
	OG0022482	ULC722_Mb3@contig_137_pilon_246	Unclassified
	OG0022483	ULC722_Mb3@contig_137_pilon_252	Unclassified
	OG0022485	ULC722_Mb3@contig_137_pilon_261	Unclassified
	OG0022494	ULC722_Mb3@contig_137_pilon_372	Unclassified
	OG0022496	ULC722_Mb3@contig_137_pilon_389	Unclassified
	OG0022501	ULC722_Mb3@contig_137_pilon_508	Unclassified
	OG0022505	ULC722_Mb3@contig_137_pilon_576	Unclassified
	OG0022506	ULC722_Mb3@contig_137_pilon_621	Unclassified
	OG0022509	ULC722_Mb3@contig_137_pilon_648	Unclassified
	OG0022510	ULC722_Mb3@contig_137_pilon_653	Unclassified
	OG0022512	ULC722_Mb3@contig_137_pilon_695	Unclassified
	OG0022516	ULC722_Mb3@contig_137_pilon_760	Unclassified
	OG0022518	ULC722_Mb3@contig_137_pilon_773	Unclassified
	OG0022519	ULC722_Mb3@contig_137_pilon_794	Unclassified
	OG0022523	ULC722_Mb3@contig_137_pilon_878	Unclassified
	OG0022525	ULC722_Mb3@contig_137_pilon_903	Unclassified
	OG0022528	ULC722_Mb3@contig_137_pilon_963	Unclassified
	OG0022532	ULC722_Mb3@contig_137_pilon_1059	Unclassified
	OG0022535	ULC722_Mb3@contig_137_pilon_1086	Unclassified
	OG0022538	ULC722_Mb3@contig_137_pilon_1111	Unclassified
	OG0022539	ULC722_Mb3@contig_137_pilon_1124	Unclassified
	OG0022540	ULC722_Mb3@contig_137_pilon_1134	Unclassified
	OG0022541	ULC722_Mb3@contig_137_pilon_1147	Unclassified
	OG0022545	ULC722_Mb3@contig_137_pilon_1207	Unclassified
	OG0022547	ULC722_Mb3@contig_137_pilon_1238	Unclassified
	OG0022548	ULC722_Mb3@contig_137_pilon_1263	Unclassified
	OG0022549	ULC722_Mb3@contig_137_pilon_1267	Unclassified
	OG0022551	ULC722_Mb3@contig_137_pilon_1358	Unclassified
	OG0022552	ULC722_Mb3@contig_137_pilon_1365	Unclassified
	OG0022553	ULC722_Mb3@contig_137_pilon_1415	Unclassified
	OG0022554	ULC722_Mb3@contig_137_pilon_1416	Unclassified
	OG0022558	ULC722_Mb3@contig_137_pilon_1457	Unclassified
	OG0022563	ULC722_Mb3@contig_137_pilon_1539	Unclassified
	OG0022569	ULC722_Mb3@contig_137_pilon_1637	Unclassified
	OG0022570	ULC722_Mb3@contig_137_pilon_1651	Unclassified
	OG0022571	ULC722_Mb3@contig_137_pilon_1668	Unclassified
	OG0022572	ULC722_Mb3@contig_137_pilon_1696	Unclassified
	OG0022575	ULC722_Mb3@contig_137_pilon_1753	Unclassified
	OG0022576	ULC722_Mb3@contig_137_pilon_1788	Unclassified
	OG0022578	ULC722_Mb3@contig_137_pilon_1846	Unclassified
	OG0022580	ULC722_Mb3@contig_137_pilon_1892	Unclassified
	OG0022581	ULC722_Mb3@contig_137_pilon_1911	Unclassified
	OG0022584	ULC722_Mb3@contig_137_pilon_1920	Unclassified
	OG0022586	ULC722_Mb3@contig_137_pilon_1929	Unclassified
	OG0022587	ULC722_Mb3@contig_137_pilon_1955	Unclassified

	OG0022589	ULC722_Mb3@contig_137_pilon_1982	Unclassified
	OG0022590	ULC722_Mb3@contig_137_pilon_2028	Unclassified
	OG0022591	ULC722_Mb3@contig_137_pilon_2031	Unclassified
	OG0022593	ULC722_Mb3@contig_137_pilon_2097	Unclassified
	OG0022596	ULC722_Mb3@contig_137_pilon_2119	Unclassified
	OG0022597	ULC722_Mb3@contig_137_pilon_2125	Unclassified
	OG0022598	ULC722_Mb3@contig_137_pilon_2133	Unclassified
	OG0022601	ULC722_Mb3@contig_137_pilon_2223	Unclassified
	OG0022603	ULC722_Mb3@contig_137_pilon_2237	Unclassified
	OG0022606	ULC722_Mb3@contig_137_pilon_2260	Unclassified
	OG0022607	ULC722_Mb3@contig_137_pilon_2291	Unclassified
	OG0022611	ULC722_Mb3@contig_137_pilon_2321	Unclassified
	OG0022616	ULC722_Mb3@contig_137_pilon_2393	Unclassified
	OG0022621	ULC722_Mb3@contig_137_pilon_2433	Unclassified
	OG0022622	ULC722_Mb3@contig_137_pilon_2484	Unclassified
	OG0022625	ULC722_Mb3@contig_137_pilon_2497	Unclassified
	OG0022627	ULC722_Mb3@contig_137_pilon_2543	Unclassified
	OG0022628	ULC722_Mb3@contig_137_pilon_2553	Unclassified
	OG0022632	ULC722_Mb3@contig_137_pilon_2598	Unclassified
	OG0022633	ULC722_Mb3@contig_137_pilon_2645	Unclassified
	OG0022635	ULC722_Mb3@contig_137_pilon_2657	Unclassified
	OG0022641	ULC722_Mb3@contig_137_pilon_2765	Unclassified
	OG0022642	ULC722_Mb3@contig_137_pilon_2778	Unclassified
	OG0022645	ULC722_Mb3@contig_137_pilon_2824	Unclassified
	OG0022646	ULC722_Mb3@contig_137_pilon_2839	Unclassified
	OG0022647	ULC722_Mb3@contig_137_pilon_2840	Unclassified
	OG0022648	ULC722_Mb3@contig_137_pilon_2846	Unclassified
	OG0022652	ULC722_Mb3@contig_137_pilon_2918	Unclassified
	OG0022654	ULC722_Mb3@contig_137_pilon_2935	Unclassified
	OG0022659	ULC722_Mb3@contig_137_pilon_3046	Unclassified
	OG0022664	ULC722_Mb3@contig_137_pilon_3126	Unclassified
	OG0022665	ULC722_Mb3@contig_137_pilon_3127	Unclassified
	OG0022666	ULC722_Mb3@contig_137_pilon_3128	Unclassified
	OG0022668	ULC722_Mb3@contig_137_pilon_3130	Unclassified
	OG0022670	ULC722_Mb3@contig_137_pilon_3192	Unclassified
	OG0022672	ULC722_Mb3@contig_137_pilon_3226	Unclassified
	OG0022674	ULC722_Mb3@contig_137_pilon_3236	Unclassified
	OG0022677	ULC722_Mb3@contig_137_pilon_3262	Unclassified
	OG0022680	ULC722_Mb3@contig_137_pilon_3315	Unclassified
	OG0022682	ULC722_Mb3@contig_137_pilon_3328	Unclassified
	OG0022694	ULC722_Mb3@contig_137_pilon_3461	Unclassified
	OG0022695	ULC722_Mb3@contig_137_pilon_3474	Unclassified
	OG0022698	ULC722_Mb3@contig_137_pilon_3526	Unclassified
	OG0022699	ULC722_Mb3@contig_137_pilon_3595	Unclassified
	OG0022700	ULC722_Mb3@contig_137_pilon_3602	Unclassified
	OG0022701	ULC722_Mb3@contig_137_pilon_3606	Unclassified

	OG0022702	ULC722_Mb3@contig_137_pilon_3617	Unclassified
	OG0022704	ULC722_Mb3@contig_137_pilon_3664	Unclassified
	OG0022705	ULC722_Mb3@contig_137_pilon_3683	Unclassified
	OG0022706	ULC722_Mb3@contig_137_pilon_3684	Unclassified
	OG0022711	ULC722_Mb3@contig_137_pilon_3742	Unclassified
	OG0022713	ULC722_Mb3@contig_137_pilon_3769	Unclassified
	OG0022715	ULC722_Mb3@contig_137_pilon_3777	Unclassified
	OG0022718	ULC722_Mb3@contig_137_pilon_3816	Unclassified
	OG0022721	ULC722_Mb3@contig_137_pilon_3846	Unclassified
	OG0022723	ULC722_Mb3@contig_137_pilon_3875	Unclassified
	OG0022733	ULC722_Mb3@contig_137_pilon_4012	Unclassified
	OG0022734	ULC722_Mb3@contig_137_pilon_4013	Unclassified
	OG0022736	ULC722_Mb3@contig_137_pilon_4051	Unclassified
	OG0022737	ULC722_Mb3@contig_137_pilon_4065	Unclassified
	OG0022740	ULC722_Mb3@contig_137_pilon_4134	Unclassified
	OG0022742	ULC722_Mb3@contig_137_pilon_4305	Unclassified
	OG0022744	ULC722_Mb3@contig_137_pilon_4400	Unclassified
	OG0022745	ULC722_Mb3@contig_137_pilon_4408	Unclassified
	OG0022747	ULC722_Mb3@contig_137_pilon_4443	Unclassified
	OG0022748	ULC722_Mb3@contig_137_pilon_4460	Unclassified
	OG0022751	ULC722_Mb3@contig_137_pilon_4512	Unclassified
	OG0022752	ULC722_Mb3@contig_137_pilon_4539	Unclassified
	OG0022754	ULC722_Mb3@contig_137_pilon_4578	Unclassified
	OG0022755	ULC722_Mb3@contig_137_pilon_4602	Unclassified
	OG0022756	ULC722_Mb3@contig_137_pilon_4636	Unclassified
	OG0022760	ULC722_Mb3@contig_137_pilon_4662	Unclassified
	OG0022763	ULC722_Mb3@contig_137_pilon_4687	Unclassified
	OG0022766	ULC722_Mb3@contig_137_pilon_4779	Unclassified
	OG0022767	ULC722_Mb3@contig_137_pilon_4790	Unclassified
	OG0022769	ULC722_Mb3@contig_137_pilon_4855	Unclassified
	OG0022771	ULC722_Mb3@contig_137_pilon_4885	Unclassified
	OG0022775	ULC722_Mb3@contig_137_pilon_4949	Unclassified
	OG0022776	ULC722_Mb3@contig_137_pilon_4959	Unclassified
	OG0022784	ULC722_Mb3@contig_137_pilon_5061	Unclassified
	OG0022786	ULC722_Mb3@contig_137_pilon_5154	Unclassified
	OG0022788	ULC722_Mb3@contig_137_pilon_5203	Unclassified
	OG0022789	ULC722_Mb3@contig_137_pilon_5212	Unclassified
	OG0022792	ULC722_Mb3@contig_137_pilon_5258	Unclassified
	OG0022793	ULC722_Mb3@contig_137_pilon_5291	Unclassified
	OG0022794	ULC722_Mb3@contig_137_pilon_5293	Unclassified
	OG0022797	ULC722_Mb3@contig_137_pilon_5341	Unclassified
	OG0022799	ULC722_Mb3@contig_137_pilon_5356	Unclassified
	OG0022801	ULC722_Mb3@contig_137_pilon_5439	Unclassified
	OG0022807	ULC722_Mb3@contig_137_pilon_5458	Unclassified
	OG0022808	ULC722_Mb3@contig_137_pilon_5462	Unclassified
	OG0022809	ULC722_Mb3@contig_137_pilon_5469	Unclassified

	OG0022811	ULC722_Mb3@contig_137_pilon_5495	Unclassified
	OG0022814	ULC722_Mb3@contig_137_pilon_5511	Unclassified
	OG0022817	ULC722_Mb3@contig_137_pilon_5595	Unclassified
	OG0022821	ULC722_Mb3@contig_137_pilon_5679	Unclassified
	OG0022824	ULC722_Mb3@contig_137_pilon_5732	Unclassified
	OG0022825	ULC722_Mb3@contig_137_pilon_5745	Unclassified
	OG0022827	ULC722_Mb3@contig_137_pilon_5762	Unclassified
	OG0022828	ULC722_Mb3@contig_137_pilon_5791	Unclassified
	OG0022829	ULC722_Mb3@contig_137_pilon_5799	Unclassified
	OG0022830	ULC722_Mb3@contig_137_pilon_5822	Unclassified
	OG0022831	ULC722_Mb3@contig_137_pilon_5845	Unclassified
	OG0022832	ULC722_Mb3@contig_137_pilon_5857	Unclassified
	OG0022843	ULC722_Mb3@contig_137_pilon_6040	Unclassified
	OG0022844	ULC722_Mb3@contig_137_pilon_6042	Unclassified
	OG0022848	ULC722_Mb3@contig_137_pilon_6090	Unclassified
	OG0022850	ULC722_Mb3@contig_137_pilon_6123	Unclassified
	OG0022851	ULC722_Mb3@contig_137_pilon_6128	Unclassified
	OG0022852	ULC722_Mb3@contig_137_pilon_6163	Unclassified
	OG0022857	ULC722_Mb3@contig_137_pilon_6258	Unclassified
	OG0022861	ULC722_Mb3@contig_137_pilon_6277	Unclassified
	OG0022863	ULC722_Mb3@contig_137_pilon_6282	Unclassified
	OG0022864	ULC722_Mb3@contig_137_pilon_6294	Unclassified
	OG0022865	ULC722_Mb3@contig_137_pilon_6310	Unclassified
	OG0022867	ULC722_Mb3@contig_137_pilon_6327	Unclassified
	OG0022868	ULC722_Mb3@contig_137_pilon_6349	Unclassified
	OG0022871	ULC722_Mb3@contig_137_pilon_6366	Unclassified
	OG0022872	ULC722_Mb3@contig_137_pilon_6399	Unclassified
	OG0022873	ULC722_Mb3@contig_137_pilon_6427	Unclassified
	OG0022874	ULC722_Mb3@contig_137_pilon_6428	Unclassified
	OG0022876	ULC722_Mb3@contig_137_pilon_6474	Unclassified
	OG0022877	ULC722_Mb3@contig_137_pilon_6487	Unclassified
	OG0022880	ULC722_Mb3@contig_137_pilon_6566	Unclassified
	OG0022885	ULC722_Mb3@contig_137_pilon_6679	Unclassified
	OG0022888	ULC722_Mb3@contig_137_pilon_6744	Unclassified
	OG0022889	ULC722_Mb3@contig_137_pilon_6804	Unclassified
	OG0022890	ULC722_Mb3@contig_137_pilon_6826	Unclassified
	OG0022891	ULC722_Mb3@contig_140_pilon_1	Unclassified
	OG0022893	ULC722_Mb3@contig_140_pilon_21	Unclassified
	OG0022894	ULC722_Mb3@contig_140_pilon_27	Unclassified
	OG0022895	ULC722_Mb3@contig_140_pilon_38	Unclassified
	OG0022897	ULC722_Mb3@contig_140_pilon_46	Unclassified
	OG0022898	ULC722_Mb3@contig_140_pilon_53	Unclassified
	OG0022899	ULC722_Mb3@contig_140_pilon_64	Unclassified
	OG0022900	ULC722_Mb3@contig_140_pilon_65	Unclassified
	OG0022901	ULC722_Mb3@contig_140_pilon_80	Unclassified
	OG0022609	ULC722_Mb3@contig_137_pilon_2309	Unknown

	OG0022467	ULC722_Mb3@contig_137_pilon_6	Unknown
	OG0022471	ULC722_Mb3@contig_137_pilon_117	Unknown
	OG0022479	ULC722_Mb3@contig_137_pilon_164	Unknown
	OG0022508	ULC722_Mb3@contig_137_pilon_639	Unknown
	OG0022531	ULC722_Mb3@contig_137_pilon_1034	Unknown
	OG0022559	ULC722_Mb3@contig_137_pilon_1496	Unknown
	OG0022562	ULC722_Mb3@contig_137_pilon_1534	Unknown
	OG0022600	ULC722_Mb3@contig_137_pilon_2188	Unknown
	OG0022623	ULC722_Mb3@contig_137_pilon_2486	Unknown
	OG0022624	ULC722_Mb3@contig_137_pilon_2490	Unknown
	OG0022638	ULC722_Mb3@contig_137_pilon_2684	Unknown
	OG0022661	ULC722_Mb3@contig_137_pilon_3119	Unknown
	OG0022684	ULC722_Mb3@contig_137_pilon_3384	Unknown
	OG0022685	ULC722_Mb3@contig_137_pilon_3386	Unknown
	OG0022707	ULC722_Mb3@contig_137_pilon_3695	Unknown
	OG0022738	ULC722_Mb3@contig_137_pilon_4075	Unknown
	OG0022746	ULC722_Mb3@contig_137_pilon_4418	Unknown
	OG0022749	ULC722_Mb3@contig_137_pilon_4464	Unknown
	OG0022764	ULC722_Mb3@contig_137_pilon_4726	Unknown
	OG0022779	ULC722_Mb3@contig_137_pilon_4983	Unknown
	OG0022805	ULC722_Mb3@contig_137_pilon_5454	Unknown
	OG0022812	ULC722_Mb3@contig_137_pilon_5498	Unknown
	OG0022815	ULC722_Mb3@contig_137_pilon_5519	Unknown
	OG0022883	ULC722_Mb3@contig_137_pilon_6598	Unknown
Antarctic strains (GCF_017313335.1, ULC096, ULC102)	OG0010098	GCF_017313335.1@NZ_JAFLQW010000170_4	Unclassified
	OG0010099	GCF_017313335.1@NZ_JAFLQW010000493_1	Unclassified
	OG0010101	GCF_017313335.1@NZ_JAFLQW010000359_5	Unclassified
	OG0010102	GCF_017313335.1@NZ_JAFLQW010000407_2	Unclassified
	OG0010103	GCF_017313335.1@NZ_JAFLQW010000676_6	Unclassified
	OG0010104	GCF_017313335.1@NZ_JAFLQW010000329_4	Unclassified
	OG0010118	GCF_017313335.1@NZ_JAFLQW010000031_1	Unclassified
	OG0010119	GCF_017313335.1@NZ_JAFLQW010000515_7	Unclassified
	OG0010120	GCF_017313335.1@NZ_JAFLQW010000580_2	Unclassified
	OG0010121	GCF_017313335.1@NZ_JAFLQW010000348_4	Unclassified
	OG0010122	GCF_017313335.1@NZ_JAFLQW010000569_9	Unclassified
	OG0010123	GCF_017313335.1@NZ_JAFLQW010000569_15	Unclassified
	OG0010124	GCF_017313335.1@NZ_JAFLQW010000569_21	Unclassified
	OG0010125	GCF_017313335.1@NZ_JAFLQW010000587_6	Unclassified
	OG0010126	GCF_017313335.1@NZ_JAFLQW010000524_6	Unclassified
	OG0010128	GCF_017313335.1@NZ_JAFLQW010000583_8	Unclassified
	OG0010129	GCF_017313335.1@NZ_JAFLQW010000583_13	Unclassified
	OG0010130	GCF_017313335.1@NZ_JAFLQW010000449_5	Unclassified
	OG0010132	GCF_017313335.1@NZ_JAFLQW010000199_1	Unclassified
	OG0010134	GCF_017313335.1@NZ_JAFLQW010000408_16	Unclassified
	OG0010139	GCF_017313335.1@NZ_JAFLQW010000632_12	Unclassified
	OG0010140	GCF_017313335.1@NZ_JAFLQW010000225_9	Unclassified

	OG0010141	GCF_017313335.1@NZ_JAFLQW010000466_3	Unclassified
	OG0010142	GCF_017313335.1@NZ_JAFLQW010000674_21	Unclassified
	OG0010143	GCF_017313335.1@NZ_JAFLQW010000617_3	Unclassified
	OG0010144	GCF_017313335.1@NZ_JAFLQW010000205_3	Unclassified
	OG0010145	GCF_017313335.1@NZ_JAFLQW010000497_7	Unclassified
	OG0010146	GCF_017313335.1@NZ_JAFLQW010000390_2	Unclassified
	OG0010147	GCF_017313335.1@NZ_JAFLQW010000090_8	Unclassified
	OG0010148	GCF_017313335.1@NZ_JAFLQW010000355_3	Unclassified
	OG0010149	GCF_017313335.1@NZ_JAFLQW010000018_9	Unclassified
	OG0010150	GCF_017313335.1@NZ_JAFLQW010000322_2	Unclassified
	OG0010151	GCF_017313335.1@NZ_JAFLQW010000278_1	Unclassified
	OG0010152	GCF_017313335.1@NZ_JAFLQW010000142_3	Unclassified
	OG0010154	GCF_017313335.1@NZ_JAFLQW010000117_7	Unclassified
	OG0010155	GCF_017313335.1@NZ_JAFLQW010000081_2	Unclassified
	OG0010156	GCF_017313335.1@NZ_JAFLQW010000514_12	Unclassified
	OG0010157	GCF_017313335.1@NZ_JAFLQW010000219_22	Unclassified
	OG0010158	GCF_017313335.1@NZ_JAFLQW010000219_25	Unclassified
	OG0010161	GCF_017313335.1@NZ_JAFLQW010000443_10	Unclassified
	OG0010164	GCF_017313335.1@NZ_JAFLQW010000640_14	Unclassified
	OG0010168	GCF_017313335.1@NZ_JAFLQW010000456_3	Unclassified
	OG0010169	GCF_017313335.1@NZ_JAFLQW010000259_2	Unclassified
	OG0010170	GCF_017313335.1@NZ_JAFLQW010000057_2	Unclassified
	OG0010172	GCF_017313335.1@NZ_JAFLQW010000057_15	Unclassified
	OG0010173	GCF_017313335.1@NZ_JAFLQW010000026_2	Unclassified
	OG0010174	GCF_017313335.1@NZ_JAFLQW010000026_9	Unclassified
	OG0010176	GCF_017313335.1@NZ_JAFLQW010000234_7	Unclassified
	OG0010177	GCF_017313335.1@NZ_JAFLQW010000234_8	Unclassified
	OG0010179	GCF_017313335.1@NZ_JAFLQW010000209_5	Unclassified
	OG0010180	GCF_017313335.1@NZ_JAFLQW010000678_13	Unclassified
	OG0010181	GCF_017313335.1@NZ_JAFLQW010000479_14	Unclassified
	OG0010184	GCF_017313335.1@NZ_JAFLQW010000139_2	Unclassified
	OG0010185	GCF_017313335.1@NZ_JAFLQW010000364_4	Unclassified
	OG0010186	GCF_017313335.1@NZ_JAFLQW010000364_5	Unclassified
	OG0010188	GCF_017313335.1@NZ_JAFLQW010000364_8	Unclassified
	OG0010194	GCF_017313335.1@NZ_JAFLQW010000473_2	Unclassified
	OG0010195	GCF_017313335.1@NZ_JAFLQW010000335_1	Unclassified
	OG0010196	GCF_017313335.1@NZ_JAFLQW010000357_2	Unclassified
	OG0010197	GCF_017313335.1@NZ_JAFLQW010000236_1	Unclassified
	OG0010198	GCF_017313335.1@NZ_JAFLQW010000236_8	Unclassified
	OG0010199	GCF_017313335.1@NZ_JAFLQW010000614_3	Unclassified
	OG0010200	GCF_017313335.1@NZ_JAFLQW010000462_1	Unclassified
	OG0010201	GCF_017313335.1@NZ_JAFLQW010000462_8	Unclassified
	OG0010203	GCF_017313335.1@NZ_JAFLQW010000283_14	Unclassified
	OG0010204	GCF_017313335.1@NZ_JAFLQW010000283_24	Unclassified
	OG0010205	GCF_017313335.1@NZ_JAFLQW010000154_1	Unclassified
	OG0010209	GCF_017313335.1@NZ_JAFLQW010000650_1	Unclassified

	OG0010212	GCF_017313335.1@NZ_JAFLQW010000270_1	Unclassified
	OG0010218	GCF_017313335.1@NZ_JAFLQW010000541_1	Unclassified
	OG0010219	GCF_017313335.1@NZ_JAFLQW010000439_10	Unclassified
	OG0010220	GCF_017313335.1@NZ_JAFLQW010000013_4	Unclassified
	OG0010222	GCF_017313335.1@NZ_JAFLQW010000303_1	Unclassified
	OG0010225	GCF_017313335.1@NZ_JAFLQW010000202_3	Unclassified
	OG0010227	GCF_017313335.1@NZ_JAFLQW010000332_20	Unclassified
	OG0010228	GCF_017313335.1@NZ_JAFLQW010000301_2	Unclassified
	OG0010229	GCF_017313335.1@NZ_JAFLQW010000301_3	Unclassified
	OG0010231	GCF_017313335.1@NZ_JAFLQW010000667_29	Unclassified
	OG0010236	GCF_017313335.1@NZ_JAFLQW010000423_4	Unclassified
	OG0010237	GCF_017313335.1@NZ_JAFLQW010000635_4	Unclassified
	OG0010238	GCF_017313335.1@NZ_JAFLQW010000609_15	Unclassified
	OG0010242	GCF_017313335.1@NZ_JAFLQW010000249_2	Unclassified
	OG0010244	GCF_017313335.1@NZ_JAFLQW010000444_3	Unclassified
	OG0010245	GCF_017313335.1@NZ_JAFLQW010000340_4	Unclassified
	OG0010246	GCF_017313335.1@NZ_JAFLQW010000619_2	Unclassified
	OG0010249	GCF_017313335.1@NZ_JAFLQW010000115_6	Unclassified
	OG0013251	ULC096_Mb2@contig_26_pilon_1848	Unclassified
	OG0013263	ULC096_Mb2@contig_26_pilon_3144	Unclassified
	OG0013303	ULC096_Mb2@contig_26_pilon_4891	Unclassified
	OG0010108	GCF_017313335.1@NZ_JAFLQW010000610_4	Unknown
	OG0010109	GCF_017313335.1@NZ_JAFLQW010000610_5	Unknown
	OG0010111	GCF_017313335.1@NZ_JAFLQW010000610_8	Unknown
	OG0010112	GCF_017313335.1@NZ_JAFLQW010000610_9	Unknown
	OG0010114	GCF_017313335.1@NZ_JAFLQW010000610_14	Unknown
	OG0010117	GCF_017313335.1@NZ_JAFLQW010000656_2	Unknown
	OG0010137	GCF_017313335.1@NZ_JAFLQW010000571_4	Unknown
	OG0010166	GCF_017313335.1@NZ_JAFLQW010000159_4	Unknown
	OG0010171	GCF_017313335.1@NZ_JAFLQW010000057_8	Unknown
	OG0010175	GCF_017313335.1@NZ_JAFLQW010000026_11	Unknown
	OG0010182	GCF_017313335.1@NZ_JAFLQW010000431_5	Unknown
	OG0010183	GCF_017313335.1@NZ_JAFLQW010000045_2	Unknown
	OG0010187	GCF_017313335.1@NZ_JAFLQW010000364_6	Unknown
	OG0010191	GCF_017313335.1@NZ_JAFLQW010000050_7	Unknown
	OG0010192	GCF_017313335.1@NZ_JAFLQW010000050_8	Unknown
	OG0010193	GCF_017313335.1@NZ_JAFLQW010000050_12	Unknown
	OG0010208	GCF_017313335.1@NZ_JAFLQW010000378_14	Unknown
	OG0010210	GCF_017313335.1@NZ_JAFLQW010000672_5	Unknown
	OG0010213	GCF_017313335.1@NZ_JAFLQW010000027_4	Unknown
	OG0010214	GCF_017313335.1@NZ_JAFLQW010000027_5	Unknown
	OG0010217	GCF_017313335.1@NZ_JAFLQW010000337_8	Unknown
	OG0010224	GCF_017313335.1@NZ_JAFLQW010000523_8	Unknown
	OG0010232	GCF_017313335.1@NZ_JAFLQW010000372_6	Unknown
	OG0010235	GCF_017313335.1@NZ_JAFLQW010000629_3	Unknown
	OG0010243	GCF_017313335.1@NZ_JAFLQW010000240_2	Unknown

	OG0013285	ULC096_Mb2@contig_26_pilon_3916	Unknown
Laspinema group	OG0005368	GCA_025370835.1@JAMXFE010000008_51	Substrate-binding protein MsmE
	OG0005359	GCA_025370835.1@JAMXFE010000007_75	Unclassified
	OG0005360	GCA_025370835.1@JAMXFE010000051_7	Unclassified
	OG0005369	GCA_025370835.1@JAMXFE010000008_91	Unclassified
	OG0005374	GCA_025370835.1@JAMXFE010000009_161	Unclassified
	OG0005377	GCA_025370835.1@JAMXFE010000010_9	Unclassified
	OG0005379	GCA_025370835.1@JAMXFE010000010_27	Unclassified
	OG0005383	GCA_025370835.1@JAMXFE010000010_113	Unclassified
	OG0005385	GCA_025370835.1@JAMXFE010000010_128	Unclassified
	OG0005386	GCA_025370835.1@JAMXFE010000010_144	Unclassified
	OG0005389	GCA_025370835.1@JAMXFE010000061_9	Unclassified
	OG0005402	GCA_025370835.1@JAMXFE010000070_4	Unclassified
	OG0005408	GCA_025370835.1@JAMXFE010000015_44	Unclassified
	OG0005414	GCA_025370835.1@JAMXFE010000078_3	Unclassified
	OG0005415	GCA_025370835.1@JAMXFE010000016_30	Unclassified
	OG0005422	GCA_025370835.1@JAMXFE010000019_32	Unclassified
	OG0005424	GCA_025370835.1@JAMXFE010000019_77	Unclassified
	OG0005425	GCA_025370835.1@JAMXFE010000020_28	Unclassified
	OG0005429	GCA_025370835.1@JAMXFE010000023_69	Unclassified
	OG0005437	GCA_025370835.1@JAMXFE010000025_34	Unclassified
	OG0005439	GCA_025370835.1@JAMXFE010000001_65	Unclassified
	OG0005441	GCA_025370835.1@JAMXFE010000001_145	Unclassified
	OG0005442	GCA_025370835.1@JAMXFE010000001_151	Unclassified
	OG0005443	GCA_025370835.1@JAMXFE010000001_292	Unclassified
	OG0005445	GCA_025370835.1@JAMXFE010000001_312	Unclassified
	OG0005446	GCA_025370835.1@JAMXFE010000001_331	Unclassified
	OG0005457	GCA_025370835.1@JAMXFE010000002_160	Unclassified
	OG0005461	GCA_025370835.1@JAMXFE010000002_250	Unclassified
	OG0005466	GCA_025370835.1@JAMXFE010000003_29	Unclassified
	OG0005470	GCA_025370835.1@JAMXFE010000003_80	Unclassified
	OG0005474	GCA_025370835.1@JAMXFE010000003_237	Unclassified
	OG0005476	GCA_025370835.1@JAMXFE010000003_270	Unclassified
	OG0005478	GCA_025370835.1@JAMXFE010000004_151	Unclassified
	OG0005482	GCA_025370835.1@JAMXFE010000004_247	Unclassified
	OG0005484	GCA_025370835.1@JAMXFE010000005_5	Unclassified
	OG0005485	GCA_025370835.1@JAMXFE010000005_86	Unclassified
	OG0005489	GCA_025370835.1@JAMXFE010000005_133	Unclassified
	OG0005493	GCA_025370835.1@JAMXFE010000005_170	Unclassified
	OG0005496	GCA_025370835.1@JAMXFE010000005_191	Unclassified
	OG0005498	GCA_025370835.1@JAMXFE010000005_230	Unclassified
	OG0005504	GCA_025370835.1@JAMXFE010000006_155	Unclassified
	OG0005507	GCA_025370835.1@JAMXFE010000006_168	Unclassified
	OG0005352	GCA_025370835.1@JAMXFE010000046_21	Unknown
	OG0005356	GCA_025370835.1@JAMXFE010000050_18	Unknown
	OG0005361	GCA_025370835.1@JAMXFE010000051_8	Unknown

	OG0005362	GCA_025370835.1@JAMXFE010000051_15	Unknown
	OG0005365	GCA_025370835.1@JAMXFE010000008_48	Unknown
	OG0005366	GCA_025370835.1@JAMXFE010000008_49	Unknown
	OG0005370	GCA_025370835.1@JAMXFE010000008_111	Unknown
	OG0005371	GCA_025370835.1@JAMXFE010000008_112	Unknown
	OG0005372	GCA_025370835.1@JAMXFE010000008_129	Unknown
	OG0005382	GCA_025370835.1@JAMXFE010000010_108	Unknown
	OG0005387	GCA_025370835.1@JAMXFE010000010_145	Unknown
	OG0005388	GCA_025370835.1@JAMXFE010000010_167	Unknown
	OG0005397	GCA_025370835.1@JAMXFE010000012_11	Unknown
	OG0005398	GCA_025370835.1@JAMXFE010000066_9	Unknown
	OG0005399	GCA_025370835.1@JAMXFE010000068_4	Unknown
	OG0005400	GCA_025370835.1@JAMXFE010000013_43	Unknown
	OG0005403	GCA_025370835.1@JAMXFE010000014_113	Unknown
	OG0005404	GCA_025370835.1@JAMXFE010000071_2	Unknown
	OG0005406	GCA_025370835.1@JAMXFE010000015_12	Unknown
	OG0005411	GCA_025370835.1@JAMXFE010000015_79	Unknown
	OG0005413	GCA_025370835.1@JAMXFE010000015_106	Unknown
	OG0005417	GCA_025370835.1@JAMXFE010000017_23	Unknown
	OG0005419	GCA_025370835.1@JAMXFE010000018_20	Unknown
	OG0005426	GCA_025370835.1@JAMXFE010000021_68	Unknown
	OG0005428	GCA_025370835.1@JAMXFE010000023_13	Unknown
	OG0005430	GCA_025370835.1@JAMXFE010000024_11	Unknown
	OG0005435	GCA_025370835.1@JAMXFE010000025_20	Unknown
	OG0005436	GCA_025370835.1@JAMXFE010000025_24	Unknown
	OG0005440	GCA_025370835.1@JAMXFE010000001_123	Unknown
	OG0005444	GCA_025370835.1@JAMXFE010000001_295	Unknown
	OG0005449	GCA_025370835.1@JAMXFE010000028_49	Unknown
	OG0005452	GCA_025370835.1@JAMXFE010000030_29	Unknown
	OG0005454	GCA_025370835.1@JAMXFE010000032_37	Unknown
	OG0005455	GCA_025370835.1@JAMXFE010000032_38	Unknown
	OG0005459	GCA_025370835.1@JAMXFE010000002_222	Unknown
	OG0005460	GCA_025370835.1@JAMXFE010000002_246	Unknown
	OG0005462	GCA_025370835.1@JAMXFE010000002_263	Unknown
	OG0005463	GCA_025370835.1@JAMXFE010000002_329	Unknown
	OG0005467	GCA_025370835.1@JAMXFE010000003_30	Unknown
	OG0005469	GCA_025370835.1@JAMXFE010000003_75	Unknown
	OG0005473	GCA_025370835.1@JAMXFE010000003_149	Unknown
	OG0005483	GCA_025370835.1@JAMXFE010000037_8	Unknown
	OG0005490	GCA_025370835.1@JAMXFE010000005_135	Unknown
	OG0005491	GCA_025370835.1@JAMXFE010000005_137	Unknown
	OG0005492	GCA_025370835.1@JAMXFE010000005_138	Unknown
	OG0005495	GCA_025370835.1@JAMXFE010000005_184	Unknown
	OG0005497	GCA_025370835.1@JAMXFE010000005_195	Unknown
	OG0005500	GCA_025370835.1@JAMXFE010000042_13	Unknown
	OG0005502	GCA_025370835.1@JAMXFE010000006_122	Unknown

	OG0005505	GCA_025370835.1@JAMXFE010000006_157	Unknown
	OG0005506	GCA_025370835.1@JAMXFE010000006_158	Unknown

7.9 Annexe 9 : Présentation des résultats de qualité de génome des Oscillatoriaceae

Assembly	QUAST RESULTS					CHECKM2 RESULTS		KRAKEN RESULTS
	# contigs	Largest contig	Total length	GC (%)	# N's per 100 kbp	Complétude (%)	CheckM2 – contamination (%)	Kraken - contamination (%)
GCA_025054815.1-abbr	34	534698	3983825	43.92	0	98.82	0.29	12.69
GCF_000332335.1-abbr	9	6929393	8594406	54.10	283.91	99.98	1.02	9.91
GCA_010672965.1-abbr	182	440762	9081372	47.90	19.68	100.0	0.32	10.21
GCA_015296065.1-abbr	317	170020	6239239	49.69	0	99.87	0.0	4.23
ULC722_Mb3-abbr	2	8014245	8097453	48.13	0	99.66	0.56	3.71
GCF_024220515.1-abbr	512	97325	7671599	47.71	0	100.0	0.36	2.72
GCF_000317105.1-abbr	3	7689443	7804270	47.61	0	100.0	0.35	0.01
GCA_025370875.1-abbr	82	658700	7031313	47.97	65.42	100.0	0.52	2.92
GCA_025370845.1-abbr	117	364837	7412545	47.88	74.20	99.99	0.96	3.13
GCA_025370835.1-abbr	106	446035	7372962	48.00	75.95	100.0	0.19	2.88
GCF_017313335.1-abbr	678	44245	5965908	47.43	0	92.88	1.0	3.06
ULC096_Mb2-abbr	1	7064422	7064422	47.31	0	100.0	0.32	3.30
ULC102_Mb1-abbr	1	7246118	7246118	47.13	0	99.99	0.69	3.46

Tableau 6 : *présentation des résultats de qualité de génomes sur les 13 souches Oscillatoriaceae.*

8 Bibliographie

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nature Methods, 11(11), 1144–1146. https://doi.org/10.1038/NMETH.3103

Alves, L. D. F., Westmann, C. A., Lovate, G. L., De Siqueira, G. M. V., Borelli, T. C., & Guazzaroni, M. E. (2018). Metagenomic Approaches for Understanding New Concepts in Microbial Science. International Journal of Genomics, 2018. https://doi.org/10.1155/2018/2312987

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barco, R. A., Garrity, G. M., Scott, J. J., Amend, J. P., Nealson, K. H., & Emerson, D. (2020). A genus definition for bacteria and archaea based on a standard genome relatedness index. MBio, 11(1). https://doi.org/10.1128/MBIO.02475-19/SUPPL_FILE/MBIO.02475-19-S0001.DOCX

Black, P. N., Dirusso, C. C., Metzger, A. K., & Heimert, T. L. (1992). Cloning, Sequencing, and Expression of the fudD Gene of Escherichia coli Encoding Acyl Coenzyme A Synthetase*. Journal of Biological Chemistry, 267(35), 25513–25520. https://doi.org/10.1016/S0021-9258(19)74070-8

Blank, C. E., & SÁnchez-Baracaldo, P. (2010). Timing of morphological and ecological innovations in the cyanobacteria – a key to understanding the rise in atmospheric oxygen. Geobiology, 8(1), 1–23. https://doi.org/10.1111/J.1472-4669.2009.00220.X

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., … Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of

bacteria and archaea. Nature Biotechnology 2017 35:8, 35(8), 725–731. https://doi.org/10.1038/nbt.3893

Buick, R. (2008). When did oxygenic photosynthesis evolve? Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1504), 2731–2743. https://doi.org/10.1098/RSTB.2008.0041

Burford, M. A., Carey, C. C., Hamilton, D. P., Huisman, J., Paerl, H. W., Wood, S. A., & Wulff, A. (2020). Perspective: Advancing the research agenda for improving understanding of cyanobacteria in a future of global change. Harmful Algae, 91, 101601. https://doi.org/10.1016/J.HAL.2019.04.004

Canfield, D. E. (2004). THE EARLY HISTORY OF ATMOSPHERIC OXYGEN: Homage to Robert M. Garrels. Https://Doi.Org/10.1146/Annurev.Earth.33.092203.122711, 33, 1–36. https://doi.org/10.1146/ANNUREV.EARTH.33.092203.122711

Cardona, T., Sánchez-Baracaldo, P., Rutherford, A. W., & Larkum, A. W. (2019). Early Archean origin of Photosystem II. Geobiology, 17(2), 127–150. https://doi.org/10.1111/GBI.12322

Castenholz, R. W., & Garcia-Pichel, F. (2013). Cyanobacterial responses to UV radiation. Ecology of Cyanobacteria II: Their Diversity in Space and Time, 481–499. https://doi.org/10.1007/978-94-007-3855-3_19/TABLES/1

Castenholz, R., & Waterbury, J. (1989). Oxygenic photosynthetic bacteria. Group I. Cyanobacteria.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics, 38(23), 5315–5316. https://doi.org/10.1093/BIOINFORMATICS/BTAC672

Chen, M. Y., Teng, W. K., Zhao, L., Hu, C. X., Zhou, Y. K., Han, B. P., Song, L. R., & Shu, W. S. (2020). Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. The ISME Journal 2020 15:1, 15(1), 211–227. https://doi.org/10.1038/s41396-020-00775-z

Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2022). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. BioRxiv, 2022.07.11.499243. https://doi.org/10.1101/2022.07.11.499243

Ciferri, O., & Tiboni, O. (1985). THE BIOCHEMISTRY AND INDUSTRIAL POTENTIAL OF SPIRULINA. Ann. Rev. Microbiol, 39, 503–529. www.annualreviews.org

Cornet, L., Ahn, A. C., Wilmotte, A., & Baurain, D. (2021). Orper: A workflow for constrained ssu rrna phylogenies. Genes, 12(11), 1741. https://doi.org/10.3390/GENES12111741/S1

Cornet, L., & Baurain, D. (2022). Contamination detection in genomic data: more is not enough. Genome Biology 2022 23:1, 23(1), 1–15. https://doi.org/10.1186/S13059-022-02619-9

Cornet, L., Bertrand, A. R., Hanikenne, M., Javaux, E. J., Wilmotte, A., & Baurain, D. (2018). Metagenomic assembly of new (Sub)polar cyanobacteria and their associated microbiome from non-axenic cultures. Microbial Genomics, 4(9), e000212. https://doi.org/10.1099/MGEN.0.000212/CITE/REFWORKS

Cornet, L., Durieu, B., Baert, F., D'hooge, E., Colignon, D., Meunier, L., Lupo, V., Cleenwerck, I., Daniel, H.-M., Rigouts, L., Sirjacobs, D., Declerck, S., Vandamme, P., Wilmotte, A., Baurain, D., & Becker, P. (2023). The GEN-ERA toolbox: unified and reproducible workflows for research in microbial genomics. 12, 1–10. https://doi.org/10.1093/gigascience/giad022

Cornet, L., Lupo, V., Declerck, S., & Baurain, D. (2022). CRitical Assessment of genomic COntamination detection at several Taxonomic ranks (CRACOT). BioRxiv, 2022.11.14.516442. https://doi.org/10.1101/2022.11.14.516442

Cornet, L., Magain, N., Baurain, D., & Lutzoni, F. (2021). Exploring syntenic conservation across genomes for phylogenetic studies of organisms subjected to horizontal gene transfers: A case study with Cyanobacteria and cyanolichens. Molecular Phylogenetics and Evolution, 162, 107100. https://doi.org/10.1016/J.YMPEV.2021.107100

Cornet, L., Meunier, L., Van Vlierberghe, M., Léonard, R. R., Durieu, B., Lara, Y., Misztak, A., Sirjacobs, D., Javaux, E. J., Philippe, H., Wilmotte, A., & Baurain, D. (2018). Consensus assessment of the contamination level of publicly available cyanobacterial genomes. PLoS ONE, 13(7). https://doi.org/10.1371/JOURNAL.PONE.0200323

Cornet, L., Wilmotte, A., Javaux, E. J., & Baurain, D. (2018). A constrained SSU-rRNA phylogeny reveals the unsequenced diversity of photosynthetic Cyanobacteria (Oxyphotobacteria). BMC Research Notes, 11(1). https://doi.org/10.1186/S13104-018-3543-Y

Couradeau, E., Benzerara, K., Gérard, E., Moreira, D., Bernard, S., Brown, G. E., & López-García, P. (2012). An early-branching microbialite cyanobacterium forms intracellular carbonates. Science (New York, N.Y.), 336(6080), 459–462. https://doi.org/10.1126/SCIENCE.1216171

Criscuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evolutionary Biology, 10(1), 1–21. https://doi.org/10.1186/1471-2148-10-210/FIGURES/9

Criscuolo, A., & Gribaldo, S. (2011). Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. Molecular Biology and Evolution, 28(11), 3019–3032. https://doi.org/10.1093/MOLBEV/MSR108

Dagan, T., Roettger, M., Stucken, K., Landan, G., Koch, R., Major, P., Gould, S. B., Goremykin, V. V., Rippka, R., De Marsac, N. T., Gugger, M., Lockhart, P. J., Allen, J. F., Brune, I., Maus, I., Pühler, A., & Martin, W. F. (2013). Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome Biology and Evolution, 5(1), 31–44. https://doi.org/10.1093/GBE/EVS117

de Vries, J., & Archibald, J. M. (2017). Endosymbiosis: Did Plastids Evolve from a Freshwater Cyanobacterium? Current Biology, 27(3), R103–R105. https://doi.org/10.1016/J.CUB.2016.12.006

Demay, J., Bernard, C., Reinhardt, A., & Marie, B. (2019). Natural Products from Cyanobacteria: Focus on Beneficial Activities. Marine Drugs 2019, Vol. 17, Page 320, 17(6), 320. https://doi.org/10.3390/MD17060320

Demoulin, C. F., Lara, Y. J., Cornet, L., François, C., Baurain, D., Wilmotte, A., & Javaux, E. J. (2019). Cyanobacteria evolution: Insight from the fossil record. Free Radical Biology and Medicine, 140, 206–223. https://doi.org/10.1016/J.FREERADBIOMED.2019.05.007

Denis Baurain. (2021). Bio-MUST-Core-0.212670 - Core classes and utilities for Bio::MUST - metacpan.org. https://metacpan.org/dist/Bio-MUST-Core

Deschamps, P., Colleoni, C., Nakamura, Y., Suzuki, E., Putaux, J. L., Buléon, A., Haebel, S., Ritte, G., Steup, M., Falcón, L. I., Moreira, D., Löffelhardt, W., Raj, J. N., Plancke, C., D'Hulst, C., Dauvillée, D., & Ball, S. (2008). Metabolic Symbiosis and the Birth of the Plant Kingdom. Molecular Biology and Evolution, 25(3), 536–548. https://doi.org/10.1093/MOLBEV/MSM280

Deusch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, K. V., Allen, J. F., Martin, W., & Dagan, T. (2008). Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Molecular Biology and Evolution, 25(4), 748–761. https://doi.org/10.1093/MOLBEV/MSN022

Dextro, R. B., Delbaje, E., Cotta, S. R., Zehr, J. P., & Fiore, M. F. (2021). Trends in Free-access Genomic Data Accelerate Advances in Cyanobacteria Taxonomy. Journal of Phycology, 57(5), 1392–1402. https://doi.org/10.1111/JPY.13200

Dextro, R. B., Delbaje, E., Freitas, P. N. N., Geraldes, V., Pinto, E., Long, P. F., & Fiore, M. F. (2023). Environmental adaptations by the intertidal Antarctic cyanobacterium Halotia branconii CENA392 as revealed using long-read genome sequencing. Limnology and Oceanography Letters. https://doi.org/10.1002/LOL2.10337

DI Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology 2017 35:4, 35(4), 316–319. https://doi.org/10.1038/nbt.3820

Duval, C., Hamlaoui, S., Piquet, B., Toutirais, G., Yéprémian, C., Reinhardt, A., Duperron, S., Marie, B., Demay, J., & Bernard, C. (2020). Characterization of cyanobacteria isolated from thermal muds of Balaruc-Les-Bains (France) and description of a new genus and species Pseudo-chroococcus couteii. BioRxiv, 2020.12.12.422513. https://doi.org/10.1101/2020.12.12.422513

Dvorák, P., Casamatta, D. A., Hašler, P., Jahodárová, E., Norwich, A. R., & Poulícková, A. (2017). Diversity of the cyanobacteria. Modern Topics in the Phototrophic Prokaryotes: Environmental and Applied Aspects, 3–46. https://doi.org/10.1007/978-3-319-46261-5_1/FIGURES/6

Dvořák, P., Hindák, F., Hašler, P., Hindáková, A., & Poulíčková, A. (2014). Morphological and molecular studies of Neosynechococcus sphagnicola, gen. et sp. nov. (Cyanobacteria, Synechococcales). Phytotaxa, 170(1), 24–34. https://doi.org/10.11646/PHYTOTAXA.170.1.3

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792–1797. https://doi.org/10.1093/NAR/GKH340

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biology, 20(1), 1–14. https://doi.org/10.1186/S13059-019-1832-Y/FIGURES/5

Eren, A. M., Esen, O. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: An advanced analysis and visualization platformfor 'omics data. PeerJ, 2015(10). https://doi.org/10.7717/PEERJ.1319/SUPP-5

Farris, J. S. (1982). Outgroups and Parsimony. Systematic Biology, 31(3), 328–334. https://doi.org/10.1093/SYSBIO/31.3.328

Garcia-Pichel, F. (1998). Solar ultraviolet and the evolutionary history of cyanobacteria. Origins of Life and Evolution of the Biosphere, 28(3), 321–347. https://doi.org/10.1023/A:1006545303412/METRICS

Gerwick, W. H., & Fenner, A. M. (2013). Drug discovery from marine microbes. Microbial Ecology, 65(4), 800–806. https://doi.org/10.1007/S00248-012-0169-9

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology, 57(1), 81–91. https://doi.org/10.1099/IJS.0.64483-0/CITE/REFWORKS

Grettenberger, C. L. (2021). Novel Gloeobacterales spp. from Diverse Environments across the Globe . MSphere, 6(4). https://doi.org/10.1128/MSPHERE.00061-21/ASSET/09067C7E-5C5E-4E7D-867A-C956BEB753EF/ASSETS/IMAGES/LARGE/MSPHERE.00061-21-F002.JPG

Groot, P. H., Scholte, H. R., & Hülsmann, W. C. (1976). Fatty Acid Activation: Specificity, Localization, and Function. Advances in Lipid Research, 14, 75–126. https://doi.org/10.1016/B978-0-12-024914-5.50009-7

Guiry, M. D. (2012). How many species of algae are there? Journal of Phycology, 48(5), 1057–1063. https://doi.org/10.1111/J.1529-8817.2012.01222.X

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics (Oxford, England), 29(8), 1072–1075. https://doi.org/10.1093/BIOINFORMATICS/BTT086

Heidari, F., Zima, J., Riahi, H., & Hauer, T. (2018). New simple trichal cyanobacterial taxa isolated from radioactive thermal springs. Fottea, 18(2), 137–149. https://doi.org/10.5507/FOT.2017.024

Hentschke, G. S., & Junior, W. A. G. (2022). Trends in Cyanobacteria: a contribution to systematics and biodiversity studies. The Pharmacological Potential of Cyanobacteria, 1–20. https://doi.org/10.1016/B978-0-12-821491-6.00001-6

Holland, H. D. (2006). The oxygenation of the atmosphere and oceans. Philosophical Transactions of the Royal Society B: Biological Sciences, 361(1470), 903–915. https://doi.org/10.1098/RSTB.2006.1838

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. Human Immunology, 82(11), 801–811. https://doi.org/10.1016/J.HUMIMM.2021.02.012

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11, 119. https://doi.org/10.1186/1471-2105-11-119

Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J. Y., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M., & Philippe, H. (2017). Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nature Ecology & Evolution 2017 1:9, 1(9), 1370–1378. https://doi.org/10.1038/s41559-017-0240-5

Jadhav, L., Phalke, V., Panse, S., Patil, S., & Bankar, A. (2022). Biodiversity of cold-adapted extremophiles from Antarctica and their biotechnological potential. Microbial Diversity and Ecology in Hotspots, 231–265. https://doi.org/10.1016/B978-0-323-90148-2.00013-4

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications 2018 9:1, 9(1), 1–8. https://doi.org/10.1038/s41467-018-07641-9

Javaux, E. (2007). Évolution de la biosphère au Précambrien et implications pour l'astrobiologie. Bulletins de l'Académie Royale de Belgique, 18(1), 15–39. https://doi.org/10.3406/BARB.2007.28580

Jungblut, A. D., Hawes, I., Mackey, T. J., Krusor, M., Doran, P. T., Sumner, D. Y., Eisen, J. A., Hillman, C., & Goroncy, A. K. (2016). Microbial mat communities along an oxygen gradient in a perennially ice-covered Antarctic lake. Applied and Environmental Microbiology, 82(2), 620–630. https://doi.org/10.1128/AEM.02699-15/SUPPL_FILE/ZAM999116849SO1.PDF

Jungblut, A. D., Hawes, I., Mountfort, D., Hitzfeld, B., Dietrich, D. R., Burns, B. P., & Neilan, B. A. (2005). Diversity within cyanobacterial mat communities in variable salinity meltwater ponds of McMurdo Ice Shelf, Antarctica. Environmental Microbiology, 7(4), 519–529. https://doi.org/10.1111/J.1462-2920.2005.00717.X

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1), 27–30. https://doi.org/10.1093/NAR/28.1.27

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ, 7(7). https://doi.org/10.7717/PEERJ.7359

Karsten, U. (2008). Defense strategies of algae and cyanobacteria against solar ultraviolet radiation. Algal Chemical Ecology, 9783540741817, 273–296. https://doi.org/10.1007/978-3-540-74181-7_13/COVER

Kasting, J. F., Pavlov, A. A., & Siefert, J. L. (2001). A coupled ecosystem-climate model for predicting the methane concentration in the archean atmosphere. Origins of Life and Evolution of the Biosphere, 31(3), 271–285. https://doi.org/10.1023/A:1010600401718/METRICS

Khomutovska, N., Sandzewicz, M., Łach, Ł., Suska-Malawska, M., Chmielewska, M., Mazur-Marzec, H., Cegłowska, M., Niyatbekov, T., Wood, S. A., Puddick, J., Kwiatowski, J., & Jasser, I. (2020). Limited Microcystin, Anatoxin and Cylindrospermopsin Production by Cyanobacteria from Microbial Mats in Cold Deserts. Toxins 2020, Vol. 12, Page 244, 12(4), 244. https://doi.org/10.3390/TOXINS12040244

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L., & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. Nature Methods, 17(11), 1103–1110. https://doi.org/10.1038/S41592-020-00971-X

Komárek, J. (2016). A polyphasic approach for the taxonomy of cyanobacteria: principles and applications. Https://Doi.Org/10.1080/09670262.2016.1163738, 51(3), 346–353. https://doi.org/10.1080/09670262.2016.1163738

Komárek, J., & Kaštovský, J. (2003). Coincidences of structural and molecular characters in evolutionary lines of cyanobacteria. Algological Studies/Archiv Für Hydrobiologie, Supplement Volumes, 109, 305–325. https://doi.org/10.1127/1864-1318/2003/0109-0305

Komárek, J., Kaštovský, J., Mareš, J., & Johansen, J. (2014). Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. Preslia. https://www.semanticscholar.org/paper/Taxonomic-classification-of-

cyanoprokaryotes-2014%2C-Kom%C3%A1rek-Ka%C5%A1tovsk%C3%BD/847a4d74565783baf38713fc151ca42124942b45

Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences of the United States of America, 102(7), 2567–2572. https://doi.org/10.1073/PNAS.0409727102/SUPPL_FILE/09727FIG6.PDF

Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Current Opinion in Microbiology, 23, 110–120. https://doi.org/10.1016/J.MIB.2014.11.014

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. PLOS ONE, 12(5), e0177459. https://doi.org/10.1371/JOURNAL.PONE.0177459

Kyrgyzov, O., Prost, V., Gazut, S., Farcy, B., & Brüls, T. (2020). Binning unassembled short reads based on k-mer abundance covariance using sparse coding. GigaScience, 9(4). https://doi.org/10.1093/GIGASCIENCE/GIAA028

Lefler, F. W., Berthold, D. E., & Dail Laughinghouse, H. (2023). CyanoSeq: a database of cyanobacterial 16S rRNA sequences with curated taxonomy. Journal of Phycology. https://doi.org/10.1111/JPY.13335

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research, 49(W1), W293–W296. https://doi.org/10.1093/NAR/GKAB301

Li, B., Lopes, J. S., Foster, P. G., Embley, T. M., & Cox, C. J. (2014). Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. Molecular Biology and Evolution, 31(7), 1697–1709. https://doi.org/10.1093/MOLBEV/MSU105

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research, 13(9), 2178. https://doi.org/10.1101/GR.1224503

Lim, S. C., Tajika, M., Shimura, M., Carey, K. T., Stroud, D. A., Murayama, K., Ohtake, A., & McKenzie, M. (2018). Loss of the Mitochondrial Fatty Acid β-Oxidation Protein Medium-Chain Acyl-Coenzyme A Dehydrogenase Disrupts Oxidative Phosphorylation Protein Complex Stability and Function. Scientific Reports 2017 8:1, 8(1), 1–17. https://doi.org/10.1038/s41598-017-18530-4

Lumian, J. E., Jungblut, A. D., Dillon, M. L., Hawes, I., Doran, P. T., Mackey, T. J., Dick, G. J., Grettenberger, C. L., & Sumner, D. Y. (2021). Metabolic Capacity of the Antarctic Cyanobacterium Phormidium pseudopriestleyi That Sustains Oxygenic Photosynthesis in the Presence of Hydrogen Sulfide. Genes 2021, Vol. 12, Page 426, 12(3), 426. https://doi.org/10.3390/GENES12030426

Mardis, E. R. (2013). Next-Generation Sequencing Platforms. Https://Doi.Org/10.1146/Annurev-Anchem-062012-092628, 6, 287–303. https://doi.org/10.1146/ANNUREV-ANCHEM-062012-092628

Mazard, S., Penesyan, A., Ostrowski, M., Paulsen, I. T., & Egan, S. (2016). Tiny Microbes with a Big Impact: The Role of Cyanobacteria and Their Metabolites in Shaping Our Future. Marine Drugs 2016, Vol. 14, Page 97, 14(5), 97. https://doi.org/10.3390/MD14050097

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P., & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics, 14(1), 1–14. https://doi.org/10.1186/1471-2105-14-60/TABLES/2

Meier-Kolthoff, J. P., Carbasse, J. S., Peinado-Olarte, R. L., & Göker, M. (2022). TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. Nucleic Acids Research, 50(D1), D801–D807. https://doi.org/10.1093/NAR/GKAB902

Meier-Kolthoff, J. P., & Göker, M. (2019). TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nature Communications 2019 10:1, 10(1), 1–10. https://doi.org/10.1038/s41467-019-10210-3

Meunier, L., Tocquin, P., Cornet, L., Sirjacobs, D., Leclère, V., Pupin, M., Jacques, P., & Baurain, D. (2020). Palantir: a springboard for the analysis of secondary metabolite

gene clusters in large-scale genome mining projects. Bioinformatics, 36(15), 4345–4347. https://doi.org/10.1093/BIOINFORMATICS/BTAA517

Mühlsteinova, R., Hauer, T., D e L e y, P., & P i e t r a s i a k, N. (2018). Seeking the true Oscillatoria: A quest for a reliable phylogenetic and taxonomic reference point. Preslia, 90(2), 151–169. https://doi.org/10.23855/PRESLIA.2018.151

Nadeau, T. L., Milbrandt, E. C., & Castenholz, R. W. (2001). EVOLUTIONARY RELATIONSHIPS OF CULTIVATED ANTARCTIC OSCILLATORIANS (CYANOBACTERIA). Journal of Phycology, 37(4), 650–654. https://doi.org/10.1046/J.1529-8817.2001.037004650.X

Nakamura, Y., Kaneko, T., Sato, S., Mimuro, M., Miyashita, H., Tsuchiya, T., Sasamoto, S., Watanabe, A., Kawashima, K., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Nakazaki, N., Shimpo, S., Takeuchi, C., Yamada, M., & Tabata, S. (2003). Complete Genome Structure of Gloeobacter violaceus PCC 7421, a Cyanobacterium that Lacks Thylakoids. DNA Research, 10(4), 137–145. https://doi.org/10.1093/DNARES/10.4.137

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. Genome Research, 27(5), 824–834. https://doi.org/10.1101/GR.213959.116/-/DC1

Ochoa de Alda, J. A. G., Esteban, R., Diago, M. L., & Houmard, J. (2014). The plastid ancestor originated among one of the major cyanobacterial lineages. Nature Communications, 5(1), 4937. https://doi.org/10.1038/ncomms5937

Orakov, A., Fullam, A., Coelho, L. P., Khedkar, S., Szklarczyk, D., Mende, D. R., Schmidt, T. S. B., & Bork, P. (2021). GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biology, 22(1), 1–19. https://doi.org/10.1186/S13059-021-02393-0/FIGURES/3

Oren, A., Arahal, D. R., Rosselló-Móra, R., Sutcliffe, I. C., & Moore, E. R. B. (2021). Emendation of general consideration 5 and rules 18a, 24a and 30 of the international code of nomenclature of prokaryotes to resolve the status of the cyanobacteria in the prokaryotic nomenclature. International Journal of Systematic and Evolutionary

Microbiology, 71(8), 004939. https://doi.org/10.1099/IJSEM.0.004939/CITE/REFWORKS

Palmer, M., Steenkamp, E. T., Blom, J., Hedlund, B. P., & Venter, S. N. (2020). All anis are not created equal: Implications for prokaryotic species boundaries and integration of anis into polyphasic taxonomy. International Journal of Systematic and Evolutionary Microbiology, 70(4), 2937–2948. https://doi.org/10.1099/IJSEM.0.004124/CITE/REFWORKS

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., & Hugenholtz, P. (2019). Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy. BioRxiv, 771964. https://doi.org/10.1101/771964

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25(7), 1043–1055. https://doi.org/10.1101/GR.186072.114

Peeters, K., Verleyen, E., Hodgson, D. A., Convey, P., Ertz, D., Vyverman, W., & Willems, A. (2012). Heterotrophic bacterial diversity in aquatic microbial mat communities from Antarctica. Polar Biology, 35(4), 543–554. https://doi.org/10.1007/S00300-011-1100-4/TABLES/5

Peretó, J. (2011). Origin and evolution of metabolisms. Origins and Evolution of Life, 270–288. https://doi.org/10.1017/CBO9780511933875.020

Pessi, I. S., Pushkareva, E., Lara, Y., Borderie, F., Wilmotte, A., & Elster, J. (2019). Marked Succession of Cyanobacterial Communities Following Glacier Retreat in the High Arctic. Microbial Ecology, 77(1), 136–147. https://doi.org/10.1007/S00248-018-1203-3/FIGURES/5

Ponce-Toledo, R. I., Deschamps, P., López-García, P., Zivanovic, Y., Benzerara, K., & Moreira, D. (2017). An early-branching freshwater cyanobacterium at the origin of plastids. Current Biology : CB, 27(3), 386. https://doi.org/10.1016/J.CUB.2016.11.056

Priscu, J. C., Wolf, C. F., Takacs, C. D., Fritsen, C. H., Laybourn-Parry, J., Roberts, E. C., Sattler, B., & Lyons, W. B. (1999). Carbon Transformations in a Perennially Ice-Covered Antarctic Lake. BioScience, 49(12), 997–1008. https://doi.org/10.1525/BISI.1999.49.12.997

Pruesse, E., Peplies, J., & Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics, 28(14), 1823–1829. https://doi.org/10.1093/BIOINFORMATICS/BTS252

Queirós, P., Delogu, F., Hickl, O., May, P., & Wilmes, P. (2021). Mantis: flexible and consensus-driven genome annotation. GigaScience, 10(6), 1–14. https://doi.org/10.1093/GIGASCIENCE/GIAB042

Quesada, A., & Vincent, W. F. (1997). Strategies of adaptation by antarctic cyanobacteria to ultraviolet radiation. European Journal of Phycology, 32(4), 335–342. https://doi.org/10.1080/09670269710001737269

Ramos, V., Morais, J., & Vasconcelos, V. M. (2017). A curated database of cyanobacterial strains relevant for modern taxonomy and phylogenetic studies. Scientific Data 2017 4:1, 4(1), 1–8. https://doi.org/10.1038/sdata.2017.54

Rippka, R., Deruelles, J., & Waterbury, J. B. (1979). Generic assignments, strain histories and properties of pure cultures of cyanobacteria. Journal of General Microbiology, 111(1), 1–61. https://doi.org/10.1099/00221287-111-1-1/CITE/REFWORKS

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017). The Origin and Diversification of Mitochondria. Current Biology, 27(21), R1177–R1192. https://doi.org/10.1016/J.CUB.2017.09.015

Roure, B., Rodriguez-Ezpeleta, N., & Philippe, H. (2007). SCaFoS: A tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evolutionary Biology, 7(SUPPL. 1), 1–12. https://doi.org/10.1186/1471-2148-7-S1-S2/FIGURES/7

Saary, P., Mitchell, A. L., & Finn, R. D. (2020). Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. Genome Biology, 21(1), 1–21. https://doi.org/10.1186/S13059-020-02155-4/TABLES/1

Sánchez-Baracaldo, P., Bianchini, G., Wilson, J. D., & Knoll, A. H. (2022). Cyanobacteria and biogeochemical cycles through Earth history. Trends in Microbiology, 30(2), 143–157. https://doi.org/10.1016/J.TIM.2021.05.008

Sánchez-Baracaldo, P., & Cardona, T. (2020). On the origin of oxygenic photosynthesis and Cyanobacteria. New Phytologist, 225(4), 1440–1446. https://doi.org/10.1111/NPH.16249

Sánchez-Baracaldo, P., Raven, J. A., Pisani, D., & Knoll, A. H. (2017). Early photosynthetic eukaryotes inhabited low-salinity habitats. Proceedings of the National Academy of Sciences of the United States of America, 114(37), E7737–E7745. https://doi.org/10.1073/PNAS.1620089114/SUPPL_FILE/PNAS.1620089114.SAPP.PDF

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Research, 50(D1), D20–D26. https://doi.org/10.1093/NAR/GKAB1112

Schirrmeister, B. E., Antonelli, A., & Bagheri, H. C. (2011). The origin of multicellularity in cyanobacteria. BMC Evolutionary Biology, 11(1), 1–21. https://doi.org/10.1186/1471-2148-11-45/FIGURES/1

Shestakov, S. V., & Karbysheva, E. A. (2017). The origin and evolution of cyanobacteria. Biology Bulletin Reviews 2017 7:4, 7(4), 259–272. https://doi.org/10.1134/S2079086417040090

Shih, P. M., Wu, D., Latifi, A., Axen, S. D., Fewer, D. P., Talla, E., Calteau, A., Cai, F., Tandeau De Marsac, N., Rippka, R., Herdman, M., Sivonen, K., Coursin, T., Laurent, T., Goodwin, L., Nolan, M., Davenport, K. W., Han, C. S., Rubin, E. M., … Kerfeld, C. A. (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proceedings of the National Academy of Sciences of the United States of America, 110(3), 1053–1058. https://doi.org/10.1073/PNAS.1217107110/-/DCSUPPLEMENTAL/SAPP.PDF

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., & Manuel, M. (2017). A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Current Biology, 27, 958–967. https://doi.org/10.1016/j.cub.2017.02.031

Simon, C., & Daniel, R. (2011). Metagenomic analyses: Past and future trends. Applied and Environmental Microbiology, 77(4), 1153–1161. https://doi.org/10.1128/AEM.02345-10/ASSET/487CD5D2-D675-4DD7-896C-6771D862062D/ASSETS/GRAPHIC/ZAM9991017850001.JPEG

Singh, J. S., Kumar, A., Rai, A. N., & Singh, D. P. (2016). Cyanobacteria: A precious bio-resource in agriculture, ecosystem, and environmental sustainability. Frontiers in Microbiology, 7(APR), 186282. https://doi.org/10.3389/FMICB.2016.00529/BIBTEX

Singh, R., Parihar, P., Singh, M., Bajguz, A., Kumar, J., Singh, S., Singh, V. P., & Prasad, S. M. (2017). Uncovering Potential Applications of Cyanobacteria and Algal Metabolites in Biology, Agriculture and Medicine: Current Status and Future Prospects. Frontiers in Microbiology, 8(APR). https://doi.org/10.3389/FMICB.2017.00515

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), 1312. https://doi.org/10.1093/BIOINFORMATICS/BTU033

Stanier, R. Y., & van Niel, C. B. (1962). The concept of a bacterium. Archiv Fur Mikrobiologie, 42(1), 17–35. https://doi.org/10.1007/BF00425185

Stanojković, A., Skoupý, S., Škaloud, P., & Dvořák, P. (2022). High genomic differentiation and limited gene flow indicate recent cryptic speciation within the genus Laspinema (cyanobacteria). Frontiers in Microbiology, 13, 3389. https://doi.org/10.3389/FMICB.2022.977454/BIBTEX

Strunecký, O., Ivanova, A. P., & Mareš, J. (2023). An updated classification of cyanobacterial orders and families based on phylogenomic and polyphasic analysis. Journal of Phycology, 59(1), 12–51. https://doi.org/10.1111/JPY.13304

Strunecký, O., Wachtlová, M., & Koblížek, M. (2021). Whole genome phylogeny of Cyanobacteria documents a distinct evolutionary trajectory of marine picocyanobacteria. BioRxiv, 2021.05.26.445609. https://doi.org/10.1101/2021.05.26.445609

Sutcliffe, I. C., Tao, L., Ferretti, J. J., & Russell, R. R. B. (1993). MsmE, a lipoprotein involved in sugar transport in Streptococcus mutans. Journal of Bacteriology, 175(6), 1853. https://doi.org/10.1128/JB.175.6.1853-1855.1993

Taton, A., Grubisic, S., Ertz, D., Hodgson, D. A., Piccardi, R., Biondi, N., Tredici, M. R., Mainini, M., Losi, D., Marinelli, F., & Wilmotte, A. (2006). POLYPHASIC STUDY OF ANTARCTIC CYANOBACTERIAL STRAINS1. Journal of Phycology, 42(6), 1257–1270. https://doi.org/10.1111/J.1529-8817.2006.00278.X

Thomas, A. M., & Segata, N. (2019). Multiple levels of the unknown in microbiome research. BMC Biology, 17(1), 1–4. https://doi.org/10.1186/S12915-019-0667-Z/FIGURES/1

van Belkum, A., Scherer, S., van Alphen, L., & Verbrugh, H. (1998). Short-Sequence DNA Repeats in Prokaryotic Genomes. Microbiology and Molecular Biology Reviews, 62(2), 275–293. https://doi.org/10.1128/MMBR.62.2.275-293.1998/ASSET/D770F695-283D-4F06-B6F7-A26B930B63AD/ASSETS/GRAPHIC/MR0280013006.JPEG

Velichko, N., Smirnova, S., Averina, S., & Pinevich, A. (2021). A survey of Antarctic cyanobacteria. Hydrobiologia, 848(11), 2627–2652. https://doi.org/10.1007/S10750-021-04588-9/FIGURES/5

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., … Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science (New York, N.Y.), 304(5667), 66–74. https://doi.org/10.1126/SCIENCE.1093857

Voorhies, A. A., Biddanda, B. A., Kendall, S. T., Jain, S., Marcus, D. N., Nold, S. C., Sheldon, N. D., & Dick, G. J. (2012). Cyanobacterial life at low O(2): community genomics and function reveal metabolic versatility and extremely low diversity in a

Great Lakes sinkhole mat. Geobiology, 10(3), 250–267. https://doi.org/10.1111/J.1472-4669.2012.00322.X

Wait, B. R., Webster-Brown, J. G., Brown, K. L., Healy, M., & Hawes, I. (2006). PChemistry and stratification of Antarctic meltwater ponds I: Coastal ponds near Bratina Island, McMurdo Ice Shelf. Antarctic Science, 18(4), 515–524. https://doi.org/10.1017/S0954102006000563

Waite, D. W., Vanwonterghem, I., Rinke, C., Parks, D. H., Zhang, Y., Takai, K., Sievert, S. M., Simon, J., Campbell, B. J., Hanson, T. E., Woyke, T., Klotz, M. G., & Hugenholtz, P. (2017). Comparative genomic analysis of the class Epsilonproteobacteria and proposed reclassification to epsilonbacteraeota (phyl. nov.). Frontiers in Microbiology, 8(APR), 264454. https://doi.org/10.3389/FMICB.2017.00682/BIBTEX

Watkins, P. A. (1997). Fatty acid activation. Progress in Lipid Research, 36(1), 55–83. https://doi.org/10.1016/S0163-7827(97)00004-0

Watkins, P. A., & Ellis, J. M. (2012). Peroxisomal acyl-CoA synthetases. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1822(9), 1411–1420. https://doi.org/10.1016/J.BBADIS.2012.02.010

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E., Stackebrandt, E., Starr, M. P., & Truper, H. G. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. International Journal of Systematic and Evolutionary Microbiology, 37(4), 463–464. https://doi.org/10.1099/00207713-37-4-463

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. A., Müller, R., Wohlleben, W., Breitling, R., Takano, E., & Medema, M. H. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research, 43(W1), W237–W243. https://doi.org/10.1093/NAR/GKV437

Whitton, B. A. (2012). Ecology of Cyanobacteria II: Their Diversity in Space and Time. Springer Science & Business Media.

https://books.google.be/books?hl=fr&lr=&id=4oJ_vi27s18C&oi=fnd&pg=PR3&ots=JG7i3PrJU_&sig=Imz1LKmlz8DDBn9ardtNodyIJIY&redir_esc=y#v=onepage&q&f=false

Wilmotte, A., Dail Laughinghouse, H. I., Capelli, C., Rippka, R., & Salmaso, N. (2017). Taxonomic Identification of Cyanobacteria by a Polyphasic Approach.

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biology, 20(1), 1–13. https://doi.org/10.1186/S13059-019-1891-0/FIGURES/2

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology, 15(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

Zakhia, F., Jungblut, A. D., Taton, A., Vincent, W. F., & Wilmotte, A. (2008). Cyanobacteria in cold ecosystems. Psychrophiles: From Biodiversity to Biotechnology, 121–135. https://doi.org/10.1007/978-3-540-74335-4_8/COVER

Zanchett, G., & Oliveira-Filho, E. C. (2013). Cyanobacteria and cyanotoxins: from impacts on aquatic ecosystems and human health to anticarcinogenic effects. Toxins, 5(10), 1896–1917. https://doi.org/10.3390/TOXINS5101896

Zimba, P. V., Shalygin, S., Huang, I. S., Momčilović, M., & Abdulla, H. (2021). A new boring toxin producer–Perforafilum tunnelli gen. & sp. nov. (Oscillatoriales, Cyanobacteria) isolated from Laguna Madre, Texas, USA. Phycologia, 60(1), 10–24. https://doi.org/10.1080/00318884.2020.1808389/SUPPL_FILE/UPHY_A_1808389_SM4589.ZIP