Recherche de séquences environnementales inconnues d'intérêt médical/biologique par l'utilisation de grands réseaux de similarité de séquences

Romain Lannes

SORBONNE UNIVERSITÉ

École doctorale Complexité du Vivant - E.D. 515

Institut de Systématique, Évolution, Biodiversité - UMR 7205

Équipe Adaptation, Intégration, Réticulation et Évolution

"RECHERCHE DE SÉQUENCES ENVIRONNEMENTALES INCONNUES D'INTÉRÊT MÉDICAL/BIOLOGIQUE PAR L'UTILISATION DE GRANDS RÉSEAUX DE SIMILARITÉ DE SÉQUENCES"

Par M. Romain Lannes

Thèse de doctorat en Biologie Évolutive

Dirigée par Dr. Éric BAPTESTE et Pr. Philippe LOPEZ

Présentée et soutenue publiquement le 19 novembre 2019

Devant un jury composé de :

Dr. Sakina-Dorothée Ayata (MC, Sorbonne Université)	Examinatrice
Dr. Éric Bapteste (DR, Sorbonne Université)	Encadrant
Dr. Chris Bowler (DR, Ecole Normale Supérieure)	Examinateur
Pr. Claudine Landès (PR, Universite d'Angers)	Rapportrice
Dr. Catherine Larose (CR, Ecole Centrale Lyon)	Rapportrice
Pr. Philippe Lopez (PR, Sorbonne Université)	Encadrant

Table des matières

Table des figures

Remerciements

Merci au jury d'avoir accepté d'évaluer ce travail.

Merci à Eric et philippe de m'avoir fait confiance, d'avoir su dégager du temps pour m'aider, discuter et me conseiller.

Cette thèse fut pour moi une aventure qui m'a beaucoup apporté, sur le plan intellectuel mais aussi personel.

A te voir corriger les copies, Philippe, je me dis qu'il est vraiment magnifique ce métier, professeur.

L'ambiance au laboratoire est vraiment exceptionelle, continuez comme ça! Je n'oubierai jamais les friday works.

Ni Bob qui m'a accompagné et à souffert à mes cotés, cet ordi me remplit de joie!

Merci à toutes les séquences qui ont été désignées volontaires pour être alignées.

Merci beaucoup à Anne-So, Juliette, Thomas, Ophélie, Emile pour les scep et soirées étudiantes, une mention spéciale à Gab pour nos discussions à la fameuse pause thé et pour la découverte du glost.

Merci aux anciens doctorants, Raphaël, Chloé et bien sur Jananan.

Jananan, ça m'a fait très plaisir de te revoir à Roscoff, ce qui ce passe à Roscoff reste à Roscoff =).

Merci beaucoup Jerome, sans toi ce manuscrit serait bien plus pauvre.

Merci à l'équipe, Charles ta thèse va être super, ne lâche rien!

Merci Louise tu as été une super stagiaire, je suis sur que tu seras une super chercheuse.

Guillaume, merci pour ton soutien qui m'a souvent aidé à relativiser et à prendre de la distance.

Andrew, grace à toi I speak english! Bien qu'un peu envieux du Watson effect, j'ai apprécié toutes nos sorties qui m'ont aidé à ne pas exploser.

Eduardo, merci pour toutes nos discussions, et ta patience pour m'expliquer les principes élémentaires en mathématiques. Tu es un didacte formidable. Avec toi, ça change des maths façon apprendre à calculer en cent leçons.

Merci à Danielle pour ton soutien, ta disponibilité, ta patience et ta bonne humeur. Sans ta bonne volonté, beaucoup plus de problèmes m'auraient pris la tête

Merci à ma famille pour son soutien, nos parties de jeu de rôles, nos coups de téléphone. Merci de vous être occupé de mon Druss et merci à lui pour sa bonne humeur inépuisable. A ma grand mère, pour tout, j'aurais tant aimé que tu sois là.

Merci à mes amis pour leur soutien. Plus particulièrement:
Merci Alicia pour nos discussions sur Whatsapp.
Merci au mumble, a.k.a. les poilus, sans vous je ne sais pas si j'aurais survécu à Paris.
Merci Léa et Cedric, d'avoir patagé avec moi ce bout de chemin, vous allez me manquer!
Merci Jeremy pour m'avoir aidé à faire de vrais breaks, SF! tout comme l'hirondelle ne craint pas la traque, on ne craint pas le rhum pirate!
En parlant de SF, merci au SSBN pour tous nos bons moments partagés et tous ces souvenirs.
Merci Gérard et Kévin pour m'avoir supporté quand je n'était pas au top et pour le reste!
Merci à la deuxième section, j'ai beaucoup appris avec vous "Audacieux dans l'action!"
Merci à Gwen, t'étais un escroc mais t'avais le coeur gros comme un maison.
Merci Alexandra, j'éspère que tu l'as retrouvé.
Askip F-J.L y'a Nabilla qui veut te causer OKLM.
Merci Sabinou, t'es la meilleure on repart en voyage quand tu veux!
Merci Emanuelle, sans vous je ne sais pas si je me serais lancé pour faire une thèse.
Merci à tous ceux que j'ai oublié et qui comptent pour moi.

C'est pas tout il faut s'y mettre, un dernier calcul et on s'en va!
"Thanks for all the fish"

Introduction

1.1 Micro-organismes: définition, importance et impact

La microbiologie est une science très dépendante des capacités techniques et technologiques. Depuis sa création, cette discipline a connu plusieurs ruptures entraînées par les progrès techniques. A chaque fois, notre perception du monde et de la vie en a été modifiée et nos sociétés impactées. Les innovations techniques, technologiques et analytiques de cette dernière décennie ont bouleversé, une fois encore, notre approche de la microbiologie. Je souhaite mettre mon travail dans la perspective de ces changements, des questions qu'ils posent et des réponses qu'ils apportent. C'est la raison pour laquelle dans la première partie de cette introduction, j'ai essayé de respecter l'ordre chronologique de l'histoire de la microbiologie.

1.1.1 Nature des micro-organismes

Les procaryotes comprennent les bactéries et les archées. Ce sont des organismes unicellulaires dont les cellules ne possèdent pas de système endomembranaire apparent bien défini (Fig. 1 A). Au contraire, les eucaryotes ont des cellules qui possèdent des compartiments facilement identifiables : un noyau, l'appareil de Golgi, le réticulum endoplasmique, etc…(Fig. 1 A). Le terme "micro-organisme" fait référence à tous les organismes vivants qui ne peuvent être observés à l'œil nu. Cela inclut les procaryotes, certains eucaryotes unicellulaires et parfois dans la littérature certains organismes pluricellulaires (nématode, tardigrade). Les procaryotes ont vraisemblablement un ancêtre commun même si ce groupe n'est pas monophylétique (Fig. 1 B). En effet, le placement phylogénétique des eucaryotes est sujet à débat : ceux ci seraient soit un groupe frère des archées(Fig. 1 B, a), soit un groupe descendant d'une archée (Fig. 1 B, b) (Lake 1988; Woese et al. 1990). Dans les deux cas, le terme procaryote fait référence respectivement à un

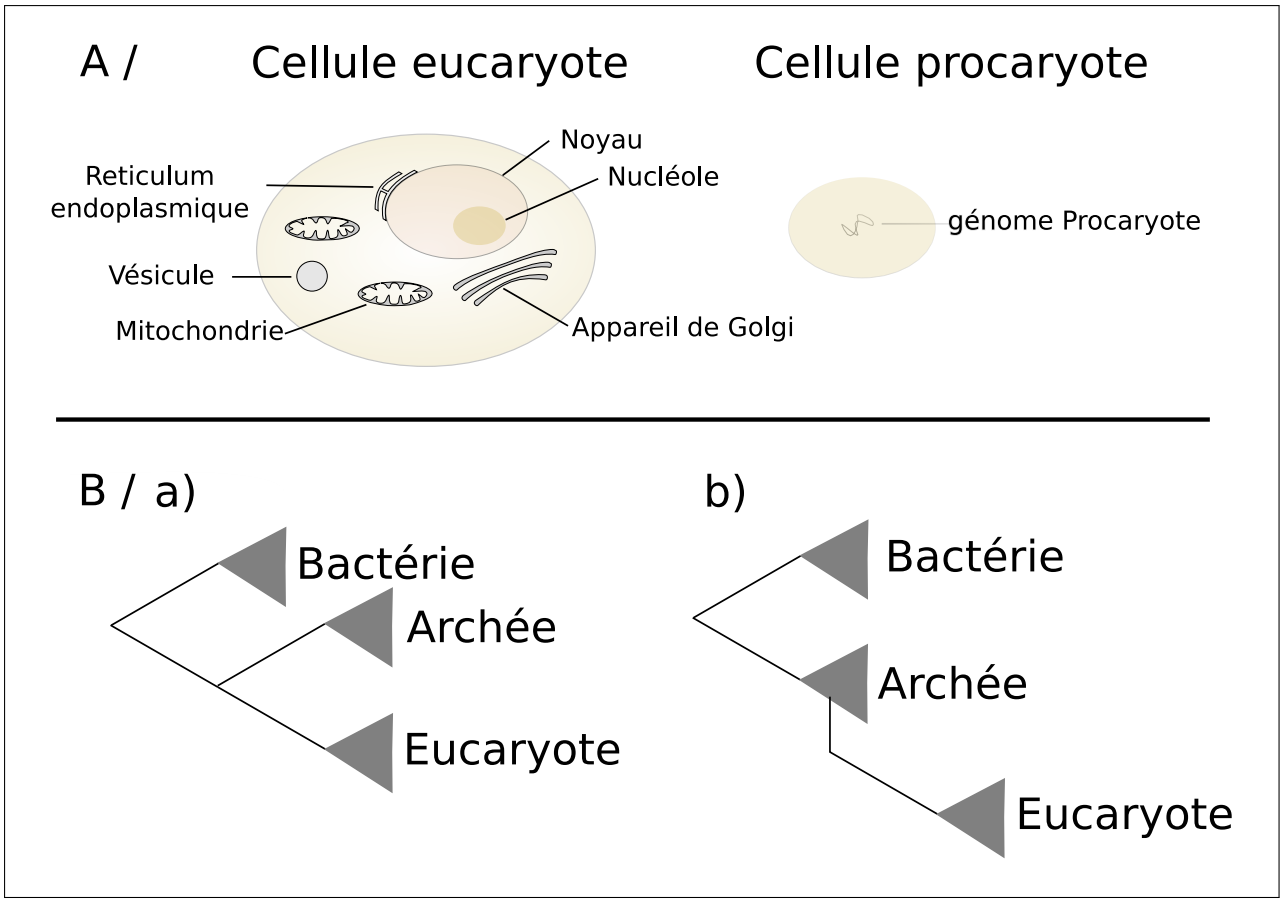


Fig. 1. Organisation de la vie cellulaire
A/ Vue d'une cellule eucaryote et procaryote. B/ Arbre de la vie. a) hypothèse : trois domaines, les Eucaryotes sont un groupe frère des archées. b) hypothèse : deux domaines, les Eucaryotes émergent des archées.

groupe paraphylétique ou à un groupe polyphylétique. Le terme procaryote ne devrait donc pas être utilisé pour classifier les organismes, cependant son utilisation perdure notamment pour dénoter les organismes cellulaires qui ne sont pas des eucaryotes (Pace 2006).

Les micro-organismes eucaryotes comprennent les protistes tels que les algues et les champignons. Lynn Margulis a proposé en 1967 que les organismes eucaryotes aient émergé de l'endosymbiose entre un hôte archée et un symbionte bactérien (Sagan 1967). Des études génétiques ont confirmé cette théorie et identifié l'ancêtre bactérien comme une proto-alphaprotéobactérie (Gray 1988; Gray et al. 1999; Yang et al. 1985). Quand à l'hôte archée, de récentes études de microbiologie environnementale ont mis en évidence un nouveau phylum d'archée, les Lokiarchaeota, qui seraient les descendants de l'hôte archée à l'origine des eucaryotes (Spang et al. 2015). De fait, l'immense majorité des eucaryotes possède une mitochondrie, un organite qui est le vestige de l'endosymbionte bactérien et qui

permet notamment la respiration cellulaire et la production d'Adénosine Tri Phosphate (ATP) (Siekevitz 1957). Chez certains eucaryotes unicellulaires, la mitochondrie a perdu ce rôle et a évolué pour devenir un organite issu d'une mitochondrie (Mitochondria Related Organelle ou MRO dans la littérature) (Landmark et al. 1973; Dyall et al. 2004; Van Der Giezen 2009). Une exception est connue, celle d'un eucaryote qui a perdu l'organite mitochondrial *Monocercomonoides sp.* (Karnkowska et al. 2016). Une autre définition plus inclusive des micro-organismes existe et inclut les entités biologiques non discernables à l'œil nu. Cette définition ajoute les virus et les prions aux micro-organismes. Dans ce manuscrit, nous utiliserons la première définition qui se réfère uniquement aux organismes cellulaires.

1.1.2 Les micro-organismes sont anciens, ubiquitaires et résistants

Les micro-organismes sont abondants et quasiment ubiquitaires sur Terre. D'après des estimations, il y aurait 4 à 6 $x10^{30}$ cellules procaryotes sur notre planète (Whitman et al. 1998), et les micro-organismes représenteraient plus de 50 % de la biomasse totale de notre planète. Les procaryotes sont les plus anciens organismes connus sur Terre et ils se sont adaptés à la plupart des environnements, si ce n'est à tous. Leur présence dans de rares environnements extrêmes fait débat (Belilla et al. 2019; Gómez et al. 2019). Mais, à ces rares exceptions près, des organismes procaryotes ont été retrouvés dans tous les environnements explorés par l'Homme sur Terre, des sources chaudes (Blöchl et al. 1997; Brock et al. 1969) aux plus secs des déserts, (Azua-Bustos et al. 2019; Uritskiy et al. 2019) et même dans des déchets radioactifs (Ferreira et al. 1997). Les procaryotes sont les seuls à avoir réussi à s'adapter à certains environnements trop extrêmes pour les autres formes de vie (Merlino et al. 2018; Pikuta et al. 2007; Pontefract et al. 2017). Par ailleurs, certains procaryotes peuvent passer dans un état dormant extrêmement résistant appelé endospore, lorsque les conditions environnementales deviennent défavorables (Errington 2003). Les endospores peuvent survivre dans des conditions extrêmes : température d'ébullition, dessiccation, rayonnements ultra-violets, froid extrême et résistent même à certains désinfectants modernes, et ce pour de très longues périodes de temps, sans nutriments (Henkin 2016). La plus ancienne endospore ramenée à une forme active en laboratoire avait entre 25 et 40 millions d'années (Cano et al.

1995).

1.1.3 Influence des micro-organismes sur la chimie de la Terre

Les micro-organismes ont, en dépit de leur petite taille, un impact exceptionnel sur l'environnement chimique de la Terre. De petites bactéries photosynthétiques furent responsables de la "Grande Oxydation" qui transforma l'environnement réducteur de la Terre en environnement oxydant il y a 2,3 milliards d'années (Schirrmeister et al. 2015; Gumsley et al. 2017). Ce changement fut la cause de la première extinction de masse connue, tuant les organismes qui n'étaient pas dans des environnements anoxiques et qui n'ont pas réussi à s'adapter à la présence de dérivés réactifs de l'oxygène. Aujourd'hui, la photosynthèse par les micro-organismes marins est responsable de 50 % de la production globale de dioxygène annuelle (Field et al. 1998; Petsch 2013). De façon remarquable, les procaryotes sont les seuls organismes capables d'oxyder le diazote (Kim et al. 1994; Nap et al. 1990; Fowler et al. 2013), un gaz très stable qui n'est pas biodisponible (que les organismes ne peuvent pas utiliser), en azote biodisponible, soutenant ainsi toutes les autres formes de vie. L'azote est un élément essentiel et constitutif des éléments moléculaires fondamentaux de la vie, notamment l'acide désoxyribonucléique (ADN) et les protéines.

1.2 Histoire de la microbiologie

Mon travail de recherche est une conséquence directe des ruptures technologiques que constitue le séquençage à haut débit appliqué à la microbiologie environnementale. Un aperçu de l'histoire de la microbiologie et de son interaction avec nos sociétés est important afin de saisir les enjeux et les questions auxquelles la microbiologie environnementale moderne essaie de répondre. Un aperçu qui a pour but d'introduire une problématique ne saurait être exhaustif. Pour une description plus précise de l'histoire de la microbiologie et des questions qu'elle soulève, le lecteur pourra se référer à l'excellent 'Philosophy of Microbiology' de Maureen O'Malley (O'Malley 2014) (uniquement disponible en anglais) entre autres.

1.2.1 Premières observations

L'Humanité a depuis le début de son histoire une relation ambivalente avec les micro-organismes, et ce sans même avoir conscience de leur existence. Les micro-organismes et les virus sont responsables de la mort de millions voire de milliards d'individus (Whitfield 2002). À l'opposé, avoir un microbiote (ensemble des symbiotes microbiens) sain est un élément essentiel pour être en bonne santé (Olivier et al. 2018; Gong et al. 2019; Round et al. 2009; Postler et al. 2017). De plus, l'Homme a utilisé les micro-organismes pour produire de nombreux produits fermentés depuis la préhistoire. La plus ancienne preuve connue de cette utilisation est la fabrication d'alcool en Chine (7000 av. J.-C.) (McGovern 2004). Il y a également de nombreuses preuves et traces archéologiques de l'utilisation de micro-organismes dans la fabrication d'autres produits comme le pain (Geller J 1993) et les produits laitiers fermentés (Miller 2016). Les micro-organismes ont été observés pour la première fois par Antoni van Leeuwenhoek au 17 ème siècle. Il fut le premier à mettre au point un microscope grossissant 250 à 300 fois, suffisant pour

observer les micro-organismes. Utilisant son invention, van Leeuwenhoek fit les premières descriptions détaillées de ce qu'il appela des animalcules. On ne peut qu'imaginer sa surprise et son émerveillement à la découverte du monde microscopique. En 1683, il écrivit : "Et si quelqu'un devait dire [...] aux gens [...] qu'il y a plus d'animaux vivant sur le dépôt dentaire d'un homme, qu'il y a d'Hommes dans tout un royaume." (citation originale : "what if one should tell [...] people [...] that we have more animals living in the scum of the teeth in a man's mouth, than there are men in a whole kingdom") (O'Malley 2014). À cette époque, la nature et le rôle écologique des animalcules étaient inconnus. Il était aussi admis que les animalcules apparaissaient spontanément à partir de rien, conformément à la théorie de la génération spontanée.

1.2.2 La révolution de la microbiologie, au 19 ème siècle

La théorie de la génération spontanée restera la théorie dominante et communément admise jusqu'aux travaux de Louis Pasteur en 1859. Pasteur mit au point une expérience qui restera célèbre (Fig. 2). Il fit bouillir deux flasques à col de cygne remplies de milieu favorable à la croissance de micro-organismes, puis il cassa le col de l'une des flasques. Dans la flasque intacte, aucun micro-organisme n'apparut, dans la flasque brisée des micro-organismes se développèrent. Cette expérience démontra deux faits: premièrement, il est possible de stériliser un milieu en le faisant bouillir un certain temps, deuxièmement, les animalcules n'apparaissent pas spontanément, sinon ils seraient aussi apparus dans la flasque intacte. Si les animalcules n'apparaissent pas spontanément, de fait ils se reproduisent.

Cette conclusion logique implique des changements importants dans la vision du monde microbien. Si les animalcules se reproduisent et sont des entités biologiques, alors il est certainement possible de les discriminer en "espèces" sur des critères morphologiques. Cela était fait par les naturalistes de l'époque pour les organismes visibles à l'œil nu: les plantes, les animaux et les champignons. Ceci fut dès lors entrepris pour les animalcules. Cohn est, entre autres, un pionnier de la classification morphologique des procaryotes. Il classe les procaryotes en quatre groupes : sphériques, bâtons courts, bâtons longs et spirales. Ce nombre peut sembler trivial mais il est presque exhaustif : la classification

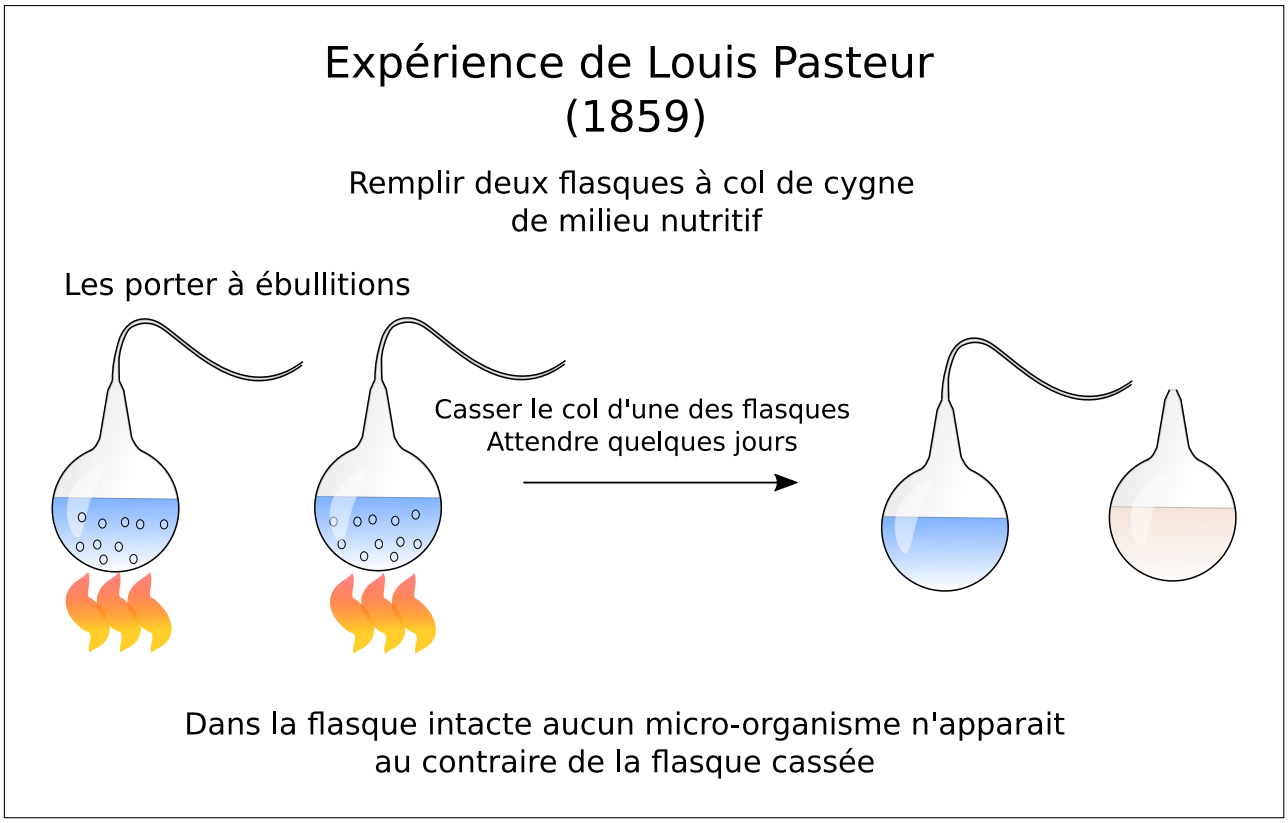


Fig. 2. Réfutation de la théorie de la génération spontanée par L. Pasteur

morphologique actuelle reconnaît cinq groupes : sphériques, bâtons et spirales, spirale longue, virgule[1]. En 1874, Bastian décrivit des bactéries (Bacillus subtilis) qui apparaissent dans un milieu isolé et bouilli, c.-à-d. stérilisé. Bastian utilisa cette observation pour défendre la théorie de la génération spontanée, précédemment discréditée. Cohn observa que les bactéries (Bacillus subtilis) présentes dans ce milieu "stérilisé" étaient toutes en forme de bâton mais comprenaient des corps intracellulaires de forme ovale. Cohn postula que ces corps faisaient partie du cycle de vie de ces bactéries et les appela spores. Cohn et Koch rapportèrent indépendamment en 1876 que ces spores sont résistantes à l'ébullition. Cette découverte fut la fin de la théorie de la génération spontanée. Ces spores sont aujourd'hui appelées endospores car, contrairement aux spores, elles ne font pas partie d'un cycle de reproduction. On ne peut pas quitter le 19 ème siècle sans parler de la contribution majeure de Robert Koch à la microbiologie. Robert Koch développa une méthode de culture pure de micro-organismes sur un milieu d'agar solide. La culture en laboratoire d'une espèce unique en conditions contrôlées amorça une ère permettant l'obtention facile de clones,

[1] https://microbiologyonline.org/about-microbiology/introducing-microbes/bacteria)

la répétabilité expérimentale (i.e. travail sur des individus qui sont des clones) mais aussi la caractérisation métabolique et l'identification des nutriments essentiels aux souches de micro-organismes isolées, en ajoutant un nouveau caractère, en plus de la morphologie, pour leur classification. Cependant, les critères morphologiques et métaboliques échouent à donner une taxonomie en accord avec les relations évolutives. Des organismes proches morphologiquement peuvent avoir des caractères métaboliques différents et inversement. Néanmoins, on comprend que la classification morphologique, au vu du faible nombre de classes, ne suffira pas à retracer les relations évolutives des groupes. Le 19 ème siècle a donc été une période de progrès majeurs pour la microbiologie, qui a changé notre perception du monde microbien, composé non d'animalcules apparaissant de façon spontanée mais d'organismes vivants se reproduisant. Ces progrès eurent de très nombreuses applications, qui transformèrent nos sociétés et notamment : la vaccination, l'hygiène, la pasteurisation et la stérilisation. La microbiologie moderne était née. Une sélection de dates majeures de la microbiologie est présentée figure 3.

La plupart des progrès majeurs de la microbiologie au 20 ème siècle ont été rendus possibles par les cultures pures des micro-organismes. Ce changement d'approche, de l'étude des micro-organismes dans leurs environnements à l'étude des micro-organismes en culture pure dans un milieu contrôlé, a laissé une empreinte profonde sur nos connaissances des micro-organismes. Les micro-organismes ont été étudié dans un environnement contrôlé sans compétition ni interaction avec d'autres micro-organismes. De plus, seuls les micro-organismes cultivables, qui peuvent croître et se multiplier dans de telles conditions ont été caractérisés. La découverte d'organismes qui ne sont pas cultivables dans un environnement contrôlé, où les nutriments essentiels sont présents présenterait un défi à notre compréhension de la vie microbienne. De tels organismes existent-ils? Ils existent et sont même majoritaires. Cette affirmation est une déduction de l'anomalie de comptage de plaques (Staley et al. 1985). Quand on isole des micro-organismes depuis un prélèvement environnemental, on commence par estimer le nombre de micro-organismes (particules dans le milieu), puis on réalise une dilution limite ; on dilue le milieu de façon à avoir en moyenne un micro-organisme par unité de volume. Enfin, chaque plaque d'agar est ensemencée avec

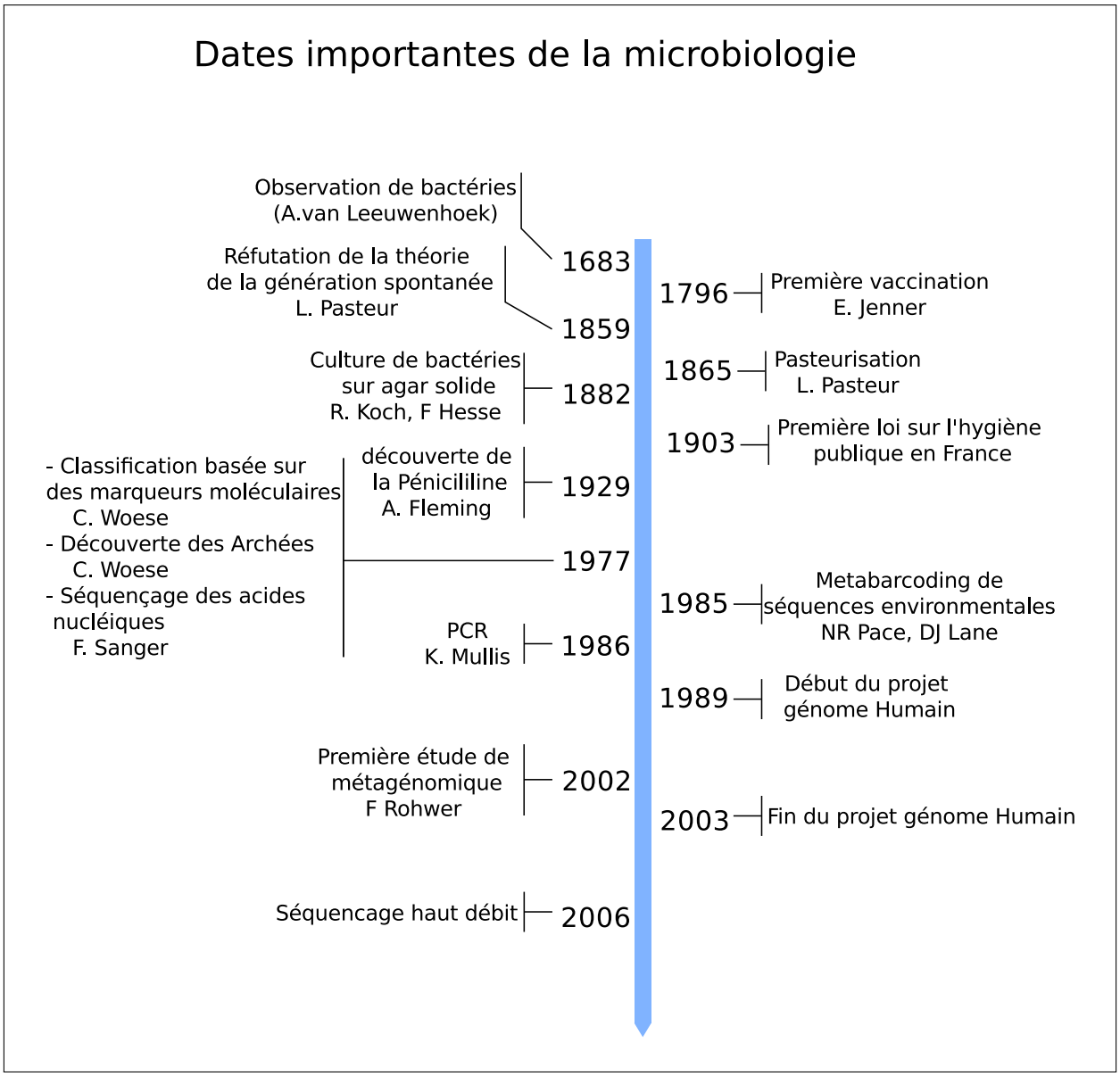


Fig. 3. Dates importantes de la microbiologie

une unité de volume, c.-à-d. une particule, puis on laisse les plaques à la température de l'environnement prélevé. Les micro-organismes isolés forment des colonies sur les plaques facilement identifiables à l'œil. Seulement 1 à 5 % des plaques présentent une colonie, c'est l'anomalie de comptage de plaques. Les micro-organismes sur les autres plaques ne se sont pas répliqués ou n'ont pas survécu: ils sont dits non cultivables. Jusqu'au début des années 2000, la plupart de nos connaissances sur les micro-organismes ont été déduites de l'étude de micro-organismes cultivables en milieu isolé et contrôlé. L'étude des micro-organismes dans l'environnement ne s'est pas arrêtée mais elle est passée en second plan. Par conséquent, notre savoir sur la biologie et les interactions des micro-organismes dans l'environnement

est probablement incomplet. Notre connaissance des micro-organismes non cultivables, qui sont majoritaires dans l'environnement, est également lacunaire et de nombreuses questions restent sans réponse : quelle est leur diversité réelle? Quel sont leurs rôles et leurs impacts écologiques?

1.3 Biologie moléculaire et microbiologie

La biologie moléculaire marque, de mon point de vue la troisième rupture technologique de l'Histoire de la microbiologie, la première étant la découverte des microbes par Leeuwenhoek (1.2.1), et la deuxième la réfutation de la génération spontanée par Pasteur et la mise en culture de micro-organismes par Koch (1.2.2). Dans cette partie, nous allons nous intéresser à l'information génétique, et à ce qu'elle peut nous apprendre sur la nature et les relations entre les organismes. La biologie moléculaire et la microbiologie sont en effet étroitement liées. Les micro-organismes isolés en laboratoire sont un formidable modèle d'étude pour la biologie moléculaire. Ils sont faciles à manipuler et leurs populations doublent environ toutes les 30 minutes (Allen et al. 2019). Les outils de la biologie moléculaire proviennent pour la majorité de micro-organismes. Ces outils permettent d'étudier, de modifier (Braman 2002; Zhang et al. 2014), de couper (Arber et al. 1969), de coller (Shuman 2009), d'amplifier des brins d'ADN (LEHMAN et al. 1958) et de les transférer d'un organisme à un autre (Rosano et al. 2014). Ils ont de nombreuses applications dans la recherche en microbiologie et notamment la production d'antibiotiques et de protéines recombinantes par des micro-organismes. Parmi les outils de la biologie moléculaire, la Taq polymérase est sans doute l'un des plus remarquables. Elle fut isolée du thermophile *Thermus aquaticus* en 1976 (Chien et al. 1976). Son fonctionnement à des températures élevées a permis la mise au point de la Réaction en Chaîne par Polymérase ou Polymerase Chain Reaction (PCR) en anglais (Mullis et al. 1989). Cette technique valut à Kary Mullis le prix Nobel en 1993. La PCR est une technique qui permet l'amplification exponentielle de brins d'ADN spécifique à partir d'amorces nucléotidiques.

1.3.1 Le séquençage, naissance de la phylogénie moléculaire

Entre 1968 et 1977, Sanger réalise de nombreux progrès et met au point des méthodes efficaces de séquençage de l'ADN, de l'ARN et des protéines (SANGER et al. 1951b; SANGER et al. 1951a; Sanger et al. 1975; Sanger et al. 1977). La méthode de Sanger consiste à interrompre la polymérisation de l'ADN tout en marquant le nucléotide terminal. Originellement, on ajoutait à l'échantillon des désoxynucléotides-triphosphates (dNTP) (pour la polymérisation de l'ADN) qui étaient ensuite divisés en quatre aliquotes recevant chacun un des quatre désoxynucléotide triphosphates (ddNTP) (Sanger et al. 1977). Les désoxynucléotides sont rajoutés à une concentration cent fois inférieure à celle des désoxynucléotides. La polymérisation est interrompue à chaque fois qu'un ddNTP est incorporé à la place d'un dNTP par la polymérase. L'étape de polymérisation va donc produire des molécules d'ADN de différentes tailles car l'interruption de la polymérisation est aléatoire. En séparant les molécules obtenues par aliquotes, et donc par nucléotide terminal ainsi que par taille, on obtient par électrophorèse un gel sur lequel on peut directement lire la séquence produite. Les premières techniques utilisant la méthode de Sanger nécessitent d'avoir accès à de grandes quantités de matériel génétique et ne sont applicables que sur des organismes isolés et cultivables en culture pure ou sur des fragments d'ADN isolés, clonés et amplifiés. Aujourd'hui, par l'utilisation de ddNTP marqué par fluorescence, on obtient un processus automatisable et sans gel. On sépare les fragments d'ADN par électrophorèse capillaire et on 'lit' la séquence par une détection de la fluorescence.

En parallèle à la méthode de séquençage de Sanger, la méthode dite de Maxam et Gilbert est publiée en 1977 (Maxam et al. 1977). Cette méthode de séquençage casse le brin d'ADN depuis son extrémité terminale à chaque occurrence d'une base nucléotidique spécifique. La taille du fragment obtenue permet de connaître la position de cette base nucléotidique. Maxam et Gilbert décrivent les réactions chimiques permettant de casser le brin d'ADN spécifiquement au niveau d'une Adénine, Cytosine, Guanine ou Thymine. Quand le produit de ces 4 réactions est résolu par taille, par électrophorèse, et par nucléotide (réactions) on

Fig. 4. Carl Woese (1928-2012)
Il identifia un nouveau domaine du vivant: les archées (Fox et al. 1977; Woese et al. 1977)[1]

peut déduire du gel obtenu la séquence ADN. Le séquençage de Maxam et Gilbert n'est plus utilisé aujourd'hui pour plusieurs raisons : l'utilisation d'hydrazine (un neurotoxique), lourd en termes de manipulations et la simplification et le développement de la méthode de séquençage de Sanger et de méthodes plus performantes (1.3.4).

Le séquençage ADN a révolutionné la microbiologie; les premières classifications basées sur des caractères moléculaires ont été produites par Carl Woese (Fig. 4) en 1977 (Woese et al. 1977). Le séquençage de génomes complets de bactéries à l'époque était encore hors de portée et Woese choisit de se servir d'ARN ribosomique. Les ARN ribosomiques sont des ARN structuraux qui ne sont pas traduits mais qui font partie du complexe ribosomique (Brimacombe et al. 1985; Lafontaine et al. 2001). Les ribosomes sont des complexes moléculaires qui permettent l'assemblage de protéines à partir de l'information génétique portée par l'ARN messager. La fonction essentielle des ribosomes est ancestrale et les ribosomes sont conservés dans tous les organismes cellulaires vivants connus: les séquences des ribosomes changent lentement au cours de l'évolution par rapport au reste du génome. Ils constituent des marqueurs permettant de retracer l'évolution des espèces de façon beaucoup plus résolutive que des gènes métaboliques ou des critères non moléculaires comme la morphologie et le métabolisme (Woese 1987; Olsen et al. 1993).

En 1977, les technologies de séquençage ne permettent pas de séquencer de longs fragments d'ADN. Cependant, en 1977, on sait isoler des séquences particulières. Woese

[1] www.pbs.org/

isole les 16rRNAs d'organismes d'intérêt, les fragmente et séquence les fragments ainsi obtenus (Fox et al. 1977). Woese compare alors la composition en fragments de chaque paire d'organismes grâce à ce qu'il appelle le "Coefficient d'Association" (S). Il définit S_{ab} le Coefficient d'association entre les séquences a et b. Soit N_a le nombre de fragments dans l'espèce a, N_{ab} le nombre de fragments communs entre les séquences a et b, le "Coefficient d'Association" entre les espèces a et b est alors $S_{ab} = \frac{2N_{ab}}{N_a+N_b}$. Il obtient donc une matrice de distances entre les espèces d'intérêt. Woese utilise alors une méthode de regroupement hiérarchique par couplage des moyennes ("clustering linkage average") pour obtenir un arbre phylogénétique. Woese fut ainsi le premier à proposer une classification des bactéries basée sur des critères moléculaires. Il proposa que les bactéries méthanogènes ne soient pas des bactéries mais forment un nouveau domaine du vivant : les archées (Fox et al. 1977; Woese et al. 1977).

1.3.2 Principe de la phylogénie moléculaire

Pour établir une classification des micro-organismes, Woese se base sur les fragments ARN similaires entre séquences de marqueurs moléculaires homologues. Deux séquences sont dites homologues si elles descendent d'une même séquence ancestrale. Il existe deux classes d'homologues : les paralogues et les orthologues. Une relation de paralogie est caractérisée par un événement de duplication ancestral au contraire d'une relation d'orthologie qui est caractérisée par une relation de descendance suite à des événements de spéciation. Une relation d'orthologie implique le plus souvent une conservation de la fonction au contraire d'une relation de paralogie qui peut permettre à une des séquences dupliquées d'évoluer vers d'autres fonctions, car elle est libérée d'une pression de sélection (Ohno 1970). La similarité entre deux séquences homologues issues d'organismes différents peut informer sur la relation évolutive entre ces organismes. En effet, l'information génétique est transmise à la descendance avec, parfois, des mutations. Des erreurs peuvent être introduites lors de la réplication de l'ADN par les polymérases (Johnson et al. 2000; Fijalkowska et al. 2012; Ganai et al. 2016). Ces erreurs peuvent être des insertions ou des délétions d'un ou plusieurs nucléotides ou la substitution d'un nucléotide par un autre.

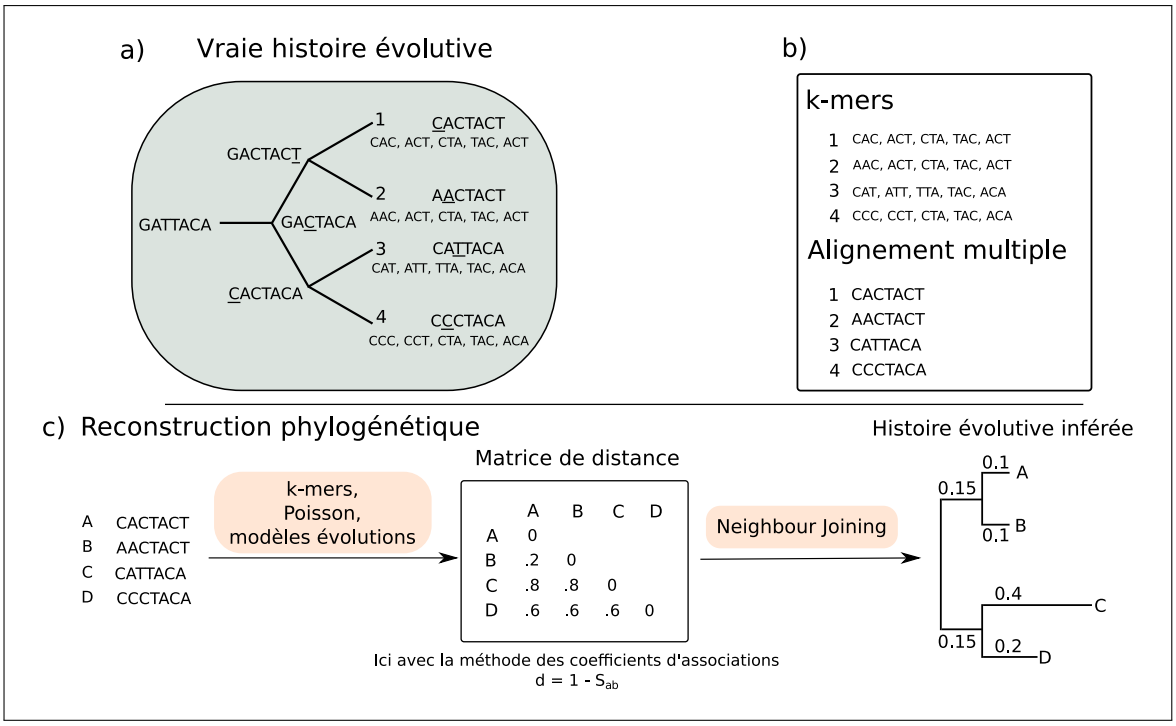


Fig. 5. Exemple d'une reconstruction phylogénétique.
a) Vraie histoire évolutive d'une séquence nucléique. b) Alignement multiple et k-mers (k=3) des séquences observées. c) Exemple de reconstruction phylogénétique. La méthode de distance utilisée ici est un hommage à Woese et n'est plus utilisée (voir texte 1.3.1). Le Neighbour Joining crée un arbre dont les distances tendent vers celles de la matrice de distance.

Les polymérases possèdent une activité de correction améliorant leur précision (Bębenek et al. 2018). L'ADN peut également être endommagé et des erreurs peuvent apparaître lors de sa réparation (Friedberg 2003). *In fine*, les mutations s'accumulent avec le temps. Théoriquement, le nombre de mutations est proportionnel au nombre de divisions cellulaires (KIMURA 1968). On parle alors d'horloge moléculaire (Kumar 2005; Donoghue et al. 2016). Les descendants d'un ancêtre commun vont ainsi petit à petit accumuler des mutations dans leurs génomes. La prise en compte de ces différences permet d'obtenir, après diverses corrections, une distance approximant la parenté entre les organismes (Zuckerkandl et al. 1965).

A partir d'une matrice de distance, une méthode comme NJ (Saitou et al. 1987) peut construire un arbre phylogénétique qui nous informe sur les relations de parenté entre organismes (Fig. 5). A son époque, Woese ne peut obtenir le génome des organismes mais il a l'idée d'identifier des séquences homologues universelles et conservées car essentielles : les

ARNs ribosomiques 16S (appelé 16S car cela correspond à leur coefficient de sédimentation de Svedberg (Timasheff et al. 1958; Lebowitz et al. 2009)). Aujourd'hui encore, même si l'on dispose de l'ensemble de l'information génomique, on se sert généralement de ces marqueurs moléculaires considérés comme conservés et informatifs (Amit Roy et al. 2014). On peut appliquer les méthodes de reconstruction phylogénétique aussi bien aux séquences nucléotidiques qu'aux séquences d'acides aminés. Une phylogénie réalisée à partir de séquence d'acides aminés est considérée plus robuste, même quand lorsque les distances évolutives sont grandes. En effet, si pour une même séquence, il y a plus de positions et donc d'information avec l'ADN, le nombre plus limité de possibilités par position fait que l'information va plus rapidement être masquée par du bruit. Malheureusement, plusieurs difficultés rendent la phylogénie moléculaire biaisée. Reconnaître ces biais permet d'appréhender les limites de cette technique. Par exemple, un même site peut muter plusieurs fois. Toutes les mutations ne sont pas équiprobables, par exemple les probabilités de transition (muter en une base de même famille i.e. de purine à purine) ou de transversion (muter en une base d'une famille différente) sont différentes (Kimura 1980). Il existe plusieurs modèles de substitution décrivant les probabilités de substitutions d'une base nucléique ou d'un acide aminé. Également, le taux de mutation n'est pas constant le long du génome ni au cours du temps.

La phylogénie moléculaire a accompagné la microbiologie environnementale et a permis de nombreux succès dans la description et la caractérisation de la diversité génétique. C'est la phylogénie moléculaire qui permit la caractérisation des archées (1.3.1) comme un nouveaux domaine du vivant. Un autre succès de la phylogénie moléculaire et l'identification d'un groupe frère des eucaryotes (voir 1.5.1). Cependant, avec l'accroissement des jeux de données, les méthodes de phylogénie moléculaire classiques atteignent leurs limites en termes de capacité de calcul.

1.3.3 Calcul de la similarité entre plusieurs séquences ADN

L'alignement peut avoir plusieurs objectifs : superposer les résidus homologues, superposer les résidus qui occupent la même position tridimensionnelle, superposer les

résidus qui ont la même fonction. Une partie des méthodes qui essaient de comprendre les relations de parenté entre des séquences utilisent le score ou le pourcentage de l'identité d'un alignement. Le pourcentage d'identité est généralement calculé comme le pourcentage de positions identiques d'un alignement entre deux séquences. De fait, l'alignement de séquences est un outil essentiel du bioinformaticien. Durant mon travail de thèse, de nombreuses séquences ont été alignées, par différents programmes d'alignement que je souhaite mentionner ici. Les deux principaux logiciels qui ont été utilisés sont BLAST (Basic Local Alignment Search Tool) (Altschul et al. 1990) et DIAMOND (Double Index AlignMent Of Next-generation sequencing Data) (Buchfink et al. 2015). Ils utilisent des heuristiques pour trouver les séquences susceptibles de s'aligner mais utilisent un algorithme d'alignement local optimal développé par Smith et Waterman en 1981 (Smith et al. 1981).

a) Algorithme de Smith et Waterman

L'algorithme de Smith et Waterman utilise la programmation dynamique pour trouver l'alignement local optimal entre deux séquences (nucléiques ou protéiques) pour un jeu de paramètres donné. La programmation dynamique est une technique qui consiste à décomposer un problème en sous-problèmes, puis à résoudre les sous-problèmes dont la résolution permet de trouver la solution au problème original. L'algorithme calcule un score pour chaque alignement et renvoie l'alignement local avec le meilleur score. L'alignement obtenu et le score dépendent fortement des paramètres (Fig. 6). L'évolution implique des événements de substitutions, de délétions et d'insertions. Ces phénomènes évolutifs sont modélisés par des paramètres lors de l'alignement. Quand deux résidus superposés sont identiques on parle de match, quand deux résidus ne sont pas identiques on parle de mismatch. Le score d'un match et d'un mistmatch peut être donné de manière générale quelque soit les résidus. Cependant, les fréquences des substitutions dans les séquences moléculaires ne sont pas équiprobables. Par exemple, il est plus probable de remplacer une base pyrimidique par une base pyrimidique que par une base purine. Dans une séquence protéique, un acide aminé a également plus de chance d'être remplacé par un acide aminé aux propriétés chimiques et d'encombrement stérique similaires. L'utilisation d'une matrice

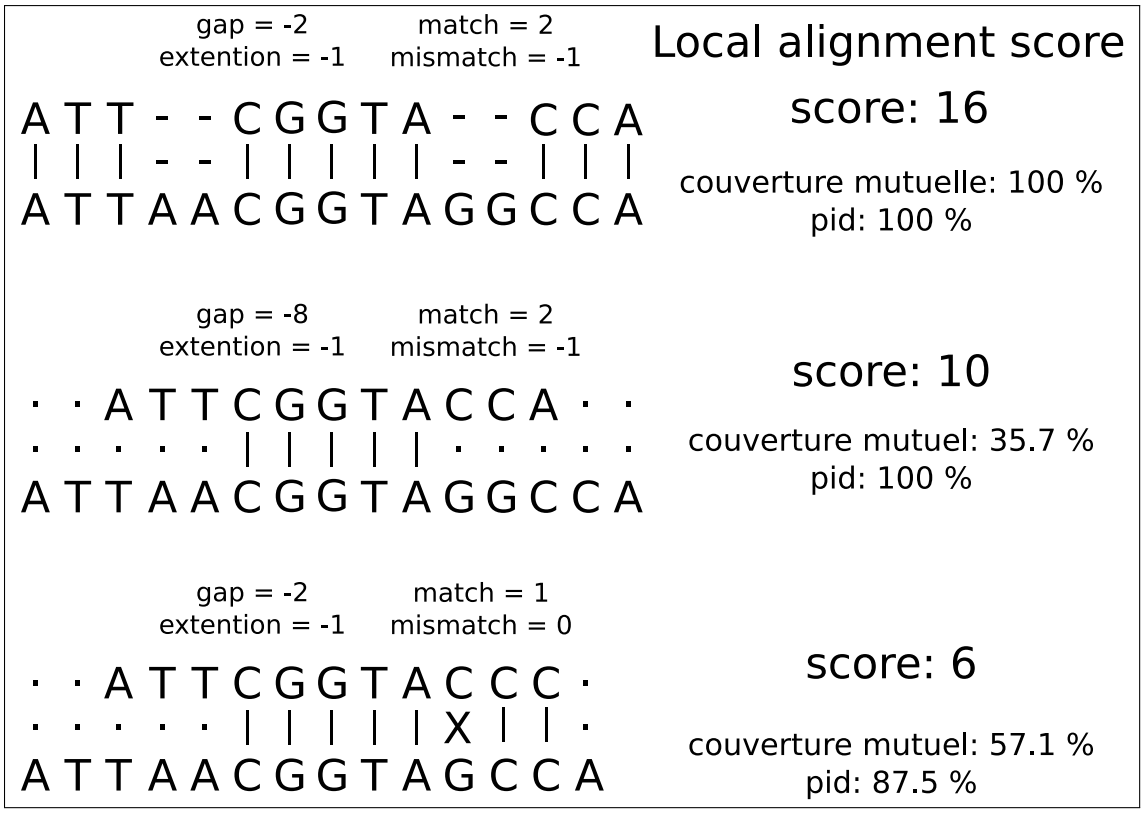


Fig. 6. Impact des paramètres de score sur un alignement local

de substitution qui indique pour chaque mutation son score, permet de prendre en compte les différentes probabilités de substitution. Une insertion ou une délétion est représenté par un trou, appelé gap, dans l'alignement. Un événement d'insertion ou de délétion est rare mais il peut inclure plusieurs résidus. Pour cette raison, on distingue l'ouverture du gap et son élongation dans l'algorithme d'alignement. Le coût de l'ouverture d'un gap et de son élongation, le coût d'un mismatch (deux résidus différents) et le score d'un match (deux résidus identiques) sont les principaux paramètres de l'algorithme. De cet alignement, nous pouvons déduire plusieurs mesures, notamment le pourcentage d'identité et la couverture mutuelle. Nous nous servirons de ces deux mesures pour déduire si deux séquences sont homologues à partir d'un alignement. La couverture d'un alignement sur une séquence correspond à la proportion de cette séquence couverte par l'alignement. La couverture mutuelle d'un alignement entre deux séquences correspond à la plus faible valeur de couverture d'alignement entre les deux séquences. Le pourcentage d'identité correspond au nombre de positions identiques le long de l'alignement.

b) Heuristique pour la comparaison d'un grand nombre de séquences

L'algorithme de Smith et Waterman trouve le meilleur alignement local pour une paire de séquences. Si l'on souhaite comparer un jeu de données contenant X séquences à une base de données contenant Y séquences, on va devoir utiliser l'algorithme de Smith et Waterman X * Y fois. En pratique, avec l'augmentation de la taille des jeux de données, cela devient rapidement très coûteux en temps de calcul. Pour cette raison, des heuristiques ont été développées comme BLAST (Altschul et al. 1990) et DIAMOND (Buchfink et al. 2015), déjà mentionnées plus haut. Le principe de ces heuristiques est d'aligner avec l'algorithme de Smith et Waterman uniquement les parties des séquences qui ont une grande probabilité de s'aligner avec un bon score.

c) BLAST

BLAST est un algorithme d'alignement local qui utilise des heuristiques pour aligner uniquement les régions des séquences qui ont une grande probabilité de s'aligner avec un score suffisant. Cela lui permet d'aligner de nombreuses séquences en limitant le nombre d'alignements à effectuer. BLAST aligne des séquences requêtes contre des séquences cibles. Une étape de pré-traitement est nécessaire pour formater les séquences cibles en base de données. BLAST découpe les séquences cibles en k-mers, c'est-à-dire en sous séquences de k lettres, k=3 pour les séquences d'acides aminés et k=11 pour les séquences de nucléotides. Pour chaque mot ainsi créé, BLAST stocke en mémoire quelles séquences possèdent ce mot et à quelle position. La base de données créée par BLAST peut donc être vue comme une table d'association qui associe à tous les mots la ou les séquences où ces mots sont présents et leurs positions. Une fois la base de données construite, BLAST peut aligner des séquences requêtes contre la base de données. Les séquences requêtes sont également découpées en k-mers. Pour chaque k-mer, BLAST va produire des k-mers dits voisins, c'est-à-dire des k-mers dont l'alignement avec le k-mer d'origine est au dessus d'un seuil, afin d'être plus sensible (Fig 5. a). Ensuite, BLAST va chercher pour chacun de ces k-mers, les k-mers identiques dans la table d'association de la base de données (Fig 7. a). BLAST se sert des k-mers communs identifiés comme graines d'alignement. La prochaine étape consiste

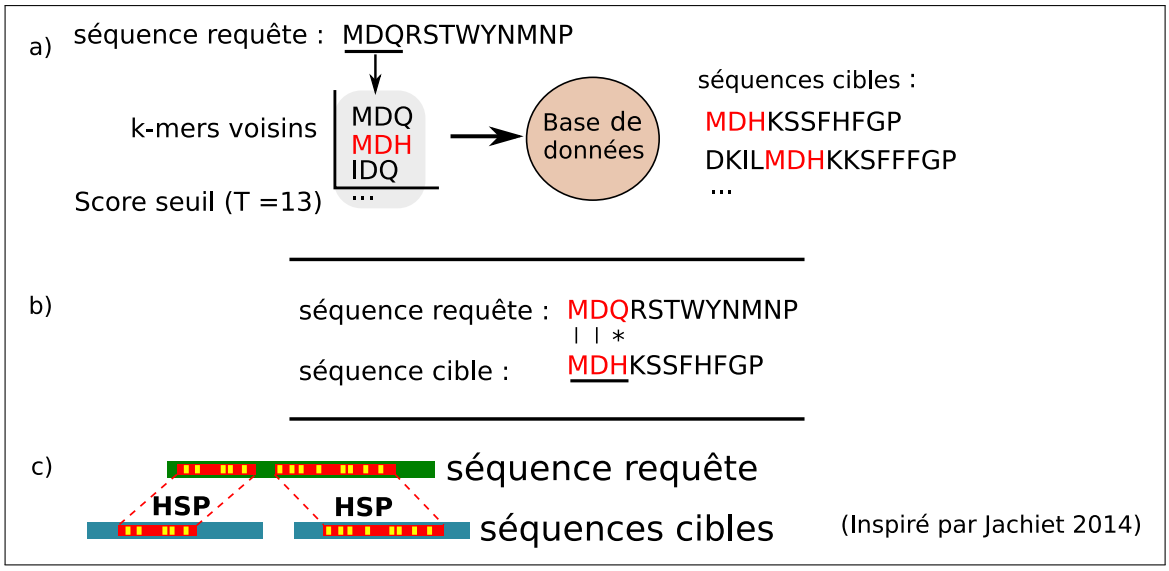


Fig. 7. BLAST, principe de fonctionnement
a) Découpage de la séquence requête en k-mers voisins et interrogation de la base de données. b) Alignement à partir des k-mers identifiés. c) Une séquence peut avoir plusieurs HSP.

à étendre ces graines d'alignement en mesurant l'augmentation ou la diminution du score d'alignement (Fig 7. b). La procédure s'arrête quand le score d'alignement décroit trop fortement. L'alignement est alors renvoyé. Ces alignements optimaux locaux sont appelés HSP (High Scoring Pair). BLAST retourne tous les HSP non inclus dans un autre HSP. L'alignement de deux séquences par BLAST peut donc donner plusieurs HSP (Fig 7. c).

a) Système de score et considérations statistiques

Le score brut d'un alignement **S** nous renseigne sur la qualité d'un alignement pour une matrice de substitution et des paramètres d'alignement donnés. On ne peut donc pas s'en servir pour comparer des alignements obtenus avec des paramètres différents. Le Score **S** ne prend pas en compte la taille du jeu de données: plus le jeu de données est grand, plus il y a de chance de trouver un alignement. Il est intéressant de pouvoir comparer la qualité et la pertinence d'alignements obtenus avec des paramètres de scores différents et sur des jeux de données différents.

Le Bit Score et la E-value ont été développés à cette fin. Le Bit Score $\mathbf{S}'$ est calculé à partir du score **S** et deux paramètres: κ, λ. Le paramètre λ permet de normaliser par rapport aux paramètres d'alignements, et κ permet de normaliser par rapport à la taille

des jeux de données. Le Bit Score est calculé comme $S' = \frac{\lambda \mathbf{S} - \ln(\kappa)}{\ln 2}$. Le Bit Score permet donc de comparer des alignements entre analyses. Cependant, le Bit Score ne nous informe pas directement sur la significativité de l'alignement, c'est-à-dire la probabilité d'obtenir un score x sachant les conditions de l'alignement et la taille du jeu de données. La E-value **E** répond exactement à cette question. On peut la calculer à partir du Bit Score **S´** ou du Score **S** avec $\mathbf{E} = mn2^{-\mathbf{S'}}$, ce qui est équivalent à $\mathbf{E} = \kappa mne^{-\lambda \mathbf{S}}$, avec n et m représentant la taille de la séquence cible et requête respectivement. On remarquera que **E** augmente proportionnellement avec la taille des séquences: plus les séquences sont longues, plus trouver un score donné est probable. De plus, elle diminue exponentiellement avec le Score: plus le Score est fort, moins il est probable qu'il soit dû au hasard.

a) Conclusion

BLAST a été un outil essentiel au développement de la comparaison de séquences *in silico*. Mais l'amélioration des techniques de séquençage et la baisse dramatique de leurs coûts ont eu pour conséquence une augmentation spectaculaire de la taille des bases de données. BLAST n'est pas assez puissant en terme de vitesse pour permettre d'exploiter convenablement ces ressources. De plus, en 1990, quand BLAST a été rendu public, l'architecture des ordinateurs était différente: la mémoire était un facteur limitant et, à l'exception des supercalculateurs les ordinateurs ne possédaient qu'un processeur (Fig. 8). Aujourd'hui la majorité des ordinateurs possèdent plusieurs processeurs et disposent de grandes quantités de mémoire physique et vive. Pour ces raisons de nouveaux programmes utilisant plusieurs processeurs de façon efficace ont fait leur apparition, notamment DIAMOND et MMSEQ2 (Fig. 5) (Buchfink et al. 2015; Steinegger et al. 2017).

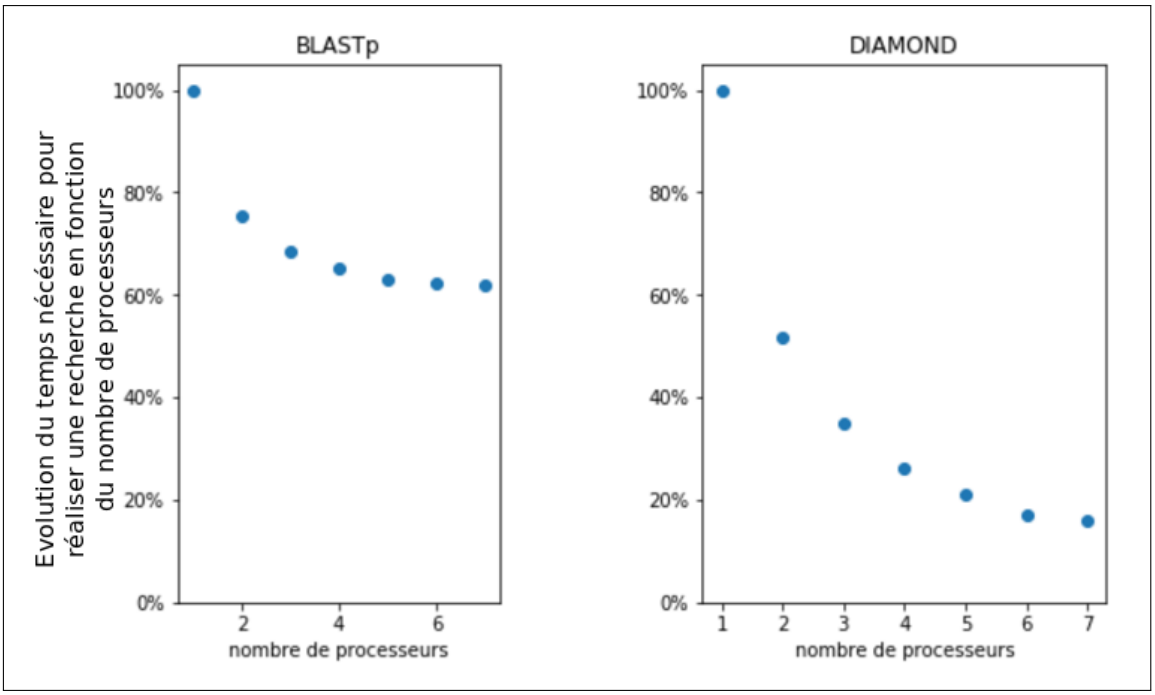


Fig. 8. Réduction du temps de calcul en fonction du nombre de processeurs

1.3.4 Le séquençage haut débit ou le déluge de données

Les technologies de séquençage vont connaître une révolution entre les années 2000 et 2006, à commencer par la formalisation du séquençage dit à l'aveugle dans les années 1980 (Staden 1979). Le séquençage à l'aveugle a été utilisé sur de petits génomes (Gardner et al. 1981) et le développement de cette technologie durant les années 1990 a permis son application au séquençage du génome humain (Venter et al. 2001) (International Human Genome Sequencing Consortium 2001). En second, l'apparition de méthodes de séquençage à haut débit (High Throughput sequencing, HTS) ou nouvelles technologies de séquençage (Next Generation Sequencing, NGS) transforme le champ des possibles. En effet, ces technologies ont dramatiquement réduit les coûts, les infrastructures et le temps de travail nécessaire au séquençage de séquences nucléotidiques (Wheeler et al. 2008). A titre d'exemple, le premier génome humain séquencé a été publié en 2001 dans une version non finalisée (la version finale est publiée en 2003) mais le projet avait commencé en 1989 et le travail était réparti entre 20 universités ou institutions internationales pour un coût de 2.7 milliards de dollars (dollars 1991), soit 5 milliards de dollars actuels. Aujourd'hui, le coût du séquençage d'un génome humain est d'un peu plus de 1000 dollars et prend environ

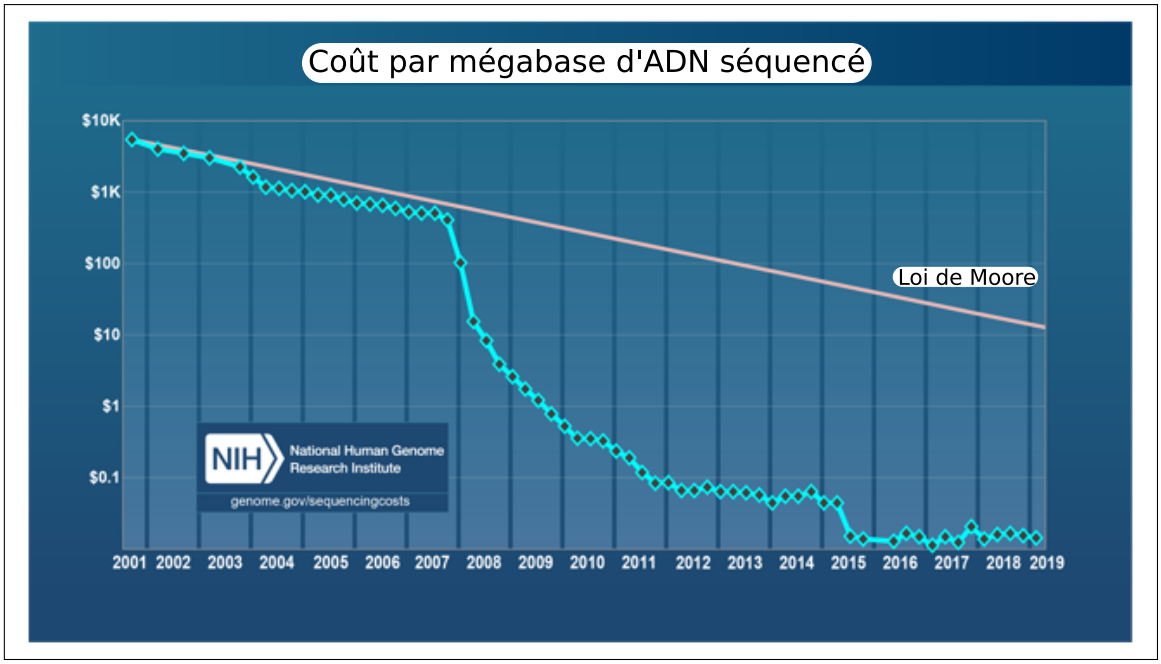


Fig. 9. Evolution du prix du séquençage d'une mégabase (inspiré du NCBI)

une semaine de travail à un technicien en biologie et à un analyste en bio-informatique. La loi de Moore est une conjecture énoncée en 1965 qui décrit l'évolution de la puissance de calcul des ordinateurs. Elle prédit un doublement de la puissance de calcul tous les 2 ans. Aujourd'hui on pense que la loi de Moore est devenue obsolète mais pendant plus de 40 ans elle s'est révélée confirmée. La technologie de séquençage a progressé plus rapidement que la loi de Moore. Cela signifie que nos capacités de traitement informatique n'évoluent pas assez rapidement pour être capables de traiter les volumes de données produits par le séquençage à haut débit. Une figure que je trouve excellente pour se rendre compte de ce phénomène est publiée tous les ans par le NIH (National Institute of Health, l'institut de santé des États-Unis). Elle compare l'évolution du coût de séquençage à la loi de Moore (Fig. 9) (l'échelle de la figure est en logarithme de base 10). Par conséquent, les progrès du séquençage ont permis l'émergence de nouvelles techniques d'étude des micro-organismes mais ils ont imposé une nouvelle limite à la biologie environnementale : notre capacité à traiter des volumes de données toujours plus grands.

1.3.5 Techniques de séquençage

Cette partie a pour objectif de présenter les avantages et inconvénients des techniques de séquençage à haut débit. Ce n'est pas une description ni une comparaison exhaustive de l'ensemble des méthodes de séquençage à haut débit. Les polymères d'ADN peuvent atteindre une très grande taille. Par exemple, alors que le génome d'une bactérie est de l'ordre du million de paires de bases, le plus grand chromosome humain possède 246 millions de paires de base (International Human Genome Sequencing Consortium 2004). A l'heure actuelle, on ne sait pas séquencer directement des fragments de cette taille. La solution est de fragmenter les chaînes d'ADN et de séquencer les fragments, que l'on appelle fragments de lecture ou reads (lectures) en anglais. En dehors du coût et du temps de préparation, les caractéristiques pour comparer des technologies de séquençage sont la taille des fragments de lecture obtenus, le taux d'erreur et le volume de sortie du séquenceur, le nombre de paires de bases que peut produire le séquenceur (Shendure et al. 2011; Ambardar et al. 2016). Les séquenceurs de seconde génération peuvent produire des fragments de lecture entre 50 et 800 paires de bases avec un volume allant jusqu'à 10 millions de millions de paires de bases (fig. 10). Une nouvelle révolution du séquençage est en cours, avec l'apparition des séquenceurs dit à longs fragments de lecture comme le Pacific Bioscience (2010) ou le MinION (2014). Ces technologies permettent l'obtention de fragments de lecture longs (de plusieurs milliers de paires de base). Nanopore a récemment dévoilé un système produisant des fragments de lectures ultra longs: mesurant jusqu'à 2 000 000 de paires de bases (Jain et al. 2018). Si la technologie des longs fragments de lectures n'est pas très précise, avec un taux d'erreur de l'ordre de 10%, elle est particulièrement utile pour l'assemblage des fragments de lecture. Plusieurs méthodes de correction des fragments de lecture longs sont développées, certaines utilisant une approche de consensus, d'autres combinant fragments de lecture courts et longs. Je suis persuadé que ces technologies et approches vont devenir des outils incontournables de la microbiologie environnementale dans un futur proche. Une étape de post-traitement est donc essentielle à tout séquençage haut débit. Elle permet d'identifier et d'enlever les

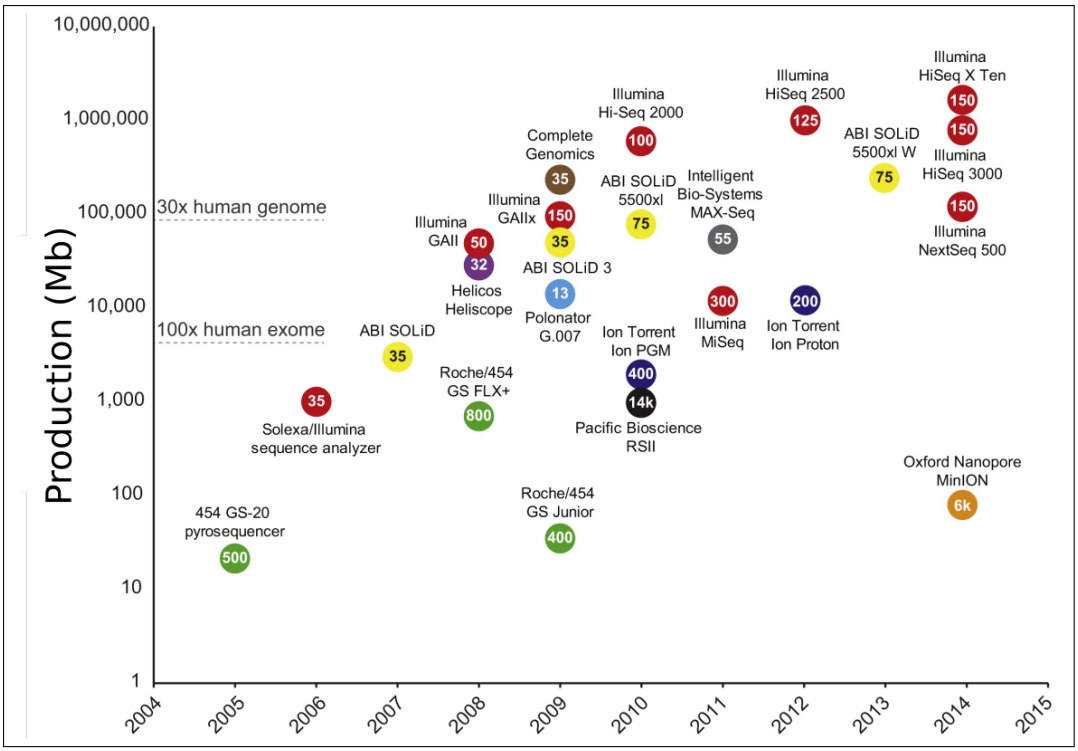


Fig. 10. Evolution des technologies de séquençage haut débit (adapté de Reuter et al. 2015) En abscisse, l'année de mise sur le marché des technologie de séquençage. En ordonnée, le nombre maximum de méga bases que peut produire une technologie donnée en un run.

fragments de lecture de faible qualité ainsi que les artefacts de séquençage.

1.3.6 Profondeur de séquençage

La profondeur de séquençage est définie comme le rapport du nombre de nucléotides produits par le séquençage sur la taille du ou des génome(s) haploïde séquencé (Sims et al. 2014). Soit N le nombre de nucléotides produits par le séquençage et L la taille des fragments uniques d'ADN, la profondeur de séquençage est $X = \frac{N}{L}$. La profondeur de séquençage permet d'obtenir la probabilité qu'une base d'ADN de l'échantillon soit représentée dans un fragment de lecture. En effet, on peut considérer que la probabilité qu'une base de l'échantillon soit séquencée dans au moins x fragments de lecture est modélisée par la loi de Poisson suivante $\frac{P^x e^{-P}}{x!}$ avec P la profondeur de séquençage. La probabilité de ne pas séquencer une base est e^{-P}, donc la probabilité de séquencer une base est $1 - e^{-P}$. Pour

une profondeur de 1, la probabilité qu'une base d'ADN soit représentée dans un fragment de lecture est de 63.21%; pour une profondeur de 10, la probabilité est de 99.99%. La profondeur de séquençage est donc une mesure qui permet d'estimer la part de l'ADN présent dans l'échantillon qui est séquencé.

1.3.7 Assemblage

Les fragments de lecture sont issus de brins d'ADN fragmenté et dans la plupart des cas ils ne couvrent qu'une infime partie du brin d'ADN. Un fragment de lecture de 200 bp couvre un dix-millionième du chromosome 1 humain par exemple. Une étape de traitement informatique des données de séquençage est nécessaire afin d'assembler les fragments de lecture pour retrouver les séquences des chromosomes originaux. Plusieurs méthodes d'assemblage existent (Benjak et al. 2015; McGrath 2007; Sohn et al. 2016), avec ou sans génome de référence.

Si aucun génome de référence n'existe, l'assemblage est dit *de novo*. L'assemblage *de novo* est un processus complexe. Pour obtenir un assemblage *de novo* de qualité, il est nécessaire d'avoir une grande couverture, un très faible taux d'erreur et des fragments de lectures de grande taille. Actuellement, aucune technologie de séquençage ne permet d'obtenir de long fragments de lectures avec une précision suffisante. Des méthodes sont développées : un traitement informatique qui va corriger les fragments de lectures longs, ou bien prendre en compte l'information d'un séquençage à fragments de lecture courts pour corriger les fragments de lecture longs. Une revue à jour qui compare les différentes méthodes de corrections de longs fragments de lecture a été publiée par Zhang (Zhang et al. 2019),

Actuellement, la majorité des méthodes d'assemblage *de novo* sont basées sur des graphes de De Bruijn (Fig. 11). Cette approche consiste à construite l'ensemble des k-mers à partir des fragments de lecture. Un graphe est ensuite construit : sous la forme d'un réseau orienté dans lequel les arrêtes sont des k-mers et les nœuds les (k-1)-mers correspondants (Medvedev et al. 2011). L'assemblage *de novo* est un processus complexe. Les méthodes d'assemblage ne sont pas optimales et il rare d'obtenir un génome complet linéaire. A la

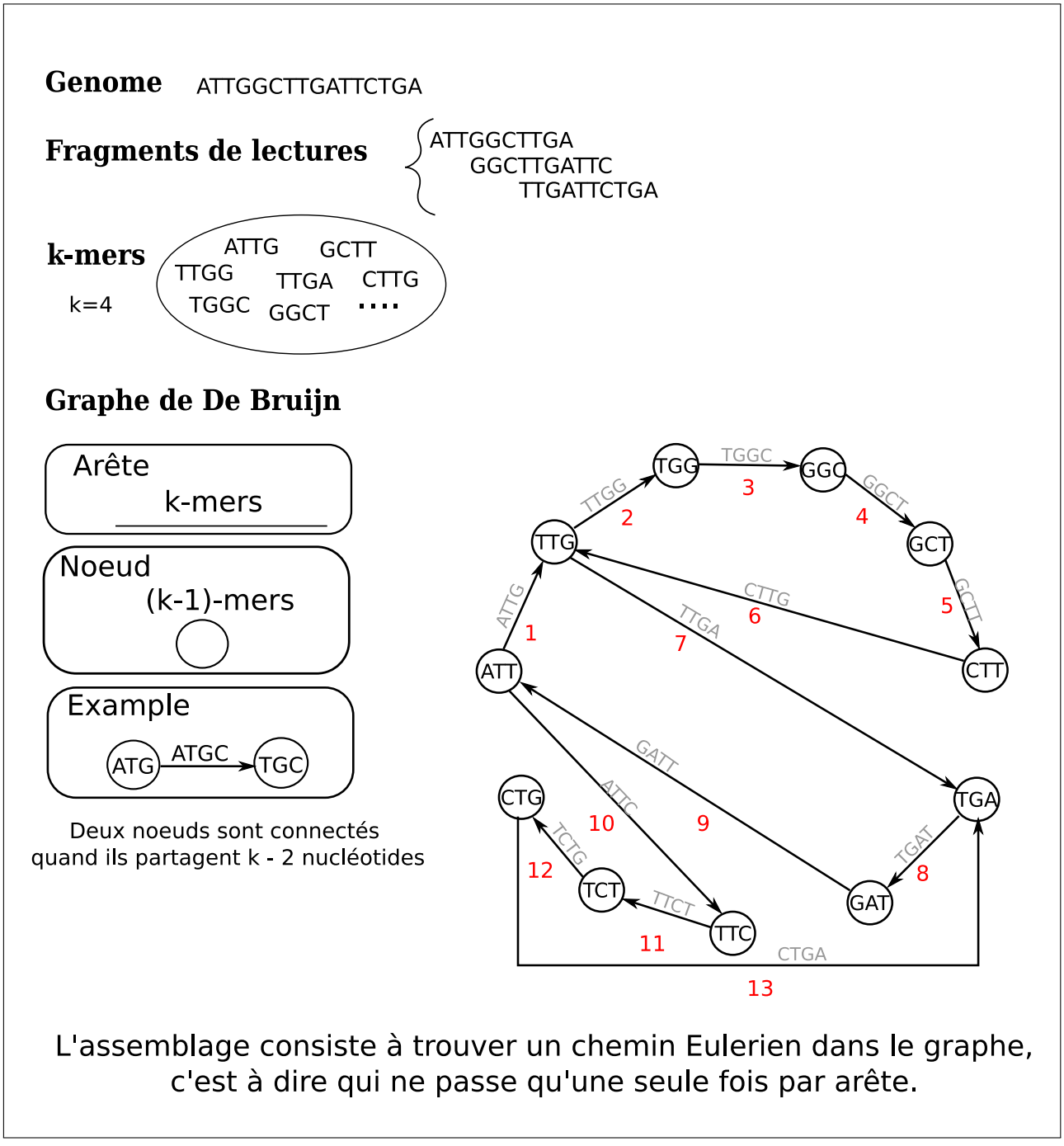


Fig. 11. Utilisation d'un graphe de De Bruijn pour l'assemblage de séquences nucléotidiques. Partant d'un génome, une expérience de séquençage donne des fragments de lecture. Ces fragments de lecture sont découpés en k-mers et un graphe de De Bruijn est construit. En rouge le chemin Eulerien qui permet de retrouver la séquence du génome initial.

place, on obtient des séquences d'ADN plus longues que l'on appelle contigs (diminutif pour fragment de lecture contigu). Cependant, ces méthodes, si elles présentent des difficultés techniques, n'ont pas de biais *a priori* et sont appliquables à des génomes qui n'ont pas de référence deja connue.

Si un génome de référence existe pour l'échantillon séquencé, l'objectif et d'identifier les variations de la séquence nucléotidique par rapport au génome de référence. La solution retenue est d'aligner les fragment de lectures sur le génome de référence. Si le génome de référence utilisé n'est pas suffisament proche, l'assemblage sera de mauvaise qualité. Pour l'exploration de la diversité microbienne qui a pour objectif l'étude d'organismes insuffisament ou non caractérisés, on preferera les méthodes d'assemblages *de novo*.

1.4 Meta-omique

Omic est un anglicisme qui désigne l'ensemble des champs biologiques que sont la génomique, la transcriptomique, la protéomique et la métabolomique (étude du métabolisme). Le terme méta fait référence à ce qui vient après. La méta-omics vient après l'omics car elle n'a plus besoin d'isoler les organismes en culture pure pour les étudier. Dans cette partie, je vais détailler les techniques de métagénomique et de métabarcoding. J'en ferai un comparatif, puis un état de l'art de la littérature et finalement je conclurai sur les questions soulevées par ces études.

1.4.1 Métabarcoding

La microbiologie environnementale a connu une rupture technologique avec les travaux de Norman Pace (Fig. 12) (Lane et al. 1985). Ce chercheur trouva un moyen d'amplifier

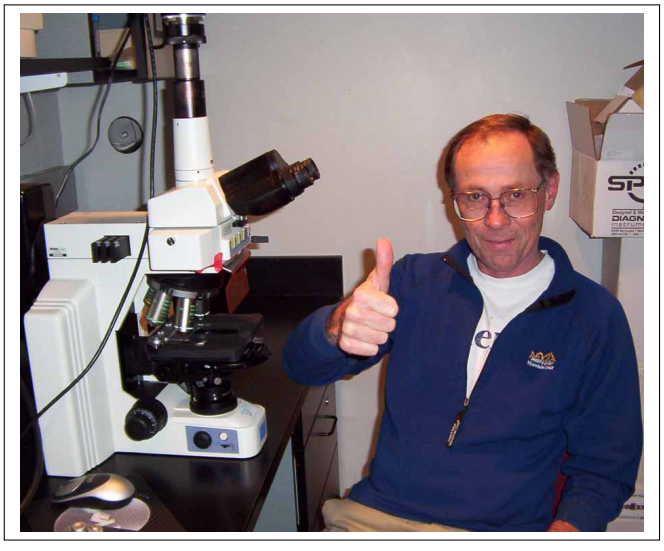


Fig. 12. Norman Pace
[1]

des séquences d'ARN ribosomaux directement depuis un prélèvement environnemental sans étape de culture pure, supprimant un des principaux biais de la recherche en

[1] http://pacelab.colorado.edu/PI$_N$*ormPace.html*

microbiologie. Les ARN ribosomaux peuvent être vus comme des marqueurs universels (1.3.1) discriminants, une étiquette spécifique de chaque espèce. C'est le principe de la méthode de métabarcoding qui consiste à amplifier un marqueur universel (i.e. commun à tous les organismes) par PCR (1.3), grâce à l'utilisation d'amorces spécifiques de ce marqueur, puis à séquencer les résultats de l'amplification (Ruppert et al. 2019).

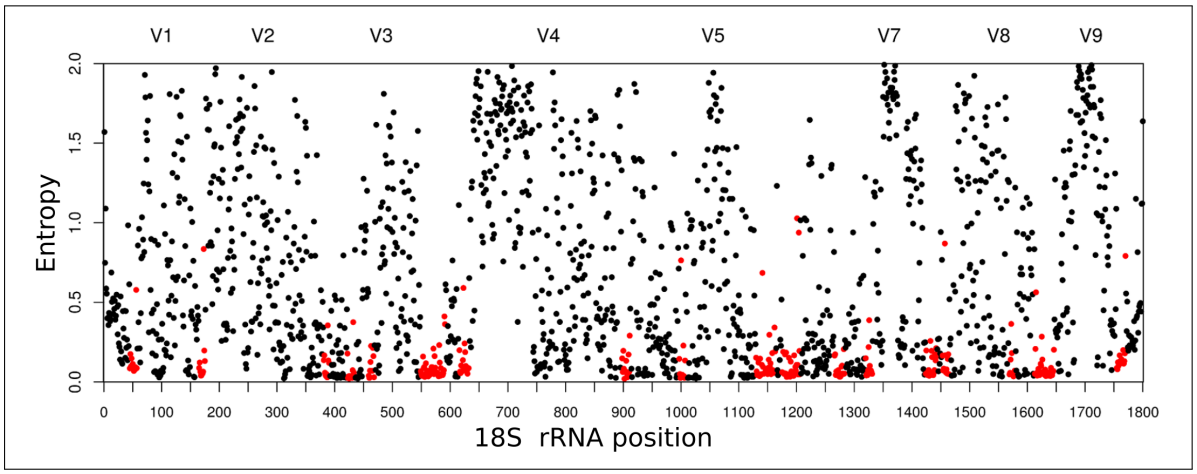


Fig. 13. Variabilité nucléotidique dans les 18S rRNA
En abscisse, la position de l'alignement de tous les 18S de la base de données SILVA. En ordonnée, l'entropie de Shannon associée à chacune des positions de l'alignement. Une entropie de Shannon élevée indique une grande variabilité. Les régions hypervariables sont indiquées au-dessus du graphique de V1 à V9. Adapté de Hadziavdic et al. 2014 [1]

On considère le marqueur moléculaire (ARN ribosomique ou gyrB) comme un code barre propre à un organisme. La technique de métabarcoding a pour objectif de caractériser la diversité microbienne d'un environnement. La majorité des études utilise des ARN ribosomaux, qui présentent des régions hypervariables et des régions conservées (Fig. 13). Les amorces sont conçues pour s'hybrider sur les régions conservées entourant la ou les régions hypervariables à amplifier. En fonction de la technologie de séquençage utilisée, une à deux régions hypervariables du 16S RNA (pour les procaryotes) ou du 18S RNA (pour les eucaryotes), sont amplifiées. Du fait de l'utilisation de la PCR, qui permet une amplification exponentielle de la séquence nucléotidique cible, le métabarcoding est une technique extrêmement sensible qui permet de détecter la présence d'organismes peu abondants. Cette approche ne nécessite pas l'assemblage des fragments de lecture puisque l'on séquence des fragments amplifiés. La technique de métabarcoding présente néanmoins

[1] image originale: https://doi.org/10.1371/journal.pone.0087624.g001

plusieurs inconvénients:

- La limite en taille de la région amplifiée. En effet, les techniques de séquençage modernes ont du mal à séquencer de longs fragments avec suffisamment de précision (voir 1.3.5). La taille de la région hypervariable amplifiée est donc directement dépendante des technologies de séquençage. L'information obtenue par de courts fragments d'ADN ne permet pas toujours d'obtenir une image précise de la diversité microbienne d'un milieu.

- La conception des amorces. Concevoir des amorces qui soient spécifiques de la région d'intérêt et qui permettent l'amplification chez tous les organismes connus est un problème non trivial et ajoute un biais, puisque les amorces sont construites à partir de nos connaissances. On peut donc manquer des organismes inconnus qui possèdent des régions conservées qui ne seront pas reconnues par les amorces (Bahram et al. 2019; Parada et al. 2016; Brown et al. 2015).

- De plus, les marqueurs 16S et 18S peuvent être présents en plusieurs copies, pas forcément identiques dans les génomes. Par conséquent le métabarcoding basé sur ces marqueurs a tendance à surestimer la diversité (Sun et al. 2013). Ils restent cependant les marqueurs de référence utilisés par la communauté scientifique.

1.4.2 Métagénomique

La métagénomique se base sur le séquençage de l'information génétique d'un échantillon environnemental. Elle permet d'obtenir une information qui n'est pas biaisée par nos connaissances de départ et qui n'est pas limitée à un seul marqueur. La métagénomique peut avoir plusieurs objectifs: caractériser la diversité microbienne d'un environnement, déterminer la capacité métabolique d'un environnement, prédire les protéines potentiellement présentes ou obtenir le génome des organismes présents dans un échantillon. On commence par effectuer un prélèvement, l'échantillon est parfois filtré pour l'enrichir en micro-organismes d'une certaine taille. Les cellules de l'échantillon sont ensuite lysées (détruites), leurs brins d'ADN récupérés puis séquencés (Quince et al. 2017).

Les micro-organismes présents dans un environnement forment une population complexe d'individus appartenant à une grande diversité de groupes phylogénétiques, présents à des abondances variées. Dans les milieux avec une grande diversité de micro-organismes, ou dans les milieux qui possèdent à la fois des micro-organismes très abondants et des organismes peu abondants, il y a de fortes chances que la profondeur de séquençage (voir 1.3.6) ne permette pas de séquencer l'ensemble des séquences, et plus particulièrement celles des organismes peu abondants. C'est la raison pour laquelle les échantillons sont souvent séparés en différentes fractions de taille par filtration. Cependant, avec l'amélioration des technologies de séquençage, l'impact de cette limitation est amenée à se réduire avec le temps.

Après l'étape de séquençage, les fragments de lecture sont annotés, en particulier pour connaître *a minima* les réactions métaboliques possibles dans le milieu et la diversité phylogénétique des organismes présents. Il existe deux classes de méthodes d'annotations: une basée sur l'alignement des fragments de lecture sur une banque de données de référence (dite sans assemblage) et une autre qui assemble les fragments de lecture en contigs (voir 1.3.5 et 1.3.7)(Quince et al. 2017). L'assemblage de données métagénomiques présente des difficultés uniques. Contrairement au séquençage de clones isolés en culture pure, c'est un milieu complexe avec plusieurs génomes présents à des abondances différentes. De plus, le volume de données produit est très grand. Ces différentes raisons expliquent le développement de méthodes dédiées à l'assemblage de données de métagénomique (Li et al. 2015; Nurk et al. 2017). L'abondance des contigs obtenus est définie par leur couverture en fragments de lecture. Il est également possible de prédire les régions ADN codantes sur les contigs en utilisant des modèles mathématiques statistiques sous la forme de Modèles de Markov Cachés (Hidden Markov Model HMM en anglais) (Hyatt et al. 2010; Delcher et al. 2007; Lukashin 1998). Les HMM permettent de décrire une séquence de caractères qui dépend d'états invisibles dit cachés. Dans le cas de la prédiction de gène, la séquence de caractères correspond à la séquence de nucléotides et les états cachés à la nature de la séquence nucléotidique, intergénique, promotrice et génique... Un HMM a besoin d'être entrainé pour reconnaître les caractéristiques des états cachés. Ces caractéristiques (biais

d'usage des codons, taux de GC des séquences codantes) varient en fonction des espèces. Le HMM doit donc être entrainé sur des génomes annotés proches du génome à annoter. Cette limitation implique de posséder un génome de référence proche et des mauvaises annotations dans le génome de référence risquent d'être propagées. Des méthodes non supervisées ont été mises au point (Hyatt et al. 2010) et améliorées pour fonctionner sur des métagénomes où les contigs sont potentiellement issus de génomes différents (Hyatt et al. 2012). Les séquences protéiques potentiellement produites par les organismes d'un milieu sont déduites des séquences codantes. Il est parfois possible de déduire une annotation fonctionnelle et/ou phylogénétique des protéines prédites en les comparant à des bases de données de référence. Il est alors possible de comparer certains marqueurs identifiés à une base de données de référence pour caractériser la diversité du milieu. De plus, à partir de l'ensemble des fonctions prédites, il est possible de prédire les métabolismes potentiellement présents dans le milieu. Cette image métabolique obtenue est sans doute surestimée. En effet, cette image ne prend pas en compte la séparation physique entre cellules. C'est une image qui correspond à la capacité métabolique de l'ensemble de l'information génétique présente dans l'échantillon.

Un contig ne représente que très rarement un génome entier. Les contigs sont souvent une représentation fragmentaire des génomes. Déterminer quels contigs proviennent du même génome est aujourd'hui un défi et une limite de la métagénomique. Néanmoins, des méthodes dites de "binning" ou de regroupement permettent de regrouper des contigs qui appartiennent potentiellement à la même souche (Lin et al. 2016; Albertsen et al. 2013; Alneberg et al. 2014; Kang et al. 2015; Namiki et al. 2012; Boisvert et al. 2012; Wu et al. 2016; Sharon et al. 2013). Les génomes obtenus par ces méthodes sont appelés génomes assemblés par métagénomique (ou GAM). Les méthodes de regroupement se basent sur des caractéristiques des contigs: l'abondance du contig, la proportion de chaque tétra nucléotide dans les contigs ou bien les deux. Bien qu'elles souffrent de limitations, les techniques de regroupement ont permis l'identification de nouveaux phyla microbiens et sont un moyen unique d'obtenir la séquence d'un génome de micro-organisme sans recourir à son isolement. Les méthodes d'analyses de métagénomes sans assemblage alignent les fragments de lecture

sur une base de donnée de référence permettant l'identification des espèces présentes dans l'échantillon même à faible abondance. Elles permettent aussi d'estimer l'abondance des espèces et également de prédire les fonctions métaboliques de l'ensemble des espèces de micro-organismes présents. La principale limitation est la caractérisation d'espèces microbiennes pas ou mal caractérisées car absentes de la base de données. De nombreux outils ont été développés pour répondre à ce problème.

1.4.3 Conclusion

Les techniques de métabarcoding et de métagénomique présentent toutes deux des avantages et inconvénients qui leur sont propres. Le métabarcoding est plus simple à mettre en place, nécessite moins de traitement informatique et est plus sensible mais il est biaisé par rapport à nos connaissances *a priori* et ne permet d'obtenir que l'information du marqueur moléculaire choisi (et non pas l'ensemble de l'information génétique). En revanche, la métagénomique permet l'amplification de l'ADN environnemental avec des *a priori* moins forts que le métabarcoding. De plus, la métagénomique apporte des informations fonctionnelles et peut permettre de reconstruire des génomes. Le séquençage à fragments de lecture longs va impacter différemment ces deux méthodes. Il va permettre de séquencer l'ensemble du 16s RNA (ou du 18s RNA pour les eucaryotes) et améliorer la description de la diversité environnementale par métabarcoding. Il va également améliorer les méthodes de regroupements et permettre d'obtenir des GAMs avec des régions plus complètess et plus spécifiques. La métagénomique et le métabarcoding sont deux méthodes de microbiologie environnementale complémentaires, qui essayent de répondre à des questions différentes.

1.5 Génomique environnementale

Dans cette section nous allons décrire très brièvement quelques récents succès de la génomique environnementale et comment ils ont transformé notre conception du monde microbien, en insistant sur l'importance biologique de ces découvertes.

1.5.1 *Prochlorococcus* ou le plus petit organisme photosynthétique

En 1979, Johnson et Sieburth décrivent la présence de bactéries extrêmement petites (moins de 0.6 μm de diamètre) oxyphototrophes dans un échantillon prélevé à 100 mètres de profondeur dans la mer des Sargasses (Johnson et al. 1979). Il faut attendre 1991 pour que, suite à la mise en culture de plusieurs souches de ces bactéries, elles soient nommées *Prochlorococcus marinus* (Fig. 14) (Chisholm et al. 1992). *Prochlorococcus marinus* possède un génome relativement petit : 1.65 Mbp et environ 1.700 gènes pour certaines souches reportées (Partensky et al. 1999). *Prochlorococcus marinus* est abondant

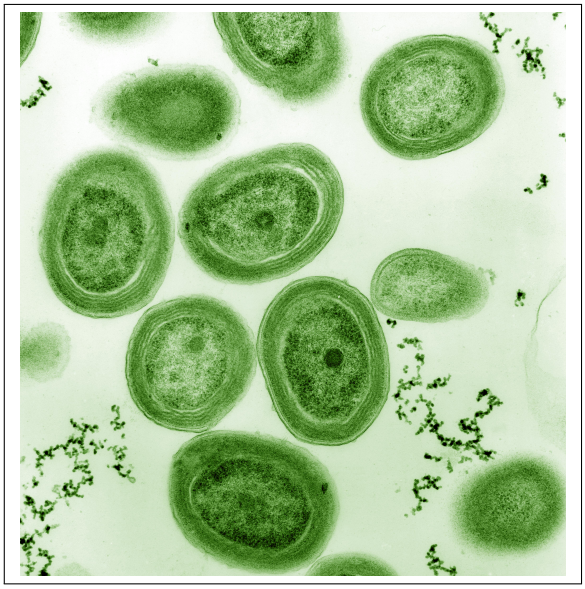


Fig. 14. *Prochlorococcus marinus*
Microscope à transmission éléctronique, coloration artificielle, MIT 2007 [1]

[1] https://www.flickr.com/photos/prochlorococcus/33750937901

et ubiquitaire entre les latitudes 40°N et 40°S à une profondeur de 100 à 200m. De plus, *Prochlorococcus marinus* est un organisme photosynthétique particulier. C'est le seul organisme phytoplanctonique à posséder des pigments de chlorophylles a et b avec un groupement divinyl (Chisholm et al. 1992). Certaines souches possèdent également des phycobiliprotéines. La combinaison de chlorophylle a et b avec la présence de phycobiliprotéines est un fait unique pour les organismes photosynthétiques oxygéniques. En raison de ces propriétés, il est estimé que *Prochlorococcus* est responsable de 8.5% de la production annuelle de carbone océanique (Partensky et al. 1999). La microbiologie environnementale a ainsi permis d'identifier un organisme abondant dans les océans qui possède une biologie inhabituelle et qui joue un rôle important dans le cycle du carbone et de l'oxygène.

1.5.2 Lokiarchaeota

En 2015, Spang et al. réalisent une étude métagénomique des fonds marins dans l'océan Arctique. En appliquant une méthode de regroupement de contigs, ils identifient les génomes d'organismes appartenant à un nouveau phylum : les Lokiarchaeota (Spang et al. 2015). Des caractéristiques métaboliques ont pu être déduites de ces génomes (Sousa et al. 2016). Ces archées possèdent des familles de gènes associées aux eucaryotes, notamment la présence d'homologues d'actine et de petite GTPase de la famille Ras. Ces protéines sont des composants structuraux du cytosquelette essentiels à la cellule eucaryote et un régulateur de la dynamique du cytosquelette, respectivement. Si, initialement, Lokiarchaeota avait été décrit comme potentiellement capable de phagocytose, des études ultérieures ont mis en doute cette conclusion (Martin et al. 2017). Une analyse phylogénétique a proposé que le phylum Lokiarchaeota soit le groupe frère des eucaryotes, suggérant qu'il soit le descendant du phylum archée ancêtre des eucaryotes (Spang et al. 2015). Ainsi, une étude sur la diversité microbienne dans les sédiments marins a permis de mieux comprendre l'apparition et l'évolution des eucaryotes. Au moment où j'écris ce manuscrit, une publication décrivant la culture d'une souche proche de Lokiarchaeota entretenant une relation syntrophique avec un organisme partenaire vient d'être déposé dans bioRxiv (Imachi et al. 2019).

Ce travail permet d'affiner le modèle de l'eucaryogènese et propose une endosymbiose phagocytose-indépendante.

1.5.3 CPR et DPANN, une biologie nouvelle

a) CPR, Radiation de Phyla Candidats (Candidate Phyla Radiation)

En 1998, Norman Pace utilise la méthode de métabarcoding qu'il a mise au point (1.5.1), et découvre un groupe de bactéries qui forme une division candidate qu'il appelle OP11 en amplifiant des 16S rRNA de sources chaudes du parc de Yellowstone (Hugenholtz et al. 1998). Une division candidate désigne un groupe de bactéries pour lesquelles il n'existe pas de représentant cultivé. En effet, l'isolation et la culture d'un clone sont considérées comme le plus haut niveau de preuve de son existence. En 2004, en utilisant des techniques de métabarcoding dans différents milieux, Harris étend considérablement la diversité de cette division candidate (Harris et al. 2004). Avec ces nouvelles données, plusieurs groupes candidats, qui potentiellement regroupent plusieurs phyla, sont identifiés au sein de la division candidate : OP11, OD1 et SR1. En 2012, Wrighton utilise des analyses de métagénomique couplées à des méthodes de regroupement et identifie 49 GAMs de bactéries appartenant à la division candidate (Wrighton et al. 2012). L'étude de ces génomes a montré qu'ils ne possèdent pas le cycle de Krebs, ni de nombreuses sous-unités des complexes de transport des électrons, ni de sous-unités de Nicotinamide adénine dinucléotide (NADH) déshydrogénases. L'absence de ces éléments indique un mode de vie anaérobie strict. C'est la première fois aussi que l'on remarque dans certaines bactéries de la division candidate la présence de la ribulose-1,5-bisphosphate carboxylase-oxygénase (RuBisCo) Erb2018. Cette enzyme est très étudiée car elle permet la fixation du CO_2 lors du cycle de Calvin-Benson-Bassham (CBB), présent chez tous les organismes photosynthétiques. Cependant, l'enzyme trouvée dans ces GAMs est un type particulier appelé RuBisCo II/III ne participant pas au cycle de CBB (Wrighton et al. 2016; Jaffe et al. 2019). La RuBisCo II/III serait utilisée pour récupérer et assimiler des composés organiques, suggérant un potentiel mode de vie de dégradeurs ou syntrophique. En 2013, Rinke valide et renforce ces résultats

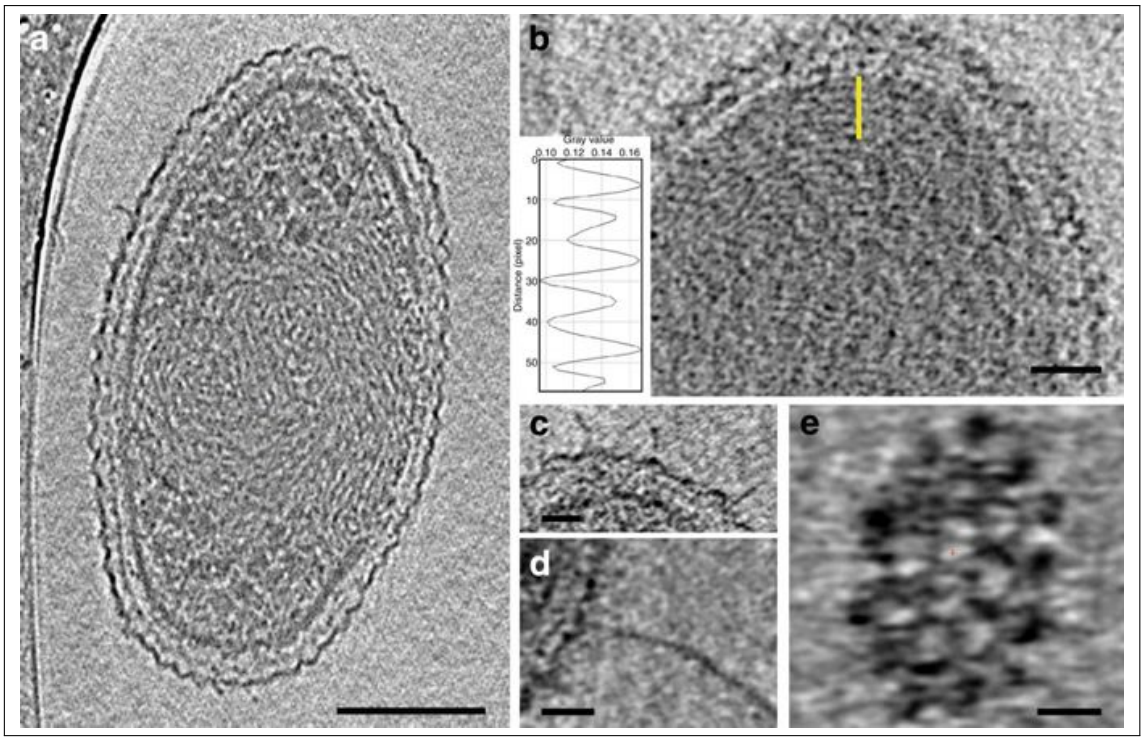


Fig. 15. Vue au microscope électronique de cellules ultra-petites échelle a) 100 nm b) 50 nm c, d, e) 20 nm. (Luef et al. 2015)

par des méthodes de séquençage de cellule unique (Rinke et al. 2013). Ces études montrent que les cellules de la division candidate sont très petites (résultats confirmés par Luef en 2015) (Fig. 15) (Luef et al. 2015); certaines ont même une taille qui approche la limite théorique inférieur de la vie avec un volume de 0.009–0.04μm^3.

En 2015, un article de Brown et al. montre que des organismes de plus de 35 phyla bactériens qui appartiennent à la division candidate échappaient à la détection par métabarcoding du 16S rRNA, du fait de la présence d'introns présents dans leurs 16 rRNA (Brown et al. 2015). L'année suivante, Hug et al. publient un arbre phylogénétique qui regroupe 1000 organismes non cultivés avec des organismes de référence pour un total de 30437 génomes intégrés à cette analyse (Fig. 16) (Hug et al. 2016). Cette publication a eu un grand impact et a été citée près de 500 fois en 3 ans. Cet arbre est remarquable car il met en évidence l'étendue des découvertes amenées par la microbiologie environnementale depuis une quinzaine d'années. Le nombre de clades récemment identifiés égale presque le nombre de clades qui étaient connus. Le terme CPR pour Candidate Phyla Radiation apparaît alors. Il désigne l'ensemble des groupes, phyla, super-phyla qui semblent former

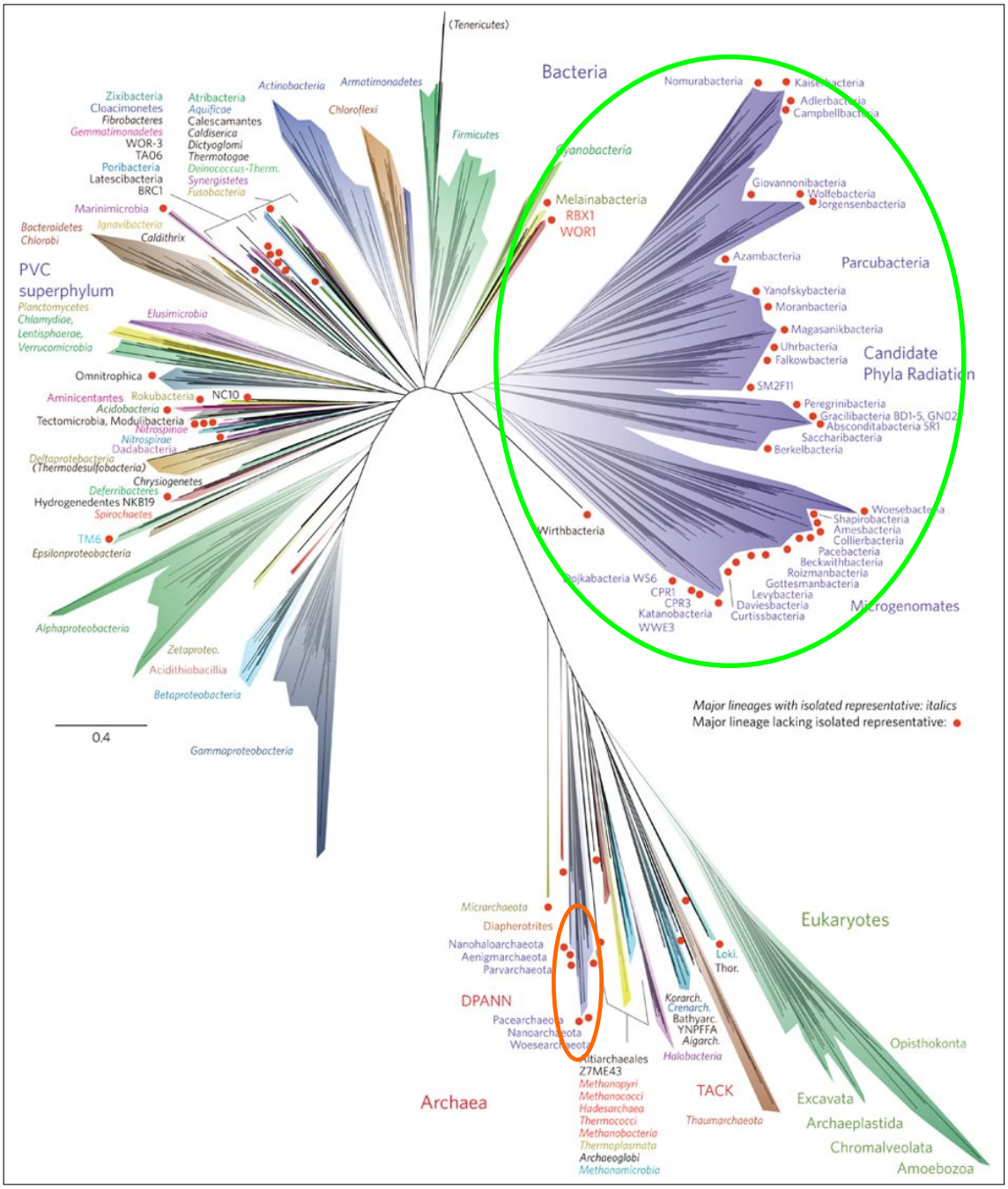


Fig. 16. Arbre phylogénétique incluant CPR et DPANN (Hug et al. 2016)
Les CPR sont entourés en vert. Les DPANN sont entourés en orange.

une radiation ou clade (Fig. 16), suite à la proposition de Brown en 2015(Brown et al. 2015). Même si quelques confusions dans son utilisation sont apparues dans la littérature, le terme CPR décrit l'ensemble de la radiation et non certains groupes en particulier. Aujourd'hui le nombre de phyla estimé dans les CPR est de 74 (Castelle et al. 2018b). Les CPR sont majoritairement trouvés dans des milieux pauvres en oxygène; la plupart ont été découverts dans des eaux souterraines, des sources chaudes, certains sont retrouvés dans l'eau douce et marine, dans le microbiome associé à des animaux y compris l'Homme; telles que *Candidatus Saccharibacteria* et *Candidatus Parcubacteria* (Dewhirst et al. 2010; Ling et al. 2014). Elles pourraient être impliquées dans certaines maladies inflammatoires du côlon (Kuehbacher et al. 2008). L'ADN de *Candidatus Parcubacteria* a même été retrouvé dans le sang humain.

b) DPANN

Le super-phylum archée DPANN a été proposé en 2013 par Rinke et al. et compte alors 5 phyla: *Diapherotrites*, *Parvarchaeota*, *Aenigmarchaeota*, *Nanoarchaeota*, *Nanohaloarchaea*. Rinke et al. décrivent comme attribut commun de ce super-phylum des tailles de génome et de cellule réduites. De plus, il met en évidence la présence de nombreux transferts latéraux dans ce clade. Il découvre ainsi le premier transfert latéral connu d'un eucaryote vers une archée (oxydoréductase d'une amibe), et la présence du facteur de transcription sigma essentiel à la transcription chez les bactéries qui n'avait été identifié qu'une seule fois chez les archées (Kirpide 1997) bien que leur rôle dans un organisme archée demeure inconnu. Membre des DPANN, *Nanoarchaeota* a été découvert avec l'isolation de l'hyperthermophile *Nanoarchaeota equitans* depuis des cheminées hydrothermales sous-marines. Ce dernier a été cultivé avec succès en association avec son hôte du genre *Ignicoccus* (Fig. 17). C'est un organisme de petite taille (400 nm), qui possède un génome extrêmement court (500 kb) et dense (95% de son génome est prédit comme codant pour des protéines ou des ARNs). De nombreuses voies de biosynthèse sont absentes de son génome: lipides, acide aminés, nucléotides, cofacteurs nécessaires au métabolisme. Il vit à la surface de son hôte. C'est donc un ecto-symbionte obligatoire.

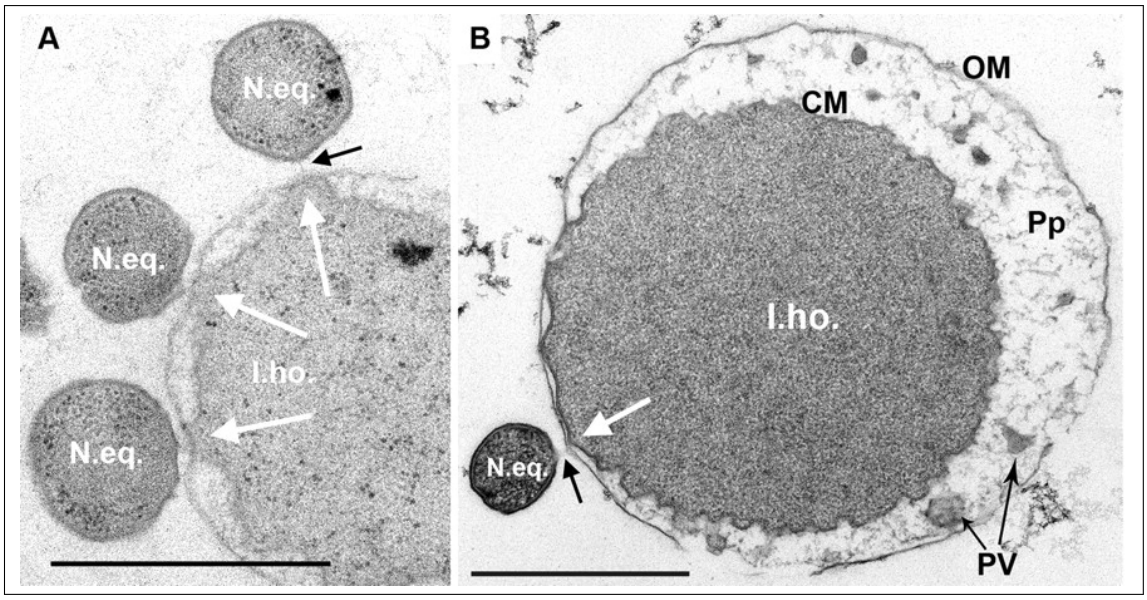


Fig. 17. *Nanoarchaeota equitans* et son hôte *Ignicoccus*
flèches blanches, points de contact; flèches noires, matière fibreuse entre les deux cellules; OM, membrane extérieure; Pp, périplasme; PV vésicule périplasmique; N.eq., N. equitans; I.ho., I. hospitalis (Jahn et al. 2008)

Des représentants du phylum *Nanohaloarchaeota* ont été identifiés dans deux lacs hypersalins, alors que ces environnements étaient considérés comme étant bien connus et échantillonnés. Mais l'application de la métagénomique sans les biais de l'isolation en culture ou de ceux du 16S rRNA a permis l'identification de génomes appartenant à ce phylum. *Nanohaloarchaeota* possède un génome relativement réduit de 1.2 Mb et les cellules font un diamètre d'environ 0.6 μm. Nanohaloarchaeota arbore également des traits inattendus, son taux de GC et la composition en acides aminés de ses protéines sont différents des Haloarchaea (archaea halophiles).

Des microorganismes identifiés dans des suintements d'eau dans une mine d'or appartiennent à un autre nouveau phylum : les *Diapherotrites*. Moins d'une dizaine de génomes sont connus (obtenus par séquençage de cellule unique). *Candidatus Iainarchaeum andersonii* en est le représentant connu car son génome est le plus complet, bien que relativement petit (1.2 Mb) et dense (90% code pour des gènes). Cet organisme possède des capacités cataboliques limitées mais théoriquement suffisantes pour un style de vie indépendant. Néanmoins, *Candidatus Iainarchaeum andersonii* est auxotrophe pour certains cofacteurs et acide aminés. Des analyses phylogénétiques ont montré que la majorité des

gènes anaboliques qu'il possède ont été acquis par transfert latéral depuis des génomes bactériens, suggérant une évolution du statut de symbiote obligatoire à celui d'organisme indépendant par acquisition de gènes bactériens.

Aujourd'hui, avec l'addition des *Altiarchaeota*, *Woesarchaeota*, *Micrarchaeota* et *Parvarchaeota* le super phylum DPANN compte 9 phyla. Le placement de ce clade dans l'arbre phylogénétique fait débat. En fonction des marqueurs choisis pour la reconstruction phylogénétique, le super phylum DPANN branche à différents endroits de l'arbre. La majorité des études décrivent néanmoins ce super phylum à la base des archées. Si le super phylum DPANN était à la base de l'arbre des archées, cela permettrait d'affiner nos prédictions sur le génome et le métabolisme de l'ancêtre commun des archées.

c) Une biologie différente

La plupart des CPR et des DPANN possèdent des génomes de petite taille (Castelle et al. 2018a) mais aussi de petits volumes cellulaires. La plupart des organismes de ces deux clades ont des capacités métaboliques et de biosynthèses limitées : absence de complexes permettant la respiration oxydative, absence de NADH déshydrogénase et cycle de Krebs incomplet (Castelle et al. 2018a). La plupart de ces organismes ne possèdent pas de voie de synthèse pour les acides nucléiques, les acides aminés, les lipides et le glucose (Castelle et al. 2018a). Certains ne possèdent pas l'information génétique pour métaboliser des lipides essentiels à leur paroi cellulaire et doivent donc les trouver dans leur environnement. De plus, les cofacteurs essentiels aux voies métaboliques prédites pour ces organismes ont également rarement été retrouvés. Si certains CPR et DPANN semblent être capables de vivre de façon indépendante, d'autres semblent être des symbiotes ou des parasites obligatoires du fait de leurs capacités métaboliques et de biosynthèse limitées (Wrighton et al. 2014; Anantharaman et al. 2016a; Probst et al. 2018). La dépendance des membres des CPR et DPANN aux autres organismes varie fortement en fonction des capacités métaboliques de chaque phylum. Les travaux sur leurs rôles et impacts sont encore récents, mais les CPR et DPANN semblent jouer un rôle important dans les cycles géochimiques (Castelle et al. 2015; Castelle et al. 2018a; Probst et al. 2017). La majorité d'entre eux semblent dégrader la matière biologique

(Wrighton et al. 2016). Par exemple, les *Altiarchaeota* seraient impliquées dans la fixation du carbone dans l'écosystème souterrain. En 2016, Anantharaman et al. ont montré que CPR et DPANN peuvent potentiellement prendre part à l'ensemble des cycles géochimiques: azote, carbone, soufre, fer, arsenic (Anantharaman et al. 2016b).

Les CPR et les DPANN avaient échappé à la détection par métabarcoding, car leurs 16S rRNA étaient trop divergents de ce qui était connu avant leur découverte par la métagénomique (Castelle et al. 2018a). La découverte des CPR et DPANN a changé notre vision et notre compréhension de la diversité microbienne et de son organisation. Il est difficile d'estimer l'impact qu'auront ces découvertes sur la microbiologie. Néanmoins, on peut affirmer que ces découvertes ont permis de valider de nouvelles méthodes pour étudier la microbiologie environnementale. Elles ont aussi bouleversé notre représentation du monde vivant, doublant la diversité des micro-organismes connue, avec une biologie particulière qui semble axée sur des relations symbiotiques ou parasitaires révélant les limites des techniques d'isolation et de culture pure pour la description et l'étude des micro-organismes dans l'environnement. Désormais, bien que l'échantillonnage continue à chercher de nouvelles diversités dans l'environnement, un effort important doit être mené pour caractériser plus finement le rôle écologique et géochimique de cette incroyable diversité afin de l'intégrer dans notre compréhension des écosystèmes microbiens. Au vu de l'impact actuel de ces découvertes, je suis persuadé qu'un tel effort fera évoluer notre compréhension des interactions entre micro-organismes.

1.5.4 Réduction de génomes, petite taille et théorie de la reine noire

Prochlorococcus marinus ainsi que la majorité des CPR et des DPANN sont des organismes de petite taille, de l'ordre du dixième de micromètre, avec un petit génome auquel il manque des gènes essentiels pour vivre de façon indépendante. Dans la littérature scientifique contemporaine, on trouve l'hypothèse que la plupart des CPR et DPANN sont des organismes parasites. Néanmoins, je m'interroge sur les possibles relations syntrophiques qui pourraient exister, notamment au sein de communautés microbiennes et de biofilms. Une relation syntrophique est une relation mutualiste dans laquelle les

partenaires doivent s'échanger des métabolites pour pouvoir assurer leurs propres voies métaboliques. Dans certains cas, il s'agit d'une adaptation, mais dans d'autres, cela peut aussi être une conséquence de la dérive génétique. Morris et al. ont présenté en 2012 l'hypothèse de la reine noire pour expliquer le phénomène de la perte de certaines fonctions métaboliques essentielles chez certains organismes (Morris et al. 2012). Cette théorie peut être résumée ainsi: si certains besoins d'un organisme peuvent être remplis par des organismes avec lesquels il est en contact, alors la perte des gènes assurant cette fonction chez le premier organisme peut être neutre voir avantageuse. En effet, l'organisme qui a perdu ses gènes ne dépense plus d'énergie ni de ressources pour maintenir cette fonction. Tant qu'un nombre suffisant de membres de cette communauté continue à assurer cette fonction pour le reste de la communauté, le système est stable mais une dépendance est créée. Dans le cadre de cette hypothèse, on parle de bénéficiaire et d'aidant. Rien n'interdit *a priori* des relations complexes où un organisme peut être un bénéficiaire pour une fonction spécifique et un aidant pour une autre fonction dans la communauté. Morris et al. prennent le cas de *Prochlorococcus marinus* qui est très sensible au stress oxydatif. Les micro-organismes des communautés auxquelles appartient *Prochlorococcus marinus* lui apportent une protection en filtrant les dérivés réactifs de l'oxygène. On pourrait considérer à première vue *Prochlorococcus marinus* comme un tricheur dans cette relation. Cependant, en tant que producteur primaire de carbone organique, il apporte une source d'énergie à la communauté microbienne à laquelle il appartient. Au vu du grand nombre de phyla récemment découverts avec des capacités métaboliques limitées, il est possible de mon point de vue que ces relations d'interdépendances soient plus communes que couramment présumé. De plus, on pourrait imaginer un niveau de dépendance encore plus élevé où les voies métaboliques sont partagées entre individus d'une communauté (Foster et al. 2012; Pande et al. 2017; Ponomarova et al. 2015).

Problématique

Notre conception du monde microbien a été altérée par les récentes découvertes réalisées grâces aux analyses méta-omics. Un potentiel descendant de l'ancêtre archaea des eucaryotes a été identifié dans le super phylum Asgardarchaeota. Plusieurs phyla de microbes extrêmement petits, probablement parasites ou symbiotes, remettent en question notre compréhension du monde microbien et particulièrement son organisation et son impact potentiel sur les cycles géochimiques et l'environnement. Cependant, des questions, des inconnues et des limites demeurent. Les capacités en puissance de calcul sont aujourd'hui insuffisantes pour utiliser des méthodes exactes et de nombreuses heuristiques sont utilisées. Une proportion considérable de séquences reste sans annotation fonctionnelle ou taxonomique. Les limites actuelles des technologies de séquençage rendent probable que des phyla de microbes peu abondants continuent d'échapper à la détection.

Dans cette thèse, je me suis intéressé aux objets les plus prometteurs pour faire de nouvelles découvertes en microbiologie, plus précisément à la matière noire microbienne. Les principales questions auxquelles j'ai essayé de répondre sont les suivantes:

- Quel est l'impact de la matière noire microbienne sur nos connaissances et comment l'étudier ?
- Que peut apporter la théorie des graphes à la microbiologie environnementale ?
- Comment travailler avec de larges jeux de données de microbiologie environnementale ?
- Comment retrouver des homologues distants dans de larges jeux de données de microbiologie environnementale ?
- Comment développer les méthodes de réseaux pour étudier la diversité dans les jeux de métabarcoding?
- Les procaryotes extrêmement petits, récemment découverts, jouent-ils un rôle dans l'écologie des communautés microbiennes et les cycles biogéochimiques ?

Au début de ma thèse, les organismes ultra-petits avaient principalement été décrits dans les aquifères, et un jeu de données originales de métagénomique possédant une fraction de taille ultra petite venait d'être publié : TARA OCEANS (Sunagawa et al. 2015). J'ai étudié le rôle possible de la matière noire microbienne ultra petite dans les cycles géochimiques

des océans et plus particulièrement la présence de voies métaboliques autotrophes. Ce travail sur les métabolismes devait être un court travail introductif avant de commencer l'application de la théorie des graphes à la microbiologie. Je me suis donc intéressé, tardivement, à la théorie des graphes qui permet l'étude des relations entre des entités. J'ai commencé par participer à la rédaction d'une revue méthodologique avec le Dr. A. Watson sur l'application de la théorie des graphes pour l'étude de la microbiologie, de l'évolution et de la diversité phylogénétique dans l'environnement. Je me suis servi de ce que je venais d'apprendre avec la théorie des graphes pour étudier la diversité d'organismes unicellulaires marins phylogénétiquement proches des animaux lors d'une collaboration avec la Dr. Alicia Arroyo-Sanchez. Finalement, j'ai également développé une nouvelle méthode pour retrouver des homologues environnementaux divergents dans de grands jeux de données.

Résultats

3.1 La matière noire microbienne

Nous avons observé dans l'introduction que la microbiologie est une discipline récente dont les progrès sont fortement dépendants des technologies disponibles. Nous avons également expliqué comment les nouvelles technologies de séquençage couplées aux progrès de la bio-informatique ont permis à la microbiologie environnementale d'apporter des résultats stupéfiants: découverte d'extrêmophiles qui repoussent les limites connues de la vie, de nouveaux groupes de micro-organismes dont certains nous éclairent sur nos origines et permettent le développement d'une nouvelle biologie fondée sur l'interdépendance. Néanmoins, dans les jeux de données de métagénomique, une grande partie des séquences moléculaires n'a pas d'annotations fonctionnelles ou taxonomiques. On parle alors de matière noire microbienne. Ce terme ne fait pas consensus au sein de la communauté scientifique et sa définition est sujette à discussion. Cette dénomination est une analogie, inspirée par la matière noire utilisée en physique pour rendre compte des observations à l'échelle astronomique. En physique, il s'agirait d'une matière qui n'interagit pas avec la lumière mais avec la gravité. On l'appelle matière noire car elle n'a jamais été observée directement, son existence est déduite d'observations aberrantes à première vue. Sans cet ajout de matière noire dans les modèles, la physique actuelle ne peut rendre compte des observations. En microbiologie environnementale c'est l'inverse, nous observons des entités dont nous ne connaissons ni la fonction ni l'origine phylogénétique. Il est probable qu'avec les moyens actuels nous n'observions pas l'ensemble des micro-organismes, c'est pourquoi il est important de continuer le développement des techniques de microbiologie environnementale et de poursuivre l'effort d'échantillonnage. L'annotation automatique de séquences est réalisée par comparaison à des bases de données de référence. On compare l'inconnu à ce qui est connu en espérant trouver des similarités et en déduire une annotation. Dans les bases de données d'organismes connus et isolés, de nombreuses séquences n'ont pas de fonction connue. Par exemple, dans la base de donnée Pfam (Finn et al. 2014),

référence pour l'annotation fonctionnelle des domaines protéiques, il existe des domaines annotés DUF, Domain of Unknown Function (domaine de fonction inconnue), retrouvés dans plusieurs organismes. On peut distinguer les séquences environnementales qui sont similaires à des séquences connues dont la fonction est inconnue des séquences environnementales qui n'ont pas de similarité avec des séquences connues. Le dernier cas est bien plus stimulant car il a le potentiel de permettre la découverte d'une biologie différente.

Dans l'article suivant, nous nous sommes intéressés à la matière noire microbienne. Qu'est-elle et quelle proportion représente-t-elle dans les données environnementale séquencées? Est-elle le résultat d'un artefact des méthodes que nous utilisons pour étudier les données de métagénomique? Que nous apprend son existence et que pouvons nous espérer y trouver?

Dans cet article, nous soutenons que l'exploration de la matière noire microbienne devrait être une priorité de la microbiologie, au vu du nombre de séquences environnementales qui ne possèdent pas d'annotations fonctionnelles ni phylogénétiques et du nombre de microbes que nous ne savons pas cultiver en laboratoire. L'analyse de la matière noire a déjà commencé à transformer nos connaissances sur l'évolution (archaea, asgard...) et remet même en cause notre conception de l'unité évolutive. Premièrement, le nombre de microbes que nous ne savons pas cultiver en culture pure représenterait *a minima* 90% de la diversité microbienne. Deuxièmement, nous avons observé de nombreuses interactions et dépendances obligatoires entre micro-organismes au sein de communautés; certains organismes ne pouvant se cultiver qu'en présence de partenaire(s) microbien(s). Troisièmement, certaines maladies humaines sont dues à une diversité anormale du microbiome; être un individu en bonne santé c'est avoir un microbiote sain. Ces trois observations contrastent avec la vision que la sélection naturelle cible les individus les plus adaptés. En effet, si pour se développer des organismes ont besoin de partenaires, peut-on les considérer comme des unités évolutives isolées? Ainsi, l'étude de la microbiologie environnementale et de la matière noire microbienne nous force à revoir notre définition d'unité évolutive.

De ce fait, la notion d'holobionte (Moran et al. 2015; Mindell 1992; Theis et al. 2016), qui correspond à une unité évolutive incluant l'hôte et son microbiome est aujourd'hui

considérée et étudiée par certains chercheurs. De plus, les micro-organismes évoluent dans des communautés dans lesquelles différents processus surviennent particulièrement: échange d'information génétique (transfert latéral)(Jaffe et al. 2019; Soucy et al. 2015), communication (quorum sensing) (Miller et al. 2001), et modification de l'environnement (construction de niche) (McNally et al. 2015). Ces processus, bien que connus, sont encore peu intégrés à nos modèles. Il est probable que comprendre le rôle de la matière noire microbienne nécessitera *a minima* d'étoffer nos modèles afin de mieux prendre en compte les interactions entre micro-organismes. Les découvertes de la microbiologie environnementale ont également permis le développement de nouvelles techniques (enzymes de restriction, PCR…). La production des enzymes recombinantes est un marché de plusieurs milliards de dollars. La découverte d'enzymes plus performantes, plus spécifiques ou de nouveaux antibiotiques représentent des enjeux sociaux et économiques importants. La microbiologie environnementale peut donc participer à découvrir, avec des approches d'évolution dirigée et de design *de novo*, de nouvelles enzymes d'intérêt. *In fine* la microbiologie environnementale et l'étude la matière noire microbienne ont le potentiel de modifier la connaissance des micro-organismes et d'améliorer nos conditions de vie.

3.1.1 Article 1, "Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery", Genome Biology and Evolution (Bernard et al. 2018)

Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery

Guillaume Bernard, Jananan S. Pathmanathan, Romain Lannes, Philippe Lopez, and Eric Bapteste*

Sorbonne Universités, UPMC Université Paris 06, Institut de Biologie Paris-Seine (IBPS), France

*Corresponding author: E-mail: eric.bapteste@upmc.fr.

Accepted: February 5, 2018

Abstract

Microbes are the oldest and most widespread, phylogenetically and metabolically diverse life forms on Earth. However, they have been discovered only 334 years ago, and their diversity started to become seriously investigated even later. For these reasons, microbial studies that unveil novel microbial lineages and processes affecting or involving microbes deeply (and repeatedly) transform knowledge in biology. Considering the quantitative prevalence of taxonomically and functionally unassigned sequences in environmental genomics data sets, and that of uncultured microbes on the planet, we propose that unraveling the microbial dark matter should be identified as a central priority for biologists. Based on former empirical findings of microbial studies, we sketch a logic of discovery with the potential to further highlight the microbial unknowns.

Key words: metagenomics, eukaryogenesis, microbial evolution, tree of life, web of life, CPR bacteria.

Introduction

Microbial studies are fascinating. Not only their findings can deeply transform knowledge in a broad range of scientific fields (from evolutionary biology to zoology and medical and environmental sciences) but also, whereas philosophers of sciences debate whether there is such thing as a logic of scientific discovery (Schickore 2014), microbial studies provide biologists with a set of empirical rules to enhance one's chances to discover novel and unexpected life forms. This unique potential of microbial studies to reshape knowledge has been recognized relatively recently, even though there is a long standing history of studies of microbial pathogens, involving famous early researchers such as Robert Koch, Louis Pasteur, or Martinus Beijerinck. If the laymen nowadays appreciate that microbes impact our everyday life (i.e., via their fermentative roles in food production), and know that microbes also impacted our recent human histories (i.e., via their contribution to major pandemics; Diamond 1997), from a scientific perspective, microbes are nonetheless rather novel objects of studies. There are both technical and conceptual reasons for this late yet broad recognition of microbes, as we will highlight below, whereas providing an empirical recipe for further insights into the microbial dark matter.

In 1619, the famous astronomer Galileo, whose observations of the moons of Jupiter had threatened the geocentric theory, modified a telescope to magnify nearby terrestrial objects. Although he clearly was a revolutionary thinker, he found these observations of the minute world of limited interest, and, only 6 years later, did his friends name *microscopio* the strange inverted telescope Galileo had invented (Falkowski 2015). By contrast, Robert Hooke, an English polymath scientist, and, later, Anton van Leeuwenhoek, who did not belong to the academic world, were much more excited by describing their microscopic observations. In 1671, van Leeuwenhoek, who had substantially changed the design of the microscope to enhance its magnifying power, initiated a series of striking findings: microscopic lifeforms are abundant and everywhere to be seen. Microbes, who had populated Earth for over 3.5 billion years, were for the first time exposed to the human eye (Falkowski 2015). Both a technical progress and an uncommon ability to delve into an unseen world were critical components of that progress. However, since biological theory at the time considered the living world was distributed into two major groups: plants and animals, van Leeuwenhoek naturally assumed he was observing populations of minute animals (with tiny organs), when microbes were mobile, rather a new kind of living beings. In that sense,

the unveiled microbiological world was first rationalized in ways that fit within preexisting theoretical categories derived from the known living world. Importantly, neither Hooke nor van Leeuwenhoek had immediate scientific successors. Arguably, it took another 200 years (Falkowski 2015), and several novel conceptual and technological developments to formulate an issue, currently at the forefront of microbial studies: « is it possible that unknown microorganisms, with different properties than those currently associated with the known living world, are thriving in nature? ».

The potential theoretical importance of such "known unknowns" and even "unknown unknowns" of the microbial world (e.g., unknown genes, genomes, functions, organisms, processes, and communities associated with uncultured microbes and viruses), that were often popularized under the catch-phrase "microbial dark matter," should not be underestimated. Interestingly, the relevance of this sentence is debatted in microbiology. Many scientists find the metaphor misleading or inaccurate, because the "microbial dark matter" does not correspond to the dark matter studied by astronomers and physicists. This latter represents a hypothetical, still unobserved, although widely accepted, kind of matter, which does not interact with light but interacts through gravity. Taking the mass of this unseen astronomic dark matter into account would explain the uncorrect predictions of the movement of galaxies by classic astronomy theories. This astronomic dark matter is thus unquestionably different from the microbial dark matter. However, other microbiologists have endorsed the analogy (Rinke et al. 2013; Lobb et al. 2015; Lok 2015; Saw et al. 2015; Bruno et al. 2017; Krishnamurthy and Wang 2017; Lewis 2017), since the sentence nonetheless conveniently stresses that, to some extent, newly discovered microbes can harbor a different biology from those that had been cultured. Although we agree that microbial and astronomic dark matter are very different notions, we also find the sentence "microbial dark matter," popularized by (Rinke et al. 2013) to be more useful than detrimental. First, it is a convenient short hand for the idea that unknown microbial life may be playing important and even dominant role in ecosystem processes. Second, it has some editorial and educational virtues, as it effectively helps raising the interest for microbiology studies beyond the field of microbiology (in which none would really conflate astronomic and microbial dark matter), surely enhancing the general interest for the unexplored diversity of microbes and their genes. We recommend however a more careful rather than sensationialistic use of the term, to describe the (overwhelming) amount of microbes, microbial genes, and microbial contributions to processes that were unknown at the time at which scientists performed their analyses.

Precisely, much of the extant knowledge in biology, that is, about biological entities and biological processes, heavily relies on analyses conducted on macro-organisms and on cultured microbes. Yet, 60–99% of the microbial diversity are not easily culturable, or are not culturable using standard techniques (Staley and Konopka 1985; Barer and Harwood 1999). Unraveling the microbial dark matter could thus led to two (nonexclusive) types of observations. Either the discovery of hidden microbes will show that microbes unveiled from the microbial dark matter are comparable in terms of genetic diversity, ecological roles, abundance, evolutionary history, and affected by processes similar to those affecting cultured microbes, in which case our current knowledge of microbes is representative of what's really going on in nature (we will simply find more of what we already knew by mining the microbial world); or the microbial dark matter will prove to host entities and processes that differ from those already described, with the major consequence that scientific knowledge will not only need to be completed but also corrected as microbiologists gain access to this still hidden microbial world in order to consider new phenomena, poorly explained in extant theories. Such significant theoretical transformations have arguably occurred when 1) microbiologists looked for life in extreme environments, 2) detected life under unexpected (i.e., very diverged) forms, and 3) unveiled new processes involving microbes, which allows us to stress some key features for the success of a scientific research oriented toward the discovery of microbiological novelty.

Searching Life in Extreme Environment: A Few Lessons

The developments of molecular markers and sequencing techniques were instrumental for the discovery of extremophiles. By unveiling the archaea, a novel early branching Domain of life, possibly sister-group to eukaryotes, Carl Woese's phylogenetic studies of the 16 S RNA revolutionized the views on the entire biological world (Woese and Fox 1977; Woese et al. 1990). Woese argued that, rather than being partitioned into two major groups, the eukaryotes and the prokaryotes, the living world encompassed a much broader microbial diversity, justifying its classification into three Domains of life. Subsequently, Woese and his colleagues (referred to as "the Woese army" by Lynn Margulis; Doolittle 2013) actively promoted this position, bringing the newly termed "archaea" into full light, while intending to ban the use of the "older" term "prokaryotes" (Pace 2006).

Importantly, this comparative approach of molecular phylogenetics was later coupled to a phase of exploratory science (Waters 2007). Exploratory science is in essence a strategy of data mining. It goes from the data to the hypotheses (Burian 2013), seeking (robust) patterns in the data or unraveling new phenomena. Although microbiology has a long history of exploratory research (O'Malley 2014), this mode of science appears in strong contrast with the more classic hypothetico-deductive strategy, heralded by Karl Popper. This deductive approach has inspired much of microbiology

and biochemistry studies, since these studies largely operated from the hypotheses to the data, that is, using data to reject preexisting hypotheses, or eventually to corroborate them. Since exploratory science is not first aimed at rejecting (or confirming) preestablished hypotheses (thus deepening current knowledge), it can potentially produce novel, unexpected knowledge, or simply fail, making the financial and scientific investment in exploratory studies especially risky.

Fortunately, the pioneering approach, first largely based on the development of 16 S rRNA gene sequencing (Schmidt et al. 1991; Barns et al. 1996; Hugenholtz et al. 1998), then on the sequencing of other makers (Beja et al. 2000), and latter on the development of metagenomics (Breitbart et al. 2002; Tyson et al. 2004; Tringe et al. 2005) and single-cell genomics, bypassed the need for culture studies, thereby lifting a blind spot imposed by culture-based investigation to comparative analyses. These studies returned a diversity of exciting findings. By the beginning of the 2000s, microbial ecologists had started characterizing the gene content, diversity, and relative abundance of environmental microbes (Venter et al. 2004). They had identified new functions of major importance in the ocean (e.g., ammonia oxidation by archaea; Francis et al. 2005), possibly affecting the global nitrogen cycle, as well as unexpected photosynthesis (and other) genes in viruses (Sullivan et al. 2005). They had also gained unprecedented insights into the survival strategies of microbes (Tyson et al. 2004), into their community structures (Tyson et al. 2004; DeLong et al. 2006), and into their niche-specific adaptations (Tringe et al. 2005), for example, by unraveling unknown iron-oxidizing and free-living diazotroph in acid mine drainage biofilms (Ram et al. 2005; Tyson et al. 2005).

Environmental genomics in particular produced remarkable results when microbiologists turned their eyes to extreme regions (in terms of temperature, pH, pressure, mineralization, radiations) that many considered a priori devoid of life (Pikuta et al. 2007). The seemingly counter-intuitive idea to sample lifeforms in environments hostile to life unveiled a broad diversity of extremophiles in the three Domains. Granted, finding DNA in extreme environments does not in itself constitute an ultimate proof that the life forms bearing this DNA existed there, but analyses of environmental DNA (be they nonassembly based, assembly based or even of genome resolved metagenomics) are nonetheless an important step in the discovery of new microbes in extreme environment. Cultivation of microbes from these extreme locations offers a much stronger evidence, that is, Karl-Otto Stetter, by this cultivation approach discovered life at the extreme temperature limits, pushing the boundaries of life as it was then known (Stetter 2013).

Using these strategies, microbiologists realized that life was possible at temperature 122 °C, at negative pH (!), and at pH$>$ 11, at pressures exceeding 1,200 atmospheres; that microbes could be resurrected after 20–40 millions of years of dormancy, survive 2.5 years of travel in space, and thrive within rocks as well as in the terrestrial stratosphere (at $>$ 44 km of altitude) (de los Rios et al. 2003; Pikuta et al. 2007) (see, e.g., https://www.slideshare.net/AnjaliMalik3/extremophiles-imp-1). Some of these statistics were so unexpected that Pikuta *et al.* (Pikuta et al. 2007), summarizing the ongoing knowledge on extremophiles drew too short axes for temperature, pH, and salinity on plots showing the physicochemical conditions compatible with life. Some environmental microbes were definitely outliers with respect to the majority of known creatures. This counter-intuitive search for extremophiles likely reaches his summit in astrobiological studies, which search for life beyond Earth, seeking to define biomarkers in exoplanetary analogs and to train to detect these biomarkers in regions of the universe that currently fit the minimal requirements for life in C, H, N, O, P, S, liquid water, and energy (Olsson-Francis and Cockell 2010). No one knows whether extraterrestrial microbes will ultimately be discovered this way, but, at least, ironically terrestrial microbes, which can grow in the International Space Station and Spacecraft Assembly Facilities (Checinska et al. 2015) have potentially increased chances to spread in space, a problem known as the issue of planetary protection (McKay and Davis 1989).

Searching for Very Divergent Homologs: A Few Lessons

In as much as environmental genomics enhance microbial dark matter studies, for example, by unraveling extremophiles, it also raises issues, since environmental genomics has its own blind spots. The selection of samples, of genes of interests (e.g., in metabarcoding projects, or more generally in targeted environmental genomics) and the many filtering decisions and heuristics in the subsequent bioinformatic treatments imposed by the wealth of environmental sequences (i.e., reads and contigs), as well as the increased standardization of the methods and questions of environmental genomics studies (a logical scientific development for a comparative science; Vigliotti et al. 2017) raise the risk that the most unexpected of life forms, even if already sequenced, remain drowned under this deluge of data. This risk has notorious roots: our observations are strongly constrained by what our theory makes us prone to expect, and therefore by former perspectives informing various criteria in the sampling process.

This limit is obvious in the process of size-fractioning associated with metagenomics analyses, such as the one conducted in the Tara expedition, which a priori optimized the net sizes of its filter to capture different taxa of marine microbes (Karsenti et al. 2011). This procedure entails the inherent risk that important players of the microbial world may be overlooked if their sizes do not satisfy these filtering conditions. For example, 10 years ago, few (or even no)

microbiologists nor virologists would have assumed that bacteria in the range of 0.2 microns and viruses >0.2 microns existed (Council 1999). This view radically changed with the discovery of ultrasmall bacteria, aka nanoorganisms, such as the CPR in 2015 (Brown et al. 2015; Luef et al. 2015) or some DPANN in 2010 (Baker et al. 2010), and with the discovery of giant viruses, such as Mimiviridae, in 2003 (La Scola et al. 2003). These taxa are now found in diverse environments, albeit at low abundance(Brown et al. 2015). CPR are remarkably phylogenetically diverse (Hug et al. 2016), representing up to 50% of the bacterial domain (Anantharaman et al. 2016), and present an unusual biology (i.e., 16 S RNA with insertion, lack of metabolic genes usually considered as essential), which suggests that CPR depend on other life forms (Kantor et al. 2013; Gong et al. 2014; Brown et al. 2015; Nelson and Stegen 2015; Danczak et al. 2017). CPR cells occupy an extremely tiny average volume of $0.009 \pm 0.002\ \mu$m3, for a spherical diameter of 253 ± 25 nm (Luef et al. 2015). Mimivirus biology is not less striking. In particular, they are hosts to yet another new kind of viruses: virophages, that is, viruses of giant viruses (Boyer et al. 2011). The phylogenetic position of these relatively newcomers, especially regarding how deep CPR and giant viruses branch (if they do) with respect to the other Domains of life, is heavily debated (Colson et al. 2012; Moreira and Lopez 2015; Hug et al. 2016), even though, regarding the phylogenetic position of CPR, Hug *et al.* did not committ themselves strongly, stressing instead that their method did not result in a well resolved phylogeny (Hug et al. 2016). Such debates illustrates that attempts to establish novel groups inevitably (and logically) arise resistances, but no one questions that an accurate picture of the microbial world and its evolution can any longer satisfactorily be achieved without including nanoorganisms and viruses, be they giant or not.

Environmental genomics has not merely unraveled new microbial lineages, it has also reported new gene families (Riesenfeld et al. 2004; Lok 2015), new CRISPR-Cas systems (Burstein et al. 2017), and unusual gene forms (i.e., very divergent homologs from known genes). In principle, newly sequenced environmental genes could fall into one of 4 groups (fig. 1). The in silico functional and taxonomical annotations of environmental genes using existing ontologies (here, applied to 339 metagenomes; Fondi et al. 2016, sampling a diversity of environments, that is, soil, seawater, inland-water, wastewater, host, air, bioremediation, biotransformation, and sludge waste) indicates that most environmental genes have unknown functions, and belong to uncharacterized microbial lineages (fig. 2). In fact, at the minimum %ID threshold of 95%, >50% of these genes are neither functionally nor taxonomically annotated, and at the minimum %ID threshold of 50%, >30% of these genes are neither functionally nor taxonomically annotated, which stresses the genuine abundance of microbial dark matter in metagenomic data.

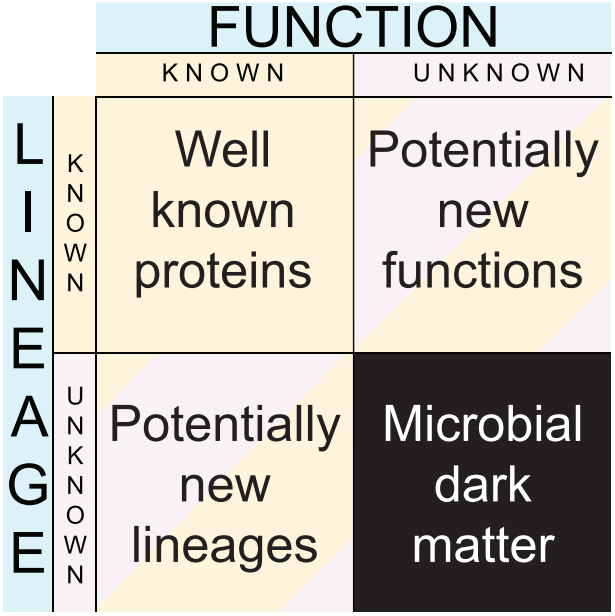


Fig. 1.—Four types of environmental sequences. Environmental sequences can be classified based on their taxonomical annotation (horizontal line) and their functional annotation (vertical column), which defines four categories. The cells in purple and black correspond to categories that are not readily explained based on current biological knowledge.

Bioinformatic developments are currently designed to associate these unknown genes to reference gene families. For example, the search for highly divergent homologs using sequence similarity networks (Lopez et al. 2015) highlighted that a large majority of the ancient gene families that are well-conserved in cultured microbes have extremely divergent homologs in nature. Lopez et al. (2015) proposed that at least some of these very divergent homologs might sign the existence of deep branching yet unseen major divisions of life. Discovering environmental deeper lineages, branching below the currently recognized prokaryotic domains, could reopen the debate on the number of Domains of life, questioning our fundamental knowledge in terms of biological classifications and regarding early life evolution. Bioinformatic studies of random environmental sequences however need to be complemented by another type of experimental evidence, that is, individual sequences of genomes from putative very early branching microbes or even isolations of these organisms. The former type of evidence typically obtains by genome resolved metagenomics, that is, genome binning from metagenomics data sets. Genome binning consists in assembling metagenomic contigs using relative abundance and/or tetra nucleotide abundance (Sedlar et al. 2017). This protocol allows to recover synteny and to identify conserved or unusual/unexpected genes for related microorganisms. This approach is invaluable to recover genomes for uncultured organisms and to study their metabolic capabilities.

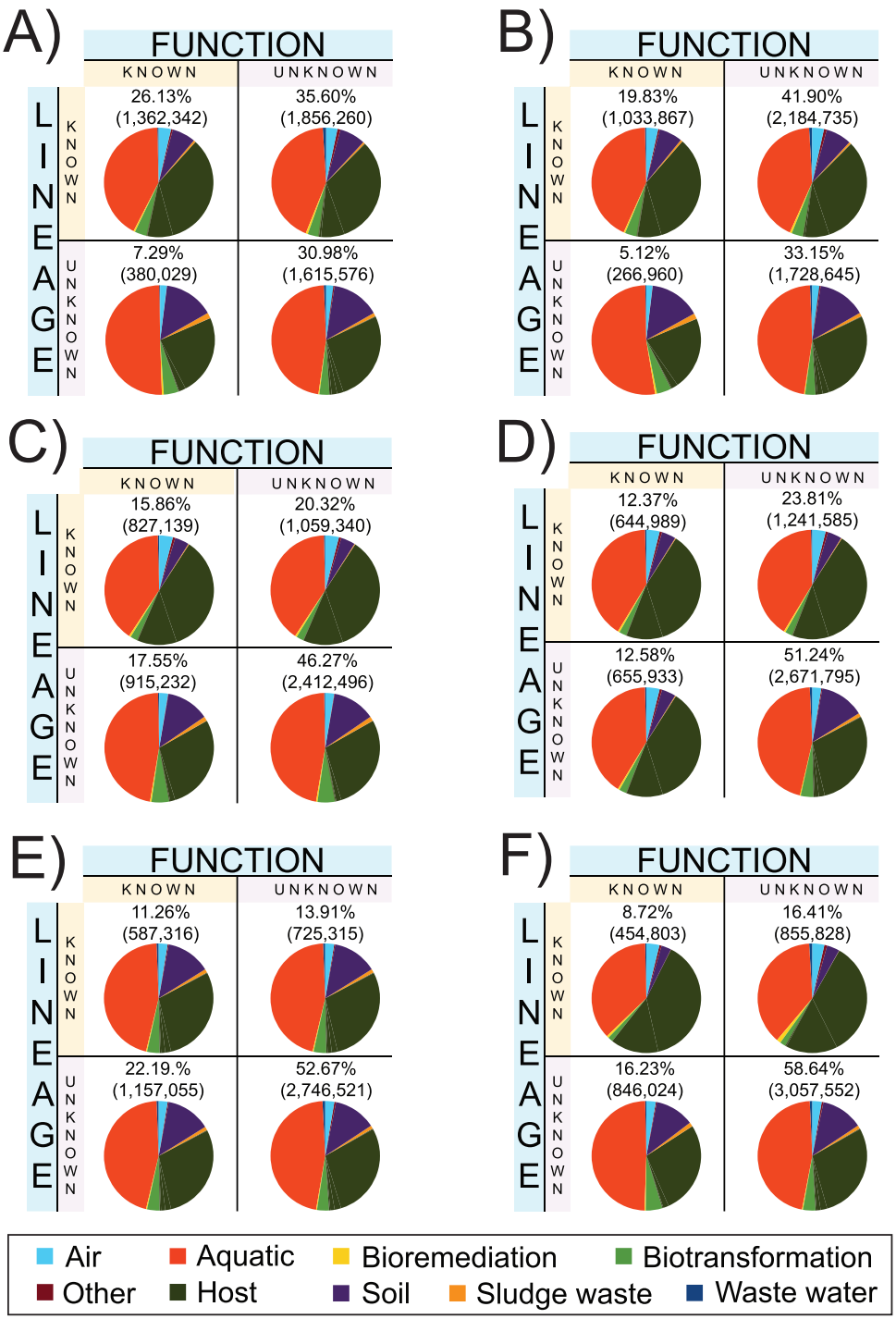


Fig. 2.—Microbial dark matter across a diversity of environmental samples. Proteins inferred (with FragGeneScan; Rho et al. 2010) based on Metagenomic sequences from (Fondi et al. 2016), clustered based on their taxonomy (using MEGAN 6; Huson et al. 2016) and functional (using EggNOG-mapper; Huerta-Cepas et al. 2017) annotation. The pie charts represent the proportion of proteins from each type of environment. The taxonomy annotation was performed using three minimum percentage of identity: 50% (panels *A* and *B*), 85% (panels *C* and *D*), and 95% (panels *E* and *F*). In panels *A*, *C*, and *E*, the proteins were clustered based on their functional annotation including the category S ("Function unknown"). Panels *B*, *D*, and *F* were clustered with the exclusion of the category S.

Moreover, within the field of environmental genomics, single cell genomics offers an additional alternative approach to produce environmental data sets, identifying genes from the same genomes. Even though these approaches are gaining popularity and data start accumulating, so far, despite the actual high number of environmental "known unknowns" no scientists (i.e., peer-reviewers) working with major scientific journals have yet been convinced that enough evidence for new candidate Domains of life is available. For example, the remarkable work by (Parks et al. 2017) did not use universally shared ribosomal proteins to build a tree of life, including simultaneously novel environmental lineages, as well as known archaeal and bacterial lineages, whereas this strategy could have identified deep branching environmental groups.

Microbial Processes as a Yet Unexhausted Source of Knowledge

At the same time that new microbes were discovered, our knowledge on processes involving or affecting microbes evolved substantially. The focus on interactions and the use of networks rather than trees to frame microbial studies is emerging as a major trend. It is becoming obvious that simple tree-based models, aiming at reconstructing the divergence of lineages from a last common ancestor, are not fully doing justice to the diversity and complexity of the processes explaining microbial evolution. For example, in nature, diversity generating retroelements contribute to rapid, targeted sequence diversification in Archaea and their viruses (Paul et al. 2015), and in CPR (Paul et al. 2017). Introgressive processes such as lateral gene transfer stress the collective dimension of microbial evolution (Doolittle 1999; Ochman et al. 2000; Bapteste et al. 2012). Likewise, the discovery of environmental microbes with genuinely incomplete genomes (i.e., lacking genes considered as essential) and of syntrophic consortia insists on the importance of metabolic, ecological, and evolutionary scaffolding in the microbial world (DeLong 2007; Morris et al. 2012; Sachs and Hollowell 2012; Caporael et al. 2013; Brown et al. 2015; Ereshefsky and Pedroso 2015). The claim that in nature microbes depend on other microbes to survive, contrasts strongly with the notion that natural selection ultimately favors individual optimized lineages via the success of the fittest cells among large and phylogenetically homogeneous microbial populations. It matches however well with the empirical observation that pure culture fails for most microbes (Staley and Konopka 1985), and in fact provides an explanation for this great plate anomaly. Microbes belong to collectives rather than they live alone. Other striking interactions are also unveiled as scientists dig further into the microbial world. For example, unheard forms of communication impact microbial and viral population dynamics (Erez et al. 2017). Microbiomes and their hosts coconstruct a broad range of animal and plant phenotypes (Gill et al. 2006; Gilbert et al. 2015), to the point that some propose to introduce holobionts (the emergent associations of hosts and microbes) as a novel kind of central evolutionary player (Bordenstein and Theis 2015; Moran and Sloan 2015; Theis et al. 2016). At an even broader scale, in the environment, microbes, most of which are unknown, are now assumed to affect the geochemical processes that shape our planet (Guidi et al. 2016) and, by a process called niche construction (Laland et al. 2016), these microbes are considered likely to impact ecosystems and the future of life. All these processes (lateral gene transfer, scaffolding, communication, microbial coconstruction, and niche construction), while widespread in the microbial world, are still rather peripheral in biological explanations. Introducing the processes to which microbial dark matter contribute within biological theory thus requires revising the relative priority currently attributed to concepts in scientific explanations, which is likely to be a slow and tedious epistemic process. For example, prokaryotic biology, especially when considering microbiomes, appears in fact so different from the biology of model eukaryotic organisms that several evolutionary biologists and theoreticians have independently suggested that key aspects of the classic Darwinian theory and of the Modern Synthesis would have been very different had microbial studies been more central during the early development of the evolutionary theory. Others however disagree that the structure and content of the evolutionary theory requires to be reshaped, even in the light of this new knowledge in microbiology (Wray 2014). Yet, debates around the gene content, nature, and phylogenetic position of Asgard archaea (Saw et al. 2015; Da Cunha et al. 2017; Zaremba-Niedzwiedzka et al. 2017) powerfully illustrates that an enhanced knowledge of the microbial dark matter has unquestionably the potential to transform central elements in the evolutionary theory. If Asgard archaea, currently only known via assemblies of environmental reads, prove to be sister-groups of eukaryotes, this should (at least) impact the very notion of a tree of life, bring further evidence regarding the number of Domains of life (since a convincing argument that the 2 domains tree is better supported than the 3 domains tree predates the discovery of Asgard; Williams et al. 2013), and, depending on the intimate structural biology and metabolisms of these Asgard, it will also help testing among competing hypotheses for the origin of eukaryotes (Koonin 2015; Sousa et al. 2016).

On a different level, newly discovered microbial genes have also impacted, and could further impact, critical societal needs. Discovering enzymes, such as lipases (Rogalska et al. 1997) or organo-phosphorus degrading enzymes (Singh 2009), with greater activity, specificity, or stability, or new antibiotics in the environment (Lok 2015), such as Teixobactin (Ling et al. 2015), is central to the development of the industrial enzymes market, which is expected to

represent up to 6.20 billion of dollars in 2020. Scientific research, as acknowledged by several Nobel Prizes, has also greatly benefited from the discovery of microbial enzymes, including restrictions enzymes, such as HindII (Smith and Wilcox 1970), or the DNA polymerases (Brock and Freeze 1969), which allowed the development of the Polymerase Chain Reaction (Saiki et al. 1988). More recently, the discovery of Crispr-cas9 systems (Jinek et al. 2012), now used for genome editing, also highlights the significant potential of microbial genes discovery to enhance the evolution of drugs, biotechnologies, and research tools.

Conclusion

The discovery of an increasing number of types of microbes has consistently shown that our planet hosts microbes with properties that were not simply identical to the ones formerly described. Studies of the microbial dark matter have brought forward the existence of novel entities (e.g., nanoorganisms, giant viruses, and virophages) and novel relationships within the microbial world (e.g., viral languages, high divergence, and scaffolding). This formerly dark microbial matter has not been unraveled randomly. To sum up its logic of discovery, it has required: to think outside the box (e.g., Woese's definition of a novel Domain), to take scientifically and financially risky decisions (e.g., sampling sites where life was unlikely), to develop novel methods pushing back the limits of detection (e.g., better microscopes, inclusive networks), to prepare one's mind to detect unknowns and unexpected forms (e.g., biomarkers), to identify and to seek to explain anomaly (e.g., the great plate count anomaly), to change perspectives (e.g., embracing the notion of nanoorganisms, or of multiple prokaryotic domains), to use analogies to uncover new microbial systems (e.g., for the study of extremophiles in space), to purposely depart from normal scientific practices and background knowledge (e.g., network studies of divergent gene forms, exploration of increasingly extreme environments), to be willing to create novel groups (e.g., Archea, CPR, Mimiviridae,...), and finally to convince (e.g., by banning competing notions, or by establishing new attractive fields, such as environmental genomics). Indeed, many of these discoveries presented in this work generated resistances. These resistances are perfectly explainable. Unraveling the unknown is especially difficult, because although we could empirically sketch a logic of scientific discovery, at the time each novel finding was made, their inventors could not yet rely on a standard method but essentially they had to convince the rest of the community that both their unusual approaches and finding were relevant. Convincing its own peers is finally essential, and possibly one of the largest and commonest challenge for microbial dark matter studies, and this seems especially difficult even for creative outsiders. Van Leeuwenhoek's pioneering example offers indeed a great reminder that extraordinary results can easily be forgotten.

Acknowledgments

R.L., G.B., J.S.P., and E.B. are funded by the European Research Council (FP7/2017-2013 Grant Agreement #615274). We thank Dr Karen Olsson-Francis, Dr Yan Boucher, and Dr Lucie Bittner for stimulating discussion.

Literature Cited

Anantharaman K, et al. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun. 7:13219.
Baker BJ, et al. 2010. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A. 107(19):8806–8811.
Bapteste E, et al. 2012. Evolutionary analyses of non-genealogical bonds produced by introgressive descent. Proc Natl Acad Sci U S A. 109(45):18266–18272.
Barer MR, Harwood CR. 1999. Bacterial viability and culturability. Adv Microb Physiol. 41:93–137.
Barns SM, Delwiche CF, Palmer JD, Pace NR. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci U S A. 93(17):9188–9193.
Beja O, et al. 2000. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289(5486):1902–1906.
Bordenstein SR, Theis KR. 2015. Host biology in light of the microbiome: ten principles of holobionts and hologenomes. PLoS Biol. 13(8):e1002226.
Boyer M, et al. 2011. Mimivirus shows dramatic genome reduction after intraamoebal culture. Proc Natl Acad Sci U S A. 108(25): 10296–10301.
Breitbart M, et al. 2002. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A. 99(22):14250–14255.
Brock TD, Freeze H. 1969. Thermus aquaticus gen. n. and sp. n., a nonsporulating extreme thermophile. J Bacteriol. 98(1):289–297.
Brown CT, et al. 2015. Unusual biology across a group comprising more than 15% of domain bacteria. Nature 523(7559):208–211.
Bruno A, et al. 2017. Exploring the under-investigated "microbial dark matter" of drinking water treatment plants. Sci Rep. 7:44350.
Burian RM. 2013. Exploratory experimentation. New York: Springer. p. 720–723.
Burstein D, et al. 2017. New CRISPR-Cas systems from uncultivated microbes. Nature 542(7640):237–241.
Caporael L, Griesemer J, Wimsatt W. 2013. Scaffolding in evolution, culture, and cognition. Massachusetts: MIT Press.
Checinska A, et al. 2015. Microbiomes of the dust particles collected from the International Space Station and Spacecraft Assembly Facilities. Microbiome 3:50.
Colson P, de Lamballerie X, Fournous G, Raoult D. 2012. Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. Intervirology 55(5):321–332.
Council NR editor. 1999. Report from the National Research Council. Washington (DC).
Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P. 2017. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLoS Genet. 13:e1006810.
Danczak RE, et al. 2017. Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. Microbiome 5(1):112.
de los Rios A, Wierzchos J, Sancho LG, Ascaso C. 2003. Acid microenvironments in microbial biofilms of antarctic endolithic microecosystems. Environ Microbiol. 5(4):231–237.
DeLong EF. 2007. Microbiology. Life on the thermodynamic edge. Science 317(5836):327–328.

DeLong EF, et al. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science 311(5760):496–503.
Diamond J. 1997. Guns, germs, and steel: the fates of human societies. New York city: W. W. Norton.
Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science 284(5423):2124–2129.
Doolittle WF. 2013. Carl R. Woese (1928–2012). Curr Biol. 23(5):R183–R185.
Ereshefsky M, Pedroso M. 2015. Rethinking evolutionary individuality. Proc Natl Acad Sci U S A. 112(33):10126–10132.
Erez Z, et al. 2017. Communication between viruses guides lysis-lysogeny decisions. Nature 541(7638):488–493.
Falkowski P. 2015. Leeuwenhoek's lucky break. Discover 1–5.
Fondi M, et al. 2016. "Every Gene Is Everywhere but the Environment Selects": global geolocalization of gene sharing in environmental samples through network analysis. Genome Biol Evol. 8(5): 1388–1400.
Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB. 2005. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proc Natl Acad Sci U S A. 102(41):14683–14688.
Gilbert SF, Bosch TC, Ledon-Rettig C. 2015. Eco-Evo-Devo: developmental symbiosis and developmental plasticity as evolutionary agents. Nat Rev Genet. 16(10):611–622.
Gill SR, et al. 2006. Metagenomic analysis of the human distal gut microbiome. Science 312(5778):1355–1359.
Gong J, Qing Y, Guo X, Warren A. 2014. "Candidatus Sonnebornia yantaiensis", a member of candidate division OD1, as intracellular bacteria of the ciliated protist *Paramecium bursaria* (Ciliophora, Oligohymenophorea). Syst Appl Microbiol. 37(1):35–41.
Guidi L, et al. 2016. Plankton networks driving carbon export in the oligotrophic ocean. Nature 532(7600):465–470.
Huerta-Cepas J, Forslund K, Pedro Coelho L, Szklarczyk D, Juhl Jensen L, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol. 34(8):2115–2122.
Hug LA, et al. 2016. A new view of the tree of life. Nat Microbiol. 1:16048.
Hugenholtz P, Goebel BM, Pace NR. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol. 180(18):4765–4774.
Huson DH, et al. 2016. MEGAN Community Edition – interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol. 12(6):e1004957.
Jinek M, et al. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337(6096):816–821.
Kantor RS, et al. 2013. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. MBio 4(5):e00708–e00713.
Karsenti E, et al. 2011. A holistic approach to marine eco-systems biology. PLoS Biol. 9(10):e1001177.
Koonin EV. 2015. Archaeal ancestors of eukaryotes: not so elusive any more. BMC Biol. 13:84.
Krishnamurthy SR, Wang D. 2017. Origins and challenges of viral dark matter. Virus Res. 239:136–142.
La Scola B, et al. 2003. A giant virus in amoebae. Science 299(5615):2033.
Laland K, Matthews B, Feldman MW. 2016. An introduction to niche construction theory. Evol Ecol. 30:191–202.
Lewis K. 2017. Antibiotics from the microbial dark matter. FASEB J. 31(Suppl 257):252.
Ling LL, et al. 2015. A new antibiotic kills pathogens without detectable resistance. Nature 517(7535):455–459.
Lobb B, Kurtz DA, Moreno-Hagelsieb G, Doxey AC. 2015. Remote homology and the functions of metagenomic dark matter. Front Genet. 6:234.
Lok C. 2015. Mining the microbial dark matter. Nature 522(7556):270–273.
Lopez P, Halary S, Bapteste E. 2015. Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life. Biol Direct. 10:64.
Luef B, et al. 2015. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun. 6:6372.
McKay CP, Davis WL. 1989. Planetary protection issues in advance of human exploration of Mars. Adv Space Res. 9(6):197–202.
Moran NA, Sloan DB. 2015. The hologenome concept: helpful or hollow? PLoS Biol. 13(12):e1002311.
Moreira D, Lopez GP. 2015. Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes? Philos Trans R Soc Lond B Biol Sci. 370(1678):20140327.
Morris JJ, Lenski RE, Zinser ER. 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio 3(2):e00036-12.
Nelson WC, Stegen JC. 2015. The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. Front Microbiol. 6:713.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405(6784):299–304.
Olsson-Francis K, Cockell CS. 2010. Experimental methods for studying microbial survival in extraterrestrial environments. J Microbiol Methods 80(1):1–13.
O'Malley MA. 2014. Philosophy of microbiology. Cambridge: Cambridge University Press.
Pace NR. 2006. Time for a change. Nature 441(7091):289.
Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8, 000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2:1533–1542.
Paul BG, et al. 2015. Targeted diversity generation by intraterrestrial archaea and archaeal viruses. Nat Commun. 6:6585.
Paul BG, et al. 2017. Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. Nat Microbiol. 2:17045.
Pikuta EV, Hoover RB, Tang J. 2007. Microbial extremophiles at the limits of life. Crit Rev Microbiol. 33(3):183–209.
Ram RJ, et al. 2005. Community proteomics of a natural microbial biofilm. Science 308(5730):1915–1920.
Rho M, Tang H, Ye Y. 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 38(20):e191.
Riesenfeld CS, Goodman RM, Handelsman J. 2004. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. Environ Microbiol. 6(9):981–989.
Rinke C, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499(7459):431–437.
Rogalska E, Douchet I, Verger R. 1997. Microbial lipases: structures, function and industrial applications. Biochem Soc Trans. 25(1):161–164.
Sachs JL, Hollowell AC. 2012. The origins of cooperative bacterial communities. MBio 3(3):e00099-12.
Saiki RK, et al. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239(4839):487–491.
Saw JH, et al. 2015. Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes. Philos Trans R Soc Lond B Biol Sci. 370(1678):20140328.
Schickore J. 2014. Scientific discovery. Stanford: The Stanford Encyclopedia of Philosophy.
Schmidt TM, DeLong EF, Pace NR. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J Bacteriol. 173(14):4371–4378.
Sedlar K, Kupkova K, Provaznik I. 2017. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Comput Struct Biotechnol J. 15:48–55.

Singh BK. 2009. Organophosphorus-degrading bacteria: ecology and industrial applications. Nat Rev Microbiol. 7(2):156–164.
Smith HO, Wilcox KW. 1970. A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties. J Mol Biol. 51(2):379–391.
Sousa FL, Neukirchen S, Allen JF, Lane N, Martin WF. 2016. Lokiarchaeon is hydrogen dependent. Nat Microbiol. 1(5):1–3.
Staley JT, Konopka A. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol. 39:321–346.
Stetter KO. 2013. A brief history of the discovery of hyperthermophilic life. Biochem Soc Trans. 41(1):416–420.
Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. 2005. Three *Prochlorococcus cyanophage* genomes: signature features and ecological interpretations. PLoS Biol. 3(5):e144.
Theis KR, Dheilly NM, Klassen JL, Brucker RM, Baines JF, Bosch TC, Cryan JF, Gilbert SF, Goodnight CJ, Lloyd EA, et al. 2016. Getting the hologenome concept right: an eco-evolutionary framework for hosts and their microbiomes. mSystems 1 (2): DOI: 10.1128/mSystems.00028-16.
Tringe SG, et al. 2005. Comparative metagenomics of microbial communities. Science 308(5721):554–557.
Tyson GW, et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428(6978):37–43.
Tyson GW, et al. 2005. Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community. Appl Environ Microbiol. 71(10):6319–6324.
Venter JC, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304(5667):66–74.
Vigliotti C, Lopez P, Bapteste E. 2017. Microbial diversity studies: the (paradoxical) challenge to have a broad view with metagenomics. In: Maurel PGMC, editor. Evolution and biodiversity. ISTE Editions. Amsterdam: Elsevier.
Waters CK. 2007. The nature and context of exploratory experimentation: an introduction to three case studies of exploratory research. Hist Philos Life Sci. 29(3):275–284.
Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504(7479):231–236.
Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 74(11):5088–5090.
Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 87(12):4576–4579.
Zaremba-Niedzwiedzka K, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541(7637):353–358.

Associate editor: Martin Embley

3.2 Métabolisme de très petits procaryotes dans les océans

Un an avant le début de ma thèse, un jeu de données métagénomiques océaniques original a été publié : TARA Océans (Sunagawa et al. 2015). Cette étude est exceptionnelle par son étendue: 153 sites ont été échantillonnés dans l'Océan Indien, l'Océan Pacifique, l'Océan Atlantique, la Mer Rouge et la Mer Méditerranée. Pour chacun des sites, des échantillons ont été prélevés à différentes profondeurs et fractionnés par taille. En particulier, 65 échantillons ont été fractionnés pour rechercher des virus, c'est-à-dire filtrés à moins à de 0.22 μm. Cette taille de filtre a longtemps été considérée comme permettant une stérilisation microbienne du milieu. Mais, en métagénomique, elle est également utilisée depuis quelques années pour rechercher des micro-organismes extrêmement petits comme les CPR ou les DPANN. Le volume de données produit par cette étude est grand: 7.2 To de fragments de lectures, 137 523 700 de séquences codantes prédites. Un regroupement par similarité des séquences a été réalisé (Sunagawa et al. 2015) pour obtenir un jeu de données non redondant de 40 000 000 de séquences représentatives de cette diversité.

Nous avons commencé par nous demander si dans cette fraction de taille dite "virale", il n'y avait pas également des organismes marins ultra-petits. Ensuite, nous nous sommes posé la question des rôles écologiques potentiels de tels micro-organismes dans l'océan. A l'origine, nous avions l'espoir de détrôner *Prochlorococcus* du titre du plus petit organisme photosynthétique. Nous n'avons pas trouvé de signal concluant par rapport à la présence de photosynthèse dans la fraction de taille ultra-petite de TARA Océans mais nous y avons détecté la présence de gènes impliqués dans la fixation autotrophique du carbone. Plus précisément, dans l'article qui suit, nous avons recherché la présence des 6 voies métaboliques de la fixation du carbone (Berg 2011) dans la fraction de taille "virale". Ce travail nous a permis de mettre en avant la présence de l'information génétique nécessaire

pour la réalisation de plusieurs des voies métaboliques de fixation du carbone dans la fraction de taille ultra-petite. A ma grande fierté, ce travail a été mis en avant par le journal qui l'a publié[1]. Avec une stagiaire de M1, Louise Cavaud, nous avons décidé d'étendre ce travail aux autres voies métaboliques autotrophiques. Ce travail est en cours de rédaction mais je peux confirmer la présence de l'information génétique nécessaire à la fixation de l'azote et du soufre dans la fraction de taille ultra petite de TARA Océans.

3.2.1 Article 2, "Carbon Fixation by Marine Ultrasmall Prokaryotes", Genome Biology and Evolution (Lannes et al. 2019)

[1] https://academic.oup.com/gbe/article/11/5/1431/5489024

Carbon Fixation by Marine Ultrasmall Prokaryotes

Romain Lannes[1], Karen Olsson-Francis[2], Philippe Lopez[3], and Eric Bapteste[3,*]

[1]Sorbonne Université, Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, EPHE, Université des Antilles, Paris, France

[2]School of Environment, Earth and Ecosystems, The Open University, Milton Keynes, United Kingdom

[3]Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, EPHE, Université des Antilles, Paris, France

*Corresponding author: E-mail: eric.bapteste@upmc.fr.

Accepted: March 4, 2019

Abstract

Autotrophic carbon fixation is a crucial process for sustaining life on Earth. To date, six pathways, the Calvin–Benson–Bassham cycle, the reductive tricarboxylic acid cycle, the 3-hydroxypropionate bi-cycle, the Wood–Ljungdahl pathway, the dicarboxylate/4-hydroxybutyrate cycle, and the 4-hydroxybutyrate cycle, have been described. Nano-organisms such as members of the Candidate Phyla Radiation (CPR) bacterial superphylum and the Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohalorchaeota (DPANN) archaeal superphylum could deeply impact carbon cycling and carbon fixation in ways that are still to be determined. CPR and DPANN are ubiquitous in the environment but understudied; their gene contents are not exhaustively described; and their metabolisms are not yet fully understood. Here, the completeness of each of the above pathways was quantified and tested for the presence of all key enzymes in nano-organisms from across the World Ocean. The novel marine ultrasmall prokaryotes were demonstrated to collectively harbor the genes required for carbon fixation, in particular the "energetically efficient" dicarboxylate/4-hydroxybutyrate pathway and the 4-hydroxybutyrate pathway. This contrasted with the known carbon metabolic pathways associated with CPR members in aquifers, where they are described as degraders (Castelle CJ, et al. 2015. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr Biol. 25(6):690–701; Castelle CJ, et al. 2018. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol. 16(10):629–645; Anantharaman K, et al. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun. 7:13219.). Our findings suggest that nano-organisms have a broader contribution to carbon fixation and cycling than currently assumed. Furthermore, CPR and DPANN superphyla are possibly not the only nanosized prokaryotes; therefore, the discovery of new autotrophic marine nano-organisms by future single cell genomics is anticipated.

Key words: metagenomics, marine ultrasmall organisms, metabolism, carbon fixation.

Introduction

Autotrophic carbon fixation is a crucial process for sustaining life on Earth as it fixes inorganic carbon, including the sequestration of atmospheric carbon dioxide (De La Rocha and Passow 2014), into organic carbon (Hügler and Sievert 2011). It is responsible for the annually net fixation of 7×10^{16} g carbon, which corresponds to the conservation of 2.8×10^{18} kJ of energy (Berg 2011). To date, there are six known pathways for autotrophic carbon fixation. This includes the Calvin–Benson–Bassham (CBB) cycle, which is quantitatively the most important mechanism of autotrophic CO_2 fixation in nature and is primarily achieved by photosynthetic organisms (Hügler and Sievert 2011). For many years, it was thought to be the only pathway for autotrophic CO_2 fixation, but more recently five additional pathways have been described. These include the reductive tricarboxylic acid cycle (rTCA), the 3-hydroxypropionate bi-cycle (HBC), the reductive acetyl-CoA pathway, which is also known as the Wood–Ljungdahl pathway (WL), the dicarboxylate/4-hydroxybutyrate cycle (DH), and the 4-hydroxybutyrate cycle (Hügler and Sievert 2011). Concurrently, an increasing number of models have been developed that highlight the role of micro-organisms in carbon fixation (Wieder et al. 2015; Dyksma et al. 2016; Guidi et al. 2016;

Guidi et al. 2016; La Cono et al. 2018). For example, *Prochlorococcus*, a small and extremely abundant photosynthetic cyanobacterium, was proposed to be a key contributor to autotrophic carbon fixation in the ocean (Partensky et al. 1999). Similarly, SAR11, one of the tiniest known photoheterotrophic organisms (cell volume of roughly 0.01 μm^3), seems to play an important ecological role as the most abundant marine planktonic organism (Rappé et al. 2002; Giovannoni 2017).

Importantly, studies of environmental microbes show that microbial diversity is still largely underexplored (Brown et al. 2015; Castelle et al. 2015; Parks et al. 2017). Recently, the number of described prokaryotic lineages doubled with the discovery of novel superphyla including some ultrasmall members: the Candidate Phyla Radiation (CPR; bacteria) and the Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohalorchaeota (DPANN; archaea) (Rinke et al. 2013; Brown et al. 2015; Luef et al. 2015; Hug et al. 2016). The physiology of these ultrasmall prokaryotes (hereafter called nano-organisms) is unusual, not only because of their reduced cell volume (these cells can pass through 0.22-μm filters, a size usually expected to exclude most micro-organisms) (Andrew et al. 1999; Luef et al. 2015) but also because of their reduced genome size and biosynthetic capability. Most of the CPR lack parts of central metabolic pathways, including nucleotide and amino acid biosyntheses (Brown et al. 2015; Castelle et al. 2015). Nano-organisms also have an incomplete tricarboxylic acid cycle and lack NADH dehydrogenase and electron transport chains (Brown et al. 2016).

Consequently, the potential role of these nano-organisms in the geochemical cycle of carbon and hydrogen (Anantharaman et al. 2016) has begun to be investigated. For example, Anantharaman et al. detected the presence of key enzymes involved in the carbon, nitrogen, sulfur, and hydrogen cycles in local metagenomic data from aquifers located in Rifle (USA, Colorado), which were assigned to the CPR superphylum. Likewise, in the same aquifers, Rubisco type II/III genes were found. These genes seemed to be active in the CPR and DPANN superphyla (Wrighton et al. 2016) suggesting the presence of the nucleotide salvaging pathway and potentially the CBB pathway. Yet, the phylogenetic and functional diversities of nano-organisms are possibly not fully appreciated and in particular their role in carbon fixation remains to be characterized. In this broad-scale study, the possible role of some known and novel candidate nano-organisms in ocean carbon fixation was investigated, specifically from sites that were sampled as part of the TARA OCEAN expedition (Sunagawa et al. 2015). First, an in silico approach was used to retrieve putative sequences of nano-organisms from the TARA OCEAN metagenome data sets and analyze their phylogenetic diversity. Second, prokaryotic carbon fixation pathways that were described in KEGG were used to identify homologs in marine nano-organisms. Finally, the completeness and geographical distributions of homologs from these autotrophic carbon fixation pathways in 65 of the TARA sampling sites were analyzed.

Materials and Methods

Selection of Sequences

The sequences used in this study were obtained from the TARA OCEAN metagenomic database (ftp://ftp.sra.ebi.ac.uk/vol1/ERA412/ERA412970/tab/OM-RGC_seq.release.tsv.gz, last accessed April 2, 2019), which is publicly available. The database consists of sequencing data from various sampling sites, depths, and fraction sizes, including an ultrasmall size fraction (<0.22 μm). About a hundred million sequences of predicted proteins have already been clustered by similarity using CD-HIT (Li and Godzik 2006; Fu et al. 2012). These clusters were sorted for two reasons: first, to decontaminate the ultrasmall size-fraction data set from TARA OCEANS. Second, to characterize the microbial dark matter in the ultrasmall size fraction (by identifying genes from candidate ultrasmall prokaryotes, increasingly different from known reference taxa). To do so, each sequence within each cluster of similarity was assigned to a size fraction of origin. Clusters without sequences from the ultrasmall size fraction were discarded from the rest of our analyses. The 6,677,440 remaining clusters included at least a representative sequence from the ultrasmall size fraction (<0.22 μm). As such clusters were not necessarily strictly associated with the ultrasmall size fraction, they were therefore called the "Potentially Ultrasmall" (PU) data set. Problematically, sequences from the PU data set were not necessarily sequenced from bona fide ultrasmall prokaryotes and may have resulted from contamination of the ultrasmall size fraction, for example, from the presence of free DNA from regular-sized prokaryotes or viruses. Therefore, a further level of stringency was used, to define UO data set (for "Ultrasmall Only"), nested in the PU data set. The UO data set included all sequences from the PU data set that were exclusively found in samples from the ultrasmall size fraction. Among the 4,586,489 clusters from UO, 1,258,638 clusters contained sequences found at more than one site. We called this latter category of widespread clusters WUO (Widespread Ultrasmall Only) (fig. 1).

The clusters from PU, UO, and WUO were further curated by detecting viral proteins through similarity searches against the NCBI nr database (March 2017) using DIAMOND (Buchfink et al. 2015; Wheeler 2007). This removed 286,388 and 130,330 potential viral proteins from the UO and WUO clusters, respectively. An additional search was performed against the sequences from the TARA ocean metavirome (project PRJEB6606, European Nucleotide Archive [https://www.ebi.ac.uk/ena; last accessed April 2, 2019]) to identify potential environmental contaminants. This resulted in the removal of 142 sequences. Notably, autotrophic carbon fixation genes returned no matches with ≥80% sequence

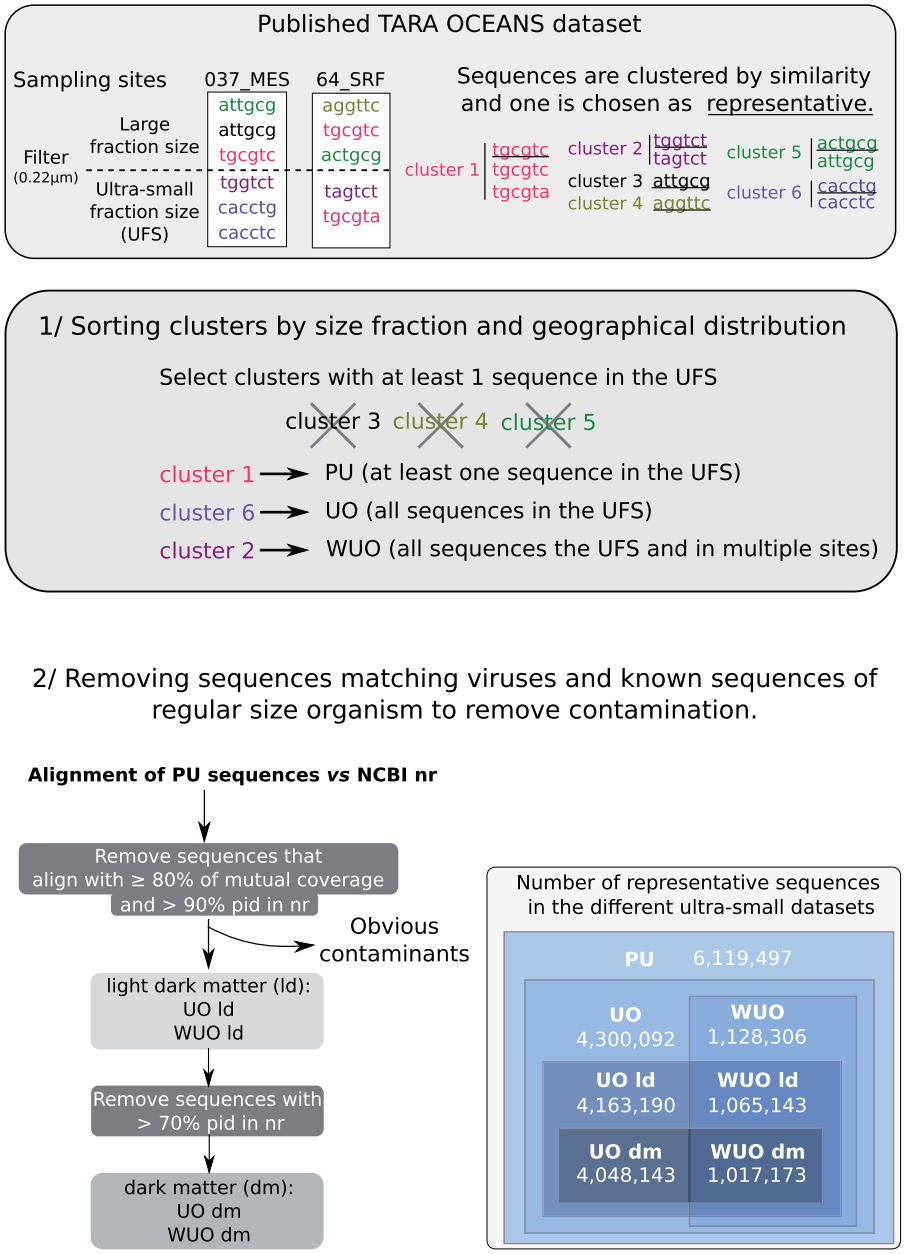


Fig. 1.—Ultrasmall data set filtration. The raw TARA Oceans data set includes more than 140 million sequences that have been clustered to 44 million sequences (Sunagawa et al. 2015). Each cluster has a representative sequence that may represent many sequences. If a cluster contains at least one sequence from the ultrasmall size fraction then this cluster was selected as part of the PU data set. If a PU cluster only contained sequences from the ultrasmall size fraction, this cluster was also assigned to the UO data set. If a UO cluster included sequences from at least two different sampling sites, this cluster was also assigned to a WUO data set. Then, PU sequences were aligned against NCBI nr in order to remove potential virus sequences. Furthermore, these alignments were used to remove potential contamination by known large micro-organisms from the UO and WUO data sets. The light dark matter and dark matter UO and WUO data sets, respectively, were defined by removing sequences with at least 80% of mutual cover and 90% %ID (light dark matter) or at least 80% of mutual cover and 70% %ID (dark matter) to sequences in the NCBI nr database. Numbers of sequences in each data set are shown in the box.

identity and a mutual alignment coverage of ≥80%, indicating that there was no positive evidence that these carbon fixation genes were carried by marine viruses.

Most annotated prokaryotes known to date, with the exception of nanosized members of the CPR and DPANN superphyla, were not expected to pass through a 0.22-µm filter. Therefore, the finding of proteins in the UO/WUO data sets that were highly similar to known regular-sized prokaryotes was likely due to contamination. For each sequence in our data set, the percentage of identity (%ID) to its best hit in

nr was considered using DIAMOND. This was carried out in order to quantify how similar the environmental sequence was to a reference sequence. The step of taxonomic annotation allowed us to classify the environmental sequences from the UO and WUO clusters into two levels of increasing divergence from a reference, looking for potential organismal dark matter in the ultrasmall size fraction. Thus, the UO and WUO data sets were split into two nested categories: "dark matter" and "light dark matter." Sequences whose best hit against nr showed a mutual coverage >80% and %ID <90% were assigned to "light dark matter" (4,300,092 sequences for UO and 1,065,606 sequences for WUO); whereas sequences that showed %ID <70% were assigned to "dark matter" (4,048,143 sequences for UO and 1,017,137 sequences for WUO). Furthermore, sequences taxonomically assigned to DPANN, unclassified bacteria, unclassified archaea, CPR, candidate or "root: unassigned" were assigned to both "light dark matter" and "dark matter," because these taxa likely correspond to bona fide ultrasmall prokaryotes. The "dark matter" clusters provided an additional perspective, but "light dark matter" clusters were a priori not more (or less) contaminated than "dark matter" clusters.

Mining the KEGG Database

Using the KEGG database (Ogata et al. 1999), a list of KEGG Orthology terms was defined, which corresponded to metabolic pathways associated with autotrophic carbon fixation (M00173, M00374, M00375, M00376, M00377, M00579, and M00620), as well as ribosomal complexes in eukaryotes (M00177), archaea (M00179), and bacteria (M00178). All corresponding proteins (179,853 proteins for carbon fixation, and 211,781 proteins for ribosomal complexes) were retrieved using the Uniprot mapping tool (http://www.uniprot.org/mapping/; last accessed April 2, 2019) or the KEGG API service (March 2017).

Homology Detection and Taxonomic Annotation

The homologs of KEGG proteins that were present in the PU, UO, and WUO data sets were identified using NCBI BLAST (version 2.6.0) (Camacho et al. 2009). The following criteria were used to assess homology: %ID > 25%, *E*-value <1e-5, and mutual alignment coverage >70% (Alvarez-Ponce et al. 2013; Haggerty et al. 2014). Using these thresholds, 20,368 sequences from the TARA ultrasmall data set were detected as homologs of proteins from autotrophic carbon fixation pathways. Additionally, using the same methodology 37,054 sequences from the PU data set were detected as homologs to proteins from ribosomal complexes.

Pathways Completeness

A KEGG pathway describes a set of reactions (modules), which require a set of enzymes (supplementary table 2, Supplementary Material online). For each sampling site, if homologs of the required enzymes existed in a data set, the module was considered present (namely, in UO, UO "light dark matter," UO "dark matter," WUO, WUO "light dark matter," and WUO "dark matter") suggesting that the ultrasmall prokaryotes could complete that step of the pathway. Optional enzymes were not considered but were reported if found. Finally, the percentage of modules present at a given site was taken as a proxy of the completeness of the pathway. Key enzymes (according to Berg [2011]) and key modules (i.e., modules that contain at least one key enzyme) were highlighted.

Correlation between Pathway Completeness and Sampling Effort

Correlations between a pathway's completeness and various assessments of the sequencing effort were computed with a Spearman correlation test, using the python3 function spearmanr from the SciPy library. Assessments of the sequencing effort for a given site were provided as the number of reads, high quality reads, predicted genes, and average read coverage per protein.

Taxonomic Enrichment or Depletion of Filtered Data Sets

For each data set, the number of proteins assigned to a taxonomic group was compared with the PU data set. A pairwise Fisher exact test (using fisher_exact function from Scipy.stats Python3 library) was used to compare each taxonomic group with the remaining groups to identify significant enrichment or depletion compared with the PU data set. Because nine taxonomic groups were tested with six data sets, the corresponding Bonferroni correction was applied to the *P* values for a 5% type I error.

Phylogenetic Analyses

Reference sequences from KEGG and their environmental homologs were aligned with DIAMOND (Buchfink et al. 2015). A sequence similarity network was built from these alignments in order to define gene families (Corel et al. 2016), with >=80% mutual coverage and >=30% %ID as thresholds for edges. Gene families were defined as connected components in this sequence similarity network. All key enzymes of the autotrophic carbon fixation pathways, as well as ribosomal proteins from connected components with more than 100 sequences, were selected for diversity and phylogenetic analyses. Homologs from all published CPR and DPANN genomes were added (2,481,154 sequences as of December 2018) using DIAMOND (>80% coverage, >30% %ID). The resulting gene families were aligned using MAFFT (Katoh et al. 2002) and the alignments were trimmed using trimAl (Capella-Gutiérrez et al. 2009) with default parameters. Maximum likelihood trees were reconstructed

Taxonomical annotation within the different datasets

data	known Bacteria	Including Proteobacteria	Including CPR	unassigned Bacteria	known Archaea	Including DPANN	unassigned Archaea	unclassified	Eukaryota	Total number of proteins
PU	22.03%	69.68%	2.07%	0.90%	1.58%	19.31%	0.39%	74.97%	0.13%	6,119,497
UO	↓ 15.08%	↓ 68.10%	↑ 2.34%	↓ 0.46%	↓ 1.06%	↑ 24.39%	↓ 0.33%	↑ 82.96%	↓ 0.11%	4,300,092
UO ld	↓ 12.34%	↓ 65.63%	↑ 2.96%	↓ 0.48%	↓ 1.08%	↑ 24.71%	↓ 0.34%	↑ 85.66%	↓ 0.11%	4,163,190
UO dm	↓ 10.00%	↓ 63.17%	↑ 3.76%	↓ 0.49%	↓ 1.00%	↑ 27.41%	↓ 0.35%	↑ 88.06%	↓ 0.11%	4,048,143
WUO	↑ 24.23%	↑ 71.27%	↓ 1.35%	↓ 0.66%	↑ 1.71%	↑ 20.93%	↓ 0.30%	↓ 72.99%	↓ 0.11%	1,128,306
WUO ld	↓ 19.82%	↓ 68.48%	↓ 1.75%	↓ 0.69%	↑ 1.80%	↑ 21.11%	↓ 0.32%	↑ 77.26%	↓ 0.11%	1,065,606
WUO dm	↓ 16.22%	↓ 66.24%	↑ 2.23%	↓ 0.73%	↑ 1.68%	↑ 23.69%	↓ 0.34%	↑ 80.92%	↓ 0.11%	1,017,137

Fig. 2.—Effect of filtration on data sets phylogenetic composition. Each row represents a data set after filtration. Known Bacteria, Archaea, and Eukaryota represent sequences that show a best hit in a BLAST search against the NCBI nr database that is referenced as Bacteria, Archaea, and Eukaryote, respectively. Unclassified sequences were environmental sequences that had no hits in the NCBI nr database or were annotated as "root; unclassified sequences". Unassigned Bacteria and Archaea sequences were closely related to sequences in the NCBI nr database that were only annotated at the domain level. "Including Proteobacteria" and "Including CPR" represented the percentage of Known Bacteria for which best hits in NCBI nr were annotated as Proteobacteria or as CPR, respectively. Including DPANN represented the percentage of Known Archaea for which best hits in NCBI nr database are annotated as a DPANN. For each data set, the effect of filtration on phyla proportion was investigated. Green and red arrows indicated phyla proportions that were significantly enriched or depleted, respectively, for a given phylum in a given data set compared with the proportion of that phylum in PU. Abbreviations: PU, Potentially Ultrasmall; UO, Ultrasmall Only; WUO, Widespread Ultrasmall Only; ld, light dark matter; and dm: dark matter.

using IQ-Tree (Nguyen et al. 2015) under the LG + G model, and 1,000 ultrafast bootstraps replicates were performed (Minh et al. 2013).

Results

Twenty thousand three hundred sixty-eight environmental homologs sequences were identified for six autotrophic carbon fixation pathways, at a threshold of sequence identity >25%, of mutual coverage >70%, and *E*-value <1e-5 (supplementary table 2, Supplementary Material online, and Materials and Methods). Some active micro-organisms can pass through a 0.22-µm filter (Hasegawa et al. 2003), particularly as "starvation forms" (Haller and Ro 1999). A screening step was added to identify potential contamination, that is, to remove sequences from organisms larger than nano-organisms and viruses. As a result, nested data sets of environmental sequences were produced, which were exclusively found in the ultrasmall fraction and defined with increasingly stringent conditions of geographic and taxonomic distributions (fig. 1). With respect to the original data sets associated with the ultrasmall size fraction of the TARA OCEANS project, the "cleaned" data sets developed in this study were significantly enriched in taxonomically unclassified sequences, and in CPR and DPANN sequences. They were also depleted in unassigned archaea and bacteria, and in known regular-sized bacterial phyla and in viruses (fig. 2, *P* values supplementary table 1, Supplementary Material online). The proportion of known archaeal lineages was unaffected by this screening process.

Our filtered data sets were phylogenetically rich in diversity of presumed ultrasmall prokaryotes. This was assessed by careful analysis of the placement of the ultrasmall prokaryotes in the maximum likelihood phylogenies of ribosomal proteins (fig. 3 and supplementary fig. 1, Supplementary Material online). In these trees, oceanic ultrasmall prokaryotes did not appear to be monophyletic. Rather, they were related to various known prokaryotic lineages, such as CPR and DPANN, but also less expectedly to Bacteroidetes and Proteobacteria. This suggested that either some contamination is retained in the filtered data sets or there are genuine ultrasmall members of these clades that are yet to be described. Moreover, some of the environmental sequences that qualified as "light dark matter" and as "dark matter" clustered in these phylogenies, hinting at undescribed ultrasmall lineages within known major prokaryotic groups. Phylogenies of key enzymes involved in carbon fixation showed similar results: sequences from the ultrasmall size fraction branched within different major prokaryotic groups, pointing to new groups within CPR, DPANN and other prokaryotic clades (fig. 4 and supplementary fig. 2, Supplementary Material online). This latter result suggests that unknown ultrasmall prokaryotes could take part in aspects of carbon fixation. For example, a widespread environmental lineage related to *Chloroflexi* and *Acidobacteria* was found to host homologs to both the malonyl-CoA reductase/3-hydroxypropionate dehydrogenase (NADP+) enzyme, fumarate hydratase, class II and the acrylyl-CoA reductase (NADPH)/3-hydroxypropionyl-CoA dehydratase/synthetase, suggesting a potential contribution to the HBC pathway (fig. 4 and supplementary fig. 2, Supplementary Material online).

To obtain a more comprehensive view of their ecological role, the geographic distribution of the environmental sequences from the ultrasmall prokaryotes and their potential to include complete autotrophic carbon fixation pathways was investigated. A heatmap (fig. 5) was produced, which represented the completeness of each of the six carbon

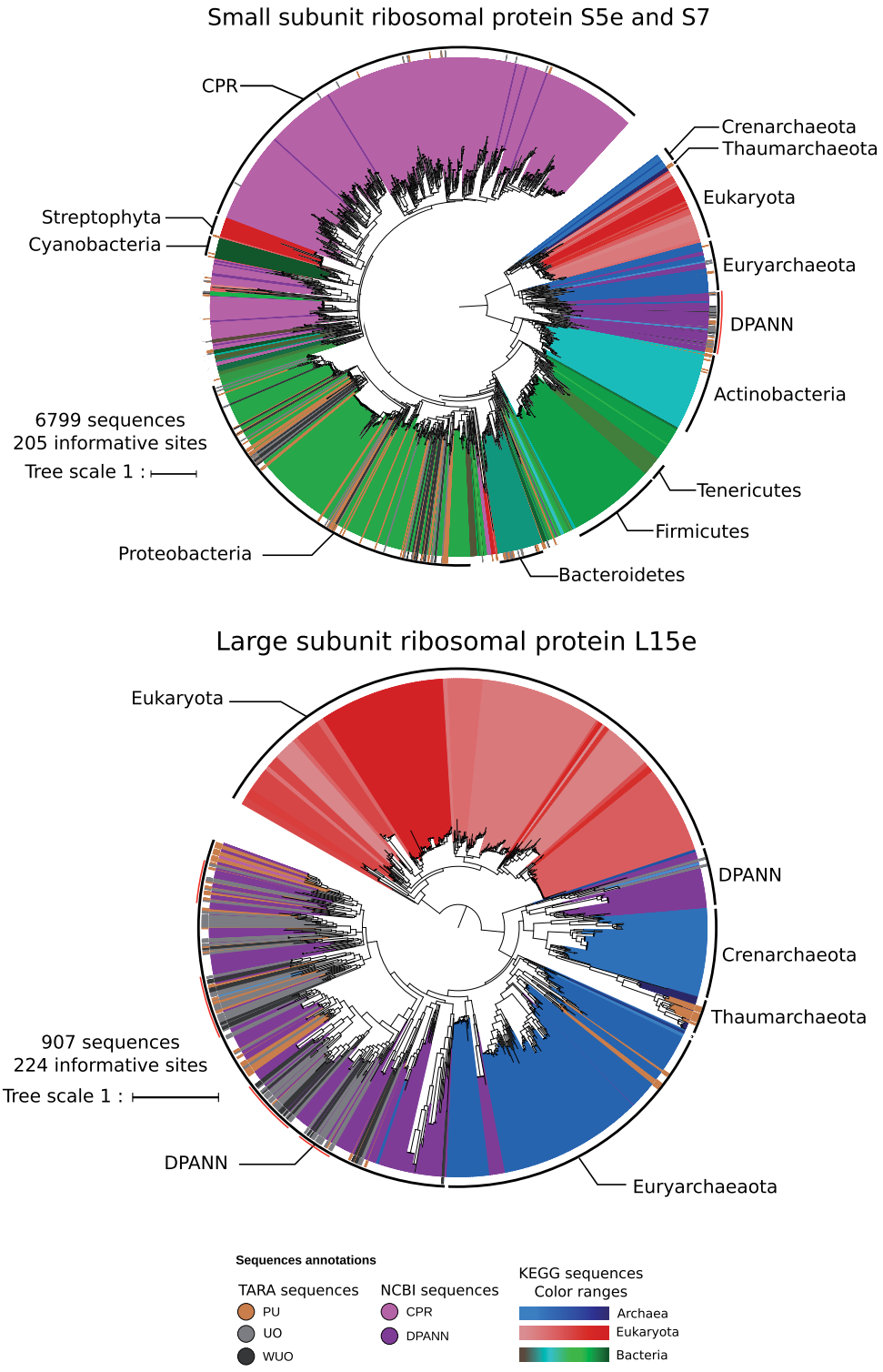


Fig. 3.—Phylogenetic trees of ribosomal proteins S5e and S7 (top) and L15e (bottom). Sequences were aligned using MAFFT in auto mode and trimmed with TrimAl. Trees were constructed using IQ-TREE with LG + G4 models and ultrafast bootstrap approximation (Minh et al. 2013). The trees were rooted between Archaea and Bacteria and branches with bootstrap values <50% were collapsed. The number of informative sites and branch length scale bars (substitutions per site) are shown. Environmental sequences are highlighted by a colored bar in the outer ring. The sequences were from three sources: 1) environmental sequences from the TARA Oceans data sets; 2) CPR and DPANN sequences from assemblies available in NCBI; and 3) other reference sequence from KEGG. Sequences are colored by taxonomic annotation. Archaeal sequences found in 037, 038, and 039 MES sampling sites are highlighted by a red arc. Abbreviations: PU, Potentially Ultrasmall; UO, Ultrasmall Only; and WUO: Widespread Ultrasmall.

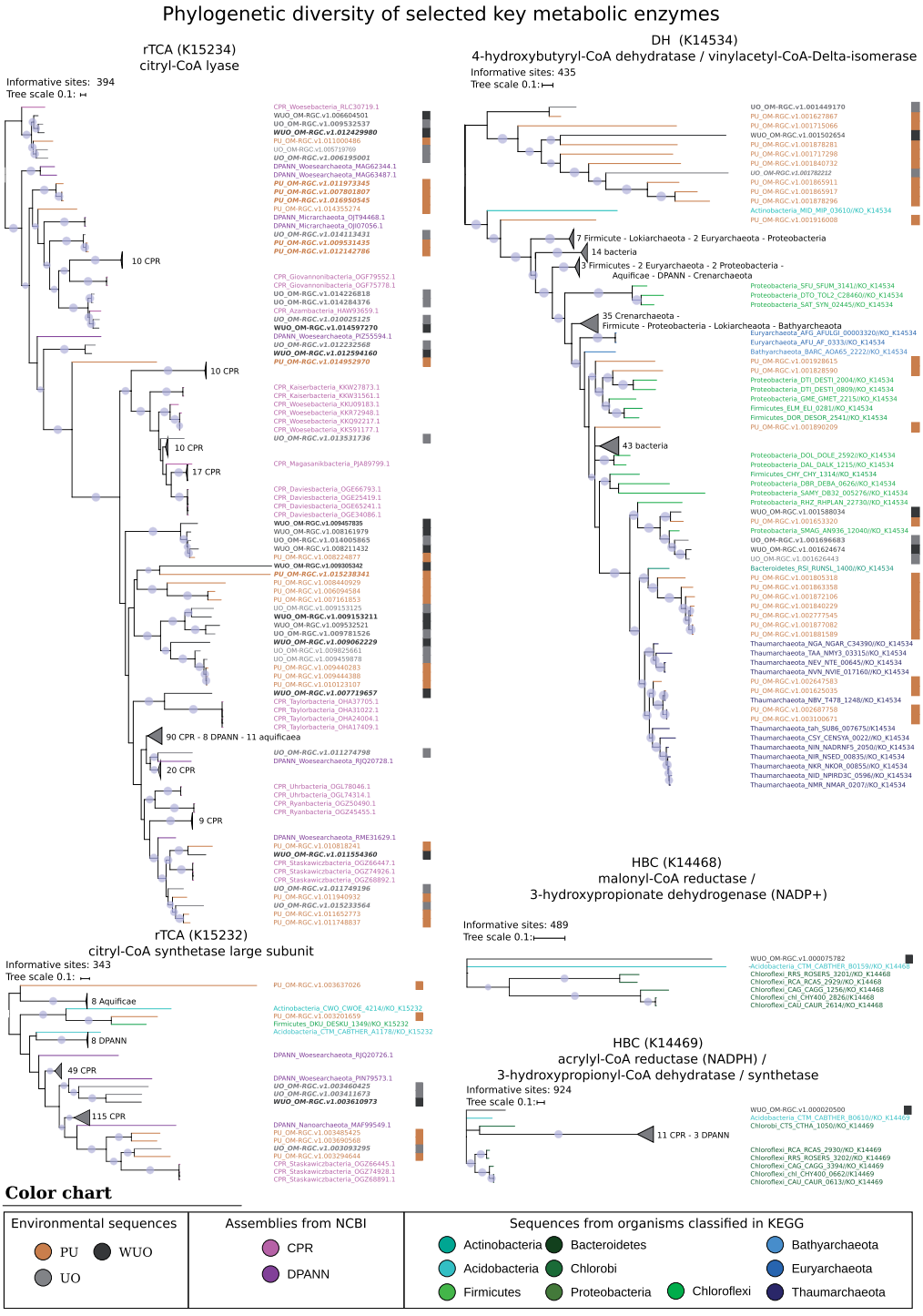


FIG. 4.—Phylogenetic trees of selected key enzymes in three carbon fixation pathways. Trees were reconstructed using maximum likelihood on trimmed alignments. The number of informative sites and branch length scale bars (substitutions per site) are shown for each tree. Sequences used are from KEGG, NCBI (CPR and DPANN) and from the TARA Oceans data set. Trees were midpoint rooted. Bootstraps were computed using 1,000 iterations of ultrafast bootstrap approximation. Branches with bootstrap <50% were collapsed, a light blue dot highlights branches with bootstrap values >80%. Environmental sequences are highlighted by a colored bar on the right of each tree. Sequence names: environmental sequences were formatted as (PU, UO, WUO)TARA identifier. KEGG sequences were formatted as phylum_KEGG_identifier. NCBI sequences were formatted as (CPR/DPANN)_phylum_proteinID. For readability, some clades were collapsed and are represented by a dark triangle with the description of the clade's sequences. Abbreviations: rTCA, reductive tricarboxylic acid cycle; DHC: dicarboxylate–hydroxybutyrate cycle; HBC, 3-hydroxypropionate bi-cycle; PU, Potentially Ultrasmall; UO, Ultrasmall Only; and WUO, Widespread Ultrasmall Only.

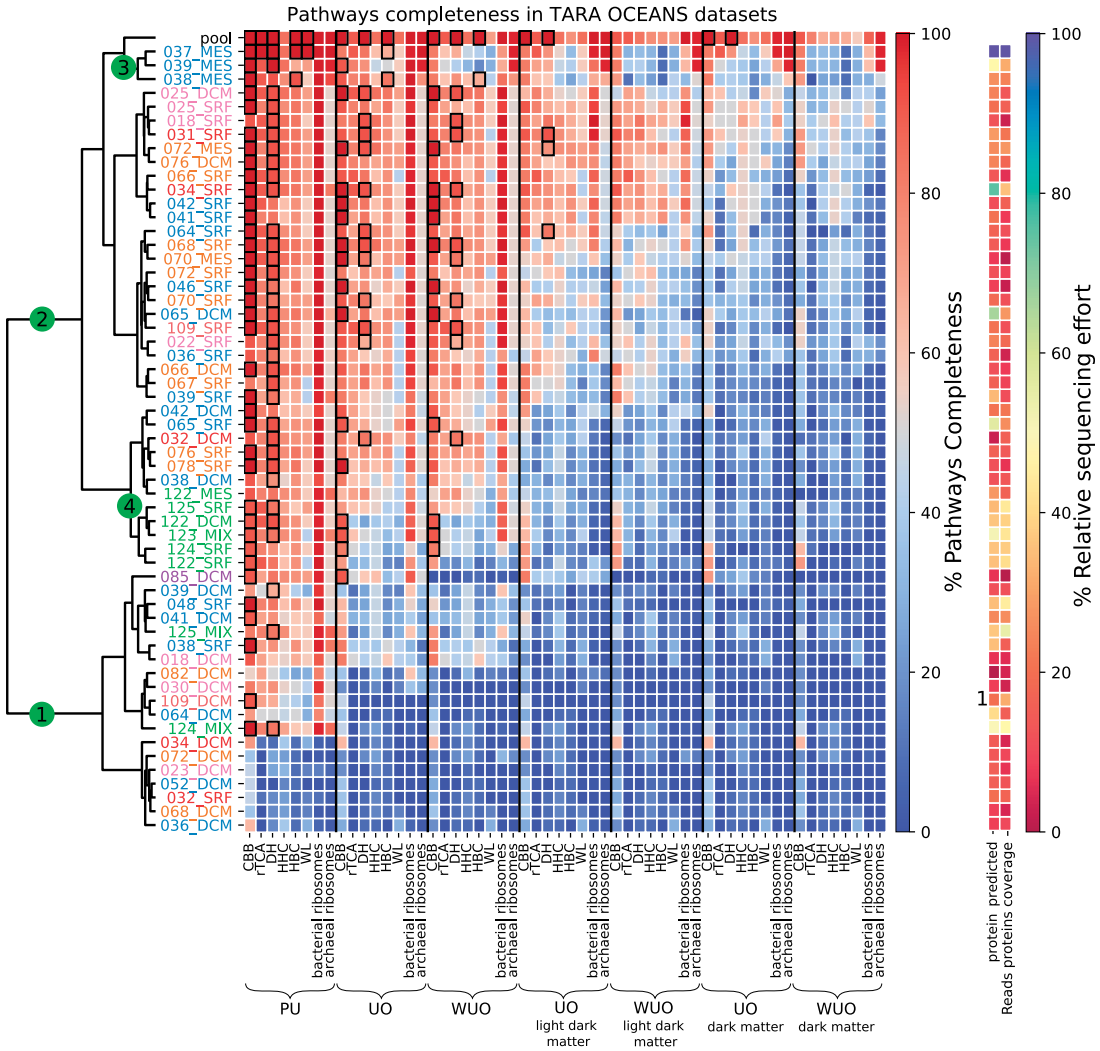


FIG. 5.—Heatmap of completeness of six carbon fixation pathways and archaeal and bacterial ribosomal complexes. The heatmap color scale shows the completeness of pathways or ribosomal complexes, with rows as sampling sites and columns as proteins sets. Black squares highlight sites with pathway completeness >60% and comprising all key enzymes. Rows were clustered using scipy.cluster.hierarchy.linkage ("ward" method). The corresponding dendrogram is shown to the left of the heatmap. Row names indicate sampling sites in the format TARA sampling site id (three digits) _ depth. Depths are: SRF (Surface), DCM (Deep Chlorophyll Maximum), MES (Mesopelagic), and MIX (mixed). Row label colors represent oceanic regions: brown for North Pacific Ocean, green for South Pacific Ocean, purple for Southern Ocean, orange for South Atlantic Ocean, dark blue for Indian Ocean, red for Red Sea, and pink for Mediterranean Sea. The "pool" row represents results for all sampling sites pooled together. Ribosomal complexes from bacteria and archaea contain 55/67 proteins, respectively, and share 31 proteins. Sequencing effort is computed as the proportion of the number of proteins found at a given site and the average number of reads per protein, relatively to the values found at 037_MES sampling site, which showed the maximum values for both indicators.

fixation pathways analyzed in this study, as well as the completeness of bacterial and archaeal ribosomal complexes. Bacterial and archaeal ribosomal complexes are composed of a comparable number of proteins to the carbon fixation pathways and were therefore used as positive controls to validate this method for detecting full-sized pathways. It was expected that ribosomal complexes would appear to be complete in a site if a sufficient sequencing effort had occurred. For each site, the number of reads and proteins predicted to be part of a ribosomal complex or an autotrophic carbon fixation pathway, and the average read coverage of these proteins, was reported. However, the information relative to the sampling effort associated with each site could warrant naïve assumptions regarding the ultrasmall prokaryotes and carbon fixation pathways, especially when undersampling could be a plausible explanation (supplementary table 3, Supplementary Material online). Conversely, the detection of homologous enzymes from carbon fixation pathways in generally undersampled sites suggested that ultrasmall prokaryotes were involved in these ecologically important pathways. The columns of the heatmap (fig. 5) represent the data sets sorted by increasing stringency; the rows

represent samplings sites and were hierarchically clustered according to the completeness of their pathways, in order to search for possible geographical trends.

Interestingly, the heatmap revealed two major clusters of sampling sites. The lower cluster (dot 1 in fig. 5) corresponded to sites in which no, or very few, homologs of the carbon fixation proteins were detected (with the exception of the CBB pathway, which appeared to be partly present at all levels of stringency). Sites sampled at the deep chlorophyll maximum depth were overrepresented in this part of the heatmap; however, the incompleteness was not due to the amount of sequencing data compared with other sites, both in terms of proportion of proteins and average read coverage per protein. In addition, there were no homologs of ribosomal proteins at these sites, which suggests that the samples associated with the lower part of the heatmap were largely viral rather than microbial (as expected for an ultrasmall size fraction). By contrast, the higher cluster (dot 2 in fig. 5) of the heatmap was enriched in sites from the surface depths. The finer-grained clustering of sampling sites within this part of the heatmap points to some local geographical patterns. First, samples 037, 038, and 039 from mesopelagic depths clustered together (dot 3 in fig. 5) corresponding to sites in the Indian Ocean, which had similar distributions of carbon fixation pathways and ribosomal complexes. These three sites presented a rich proportion of archaeal complexes, even in data sets with very stringent thresholds. This hinted at the presence of still undescribed ultrasmall archaea in the Indian Ocean, which were indeed detected in 29 individual phylogenies of ribosomal proteins (fig. 2 and supplementary information, Supplementary Material online). These potentially new ultrasmall archaea were generally polyphyletic, and some were often related to an archaeon GW2011 AR20, assigned to the DPANN superphyla (Castelle et al. 2015). A similar cluster was also detected from sites from the South Pacific Ocean (dot 4 in fig. 5).

In terms of pathways completeness, the CBB and DH pathways were the most commonly complete carbon fixation pathways, even with the requirement that homologous enzymes should be found at multiple sites. This suggests that ultrasmall prokaryotes are primary involved in these two pathways. The presence of the DH pathway is particularly noteworthy because several enzymes of the pathway are sensitive to oxygen and this rare pathway is strictly anaerobic (Berg, Kockelkorn, et al. 2010). This is consistent with the DH pathway being found in anaerobic crenarchaeal orders *Thermoproteales* and *Desulfurococcales* (Berg, Ramos-Vera, et al. 2010), and possibly present in "marine group I archaea" *Thaumarchaeota* (Könneke et al. 2014).

Moreover, the four remaining pathways were also found with more than 50% completeness in multiple sampling sites, especially in the top portion of the heatmap. This observation was particularly interesting as the complete HBC pathway uses dissolved bicarbonate HCO_3^- as a starting substrate (Zarzycki et al. 2009). However, complete or even rudimentary HBC can co-assimilate trace amounts of organic compounds such as fermentation products (acetate, propionate and succinate) and numerous other compounds that are metabolized through acetyl-CoA and propionyl CoA (Zarzycki and Fuchs 2011). Such characteristics make HBC well suited for a parasitic or symbiotic lifestyle, because a nanoparasite with HBC could in principle fix (in)organic carbon and share organic carbon with its host. Whereas bacteria from the CPR superphylum have been described as likely symbionts or parasites, no CPR members harboring a HBC pathway have been described thus far (Kantor et al. 2013; Gong et al. 2014; Brown et al. 2015; Nelson and Stegen 2015). This suggests that nano-organisms may use either the complete or partial HBC pathway, even if this pathway was not observed in newly assembled genomes from TARA OCEANS (Tully et al. 2018).

Finally, when sequences from all sites were pooled together to produce an overall picture of the metabolic potential of ultrasmall prokaryotes, sequences associated with ultrasmall size fraction encoded a large fraction of the autotrophic carbon fixation pathways. The completeness of both carbon fixation pathways and ribosomal complexes decreases as data sets become more stringent, likely because of the reduction in the overall size of the data set. However, six sites (in majority from the surface or SRF) still included more than 50% of the enzymes involved in the HHC pathway within the set of sequences associated with the ultrasmall microbial "dark matter." By contrast, little evidence of a complete WL pathway in the ultrasmall "light dark matter" and in the ultrasmall "dark matter" was found, although the WL pathway is thought to be the ancestral and the most energetically efficient autotrophic carbon fixation pathway.

In sampling sites with high pathway completeness ($\geq$60%), further investigations were carried out to identify key enzymes, that is, enzymes that were specific to a metabolic pathway and thought to have appeared once during evolution (Berg 2011). The presence of all key enzymes of a metabolic pathway, together with a high completeness, strongly suggested the occurrence of that metabolic pathway in the environment. In the PU data set, the key enzymes for the CBB, rTCA, DH, HBC, and WL pathways were identified in some sites (Berg, Ramos-Vera, et al. 2010; Berg 2011). The distribution of CBB and DH pathways appeared widespread, whereas rTCA and WL were only found in the Indian Ocean cluster (dot 3 in fig. 5). Of note, the rTCA pathway is the second least expensive cycle after WL, using two ATPs, making it suitable for fermenting organisms to utilize. Several rTCA enzymes are sensitive to oxygen, restricting rTCA activity to anaerobic or low oxygen environments. In the UO data sets, the key enzymes for the CBB, DH, and HBC pathways were detected, but the HBC pathway was restricted to two sampling sites (dot 3 in fig. 5). In the WUO data sets, key enzymes for the CBB pathway were found in 12 sites and the DH pathway was found in 10 sites, whereas the HBC

was only found in 1 (038 at mesopelagic depth). However, the presence of the HBC pathway in nano-organisms deserves further investigation. Recent articles (Shih et al. 2017) suggest different key enzymes for HBC than those used here (Berg 2011), and homologs of some of these alternative enzymes have been found in our most stringent data set WUO "dark matter" (K08691 28 sequences, K09709 19 sequences, and K14449 7 sequences), albeit with a rather low number of occurrences.

In the UO "light dark matter" data set, the DH pathway was still found in three sites but the CBB pathway was only complete in the pool data set; whereas in the UO "dark matter" data set, the CBB and DH pathways were both only complete in the pool data set.

Discussion

Ultrasmall prokaryotes have only recently been discovered, but what is known to date about their physiology highlights their uniqueness. Members of the CPR and DPANN superphyla from aquifers have recently been described as able to perform reactions related to carbon fixation, although they are usually described as degraders rather than carbon fixing (Castelle et al. 2015, 2018; Anantharaman et al. 2016). Although aquifers represent a fraction of the aquatic environments on Earth, oceans represent a different and larger type of aquatic environment; therefore, the conclusions obtained from studying aquifers may not be applicable to oceans. In particular, we postulate that a broader diversity of microbes, including ultrasmall ones, would thrive in the oceans (although the ultrasmall size fraction has not been extensively studied). This reasoning is in agreement with our hypothesis that new, unidentified lineages of ultrasmall prokaryotes may play a role in autotrophic carbon fixation in the oceans. Using the broad TARA oceans data set, the aim of this work was to determine if members of the CPR and DPANN superphyla, and potentially additional ultrasmall prokaryotes, could contribute to (and eventually complete) pathways of carbon fixation in the oceans.

The diversity of nano-organisms is probably still under appreciated because few studies (Brown et al. 2015; Castelle et al. 2015; Luef et al. 2015; Anantharaman et al. 2016; Paul et al. 2017) have focused on the ultrasmall size fraction of publicly available metagenomes. In our study, for example, analyses of ribosomal markers suggested the existence of at least one large clade of tiny archaea, restricted to two sites that were geographically close in the Indian Ocean (TARA sampling sites 037 MES, 038 MES, and 039 MES). The phylogenetic analyses also hinted at a diversity of novel minute bacteria. Unraveling these additional actors suggests that the ecological and evolutionary roles of microbial diversity within the ocean remain to be fully described. In particular, nano-organisms could deeply impact carbon cycling and carbon fixation; while also contributing to trophic chains and the dynamics of microbial communities (Morris et al. 2012; Biller et al. 2015; Ponomarova and Patil 2015; Zelezniak et al. 2015) in ways that are still to be modeled. Abundant, ubiquitous taxa, such as *Prochlorococcus* and *SAR11* (Partensky et al. 1999; Giovannoni 2017), have already been proposed to affect geochemical cycles and biotic communities at a very large (planetary) scale. Populations of less abundant nano-organisms may also have an influence, at a scale which remains to be determined. Rate measurements will be needed (possibly in simple ecosystems) to test this hypothesis.

In this study, we were able to detect genes involved in the six known autotrophic carbon fixation pathways among those unassigned taxa, exclusive to the ultrasmall size fraction of the TARA OCEAN project. In spite of the limited sequencing depth at each site, these pathways were more than 50% complete at some sites. Moreover, in our stringent data sets (WUO) the anaerobic and energetically efficient DH pathway was more than 50% complete at 33 sampling sites. Interestingly, this in contrast to the carbon fixation pathways associated with CPR and DPANN superphyla in aquifers (Probst et al. 2017), which suggest that nano-organisms may have a broader contribution to carbon fixation than currently assumed. It is possible that some carbon fixation genes are carried by viral particles (although our analyses did not find any signal for this).

Assuming microbial communities were sufficiently well sampled, the detection of partial metabolic pathways and associated key enzymes raises the question of the actual contribution of these genes to carbon fixation and cycling in the environment. These genes may play an effective role under two distinct conditions. First, the genomes hosting the partial pathways may also host alternative genes encoding for unknown enzymes that can perform the missing steps for carbon fixation. Second, alternative genes encoding unknown enzymes would perform the missing steps, which may be distributed across phylogenetically diverse community members and interacting *via* metabolic hand-offs (Embree et al. 2015; Tsoi et al. 2018; Rubin-Blum et al. 2019). The contribution of marine nano-organisms to carbon fixation might therefore be a collective property, in which different microbes contribute to different steps of carbon fixation. Such metabolic cooperation in microbial communities has been described (DeLong 2007; Stams and Plugge 2009), but in the ocean such interactions might be rare except for communities associated with floating particles and sediments. Under the first hypothesis, transporters for some of the metabolic intermediates should exist in nature. We indeed found transporter candidates in the WUO "dark matter" data set, including a putative citrate/succinate antiporter (COG0471), both molecules being present in rTCA, and numerous ATPase components of ABC transporters (COG0488). The alternative hypothesis, that is, the contribution of specific novel lineages to carbon fixation, could lead to the discovery of new autotrophic nano-organisms, which are of similar importance to

Prochlorococcus or SAR11, currently the smallest described carbon fixing organism.

Under both hypotheses, our study encourages single cell genome analyses and/or the binning of metagenomes into genomes of nanosized micro-organisms. This would allow further characterization of the precise mechanisms by which the organisms contribute to carbon fixation.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank two anonymous reviewers for critical comments, and Dr S. Sunagawa, Dr S. Chaffron, Dr L. Bittner, Dr C. de Vargas, and Dr L.P. Coelho for giving access upon request to the abundance matrix and cd-hit output file of TARA OCEAN metagenomics. We also thank J. Pathmanathan for his help in experimental design, A.K. Watson and D. Bhattacharya for stimulating discussions. This work was granted access to the HPC resources of the Institute for Computing and Data Sciences (ISCD) at Sorbonne Université. R.L. and E.B. were supported by FP7/2007-2013 grant agreement 615274. K.O.-F. is funded by the UK Space Agency.

Literature Cited

Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO. 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Natl Acad Sci U S A. 110(17):E1594–E1603.

Anantharaman K, et al. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun. 7:13219.

Andrew K, et al. 1999. Size limits of very small microorganisms: proceedings of a workshop. Washington (DC): National Academies Press.

Berg IA. 2011. Ecological aspects of the distribution of different autotrophic CO_2 fixation pathways. Appl Environ Microbiol. 77(6):1925–1936.

Berg IA, et al. 2010. Autotrophic carbon fixation in archaea. Nat Rev Microbiol. 8(6):447–460.

Berg IA, Ramos-Vera WH, Petri A, Huber H, Fuchs G. 2010. Study of the distribution of autotrophic CO_2 fixation cycles in Crenarchaeota. Microbiology 156(Pt 1):256–269.

Biller SJ, Berube PM, Lindell D, Chisholm SW. 2015. Prochlorococcus: the structure and function of collective diversity. Nat Rev Microbiol. 13(1):13–27.

Brown CT, Olm MR, Thomas BC, Banfield JF. 2016. Measurement of bacterial replication rates in microbial communities. Nat Biotechnol. 34(12):1256–1263.

Brown CT, et al. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523(7559):208–211.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12(1):59–60.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.

Castelle CJ, et al. 2015. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr Biol. 25(6):690–701.

Castelle CJ, et al. 2018. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol. 16(10):629–645.

Corel E, Lopez P, Méheust R, Bapteste E. 2016. Network-thinking: graphs to analyze microbial complexity and evolution. Trends Microbiol. 24(3):224–237.

De La Rocha CL, Passow U. 2014. The Biological Pump. In: Turekian HDHK, editor. Treatise on Geochemistry. 2nd ed. Oxford: Elsevier.

DeLong EF. 2007. Life on the thermodynamic edge. Science 317(5836):327–328.

Dyksma S, et al. 2016. Ubiquitous gammaproteobacteria dominate dark carbon fixation in coastal sediments. ISME J. 10(8):1939–1953.

Embree M, Liu JK, Al-Bassam MM, Zengler K. 2015. Networks of energetic and metabolic interactions define dynamics in microbial communities. Proc Natl Acad Sci U S A. 112(50):15450–15455.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23): 3150–3152.

Giovannoni SJ. 2017. SAR11 bacteria: the most abundant plankton in the oceans. Ann Rev Mar Sci. 9:231–255.

Gong J, Qing Y, Guo X, Warren A. 2014. 'Candidatus Sonnebornia yantaiensis', a member of candidate division OD1, as intracellular bacteria of the ciliated protist *Paramecium bursaria* (Ciliophora, Oligohymenophorea). Syst Appl Microbiol. 37(1):35–41.

Guidi L, et al. 2016. Plankton networks driving carbon export in the oligotrophic ocean. Nature 532(7600):465–470.

Haggerty LS, et al. 2014. A pluralistic account of homology: adapting the models to the data. Mol Biol Evol. 31(3):501–516.

Haller CM, Rölleke S, Vybiral D, Witte A, Velimirov B. 1999. Investigation of 0.2 μm filterable bacteria from the Western Mediterranean Sea using a molecular approach: dominance of potential starvation forms. FEMS Microbiol Ecol. 31:153–161.

Hasegawa H, Naganuma K, Nakagawa Y, Matsuyama T. 2003. Membrane filter (pore size, 0.22–0.45 μm; thickness, 150 μm) passing-through activity of *Pseudomonas aeruginosa* and other bacterial species with indigenous infiltration ability. FEMS Microbiol Lett. 223:41–46.

Hug LA, et al. 2016. A new view of the tree and life's diversity. Nat Microbiol. 1:16048.

Hügler M, Sievert SM. 2011. Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. Ann Rev Mar Sci. 3:261–289.

Kantor RS, et al. 2013. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. MBio 4(5):e00708–e00713.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066.

Könneke M, et al. 2014. Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO_2 fixation. Proc Natl Acad Sci U S A. 111(22):8239–8244.

La Cono V, et al. 2018. Contribution of bicarbonate assimilation to carbon pool dynamics in the deep Mediterranean Sea and cultivation of actively nitrifying and CO_2-fixing bathypelagic prokaryotic consortia. Front Microbiol. 9:3.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Luef B, et al. 2015. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun. 6:6372.

Minh BQ, Nguyen MAT, Von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 30(5):1188–1195.

Morris JJ, Lenski RE, Zinser ER. 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio 3: e00036–12.
Nelson WC, Stegen JC. 2015. The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. Front Microbiol. 6:713.
Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274.
Ogata H, et al. 1999. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 27(1):29–34.
Parks DH, et al. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2:1533–1542.
Partensky F, Hess WR, Vaulot D. 1999. Prochlorococcus, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev. 63(1):106–127.
Paul BG, et al. 2017. Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. Nat Microbiol. 2:17045.
Ponomarova O, Patil KR. 2015. Metabolic interactions in microbial communities: untangling the Gordian knot. Curr Opin Microbiol. 27:37–44.
Probst AJ, et al. 2017. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO_2 concentrations. Environ Microbiol. 19:459–474.
Rappé MS, Connon SA, Vergin KL, Giovannoni SJ. 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418(6898):630–633.
Rinke C, et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499(7459):431–437.
Rubin-Blum M, Dubilier N, Kleiner M. 2019. Genetic evidence for two carbon fixation pathways (the Calvin–Benson–Bassham cycle and the reverse tricarboxylic acid cycle) in symbiotic and free-living bacteria. mSphere 4:e00394–18.
Shih PM, Ward LM, Fischer WW. 2017. Evolution of the 3-hydroxypropionate bicycle and recent transfer of anoxygenic photosynthesis into the Chloroflexi. Proc Natl Acad Sci U S A. 114(40):10749–10754.
Stams AJM, Plugge CM. 2009. Electron transfer in syntrophic communities of anaerobic bacteria and archaea. Nat Rev Microbiol. 7(8):568–577.
Sunagawa S, et al. 2015. Structure and function of the global ocean microbiome. Science 348(6237):1261359.
Tsoi R, et al. 2018. Metabolic division of labor in microbial systems. Proc Natl Aacd Sci U S A. 115:2526–2531.
Tully BJ, Graham ED, Heidelberg JF. 2018. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 5:162503.
Wheeler DL, et al. 2007. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 36:D13–D21.
Wieder WR, Cleveland CC, Lawrence DM, Bonan GB. 2015. Effects of model structural uncertainty on carbon cycle projections: biological nitrogen fixation as a case study. Environ Res Lett. 10. doi:10.1088/1748-9326/10/4/044016.
Wrighton KC, et al. 2016. RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria. ISME J. 10(11):2702–2714.
Zarzycki J, Brecht V, Müller M, Fuchs G. 2009. Identifying the missing steps of the autotrophic 3-hydroxypropionate CO_2 fixation cycle in *Chloroflexus aurantiacus*. Proc Natl Acad Sci U S A. 106(50):21317–21322.
Zarzycki J, Fuchs G. 2011. Coassimilation of organic substrates via the autotrophic 3-hydroxypropionate bi-cycle in *Chloroflexus aurantiacus*. Appl Environ Microbiol. 77(17):6181–6188.
Zelezniak A, et al. 2015. Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proc Natl Acad Sci U S A. 112(20):6449–6454.

Associate editor: Laura A. Katz

3.3 Procaryotes très petits et communautés microbiennes

Le quorum sensing désigne un système de communication par lequel les micro-organismes synchronisent l'émergence d'un comportement collectif (Miller et al. 2001). Les micro-organismes sécrètent des molécules de signalisation. Quand la concentration de ces molécules atteint un seuil (quorum), cela déclenche une réponse de la part des organismes équipés de récepteurs spécifiques à la molécule signal. Cela permet aux micro-organismes d'estimer leur densité de population et de déclencher une action quand celle-ci dépasse un certain seuil (quorum). Le système de quorum sensing est composé de trois éléments: une synthase, un récepteur et un régulateur de réponse qui sont responsables respectivement de la synthèse de la molécule signal, de sa détection et de la réponse transcriptionnelle associée à la détection de la molécule signal. Le système de quorum sensing interspécifique a notamment été décrit comme un système de bio-communication pour la mise en place et le maintien d'interactions symbiotiques ou parasitaires entre micro-organismes (Bedree et al. 2018). Avec la découverte récente des CPR et des DPANN, qui sont pour beaucoup décrits comme des symbiotes/parasites obligatoires, il nous est apparu opportun d'enquêter sur la présence de systèmes de quorum sensing chez ces organismes. Le manque de voies de biosynthèses chez les CPR, DPANN est peut-être compensé par la présence de gènes "sociaux". En effet, la présence d'un système de quorum sensing même partiel chez les CPR et des DPANN nous aiderait à comprendre comment ils s'intègrent dans les communautés microbiennes et interagissent avec leurs hôtes pour mettre en place d'éventuels modes de vie symbiotiques.

3.3.1 Article 3, "Rich repertoire of quorum sensing systems in CPR and DPANN associated with inter-species communication", International Society for Microbial Ecology Journal, 2019 submitted

Rich repertoire of quorum sensing systems in CPR and DPANN associated with inter-species and inter-kingdom communication

RUNNING TITLE

Diverse communication systems in CPR and DPANN

AUTHORS

Charles Bernard[1,2], Romain Lannes[1], Yanyan Li[2], Eric Bapteste[1] and Philippe Lopez[1]

AFFILITATIONS

[1] *Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, Campus Jussieu, Bâtiment A, 4eme et. Pièce 429, 75005 Paris, France*

[2] *Unité Molécules de Communication et Adaptation des Micro-organismes (MCAM), CNRS, Museum National d'Histoire Naturelle, CP 54, 57 rue Cuvier, 75005 Paris, France*

CORRESPONDING AUTHOR

Correspondence to Philippe Lopez:
- E-mail address: philippe.lopez@upmc.fr
- Telephone number: +33 (0)1 44 27 34 70
- Postal adress: Campus Jussieu, Bâtiment A – 4e étage – pièce 429, 75005 Paris, FRANCE

COMPETING INTERESTS

The authors declare that they have no competing interest.

ABSTRACT

The bacterial Candidate Phyla Radiation (CPR) and the archaeal DPANN superphylum are two novel lineages that have substantially expanded the tree of life due to their large phylogenetic diversity. Because of their ultrasmall cell size, reduced genome and lack of core biosynthetic capacities such as amino acids or nucleotides de novo synthesis, CPR and DPANN members are believed to be sustained through their interactions with other organisms. How they achieve such interactions is, however, little understood. Here, we hypothesized that CPR and DPANN species might rely on chemical communication via quorum sensing (QS) to interact with other species. This hypothesis motivated our *in silico* analysis to identify whether CPRs and DPANNs had homologs of reference proteins involved in 37 well known QS systems. Our survey shows that many CPR and DPANN species harbor QS proteins homologous to those used by *Proteobacteria* to either signal their presence to other prokaryotes, sense the presence of other prokaryotes, manipulate host motility or eavesdrop on inter-kingdom signals. Our predictions therefore give more insights into the underlying functions supporting the infered symbiotic lifestyles of CPR and DPANN and opens a perspective towards significantly expanding our knowledge of microbial communication across the tree of life.

INTRODUCTION

The recent efforts at sequencing the DNA extracted from diverse environments enabled access to genomes of uncultivated microorganisms with no isolated representatives, which have expanded our vision of life's diversity (1). Most of this expansion is attributable to the discovery of two novel microbial lineages, the Candidate Phyla Radiation (CPR), estimated to account for more than 26% of the currently known bacterial diversity (2), and the archaeal DPANN (for Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota) superphylum (3). Although little is known about these lineages, they already challenge our perspectives on the biology of prokaryotes: CPR and DPANN microorganisms have small to ultra-small cell sizes (some can pass through 0.22 µm filters (4)), reduced genome sizes and most of them lack core genes in pathways considered as essential in other prokaryotic lineages, such as nucleotides, amino acids and lipids biosyntheses (3,5). These singularities suggest that the majority of these ultrasmall species might depend on other organisms to survive (6) or may even be obligate symbionts (5), a suggestion supported by the few endobiotic (7) and epibiotic (8,9) relationships uncovered between a CPR or a DPANN and other microorganism(s). However, the biological functions by which these interactions are driven are currently little understood.

Here, we hypothesized that one of these functions could be cell-cell communication *via* quorum sensing (QS), which is known to be a central feature for many microorganisms to interact and adapt to their environments, and that is particularly important in the development and the maintenance of symbiotic or parasitic relationships with hosts (10–14). Specifically, QS involves the production and the perception of diffusible signaling molecules that translates into the emergence of a synchronous collective behaviour in individuals upon reaching a sufficiently high population density. Some examples of behaviors controlled by QS, known to promote symbiotic or parasitic interactions, include bioluminescence in *Aliivibrio fischeri*, beneficial for the hunting performance of its host, the squid *Euprymna scolopes (15)*, biofilm formation in *Sinorhizobium freedi,* required for its successfull symbiosis with the roots of *Glycine max (16)*, and virulence in pathogens such as *Pseudomonas aeruginosa (17)* or *Vibrio cholerae (18),* crucial for an effective host invasion.

QS presupposes that their users occasionally encounter high density and a few studies have reported that some CPR species are found to be abundant under certain conditions. For example, the parcubacterium *Candidatus Sonnebornia yantaiensis* has been described as a rare taxon that is sometimes found by the thousands in the cytoplasm of paramecia isolated from a freshwater pond in Yuntai, China (7). Again, *Saccharibacteria* are known to be present at a relative abundance of ~1% in healthy human oral cavities, but this can increase up to 21% of the whole microbial community in case of periodontal diseases (19,20). To our knowledge, no studies have yet reported events of occasional density in the DPANN superphylum.

Taken together, the initial studies on CPR and DPANN species lend therefore some credibility to the hypothesis that their reduced genomes might retain or expand genes for QS, which they would likely rely upon to develop and maintain crucial interactions with other species. Hence, to test this hypothesis, we performed an *in silico* analysis to identify homologs of known QS systems in their genomes. Here, we show that many CPR and DPANN members harbor putative homologs of proteins responsible for either the synthesis of QS signals or their integration, and associated with inter-species and inter-kingdom communication (Figure 1). Finally, the majority of the putative QS components in CPR and DPANN lineages are divergent from the proteins predicted from all the complete genomes of prokaryotes available, representing therefore good candidates for the discovery of new signaling molecules and circuits.

MATERIAL AND METHODS

Construction of a reference database of sequences of QS synthases, receptors and response regulators

We carefuly mined the literature related to QS to establish a nearly comprehensive list of QS systems. This list is available as a tabular file (Supplementary Table 1) which notably records, for each QS system, the species in which it has been characterized, the induced QS response, the chemical characteristics of the QS signal, as well as the function, the genomic coordinates and the NCBI protein identifier(s) of each of its components. The protein and the coding sequences corresponding to each entry were retrieved from the NCBI protein and nucleotide databases (21). Additionally, the protein and coding sequences corresponding to the *luxI* and *luxR* gene families were accessed through the Sigmol database (22).

Retrieval of CPR and DPANN proteomes and genomes

All the genomes and the corresponding predicted proteomes of CPR and DPANN species were downloaded from the NCBI taxonomy database in June 2018 (21). At the time of writing, these CPR genomes/proteomes can be retrieved from the lineages corresponding to the following taxonomic identifiers: txid74243, txid95818, txid221235, txid363464, txid422282, txid1618330, txid1618338, txid1618339, txid1618340, txid1619053, txid1794810, txid1794811 and txid1817799, whereas DPANN genomes/proteomes can be retrieved from the followings: txid1462430, txid1783276, txid1803511.

Detection of homologs of reference proteins in CPR and DPANN predicted proteomes

The homologs of reference QS-related proteins and of *L. pneumophila* and *R. prowazeckii* effectors of parisitism in CPR/DPANN proteomes were identified using the DIAMOND sequence aligner (version 0.9.19) (23). Classically, homology was assessed according to the following thresholds: sequence identity ≥ 30%, E-value < 1e-5 and mutual alignment coverage ≥ 80% (24–26).

Detection of homologs of putative QS-related CPR/DPANN proteins in the available complete proteomes of prokaryotes

The predicted proteomes corresponding to the available complete genomes of 279 archaea and 12,762 bacteria were retrieved from the NCBI assembly database on 07/03/2019. Homologs of CPR/DPANN putative QS-related proteins in these complete prokaryotic proteomes were identified using the same method as described above. The comprehensive list of all the CPR/DPANN QS-related homologs, the reference QS component to which they correspond to as well as their best homolog found in the complete prokaryotic proteomes is given in Supplementary Table 2.

HMM (Hidden Markov Model) search of LuxI profiles in CPR and DPANN proteomes

Assessment of the mutal homology of 76 experimentally validated LuxI synthases led to the identification of 5 clusters (Supplementary Figure 2A). The list of reference sequences

composing each of these clusters was further extended to homologs present in the available complete proteomes of prokaryotes (sequence identity > 60% and mutual alignment coverage > 80%). Multiple sequence alignment of the protein sequences of each extended cluster was performed by MUSCLE (version 3.8.31) (27) and further trimmed on the N-terminal and C-terminal extremities by trimAl (version 1.4.rev22) (28) *via* the options "-terminalonly -gappyout". The HMM profiles of each of the 5 clusters were constructed with HMMbuild (from the HMMER suite version 3.2.1) (29) based on these multiple alignments. The CPR/DPANN proteins which matched these HMM profiles were identified by HMMsearch and only the hits with an HMMscore > 20 and an E-value < 0.01 were retained.

Multiple sequence alignment of the CPR/DPANN families of homologs

Likewise, the multiple protein sequence alignment of each CPR/DPANN family of homologs (corresponding to a unique reference QS-related protein) was performed by MUSCLE (version 3.8.31) (27) with the option "-maxiters 50".

For the LuxR and the LqsA protein families, an additional alignment which, this time, included the sequences of reference proteobacterial proteins was undertaken to assess the conservation of key residues (as described in the literature) in CPR/DPANN sequences. But prior to these two alignments, a clustering of the CPR/DPANN sequences on the basis of sequence identity (% identity cutoff > 90%) was performed by the CDHIT online suite (30) to retain only one representative sequence per cluster. The clustering step was meant to not overload the LuxR or LqsA alignments and facilitate visualization. Eventually, the positions which gave rise to more than 80% gaps were removed from these two alignments. The Jalview program (31) was used to visualize the results.

dN/dS ratio computation for each family of homologs

Starting from the multiple protein sequence alignment of each CPR/DPANN family of homologs, we generated the corresponding alignment for their coding sequences (CDS). Hence the amino-acids of each sequence in the alignment were substituted by the codons corresponding to their position in the CDS and any gap character was substituted by three gap characters. Importantly, no stop codons were present in these multiple CDS alignments. Then, we introduced a slight modification in the source code of the SNAP perl script, a tool designed to calculate synonymous and nonsynonymous values for an alignment of CDS (32). Namely, in the vector which associates an amino-acid letter to a cognate codon, we substituted the dummy "Z" character associated with the opale stop codon (UGA) by a "G". This was meant to take into account the alternative genetic code of Gracilibacteria and Absconditabacteria (SR1) CPR phyla in which the opale codon encodes a glycine (NCBI:transl_table=25). The dN/dS ratio of each family of homolog was then output by the

modified SNAP program as the average of all the dN/dS ratios computed for each possible pairwise comparison of sequences in the alignment. Nevertheless, whenever a CPR/DPANN family of homologs did not comprise more than two sequences, the dN/dS ratio was not computed.

Phylogenetic tree inference for LuxS proteins

The LuxS tree was built from the CPR/DPANN putative LuxS synthases as well as their detected homologs in the available complete proteomes of prokaryotes. The pipeline to infer the phylogeny of each of the LuxS protein family was initiated by a multiple sequence alignment with MUSCLE (27), followed by a trimming step undertaken by trimAl (28) with the option “-automated1”. Finally, each of the resulting trimmed alignments was given as input to IQ-Tree (version multicore 1.6.10) (33) to build a maximum likelihood phylogenetic tree under the LG+G model with 1000 ultrafast bootstraps (34).

RESULTS

The predicted proteomes of CPRs and DPANNs comprise homologs to reference proteins involved in diverse, mainly proteobacterial, QS systems

A QS system comprises three main components: a synthase, a receptor and a response regulator which are responsible for the signal synthesis, its detection and the subsequent induction of the QS response at a transcriptional level, respectively. The quorum is thereby materialized by the threshold concentration at which the collectively produced signal is effectively sensed by the receptors and relayed to the response regulators. Out of the 37 reference QS systems that we identified in the literature (Table 1 and Supplementary Table 1), 21 had at least one component other than a response regulator detected as homologous in at least one CPR or DPANN species (Figure 2). Here, a homolog is defined as a protein whose sequence identity is no less than 30% to a reference protein, with over 80% of mutual coverage and with a E-value below 1e-5 in a DIAMOND search (Material and Methods). These thresholds offer a good trade-off between functional reliability and permissive stringency, according to their application on a set of 76 experimentally validated LuxI synthases retrieved from the Sigmol database (22). (Supplementary Figure 2A).

Unsurprisingly, the intra-species and therefore isolated mode of communication of the *Firmicutes* phylum, relying on the production of autoinducer peptides (AIP) is not well represented in CPR and DPANN lineages: only 3 out of the 21 QS systems correspond to reference AIP signaling, and are moreover found to be largely incomplete (Figure 2). The vast majority of the QS systems found in CPRs and DPANNs (18/21) correspond indeed to those of the *Proteobacteria* phylum, which synthetize and sense small diffusable molecules

rather than peptides. This finding is in agreement with the broader spectrum of specificites reported for the proteobacterial QS molecules (QSM), ranging from intra-species to inter-kingdom levels of secrecy (35,36), and implying that some QSM are proned to be produced or recognized by phylogenetically distant species. Of note, the only known QS system of the *Actinobacteria* phylum (relying on the gamma-butyrolactone QSM) is not found in CPR/DPANN lineages.

The vast majority of the sequences of the QS homologs are divergent and under strong selective pressure

In order to better characterize these QS homologs, we sought to identify what was their closest protein, in term of sequence, amongst the available complete proteomes of prokaryotes (Supplementary Table 2). For each CPR/DPANN homolog, its sequence identity to its reference QS protein and to its closest prokaryotic protein detected were hereafter plotted against each other in a scatterplot (Supplementary Figure 1A). It appears that the overwhelming majority of these homologs have no more than 60% of identity to any protein predicted from the complete genome of 13,041 prokaryotes, highlighting thereby how divergent their sequence are from those of well studied organisms. On another hand, the comparison of the functional annotations of, respectivily, each reference QS protein, the best corresponding CPR/DPANN homolog and its closest prokaryotic protein (Supplementary Table 3) does not reveal clear contradictions, with the exception of the CcfA synthases, whose CPR/DPANN homologs turn out to be more likely YidC translocases according to the functional annotation of their closest proteins identified in the complete proteomes of prokaryotes.

Although the CPR/DPANN QS-related homologs are divergent, selection is inferred to act against changes in their protein sequence. The fact that the dN/dS ratio, a metric relying on the ratio of the number of non-synonymous mutations to the number of synonymous mutations along distinct coding sequences, falls systematically below 1 for each family of homologs is indeed indicative of a purifying selection acting on these genes (37) (Figure 2). This finding highlights the importance of the functions that the CPR/DPANN putative QS-related proteins actually support.

The QS systems found in CPR and DPANN species exhibit different levels of completeness

Interestingly, no complete QS genetic circuits (synthase + receptor + response regulator) have been identified in CPR and DPANN species. Nevertheless, it must be said that not all of the 37 reference QS systems considered in this study are fully characterized and some await

the experimental identification of all their constitutive components. Two of such partially characterized reference QS systems have been found in certain CPR or DPANN species and could hopefully be part of a full QS circuit. These two systems are the AsgA/SasS system of *Myxococcus xanthus* and the QseC/QseB two-component sensor system of the pathogenic *Escherichia coli* O157:H7 (Figure 2). The former system is singular in the sense that it does not rely on a signaling molecule/peptide but rather on a mixture of proteases, peptides and amino acids, named A-signal, and of which the biosynthesis remains to be characterized (38). As it seems likely that the complex composition of the A-signal is specific to *M. xanthus*, the analogous QS system detected in some CPR and DPANN species might support other functions than QS. The latter system, however, is more reliable and is discussed later in this paper.

As no QS systems can yet be said to be complete in CPR/DPANN, we next sought to systematically check the presence in CPR/DPANN genomes of i) any entire biosynthetic pathway required for the production of a given QS signal ii) any receptor / response regulator QS sensor required for eavesdropping on exogenous signals. In agreement with what the poor anabolic capacities of CPRs and DPANNs could suggest, whenever a QS signal requires a whole cluster of genes for its production (several synthases), this cluster is never found to be complete in CPR/DPANN genomes at the sole exception of the RpfF and RpfB coupled synthases. Adjacently positioned in the reference genome of *Xanthomonas campestris*, *rpfF* and *rpfB* are also found to be adjacent in the genome of a *Nanoarchaeota* DPANN (protein ids RLG18487.1 and RLG18486.1) and spaced by only three genes in the genome of a *Niyogibacteria* CPR (protein ids PIR69995.1 and PIR69999.1), indicating that these two species might produce diffusible signal factors (DSF). On the other hand, the detected homologs of PpyS, CqsA, LqsA, FilI and to a lesser extent, Tdh, automatically *a priori* satisfy the condition for an effective signal production (of Photopyrone (PPY), Cholera Autoinducer-1 (CAI-1), Legionella Autoinducer-1 (LAI-1), Carboxylated Homoserine Lactone and 3,5-dimethylpyrazin-2-ol (DPO), respectively), since each of these reference proteins represents the sole synthase of a QS system (Figure 2, Supplementary Table 1). As for the copresence of a receptor with a response regulator inside a CPR or a DPANN genome, it must be noted that some transcription factors possess both a ligand binding domain for signal detection and a DNA binding domain for subsequent transcriptional regulation (39). Such proteins are refered to as « one-component systems » and comprise in this study the reference proteins PauR, PluR, PgaR, VqmA, as well as those of the LuxR family. Only homologs of the luxR one-component systems have been identified in CPR/DPANN and it therefore follows that these can be considered *a priori* as functional signal integrating apparatuses (Figure 2). Because the one-component systems gather the sensor and the

transcription factor domains on the same protein, they are believed to be circumscribed to the cytoplasm and to sense only intracellular signals (imported or endogenously produced) (39). Later on during evolution, this physical link would have been uncoupled into two distinct proteins, a membrane receptor (most of the time an histidine kinase) and its associated intracellular response regulator, thereby allowing the integration of extracellular signals (39,40). These two distinct proteins are refered to as a « two-component system » (TCS) and concern here the SpaK/SpaR, CylR1/CylR2, NwsA/NwsB, RpfC/RpfG, LqsS/LqsR, LsrB/LsrR, QseC/QseB, PhcS/PhcR couples. Out of these 8 TCS, only the RpfF/RpfG, SpaK/SpaR, QseC/QseB are found to be complete in 1, 4 and 71 CPR species, respectively (Table 2).

Intra-species communication through autoinducer-1 (acyl-homoserine lactone) is unlikely in CPR/DPANN

Acyl-homoserine lactone (AHL or autoinducer-1 (AI-1))-based QS system is common and wide-spread in Gram-negative bacteria. The canonical gene organization comprises adjacent *luxI* and *luxR* genes that encode an AHL synthase and a one-component receptor/regulator (35). It is interesting to note that homologs of the LuxR AHL one-component system are found across 70 out of 76 CPR/DPANN phyla, while no homologs of their cognate LuxI, HtdS or LuxM bacterial synthases have been detected (Figure 2, see LuxI family). A closer look at each sequence of these LuxR homologs revealed that none of them indeed contains the WYPDWG motif required for AHL binding (41,42). This observation suggest that the LuxR homologs are probably false positives for AI-1 recognition but might perhaps bind other QS signals than AHLs (43), by analogy with the LuxR homologs PauR and PluR that do not contain either the WYDPWG motif, nor are adjacent to their cognate synthase but do sense nonetheless their photopyrone and dialkylresorcinol/cyclohexanedione QS signals (41). Such solo or orphan luxR have been suggested to participate in inter-species or inter-kingdom communication (44). To ensure that AI-1 based QS is indeed unlikely in CPR and DPANN lineages, we built 5 HMM profiles from the clusters of 76 experimentally validated AHL synthases at our disposal (Supplementary Figure 2A). These HMM profiles matched only 20 CPR proteins (Supplementary Figure 3B), none of which having a coding sequence adjacent to a *luxR* homolog. In fact, these 20 matches might rather correspond to GNAT N-acetlytransferases (with which the LuxI-type AHL synthases share similar mechanisms (45)), according to the functional annotation of their best homologs in the available complete proteomes of prokaryotes (Supplementary Table 4). Worthy of note, the only known archaeal AHL synthase (46), named FilI, has 51 detected homologs in CPRs and 2 in DPANN altiarchaeales. Nevertheless, FilI is annotated as a multisensor histidine kinase and the homology with the 51 CPR/DPANN similar proteins applies only on its C-terminal domains

linked to the histidine kinase activity and not on its N-terminal CHASE4 domain responsible for carboxylated AHL synthesis (46). These FilI homologs are then probably just multisensor kinases. However, we incidentally observed that when *filI* and *luxR* homologous genes are adjacent in a CPR genome, they are systematically found upstream from the gene cluster responsible for pilus type-IV assembly. This pilus system is a cell surface structure in prokaryotes involved in cell mobility, adhesion to the surface, DNA uptake and biofilm formation. The potential link of FilI/LuxR signaling with type-IV pili biosynthesis could therefore have an implication in the episymbiotic lifestyle of certain CPRs.

Biosynthesis of the autoinducer-2 (AI-2) inter-species signal seems specific to *Gracilibacteria* from Crystal Geyser, Utah

Unlike AI-1/AHL signaling, ranging from intra-species to inter-species specificities of cell-cell communication, autoinducer-2 (AI-2) is believed to be a universal mechanism that mediates inter-species communication in bacteria. Surprinsingly, the 5 CPR homologs of the AI-2 synthase LuxS (PIQ41187.1, OIO77452.1, PIZ01540.1, PJC56868.1 and PIQ10870.1) constitute the unique family of homologs for which no representative protein is divergent from the proteins of fully-sequences prokaryotes (Supplementary Figure 1B). They are all identical to each others, and belong to species of the *Gracilibacteria* CPR phylum sampled from groundwater in Crystal Geyser, Green River, Utah (47). The LuxS phylogenetic tree (Figure 3C), resulting from the alignment of the *Gracilibacteria* AI-2 putative synthases with 351 bacterial homologs show that these CPR proteins are included in a homogeneous group with LuxS proteins from the *Lachnospiraceae* family of *Firmicutes*, hinting at a possible lateral transfer. Interestingly, no proteins from the 5 *Gracilibacteria's* predicted proteomes had more than 30% of sequence identity over 80% query cover with Pfs, the AI-2 precursor synthase, but the RKW20012.1 protein of another *Gracilibacterium* isolated from the human oral microbiome did (Figure 3A-B). Interestingly, RKW20012.1 shares 54.6% sequence identity with PIQ41186.1, OIO077453.1, PIZ01541.1, PJC56867.1 and PIQ10871.1, whose coding sequences are all directly upstream from those of the previously identified LuxS synthases (Figure 3B). The fact that these distant *pfs* homologs are adjacent with the *luxS* homologs in the genomes of the *Gracilibacteria* from Crystal Geyser is perhaps the genomic signature of a functional linkage related to AI-2 biosynthesis.

Putative eavesdropping on eukaryotic communication might be mediated by the QseC/QseB autoinducer-3 (AI-3) sensor across 19 CPR phyla

As already mentioned above, QseC/QseB is the most prevalent QS sensor in CPRs, since it is found complete in 71 species, across 19 phyla, most of which belonging to the

Microgenomates and Parcubacteria superphyla (Figure 2). In *Enterobacteria*, the QseC/QseB TCS functions as an adrenergic receptor, able to sense the autoinducer-3 QS signal (AI-3) as well as the epinephrine and norepinephrine inter-kingdom signaling molecules (48). In *Aggregatibacter actinomycetemcomitans*, QseC is activated by a combination of iron and epinephrine/norephrine, whereas only zinc and ferrous ions activate the *Haemophilus influenzae* sensor (49). Epinephrine and norepinephrine are stress hormones called catecholamines that are produced by animals (48) and by some protozoans (50,51). Hence, the QseC/QseB sensor offers the means for various pathogens, including *Escherichia coli* O157:H7 or *Salmonella enterica* serovar Typhimurium to eavesdrop the state of stress of their hosts and to respond by upregulating the expression of virulence and motility effectors (48,52). In species where the QseC/QseB system functions also as an iron sensor, it also allows to activate genes related to anaerobic metabolism (49). In *Escherichia coli* O157:H7, the *qseC* and *qseB* genes form an operon induced by the QseB response regulator upon AI-3 induction (48). Likewise, this co-directional synteny is found to be conserved in 48 CPR species (Table 2). Worthy of note, the QseC/QseB homologous TCS in CPR might probably not be used to eavesdrop on the acute stress response of animals since no complete QseC/QseB sensors were found in *Gracilibacteria*, *Saccharibacteria* or *Absconditabacteria*, the 3 CPR phyla that are known to live within animals (53–56) although homologs of the QseC and of the QseB proteins were identified in distinct *Saccharibacteria* (Figure 2). Then, QseC/QseB homologs could be rather used for recognition of the AI-3 QS signal and/or metals, or for eavesdropping on the catecholamines synthesized by protozoans. However, since the biosynthetic pathway of the AI-3 signal has not yet been identified yet, it is impossible to tell whether CPRs communicate with themselves or with their putative hosts *via* an endogenous synthesis of the AI-3 molecule.

CPR and DPANN members share effectors of the parasitic lifestyle of *Legionella pneumophila* and *Rickettsia prowazeckii*, including host-pathogen communication systems

Similarly to *E. coli* O157.H7 with QseC/QseB*, S. enterica* serovar Typhimurium is known to possess two TCSs which participate to host-pathogen interactions through neuroendocrine hormones: PmrB/PmrA and CpxA/CpxR (52). Together with the LetA/LetS and LqsS/LqsR, the analogous systems of PmrB/PmrA and CpxA/CpxR in *Legionella pneumophila* are known to be the key regulators of its parasitic life cycle (57). Interestingly, the LqS/LqsR system senses the Legionella Autoinducer 1 QS signal (LAI-1), metabolized by the LqsA synthase, a protein which has 65 detected CPR/DPANN homologs. The *Legionella pneumophila* LqsA and the analogous *Vibrio cholerae* CqsA reference synthases are pyridoxal-5'-phosphate

(PLP)-dependent aminotransferases-like enzymes and the sequences of these CPR/DPANN homologs exhibit indeed the 7 key residues necessary for PLP binding and catalysis (58) (Supplementary Figure 3). The LAI-1 signal is a hydroxyketone molecule that is not only produced as a QS signal by the intracellular parasites to repress their replication into the cytoplasm of their hosts, but also as an inter-kingdom signaling molecule that modulates host motility (57). In CPR and DPANN species, the whole LAI-1 based QS system is never found to be complete (Figure 2), rather hinting at an inter-kingdom signaling role of the LqsA homologs.

Altogether, these observations suggest that some CPR members might detect and produce inter-kingdom signaling molecules, usually associated with pathogenicity and parasitism in bacteria. This motivated our approach to assess whether the ultra-small CPR and DPANN prokaryotes possess analogous effectors and regulators to that used by two well characterized endoparasites, *L. pneumophila* and *Rickettsia prowazeckii,* to enter and exit their eukaryote hosts (57,59,60) (Figure 4 and Supplementary Table 5). We found that homologs of the phagocytosis escape effectors TlyC (73 CPR / 10 DPANN) and Pld (9 DPANN) are present inside the predicted proteomes of CPR/DPANN. Furthermore, systems upon which the decision to exit the host is made, like the RelA and SpoT indicators of amino-acids and lipids exhaustion or the *Legionella* PmrB/PmrA and CpxA/CpxR putative receptors of eukaryotic signals are widespread in CPRs and DPANNs. Last but foremost, the pleiotropic RtxA protein of *L. pneumophila*, involved firstly in host entry and adherence and further in cytotoxicity and pore-formation is found in 214 CPR and 6 DPANN species (61) (Supplementary Table 5). These findings suggest that the presence of the *Candidatus Sonnebornia yantaiensis* CPR inside a *ciliate(7)* is probably not an isolated case and that eukaryotic endobiosis could actually be more frequent as previously thought in the CPR and the DPANN lineages.

DISCUSSION

CPR and DPANN members alter the canonical view of the survival of the fittest: they are the living proof that intrinsically weak unicellulars (as reflected by their reduced genome and poor biosynthetic capacities) can be sustained through their interactions with other organisms. In this respect, our analysis provides some insights into the underlying functions that could serve as support for the promotion and the maintenance of these crucial interactions. Indeed, we predicted for the first time the presence of genes related to QS in genomes belonging to the CPR and DPANN novel expansions of the tree of life. Specifically, we showed the distribution, across the different CPR and DPANN phyla, of homologs of reference proteins involved in 21 different QS systems, out of 37 tested. Finally, we found that the most reliable

QS systems in CPR/DPANN lineages appeared to be related to inter-species and inter-kingdom communication (Figure 1). Since most CPR and DPANN species are suggested to have undergone genome reduction, the persistence of genes which presumably allow them to signal their presence, influence their neighbors and collect social cues might underline the prime importance of the role played by social traits in their survival. Our predictions could also convey the strong message that social traits are more critical for the survival of certain species that many traits considered as essential in others such as nucleotides, lipids or amino acids *de novo* synthesis.

Nevertheless, it must be admitted that no complete QS genetic circuits, from the synthesis to the integration of the signal, have been identified within CPR or DPANN genomes. Although this apparent incompleteness might indeed reflect a biological reality, one must keep in mind that it could also be explained by the high stringency of the thresholds that we have defined to detect an homology, by the poor amount of CPR/DPANN complete genomes available, or even by the fact that some components (synthase, receptor or response regulator) of certain reference QS systems still await to be characterized. However, should these QS genetic circuits be really incomplete (only either one of the signal synthase or of the signal sensor being selected), they would then not support intra-species communication but would still enable CPR and DPANN species to include themselves in inter-species and inter-kingdom communication networks. For instance, the QseC/QseB, CpxA/CpxR and CqsS homologs in CPR and DPANN species would still allow these species to eavesdrop on inter-kingdom signals, and would be anyway of likely upmost importance with respect to their suggested lifestyle associated with eukaryote hosts. With this respect, it has been already reported that the presence of a sole sensor without its cognate synthase within a genome offered the means for viruses and bacteria to eavesdrop on the signals produced by their neighbors or their hosts and adapt their physiology correspondingly (52,62). In the other way around, a sole signal production without recognition would raise an evolutionary issue, namely paying the cost of a signal production without any further pay-off in term of fitness. Nevertheless, this issue could be resolved if the cost of expression of an orphan synthase could be alleviated by the advantageous influence that their synthesized signal would have on the behaviour of other species in the neighborhood. For instance, CPR/DPANN homologs of the *L. pneumophila* LsqA synthase might produce an inter-kingdom signal that could manipulate the host motility to their benefit. This idea that some CPR/DPANN proteins might influence the behaviour of other species is supported by the recent discovery of systems to manipulate an eukaryotic host in a member of the TM6 phylum, a phylum phylogenetically close from the CPR radiation (63).

Hence, our predictions of the presence of certain partial QS systems in some CPR/DPANN genomes give insights into the means by which CPR and DPANN achieve critical interactions with other species. Nevertheless, understanding which processes these systems do regulate in CPR and DPANN species, whether involved in symbiosis, commensalism or parasitism awaits further functional studies. Nowadays, such studies appear difficult due to the low number of cultivable CPR and DPANN species and by the incovenience for genetic engineering imposed by their dependency on other microbial species but we hope that these obstacles will soon be overcome. Meanwhile, our study opens exciting perspectives in prokaryotic QS research in a foreseeable future. Specifically, the divergence of the sequences of CPR/DPANN putative QS synthases to well characterized ones is a great promise for the discovery of new molecules of communication *via* heterologous expression. Given that they may act as antagonists of known QS receptors in pathogens, this could notably lead to new anti-infective strategies (64). In the long term, deciphering experimentally the nature of the QS processes that are likely happening in the newly discovered CPR and DPANN lineages could greatly expand our knowledge of microbial communication across the tree of life, since QS has so far only been demonstrated in some well known phyla of bacteria and archaea, diatoms, unicellular fungi and some viruses.

ACKNOWLEDGMENTS

We acknowledge the “Interface pour le vivant” PhD Program of the Sorbonne University for the fundings of this study.

COMPETING INTERESTS

The authors declare that they have no competing interest to disclose.

DATA AVAILABILITY

All data generated or analysed during this study are included in this published article and its supplementary information files. IDs of all the reference QS proteins used as query in this study are given in Supplementary Table 1. IDs of reference proteins, corresponding CPR/DPANN homologs and best homologs found in the complete proteomes of prokaryotes are given in Supplementary Table 2. IDs of the CPR/DPANN proteins that match the different luxI HMM profiles are given in Supplementary Table 4. IDs of reference proteins of *R. prowazeckii* and *L. pneumophila* and CPR/DPANN homologs are given in Supplementary Table 5. Taxonomic names of the CPR/DPANN species considered in this study are given in Supplementary Table 6.

AUTHORS’ CONTRIBUTION

C.B., Y.L., E.B. and P.L. conceived the study. R.L. constructed the dataset of CPR/DPANN genomes/proteomes. C.B. and Y.L. constructed the database of QS systems. C.B. performed the analyses. C.B., Y.L. and E.B. wrote the manuscript with input from all authors. All documents were edited and approved by all authors.

REFERENCES

1. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016 May 11;1(5):16048.

2. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017 Nov 11;2(11):1533–42.

3. Castelle CJ, Banfield JF. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell. 2018 Mar 8;172(6):1181–97.

4. Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC, et al. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun. 2015 Dec 27;6(1):6372.

5. Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol. 2018 Oct 4;16(10):629–45.

6. Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, et al. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. MBio. 2013 Oct 22;4(5):e00708-13.

7. Gong J, Qing Y, Guo X, Warren A. “Candidatus Sonnebornia yantaiensis”, a member of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium bursaria (Ciliophora, Oligohymenophorea). Syst Appl Microbiol. 2014 Feb;37(1):35–41.

8. He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu S-Y, et al. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc Natl Acad Sci U S A. 2015 Jan 6;112(1):244–9.

9. Dombrowski N, Lee J-H, Williams TA, Offre P, Spang A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. FEMS Microbiol Lett. 2019 Jan 1;366(2).

10. Kai K. Bacterial quorum sensing in symbiotic and pathogenic relationships with hosts. Biosci Biotechnol Biochem. 2018 Mar 4;82(3):363–71.

11. Mukherjee S, Bassler BL. Bacterial quorum sensing in complex and dynamically changing environments. Nat Rev Microbiol. 2019 Jun 3;17(6):371–82.

12. Asfahl KL, Schuster M. Social interactions in bacterial cell–cell signaling. Gibbs K, editor. FEMS Microbiol Rev. 2017 Jan;41(1):92–107.

13. Silpe JE, Bassler BL. A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. Cell. 2019 Jan 10;176(1–2):268–280.e13.

14. van de Water JAJM, Allemand D, Ferrier-Pagès C. Host-microbe interactions in octocoral holobionts - recent advances and perspectives. Microbiome. 2018 Dec 2;6(1):64.

15. Verma SC, Miyashiro T. Quorum sensing in the squid-Vibrio symbiosis. Int J Mol Sci. 2013 Aug 7;14(8):16386–401.

16. Pérez-Montaño F, Jiménez-Guerrero I, Del Cerro P, Baena-Ropero I, López-Baena FJ, Ollero FJ, et al. The symbiotic biofilm of Sinorhizobium fredii SMH12, necessary for successful colonization and symbiosis of Glycine max cv Osumi, is regulated by Quorum Sensing systems and inducing flavonoids via NodD1. PLoS One. 2014;9(8):e105901.

17. Smith RS, Iglewski BH. P. aeruginosa quorum-sensing systems and virulence. Curr Opin Microbiol. 2003 Feb;6(1):56–60.

18. Jung SA, Chapman CA, Ng W-L. Quadruple quorum-sensing inputs control Vibrio cholerae virulence and maintain system robustness. PLoS Pathog. 2015 Apr;11(4):e1004837.

19. Liu B, Faller LL, Klitgord N, Mazumdar V, Ghodsi M, Sommer DD, et al. Deep Sequencing of the Oral Microbiome Reveals Signatures of Periodontal Disease. Highlander SK, editor. PLoS One. 2012 Jun 4;7(6):e37919.

20. Rylev M, Bek-Thomsen M, Reinholdt J, Ennibi O-K, Kilian M. Microbiological and immunological characteristics of young Moroccan patients with aggressive periodontitis with and without detectable Aggregatibacter actinomycetemcomitans JP2 infection. Mol Oral Microbiol. 2011 Feb;26(1):35–51.

21. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016 Jan 4;44(D1):D7–19.

22. Rajput A, Kaur K, Kumar M. SigMol: repertoire of quorum sensing signaling molecules in prokaryotes. Nucleic Acids Res. 2016 Jan 4;44(D1):D634–9.

23. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan 17;12(1):59–60.

24. Pathmanathan JS, Lopez P, Lapointe F-J, Bapteste E. CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection. Mol Biol Evol. 2018;35(1):252–5.

25. Lannes R, Olsson-Francis K, Lopez P, Bapteste E. Carbon Fixation by Marine

Ultrasmall Prokaryotes. Genome Biol Evol. 2019 Apr 1;11(4):1166–77.

26. Jaffe AL, Corel E, Pathmanathan JS, Lopez P, Bapteste E. Bipartite graph analyses reveal interdomain LGT involving ultrasmall prokaryotes and their divergent, membrane-related proteins. Environ Microbiol. 2016 Dec;18(12):5072–81.

27. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

28. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009 Aug 1;25(15):1972–3.

29. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Comput Biol. 2011 Oct 20;7(10):e1002195.

30. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010 Mar 1;26(5):680–2.

31. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009 May 1;25(9):1189–91.

32. Korber-Irrgang B. HIV Signature and Sequence Variation Analysis. In: Computational Analysis of HIV Molecular Sequences. 2000. p. 55–72.

33. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. 2015 Jan 1;32(1):268–74.

34. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018 Feb 1;35(2):518–22.

35. Papenfort K, Bassler BL. Quorum sensing signal-response systems in Gram-negative bacteria. Nat Rev Microbiol. 2016;14(9):576–88.

36. Khan F, Javaid A, Kim Y-M. Functional Diversity of Quorum Sensing Receptors in Pathogenic Bacteria: Interspecies, Intraspecies and Interkingdom Level. Curr Drug Targets. 2019 Mar 29;20(6):655–67.

37. Yang, Bielawski. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000 Dec 1;15(12):496–503.

38. Kaplan HB, Plamann L. A *Myxococcus xanthus* cell density-sensing system required for multicellular development. FEMS Microbiol Lett. 1996 Jun 1;139(2–3):89–95.

39. Ulrich LE, Koonin E V, Zhulin IB. One-component systems dominate signal transduction in prokaryotes. Trends Microbiol. 2005 Feb;13(2):52–6.

40. Wuichet K, Cantwell BJ, Zhulin IB. Evolution and phyletic distribution of two-component signal transduction systems. Curr Opin Microbiol. 2010 Apr;13(2):219–25.

41. Brameyer S, Kresovic D, Bode HB, Heermann R. LuxR solos in Photorhabdus species. Front Cell Infect Microbiol. 2014 Nov 18;4:166.

42. Covaceuszach S, Degrassi G, Venturi V, Lamba D. Structural insights into a novel interkingdom signaling circuit by cartography of the ligand-binding sites of the homologous quorum sensing LuxR-family. Int J Mol Sci. 2013 Oct 15;14(10):20578–

96.

43. Hudaiberdiev S, Choudhary KS, Vera Alvarez R, Gelencsér Z, Ligeti B, Lamba D, et al. Census of solo LuxR genes in prokaryotic genomes. Front Cell Infect Microbiol. 2015 Mar 12;5:20.

44. González JF, Venturi V. A novel widespread interkingdom signaling circuit. Trends Plant Sci. 2013 Mar;18(3):167–74.

45. Churchill MEA, Chen L. Structural basis of acyl-homoserine lactone-dependent signaling. Chem Rev. 2011 Jan 12;111(1):68–85.

46. Zhang G, Zhang F, Ding G, Li J, Guo X, Zhu J, et al. Acyl homoserine lactone-based quorum sensing in a methanogenic archaeon. ISME J. 2012 Jul;6(7):1336–44.

47. Probst AJ, Ladd B, Jarett JK, Geller-McGrath DE, Sieber CMK, Emerson JB, et al. Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. Nat Microbiol. 2018 Mar 29;3(3):328–36.

48. Moreira CG, Sperandio V. The Epinephrine/Norepinephrine/Autoinducer-3 Interkingdom Signaling System in Escherichia coli O157:H7. Adv Exp Med Biol. 2016;874:247–61.

49. Weigel WA, Demuth DR. QseBC, a two-component bacterial adrenergic receptor and global regulator of virulence in Enterobacteriaceae and Pasteurellaceae. Mol Oral Microbiol. 2016;31(5):379–97.

50. Janakidevi K, Dewey VC, Kidder GW. The biosynthesis of catecholamines in two genera of protozoa. J Biol Chem. 1966 Jun 10;241(11):2576–8.

51. Ud-Daula A, Pfister G, Schramm K-W. Growth inhibition and biodegradation of catecholamines in the ciliated protozoan *Tetrahymena pyriformis*. J Environ Sci Heal Part A. 2008 Nov 5;43(14):1610–7.

52. Karavolos MH, Winzer K, Williams P, Khan CMA. Pathogen espionage: multiple bacterial adrenergic sensors eavesdrop on host communication systems. Mol Microbiol. 2013 Feb 1;87(3):455–65.

53. Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, et al. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol. 2012;13(6):R42.

54. Espinoza JL, Harkins DM, Torralba M, Gomez A, Highlander SK, Jones MB, et al. Supragingival Plaque Microbiome Ecology and Functional Potential in the Context of Health and Disease. MBio. 2018 Dec 27;9(6):e01631-18.

55. Baker JL, Bor B, Agnello M, Shi W, He X. Ecology of the Oral Microbiome: Beyond Bacteria. Trends Microbiol. 2017 May 1;25(5):362–74.

56. Bik EM, Costello EK, Switzer AD, Callahan BJ, Holmes SP, Wells RS, et al. Marine mammals harbor unique microbiotas shaped by and yet distinct from the sea. Nat Commun. 2016 Apr 3;7(1):10516.

57. Hochstrasser R, Hilbi H. Intra-Species and Inter-Kingdom Signaling of Legionella pneumophila. Front Microbiol. 2017;8:79.

58. Spirig T, Tiaden A, Kiefer P, Buchrieser C, Vorholt JA, Hilbi H. The Legionella

autoinducer synthase LqsA produces an alpha-hydroxyketone signaling molecule. J Biol Chem. 2008 Jun 27;283(26):18113–23.

59. Uchiyama T. Tropism and Pathogenicity of Rickettsiae. Front Microbiol. 2012 Jun 25;3:230.

60. Eisenreich W, Heuner K. The life stage-specific pathometabolism of *Legionella pneumophila*. FEBS Lett. 2016 Nov 1;590(21):3868–86.

61. Cirillo SL, Bermudez LE, El-Etr SH, Duhamel GE, Cirillo JD. Legionella pneumophila entry gene rtxA is involved in virulence. Infect Immun. 2001 Jan;69(1):508–17.

62. Silpe JE, Bassler BL. A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. Cell. 2019 Jan 10;176(1–2):268–280.e13.

63. Deeg CM, Zimmer MM, George EE, Husnik F, Keeling PJ, Suttle CA. Chromulinavorax destructans, a pathogen of microzooplankton that provides a window into the enigmatic candidate phylum Dependentiae. McGraw EA, editor. PLOS Pathog. 2019 May 31;15(5):e1007801.

64. Geske GD, O'Neill JC, Blackwell HE. Expanding dialogues: from natural autoinducers to non-natural analogues that modulate quorum sensing in Gram-negative bacteria. Chem Soc Rev. 2008 Jun 23;37(7):1432.

FIGURE LEGENDS

Figure 1: **Summary of the *in silico* analysis.**

The top section explains the design and the rationale of the study, namely the identification of CPR/DPANN homologs (target) of proteins constitutive of reference QS systems (query). Subsequently, the closest protein sequence to each CPR/DPANN homolog amongst the available complete proteomes of prokaryotes is identified (functional consistency). Eventually, the functional annotations of each trio of proteins (reference, CPR/DPANN homolog, closest prokaryotic sequence) are crosschecked to assess the reliability of each CPR/DPANN putative QS proteins. The bottom section shows hypotheses about the putative roles in inter-species and inter-kingdom communication of chosen CPR/DPANN homologs, based on the characterized functions of their reference proteins. Color code of cells: blue = bacteria, red = archaea, green = eukaryotes; Scale of cells: small cells = CPR/DPANN, normal cells = other unicellulars; Color code of synthesized signals: orange = bacterial signal, pink = eukaryotic signal; Abbreviations of QS signals: AI-2 = autoinducer-2, DPO = 3,5-dimethylpyrazin-2-ol, DSF = diffusible small factor, AI-3 = autoinducer-3, LAI-1 = legionella autoinducer-1.

Figure 2: **Distribution and completeness of QS systems across CPR and DPANN phyla**

Heatmap of completeness of the 21 QS systems detected in CPR and DPANN phyla out of the 37 tested. Each column represents a phylum, and the histogram on the top displays the number of species per phylum within which at least one homolog of a reference QS-related protein has been identified. Rows represent reference proteins and are grouped by QS systems, which are labelled, on the left, according to the name of their associated QS signal. The label on the right of the first row of each QS system indicates the species within which the system has been characterized. The symbol adjacent to the name of each reference protein indicates whether it is a QS synthase, receptor or response regulator. Rows are duplicated if the protein is both a receptor and a response regulator (one-component

system). The background in grayscale at each intersection of the heatmap indicates the number of homologs of a reference protein detected in a CPR or DPANN phylum, normalized by the number of species in the phylum. The color circle in the foreground gives the percentage of sequence identity between a reference protein and the best of its detected homologs in a CPR or DPANN phylum. Black rectangles highlight phyla within which at least one species harbors a complete QS system. The plot on the right panel of the heatmap displays the dN/dS ratio of each protein family, computed from the totality of the CPR and DPANN homologous sequences of a reference QS component.

Figure 3: **AI-2 biosynthetic pathway in Gracilibacteria from Crystal Geyser, Utah**

A. Autoinducer AI-2 biosynthesis in *Vibrio cholerae*, involving two enzymes: Pfs and LuxS. The last step of the reaction, namely the conversion of the (4S)-4,5-dihydroxy-2,3-pentanedione (DPD) to AI-2, is spontaneous and requires a borate anion. **B.** Undirect and direct homology detection of *V. cholerae* Pfs and LuxS proteins in Gracilibacteria from Crystal Geyser. The dotted edge connecting two proteins is labelled according to the percentage of sequence identity between them, over more than 80% mutal coverage. Proteins are positionned according to the location of their coding sequences in their respective genomes. **C.** Midpoint-rooted phylogenetic tree of LuxS. Branch length scale bar is displayed on the left of the tree. A black dot marks branches with bootstrap values > 95%. Leaves are labelled according to the NCBI identifier of the LuxS proteins. Branches are colored according to the taxonomic classification of the species to which the LuxS proteins belong to. Specifically, Gracilibacteria's LuxS homologs are colored in red and are included in a homogeneous group with the LuxS protein of *Lachnoclostridium phytofermentans* (protein id WP_012198920.1). The point of this tree was not to reconstruct the evolutionary history of LuxS but just to see whether or not Gracilibacteria's LuxS homologs were included in a homogeneous clan with other proteins.

Figure 4: **Distribution of effectors of *Rickettsia* and *Legionella* parasitic lifestyle across CPR and DPANN phyla**

Heatmap of completeness of the lists of effectors involved in the different phases of the lifestyle of *Rickettsia prowazeckii* and *Legionella pneumophila* in CPR and DPANN phyla. These lists are gathered under three main groups: the replicative phase (host entry and host adherence), the transmissive phase (host exit) and the biphasic switch (decision making about whether to remain in or exit the host). Each row represents a reference protein in *R. prowazeckii* or *L. pneumophila* and the color code adjacent to the name of each reference protein indicates its species origin. Each column represents a CPR or DPANN phylum and the histogram on the top displays the number of species per phylum within which at least one homolog of a reference protein has been identified. The background in grayscale at each intersection of the heatmap indicates the number of homologs of a reference protein detected in a CPR or DPANN phylum, normalized by the number of species in the phylum. The color circle in the foreground gives the percentage of sequence identity between a reference protein and the best of its detected homologs in a CPR or DPANN phylum.

Figure 1

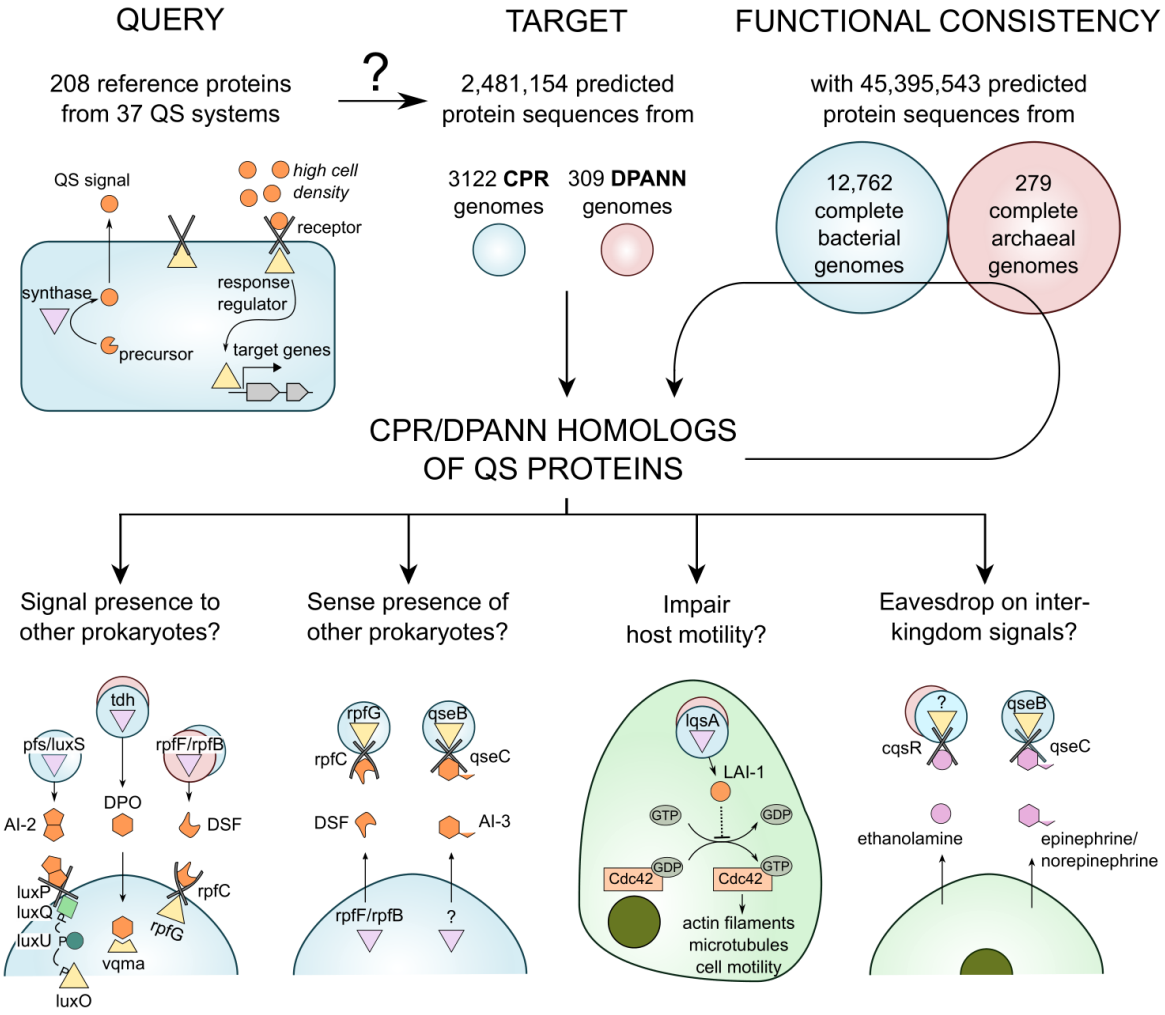

QUERY
TARGET
FUNCTIONAL CONSISTENCY
208 reference proteins from 37 QS systems
?
2,481,154 predicted protein sequences from
3122 CPR genomes
309 DPANN genomes
with 45,395,543 predicted protein sequences from
12,762 complete bacterial genomes
279 complete archaeal genomes
QS signal
high cell density
receptor
synthase
response regulator
precursor
target genes
CPR/DPANN HOMOLOGS OF QS PROTEINS
Signal presence to other prokaryotes?
Sense presence of other prokaryotes?
Impair host motility?
Eavesdrop on inter-kingdom signals?
tdh
pfs/luxS
rpfF/rpfB
DPO
AI-2
DSF
luxP
luxQ
luxU
luxO
rpfC
rpfG
vqma
rpfG
rpfC
qseB
qseC
DSF
AI-3
rpfF/rpfB
?
lqsA
LAI-1
GTP
GDP
Cdc42
GDP
GTP
Cdc42
actin filaments
microtubules
cell motility
?
cqsR
qseB
qseC
ethanolamine
epinephrine/
norepinephrine

Figure 2

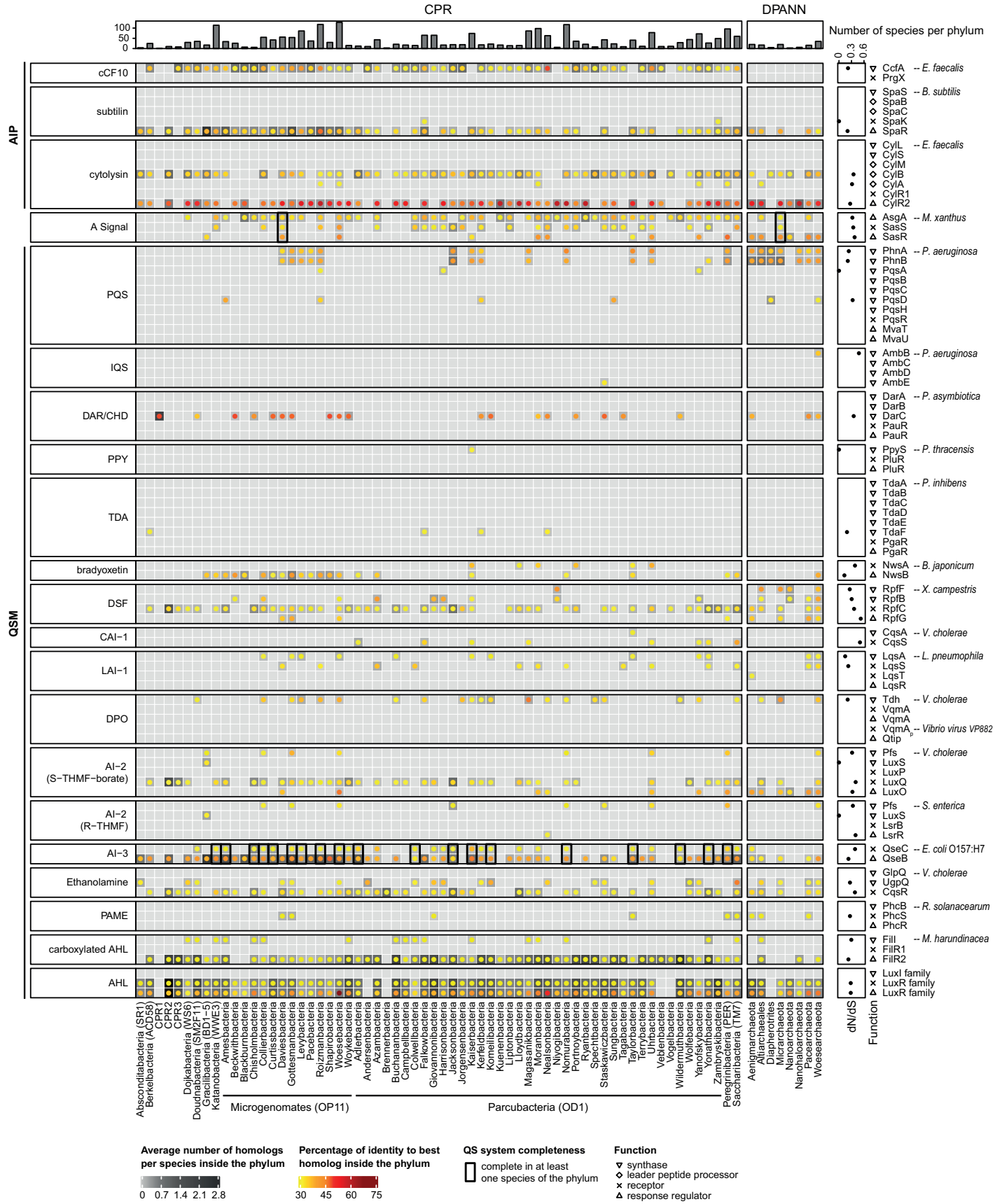

CPR
DPANN
Number of species per phylum
AIP
QSM
cCF10
subtilin
cytolysin
A Signal
PQS
IQS
DAR/CHD
PPY
TDA
bradyoxetin
DSF
CAI−1
LAI−1
DPO
AI−2 (S−THMF−borate)
AI−2 (R−THMF)
AI−3
Ethanolamine
PAME
carboxylated AHL
AHL
CcfA -- E. faecalis
PrgX
SpaS -- B. subtilis
SpaB
SpaC
SpaK
SpaR
CylL -- E. faecalis
CylS
CylM
CylB
CylA
CylR1
CylR2
AsgA -- M. xanthus
SasS
SasR
PhnA -- P. aeruginosa
PhnB
PqsA
PqsB
PqsC
PqsD
PqsH
PqsR
MvaT
MvaU
AmbB -- P. aeruginosa
AmbC
AmbD
AmbE
DarA -- P. asymbiotica
DarB
DarC
PauR
PauR
PpyS -- P. thracensis
PluR
PluR
TdaA -- P. inhibens
TdaB
TdaC
TdaD
TdaE
TdaF
PgaR
PgaR
NwsA -- B. japonicum
NwsB
RpfF -- X. campestris
RpfB
RpfC
RpfG
CqsA -- V. cholerae
CqsS
LqsA -- L. pneumophila
LqsS
LqsT
LqsR
Tdh -- V. cholerae
VqmA
VqmA
VqmA -- Vibrio virus VP882
Qtip
Pfs -- V. cholerae
LuxS
LuxP
LuxQ
LuxO
Pfs -- S. enterica
LuxS
LsrB
LsrR
QseC -- E. coli O157:H7
QseB
GlpQ -- V. cholerae
UgpQ
CqsR
PhcB -- R. solanacearum
PhcS
PhcR
FilI -- M. harundinacea
FilR1
FilR2
LuxI family
LuxR family
LuxR family
dN/dS
Function
Absconditabacteria (SR1)
Berkelbacteria (ACD58)
CPR1
CPR2
CPR3
Dojkabacteria (WS6)
Doudnabacteria (SM2F11)
Gracilibacteria (BD1−5)
Katanobacteria (WWE3)
Amesbacteria
Beckwithbacteria
Blackburnbacteria
Chisholmbacteria
Collierbacteria
Curtissbacteria
Daviesbacteria
Gottesmanbacteria
Jacobsbacteria
Peabacteria
Roizmanbacteria
Shapirobacteria
Woesebacteria
Woykebacteria
Adlerbacteria
Andersenbacteria
Azambacteria
Brennerbacteria
Buchananbacteria
Campbellbacteria
Colwellbacteria
Falkowbacteria
Giovannonibacteria
Harrisonbacteria
Jacksonbacteria
Jorgensenbacteria
Kaiserbacteria
Kerfeldbacteria
Komeilibacteria
Kuenenbacteria
Liptonbacteria
Lloydbacteria
Magasanikbacteria
Moranbacteria
Nealsonbacteria
Nyvogbacteria
Nomurabacteria
Portnoybacteria
Roizmanbacteria
Spechtbacteria
Staskawiczbacteria
Sungbacteria
Tagabacteria
Taylorbacteria
Terrybacteria
Uhrbacteria
Veblenbacteria
Vogelbacteria
Wildermuthbacteria
Wolfebacteria
Yanofskybacteria
Yonathbacteria
Zambryskibacteria
Peregrinibacteria (PER)
Saccharibacteria (TM7)
Aenigmarchaeota
Altiarchaeales
Diapherotrites
Micrarchaeota
Nanoarchaeota
Nanohaloarchaeota
Pacearchaeota
Woesearchaeota
Microgenomates (OP11)
Parcubacteria (OD1)
Average number of homologs per species inside the phylum
0 0.7 1.4 2.1 2.8
Percentage of identity to best homolog inside the phylum
30 45 60 75
QS system completeness
complete in at least one species of the phylum
Function
synthase
leader peptide processor
receptor
response regulator

Figure 3

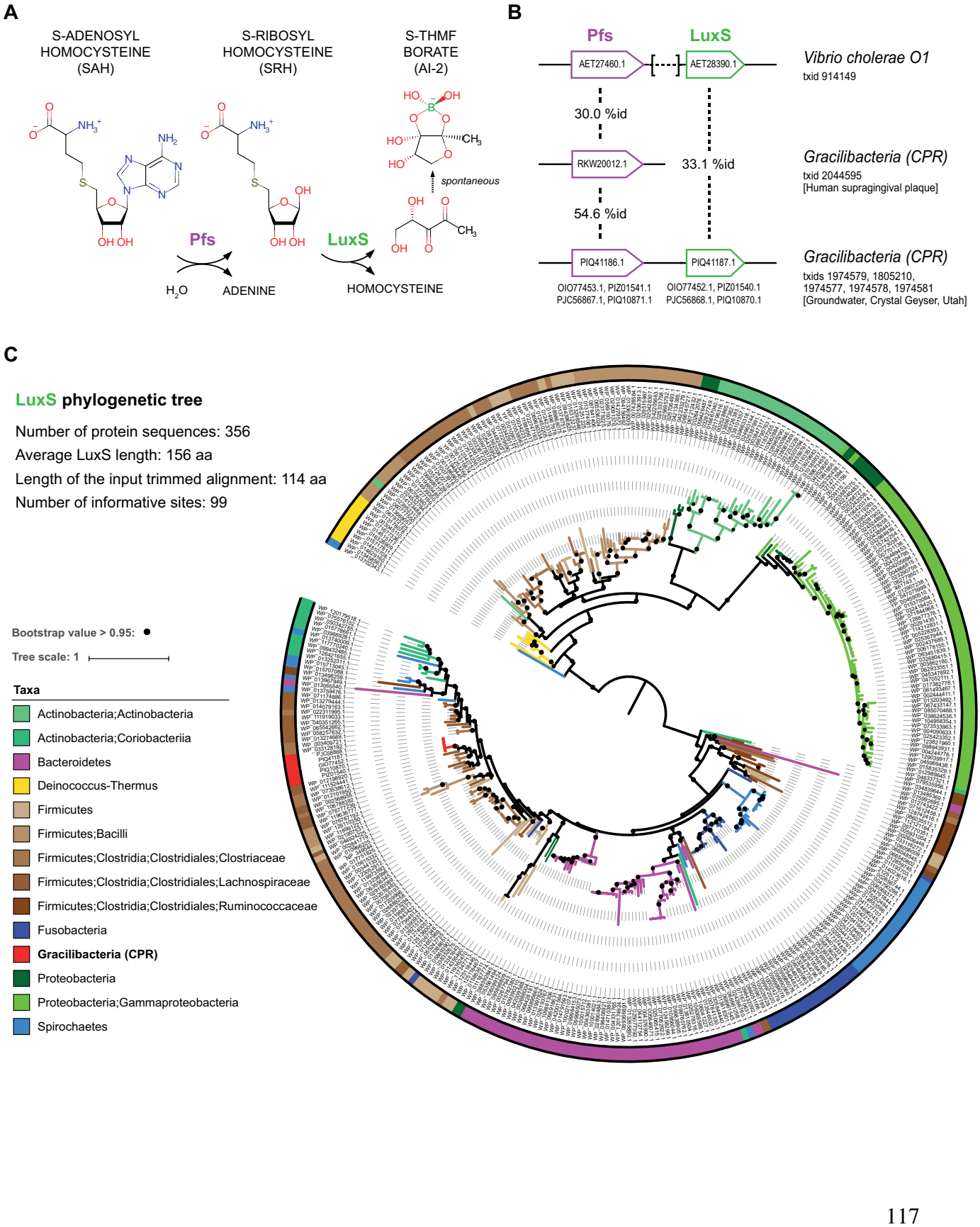

A
S-ADENOSYL HOMOCYSTEINE (SAH)
S-RIBOSYL HOMOCYSTEINE (SRH)
S-THMF BORATE (AI-2)
Pfs
LuxS
spontaneous
H2O
ADENINE
HOMOCYSTEINE
B
Pfs
LuxS
AET27460.1
AET28390.1
Vibrio cholerae O1
txid 914149
30.0 %id
RKW20012.1
33.1 %id
Gracilibacteria (CPR)
txid 2044595
[Human supragingival plaque]
54.6 %id
PIQ41186.1
PIQ41187.1
Gracilibacteria (CPR)
txids 1974579, 1805210, 1974577, 1974578, 1974581
[Groundwater, Crystal Geyser, Utah]
OIO77453.1, PIZ01541.1 PJC56867.1, PIQ10871.1
OIO77452.1, PIZ01540.1 PJC56868.1, PIQ10870.1
C
LuxS phylogenetic tree
Number of protein sequences: 356
Average LuxS length: 156 aa
Length of the input trimmed alignment: 114 aa
Number of informative sites: 99
Bootstrap value > 0.95:
Tree scale: 1
Taxa
Actinobacteria;Actinobacteria
Actinobacteria;Coriobacteriia
Bacteroidetes
Deinococcus-Thermus
Firmicutes
Firmicutes;Bacilli
Firmicutes;Clostridia;Clostridiales;Clostriaceae
Firmicutes;Clostridia;Clostridiales;Lachnospiraceae
Firmicutes;Clostridia;Clostridiales;Ruminococcaceae
Fusobacteria
Gracilibacteria (CPR)
Proteobacteria
Proteobacteria;Gammaproteobacteria
Spirochaetes

Figure 4

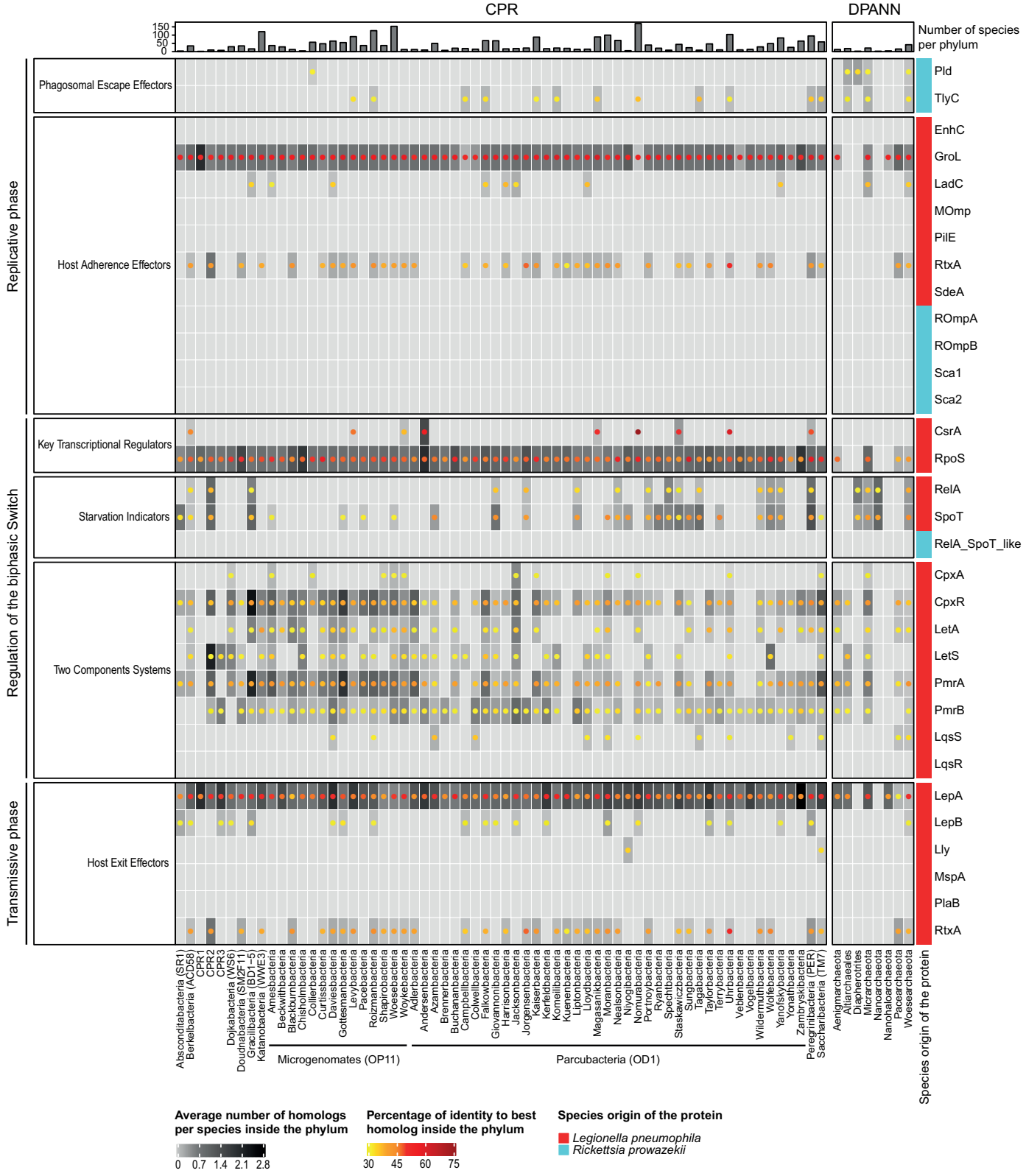

CPR
DPANN
Number of species per phylum
150
100
50
0
Replicative phase
Regulation of the biphasic Switch
Transmissive phase
Phagosomal Escape Effectors
Host Adherence Effectors
Key Transcriptional Regulators
Starvation Indicators
Two Components Systems
Host Exit Effectors
Pld
TlyC
EnhC
GroL
LadC
MOmp
PilE
RtxA
SdeA
ROmpA
ROmpB
Sca1
Sca2
CsrA
RpoS
RelA
SpoT
RelA_SpoT_like
CpxA
CpxR
LetA
LetS
PmrA
PmrB
LqsS
LqsR
LepA
LepB
Lly
MspA
PlaB
RtxA
Absconditabacteria (SR1)
Berkelbacteria (ACD58)
CPR1
CPR2
CPR3
Dojkabacteria (WS6)
Doudnabacteria (SM2F11)
Gracilibacteria (BD1−5)
Katanobacteria (WWE3)
Amesbacteria
Beckwithbacteria
Blackburnbacteria
Chisholmbacteria
Collierbacteria
Curtissbacteria
Daviesbacteria
Gottesmanbacteria
Levybacteria
Pacebacteria
Roizmanbacteria
Shapirobacteria
Woesebacteria
Woykebacteria
Adlerbacteria
Andersenbacteria
Azambacteria
Brennerbacteria
Buchananbacteria
Campbellbacteria
Colwellbacteria
Falkowbacteria
Giovannonibacteria
Harrisonbacteria
Jacksonbacteria
Jorgensenbacteria
Kaiserbacteria
Kerfeldbacteria
Komeilibacteria
Kuenenbacteria
Liptonbacteria
Lloydbacteria
Magasanikbacteria
Moranbacteria
Nealsonbacteria
Nyoguchibacteria
Nomurabacteria
Portnoybacteria
Ryanbacteria
Spechtbacteria
Staskawiczbacteria
Sungbacteria
Taylorbacteria
Terrybacteria
Ulrichbacteria
Veblenbacteria
Vogelbacteria
Wildermuthbacteria
Wolfebacteria
Yanofskybacteria
Yonathbacteria
Zambryskibacteria
Peregrinibacteria (PER)
Saccharibacteria (TM7)
Microgenomates (OP11)
Parcubacteria (OD1)
Aenigmarchaeota
Altiarchaeales
Diapherotrites
Micrarchaeota
Nanoarchaeota
Nanohaloarchaeota
Pacearchaeota
Woesearchaeota
Species origin of the protein
Average number of homologs per species inside the phylum
0 0.7 1.4 2.1 2.8
Percentage of identity to best homolog inside the phylum
30 45 60 75
Species origin of the protein
Legionella pneumophila
Rickettsia prowazekii

Table 1

Reference quorum sensing systems considered in this study

AIP or QSM	Signal family	QS system: (last synthase / receptor [species])	Reference (Pubmed id)
AIP	CSP (Competence Stimulating Peptide)	ComX/ComP [*Bacillus subtilis*] ; BlpC/BlpH, ComC/ComD [*Streptococcus pneumoniae*]	11544353, 28067778
AIP	cyclic peptide	FsrD/FsrC [*Enterococcus faecalis*] ; AgrD/AgrC [*Staphylococcus aureus*]	28467378, 11544353
AIP	eukaryotic AIP	Qsp1/? [*Cryptococcus neoformans*]	27212659
AIP	lantibiotic	SpaS/SpaK [*Bacillus subtilis*] ; CylS/CylR1 [*Enterococcus faecalis*] ; NisA/NisK [*Lactococcus lactis sp. lactis*]	15374645, 28467378
AIP	RNPP (Rap-Npr-PlcR-prgX genes)	NprX/NprR, PhrC/RapC [*Bacillus subtilis*] ; PapR/PlcR [*Bacillus thuringiensis*] ; CcfA/PrgX, PrgQ/PrgX [*Enterococcus faecalis*]	20502894
AIP	Thermogata maritima QS peptide	TM0504/? [*Thermogata maritima MSB8*]	15660994
AIP	viral AIP	AimP/AimR [*Bacillus phage phi3T*]	28099413
?	A-Signal	?/SasS [*Myxococcus xanthus*]	11544353
QSM	AHK (Alpha Hydroxy Ketone)	LqsA/LqsS [*Legionella pneumophila*] ; CqsA/CqsS [*Vibrio cholerae*]	17614967, 21219472
QSM	AI-2 (Auto-Inducer 2)	LuxS/LsrB [*Salmonella enterica* serovar Typhimirum] ; LuxS/LuxP [*Vibrio cholerae*]	19494577, 11823863
QSM	AI-3 (Auto-Inducer 3)	QseC/QseB [*Escherichia coli* O157:H7]	16803956
QSM	BL (Butyrolactone)	AfsA/ArpA [*Streptomyces griseus*]	17277085
QSM	Bradyoxetin	?/NwsA [*Bradyrhizobium japonicum*]	12393811
QSM	AHL (acyl/aryl homoserine lactones)	LuxI_family/LuxR_family (76 couples) [*Proteobacteria phylum*]	26490957, 29967162, 29977398, 24273537
QSM	carboxylated AHL	FilI/FilR [*Methanosaeta harundinacea*]	22237544
QSM	DAR/CHD (Dialkylresorcinol/Cyclohexanedione)	DarA/PauR [*Photorabdus asymbiotica subsp. asymbiotica*]	25550519
QSM	DPO (Dimethyl Pyrazinol)	Tdh/Vqma [*Vibrio cholerae*]	28319101
QSM	QSM DSF (Diffusible Signal Factor)	RpfB/RpfC [*Xanthomonas campestris*]	18049456
QSM	Ethanolamine	UgpQ/CqsR [*Vibrio cholerae*]	
QSM	Eukaryotic QS	DPP3/? [*Candida albicans*]	12954333
QSM	HAQ (Hydroxy Alkyl Quinoline)	PqsD/PqsR, PqsH/PqsR [*Pseudomonas aeruginosa*]	22390972
QSM	IQS (Hydroxyphenyl thiazole carbaldehyde)	AmbE/? [*Pseudomonas aeruginosa*]	23542643
QSM	Methyl Ester	PhcB/PhcS [*Ralstonia solanacearum*]	28642776
QSM	PPY (Photopyrone)	PpyS/PluR [*Photorabdus thracensis*]	23851573
QSM	TDA (Tropodithietic Acid)	TdaF/PgaR [*Phaeobacter inhibens DSM 17395*]	28389641

AIP=Autoinducer Peptide, QSM=Quorum Sensing Molecule. The comprehensive list of components of each QS system is given in Supplementary Table 1.

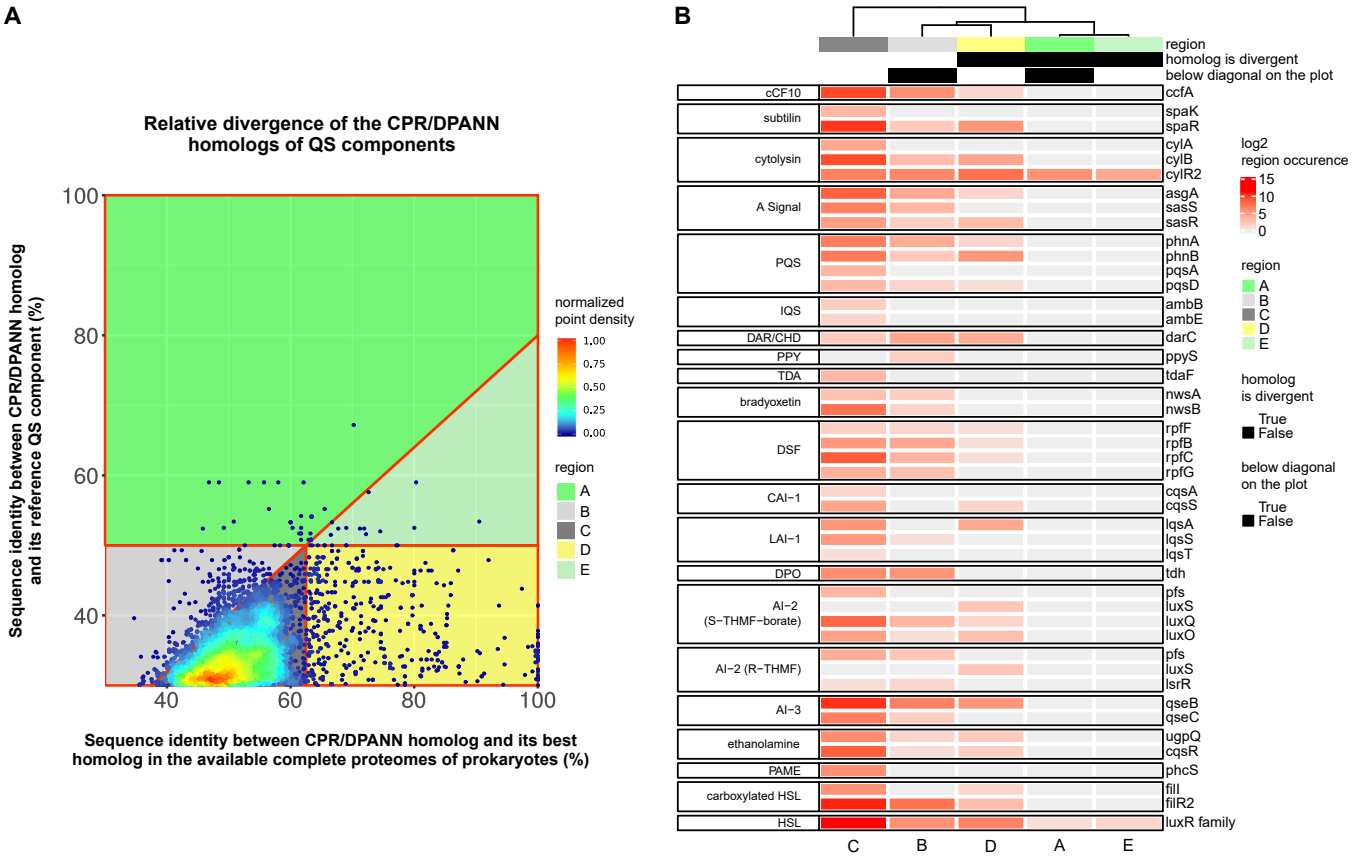
Table 2

Homologs of quorum sensing two-component systems (TCS) in CPR and DPANN phyla

reference TCS	**Signal**	**CPR and DPANN homologous TCS** (receptor_NCBI_ID/response_regulator_NCBI_ID [*CPR/DPANN phylum*])
SpaK/SpaR	subtilin	PIP33396.1/PIP34197.1, PJA09183.1/PJA09447.1 [*Falkowbacteria*] OHB14417.1/OHB14415.1, OHB16813.1/OHB16815.1 [*Zambryskibacteria*]
RpfC/RpfG	DSF	RLC29999.1/RLC32564.1 [*Woesebacteria*]
LuxQ/LuxO	AI-2	ODS40432.1/ODS41154.1, OYT53618.1/OYT53614.1 [*Altiarchaeales*]
QseC/QseB	AI-3 /epinephrine	OGC84887.1/OGC84889.1 [*Adlerbacteria*] HBC73015.1/HBC73014.1*, KKT58670.1/KKT58669.1*, KKU30987.1/KKU30986.1*, KKU67145.1/KU67144.1*, OGD05499.1/OGD05498.1* [*Amesbacteria*] OGY17956.1/OGY17957.1*, OGY19605.1/OGY19606.1*, OGY22427.1/OGY22428.1* [*Chrisholmbacteria*] KKT34505.1/KKT35709.1, KKT44870.1/KKT46785.1, KKU30217.1/KKU30365.1 [*Collierbacteria*] OGY60012.1/OGY59557.1 [*Colwellbacteria*] ; OGD93365.1/OGD93364.1*, OGD95038.1/OGD95037.1*, OGE01676.1/OGE01677.1*, OGE04685.1/OGE04684.1*, OGE08670.1/OGE08669.1*, OGE10927.1/OGE10928.1*, OGE13022.1/OGE13023.1*, OGE15756.1/OGE15757.1*, OGE16749.1/OGE16748.1*, KKR56511.1/KKR56510.1*, KKR59411.1/KKR59410.1*, KKR64463.1/KKR64462.1*, KKS02393.1/KKS02392.1* [*Curtissbacteria*] OGG08665.1/OGG08664.1*, OGG15737.1/OGG15476.1, OGG22126.1/OGG22127.1* [*Gottesmanbacteria*] HAZ16907.1/HAZ16810.1, OGY71026.1/OGY71212.1 [*Jacksonbacteria*] OGG51734.1/OGG51733.1*, OGG52928.1/OGG52929.1*, OGG65233.1/OGG66263.1, PIR84737.1/PIR84736.1* [*Kaiserbacteria*] KKS17028.1/KKS17615.1, KKS21335.1/KKS20747.1, KKS22658.1/KKS22850.1, OGC50360.1/OGC50411.1, OGC55057.1/OGC55107.1 [*Katanobacteria*] OGY90737.1/OGY90708.1 [*Komeilibacteria*] OGH11958.1/OGH11794.1, OGH38431.1/OGH38758.1 [*Levybacteria*] KKP29440.1/KKP29961.1 [*Nomurabacteria*] ALM10462.1/ALM10461.1*, ALM11565.1/ALM11564.1*, ALM12667.1/ALM12666.1*, ALM13768.1/ALM13767.1*, ALM14871.1/ALM14870.1*, HAI98808.1/HAI98809.1*, HAS34070.1/HAS34069.1*, HBH19536.1/HBH19535.1*, HBU09384.1/HBU09385.1*, OGJ61169.1/OGJ61168.1*, OGJ70245.1/OGJ70244.1*, OGJ77923.1/OGJ77924.1*, OGJ84078.1/OGJ84079.1*, OIO55862.1/OIO55861.1*, PIR53047.1/PIR53046.1* [*Peregrinibacteria*] OGK56751.1/OGK56750.1*, PIS15696.1/PIS15697.1* [*Roizmanbacteria*] OHA21004.1/OHA21006.1 [*Taylorbacteria*] OHA63067.1/OHA63545.1, OHA69181.1/OHA68948.1, OHA72907.1/OHA73420.1 [*Wildermuthbacteria*] KKP47001.1/KKP47000.1*, KKP47894.1/KKP47895.1*, KKP52047.1/KKP52048.1* [*Woesebacteria*] OHA83707.1/OHA83708.1*, OHA86265.1/OHA86266.1* [*Yonathbacteria*]

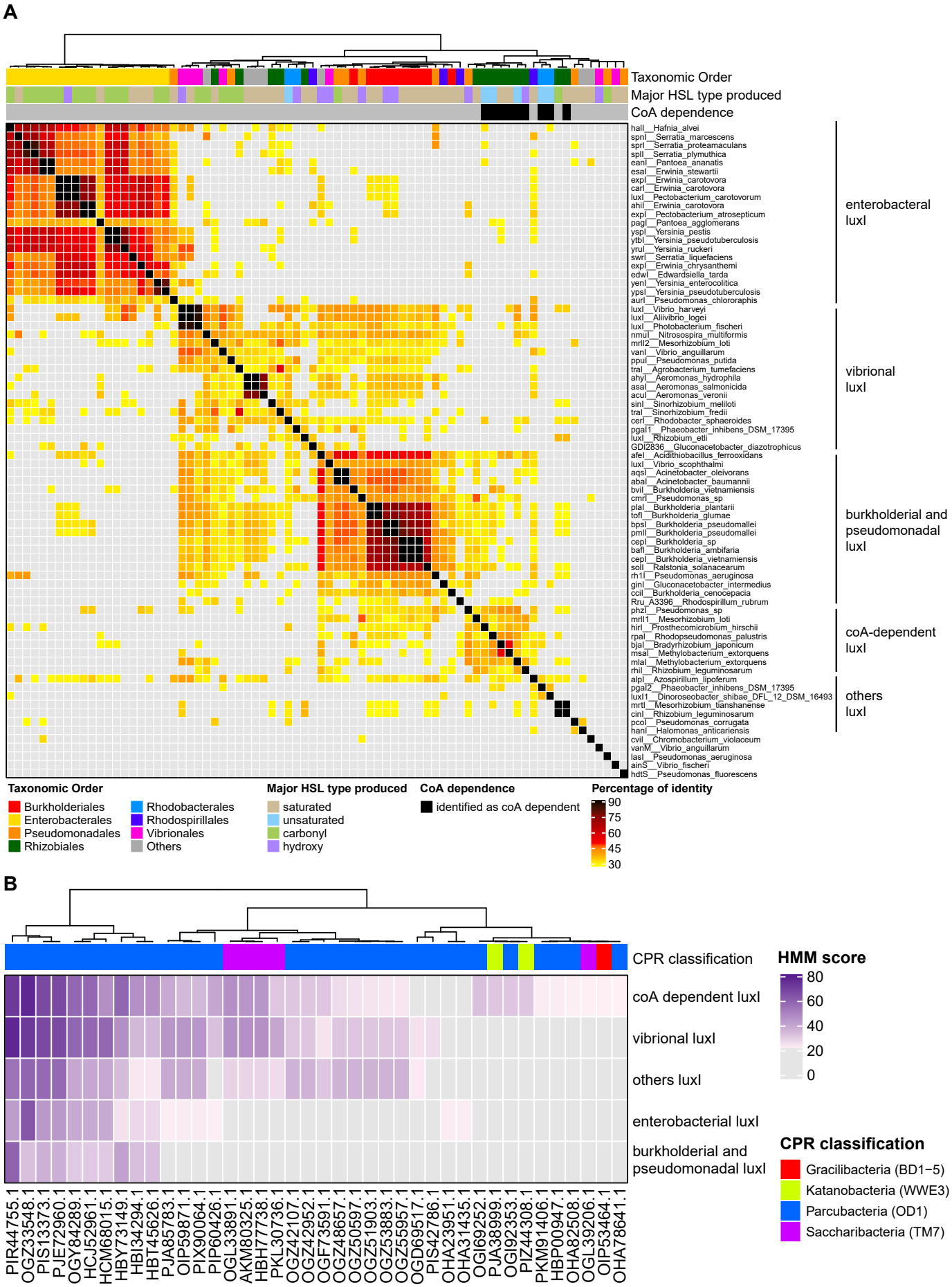
* = TCS for which the coding sequences of the receptor and the response regulator are adjacent to each other in a CPR genome

Supplementary Figure 1



Supplementary Figure 1

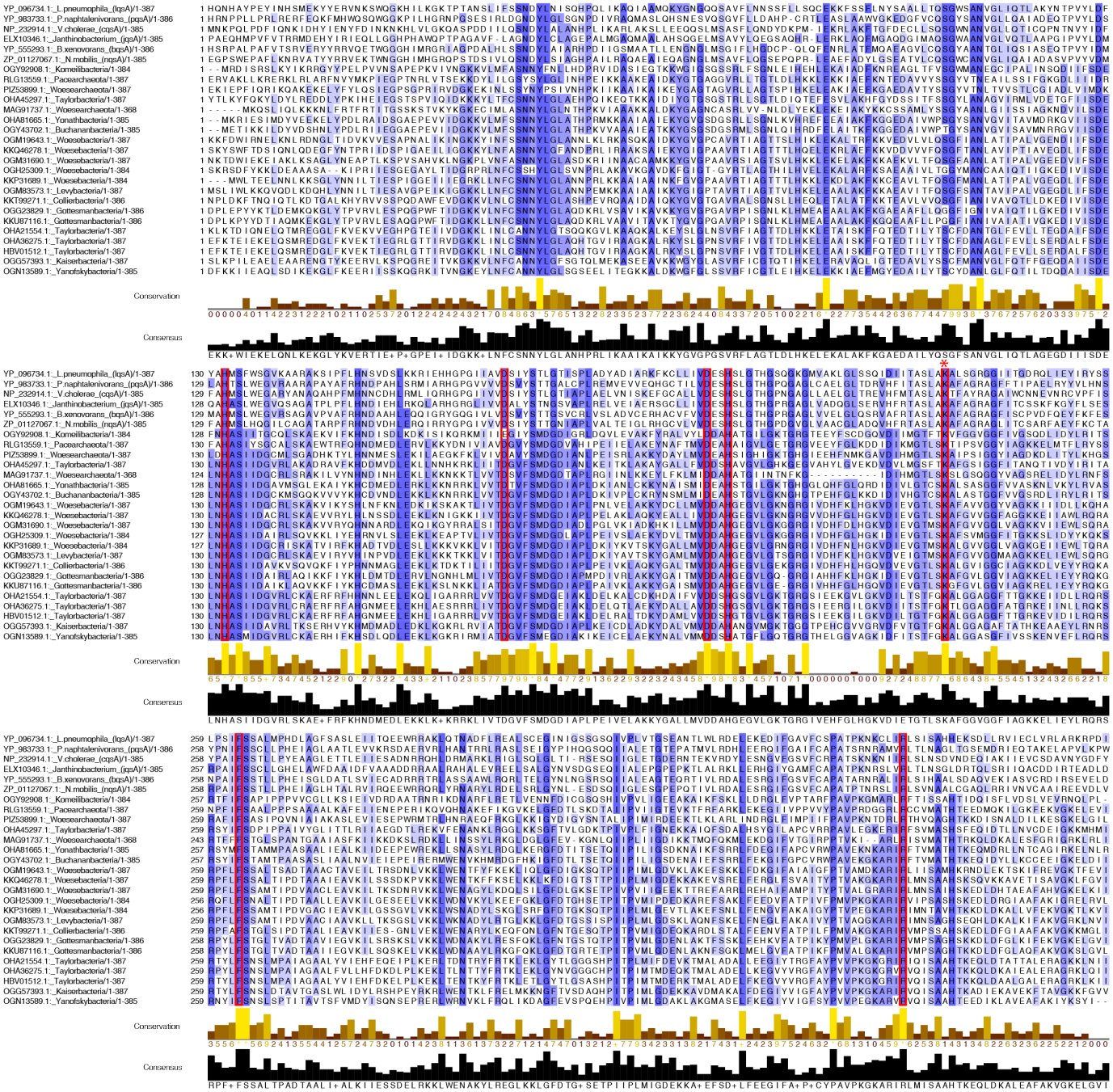
A. Scatterplot of the sequence identities of each CPR/DPANN homolog with: its corresponding reference QS-related protein (Y axis), its closest homolog identified in the predicted proteomes of fully sequenced prokaryotes (X axis). Since the majority of CPR/DPANN homologs are expected to be closer to one among all the protein sequences predicted from 12,941 complete prokaryotic genomes than to one single reference QS component, a diagonal representative of the function Y=0.8X was arbitrary drawn to visualize the few CPR/DPANN proteins whose sequence is remarkably nearly as identical to that of its reference QS protein as to that of its best homolog in the available complete proteomes of prokaryotes. An horizontal and a vertical lines intersect the diagonal at a corresponding value of 50% sequence identity of a CPR/DPANN homolog with a QS component and divide the scatterplot into 5 regions. **B**. Heatmap of the occurence of the different CPR/DPANN family of homologs into each of the 5 regions of the scatterplot. The top annotation of the heatmap tells whether a region of the scatterplot is located below or above the diagonal and whether or not the region is associated with a divergence of the sequences of CPR/DPANN homologs to the ones of well studied prokaryotes.


A
Taxonomic Order
Major HSL type produced
CoA dependence
hall__Hafnia_alvei
spnl__Serratia_marcescens
sprl__Serratia_proteamaculans
spll__Serratia_plymuthica
eanl__Pantoea_ananatis
esal__Erwinia_stewartii
expl__Erwinia_carotovora
carl__Erwinia_carotovora
luxl__Pectobacterium_carotovorum
ahil__Erwinia_carotovora
expl__Pectobacterium_atrosepticum
pagl__Pantoea_agglomerans
yspl__Yersinia_pestis
ytbl__Yersinia_pseudotuberculosis
yrul__Yersinia_ruckeri
swrl__Serratia_liquefaciens
expl__Erwinia_chrysanthemi
edwl__Edwardsiella_tarda
yenl__Yersinia_enterocolitica
ypsl__Yersinia_pseudotuberculosis
aurl__Pseudomonas_chlororaphis
luxl__Vibrio_harveyi
luxl__Aliivibrio_logei
luxl__Photobacterium_fischeri
nmul__Nitrosospira_multiformis
mrll2__Mesorhizobium_loti
vanl__Vibrio_anguillarum
ppul__Pseudomonas_putida
tral__Agrobacterium_tumefaciens
ahyl__Aeromonas_hydrophila
asal__Aeromonas_salmonicida
acul__Aeromonas_veronii
sinl__Sinorhizobium_meliloti
tral__Sinorhizobium_fredii
cerl__Rhodobacter_sphaeroides
pgal1__Phaeobacter_inhibens_DSM_17395
luxl__Rhizobium_etli
GDI2836__Gluconacetobacter_diazotrophicus
afel__Acidithiobacillus_ferrooxidans
luxl__Vibrio_scophthalmi
aqsl__Acinetobacter_oleivorans
abal__Acinetobacter_baumannii
bvil__Burkholderia_vietnamiensis
cmrl__Pseudomonas_sp
plal__Burkholderia_plantarii
tofl__Burkholderia_glumae
bpsl__Burkholderia_pseudomallei
pmll__Burkholderia_pseudomallei
cepl__Burkholderia_sp
bafl__Burkholderia_ambifaria
cepl__Burkholderia_vietnamiensis
solI__Ralstonia_solanacearum
rh1I__Pseudomonas_aeruginosa
ginl__Gluconacetobacter_intermedius
ccil__Burkholderia_cenocepacia
Rru_A3396__Rhodospirillum_rubrum
phzl__Pseudomonas_sp
mrll1__Mesorhizobium_loti
hirl__Prosthecomicrobium_hirschii
rpal__Rhodopseudomonas_palustris
bjal__Bradyrhizobium_japonicum
msal__Methylobacterium_extorquens
mlal__Methylobacterium_extorquens
rhil__Rhizobium_leguminosarum
alpl__Azospirillum_lipoferum
pgal2__Phaeobacter_inhibens_DSM_17395
luxl1__Dinoroseobacter_shibae_DFL_12_DSM_16493
mrtl__Mesorhizobium_tianshanense
cinl__Rhizobium_leguminosarum
pcol__Pseudomonas_corrugata
hanl__Halomonas_anticariensis
cvil__Chromobacterium_violaceum
vanM__Vibrio_anguillarum
lasI__Pseudomonas_aeruginosa
ainS__Vibrio_fischeri
hdtS__Pseudomonas_fluorescens
enterobacteral luxI
vibrional luxI
burkholderial and pseudomonadal luxI
coA-dependent luxI
others luxI
Taxonomic Order
Burkholderiales
Enterobacterales
Pseudomonadales
Rhizobiales
Rhodobacterales
Rhodospirillales
Vibrionales
Others
Major HSL type produced
saturated
unsaturated
carbonyl
hydroxy
CoA dependence
identified as coA dependent
Percentage of identity
90
75
60
45
30
B
CPR classification
coA dependent luxI
vibrional luxI
others luxI
enterobacterial luxI
burkholderial and pseudomonadal luxI
HMM score
80
60
40
20
0
CPR classification
Gracilibacteria (BD1−5)
Katanobacteria (WWE3)
Parcubacteria (OD1)
Saccharibacteria (TM7)
PIR44755.1
OGZ33548.1
PIS13373.1
PJE72960.1
OGY84289.1
HCJ52961.1
HCM68015.1
HBY73149.1
HBI34294.1
HBT45626.1
PJA85783.1
OIP59871.1
PIX90064.1
PIP60426.1
OGL33891.1
AKM80325.1
HBH77738.1
PKL30736.1
OGZ42107.1
OGZ42952.1
OGF73591.1
OGZ48657.1
OGZ50597.1
OGZ51903.1
OGZ53883.1
OGZ55957.1
OGD69517.1
PIS42786.1
OHA23951.1
OHA31435.1
OGI69252.1
PJA38999.1
OGI92353.1
PIZ44308.1
PKM91406.1
HBP00947.1
OHA82508.1
OGL39206.1
OIP53464.1
OHA78641.1

Supplementary Figure 2

A. Heatmap representing the mutual homology of 76 experimentally validated acyl/aryl homoserine lactone synthases retrieved from the Sigmol database. The matrix is symetrical and each row/column represents a synthase. Each row is labelled according to the name of a synthase, followed by the species within which it has been characterized. The color intensity of the heatmap represents the percentage of identity between two synthases over more than 80% mutual coverage. When the color is grey, it means that the two synthases did not pass our thresholds to assess an homology (sequence identity ≥ 30%, mutual coverage ≥ 80%, E-value < 1^{e}-5). The top annotation of the heatmap displays three information for each synthase: i) the taxonomic order of the species to which it belongs to, ii) the major chemical type of acyl homoserine lactones it produces according to Sigmol, iii) whether it is known to preferentially use coenzyme A (CoA) over acyl carrier protein (ACP) for the biosynthesis of acyl/aryl homoserine lactones. The vertical lines on the right of the heatmap delineates the 5 clusters of synthases used as starting points to build the 5 different HMM profiles of luxI (Material and methods). **B**. Heatmap of the best HMM scores for the 5 different profiles of acyl/aryl homoserine lactone synthases in CPR proteomes. Each row represents an HMM profile and each column a CPR protein. The top annotation gives the taxonomic annotation of the species to which a CPR protein belongs to. Any grey intersection in the heatmap indicates that a CPR protein did not pass the thresholds of significance (HMM score >= 20 and E-value < 0.01) for a given HMM profile.

Supplementary Figure 3



Supplementary Figure 3

Trimmed multiple sequence alignment of reference hydroxyketone synthases (6 first sequences) with each of the representative sequences of CPR/DPANN clusters of homologs (Material and methods). Each sequence is labelled on the left according to its NCBI protein identifier followed by the species or the CPR/DPANN phylum to which it belongs to. The residues are colored according to their conservation rate. The columns in red highlight the HDDHKFF residues forming the active site of the reference enzymes and the red asterix on the top of a column denotes the conserved lysine which binds the pyridoxal-5'-phosphate (PLP) substrate (Spirig, T. et *al*. *J. Biol. Chem.* **283**, 18113-23 (2008)).

3.4 Les graphes pour analyser l'évolution et la complexité microbienne

La théorie des graphes est un outil performant pour analyser des relations (arêtes) entre objets (nœuds). La théorie des graphes est donc adaptée à l'étude des relations entre séquences biologiques, qu'il s'agisse de relations de parenté ou de co-occurrence. Par exemple, un arbre phylogénétique, qui permet l'étude des relations de parenté verticales entre séquences ou organismes est un réseau dirigé acyclique avec une contrainte sur le nombre de voisins possible. Un réseau phylogénétique dirigé cyclique est une généralisation des arbres phylogénétiques qui permet d'étudier les relations de parenté verticales mais aussi horizontales. Le réseau de similarité de séquence (SSN[1]) qui est détaillé dans la publication qui suit est une autre stratégie non phylogénétique pour généraliser les approches en arbres. La théorie des graphes existe depuis 1736, initiée par les travaux de Leonhard Euler qui s'en servit pour montrer qu'il n'existe pas de chemin permettant de traverser une fois et une seule tous les ponts de la ville de Königsberg. Elle est enrichie par les travaux de nombreux mathématiciens et est aujourd'hui utilisée dans un grand nombre de disciplines théoriques mais aussi appliquées (par exemple: l'analyse des réseaux sociaux, les applications de cheminement). De nombreuses extensions et généralisations des graphes existent: réseaux de Petri, hypergraphes, graphes bipartites. De fait, de nombreux outils existent pour étudier un réseau, que ce soit l'identification de groupes (par exemple: des familles de gènes dans des SSN), la détection de structures et d'associations préférentielles ainsi que l'identification de séquences qui occupent une place particulière dans le réseau. La publication qui suit est une revue méthodologique sur la création et l'analyse de graphes basé sur la similarité de séquence pour l'étude des communautés microbiennes et des relations évolutives entre

[1] Sequence Similarity Network

séquences. Nous y détaillons la construction des réseaux sur la base de données de séquences moléculaires: réseaux de similarité de séquences, réseaux de génomes de partage de gènes, réseaux bipartites. Nous y expliquons aussi comment analyser ces graphes en vue de la détection de gènes composites, de familles de gènes, de séquences centrales et d'associations préférentielles entre les séquences de taxa de l'environnement.

3.4.1 Article 4, "The Methodology Behind Network Thinking: Graphs to Analyze Microbial Complexity and Evolution", (Watson et al. 2019)

Chapter 9

The Methodology Behind Network Thinking: Graphs to Analyze Microbial Complexity and Evolution

Andrew K. Watson, Romain Lannes, Jananan S. Pathmanathan, Raphaël Méheust, Slim Karkar, Philippe Colson, Eduardo Corel, Philippe Lopez, and Eric Bapteste

Abstract

In the post genomic era, large and complex molecular datasets from genome and metagenome sequencing projects expand the limits of what is possible for bioinformatic analyses. Network-based methods are increasingly used to complement phylogenetic analysis in studies in molecular evolution, including comparative genomics, classification, and ecological studies. Using network methods, the vertical and horizontal relationships between all genes or genomes, whether they are from cellular chromosomes or mobile genetic elements, can be explored in a single expandable graph. In recent years, development of new methods for the construction and analysis of networks has helped to broaden the availability of these approaches from programmers to a diversity of users. This chapter introduces the different kinds of networks based on sequence similarity that are already available to tackle a wide range of biological questions, including sequence similarity networks, gene-sharing networks and bipartite graphs, and a guide for their construction and analyses.

Key words Sequence similarity network, Evolution, Lateral gene transfer (LGT), Metagenomics, Gene remodeling, Ecology

1 Introduction

An evolutionary biologist is interested in how processes governing evolution have produced the diversity of genes, genomes, organisms, species, and communities that are observed today. For example, a biologist interested in the eukaryotes may wonder what symbiotic partners have contributed to their origins and evolution. Eukaryotic nuclear genomes are chimeric in nature, encoding many genes acquired from their alphaproteobacterial endosymbiont [1–3]. However, in recent years, it has been proposed that the ongoing gain of genes by both microbial [4–6] and multicellular eukaryotes [7, 8] via lateral gene transfer (LGT) has continued to contribute to eukaryotic evolution, though to a lesser extent than

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_9,

prokaryotes [9]. A biologist interested in prokaryotes may wish to investigate lateral gene transfer to explore the numbers and kinds of genes transferred between bacteria, archaea, and their mobile genetic elements [10–14]. These transfers are important for understanding the accessory genomes of prokaryotes [15–17]. Further, studying gene transfers in real bacterial communities from different environments can help to test the effect of LGT on ecology and evolution of communities [18]. Given the prevalence of introgression [9–11, 19], one interesting question is whether gene transfer has led to the formation of novel fusion genes that combine parts of genes originating from separate domains of life [20]. An ecologist may wish to analyze the distribution of genes and species in the environment [21]. A metagenome analyst may need to overcome an additional challenge exploring the nature of the large proportion of sequences in metagenome datasets that have little or no detectable similarity to characterize sequences and to study the "microbial dark matter" [22].

High-throughput sequencing technologies present new opportunities to investigate these diverse kinds of questions with molecular data; however, they also present challenges in terms of the scale of the analyses. Consequently, a number of network-based methods have recently been developed to expand the toolkit available to molecular biologists [23], and these have already made major contributions to our understanding of molecular evolution. Networks have been used to shed light on the nature of the "microbial dark matter" [24] and used in ecological studies to explore the geographical distribution of organisms or genes [25, 26] or the evolution of different lifestyles [27]. Their suitability for investigating introgressive events has been used to enhance our understanding of the chimeric origin of genes in the eukaryotic proteome [28, 29], the flow of genes between prokaryotes and their mobile genetic elements [30–35], and gene sharing across mobile elements to study the transfer of resistance factors [14, 36]. Networks have also been used to classify highly mosaic viral genomes [37, 38] and identify gene families [39, 40]. These approaches are highly complementary to traditional phylogenetic approaches, highlighted by the development of hybrid approaches and phylogenetic and phylogenomic networks [34, 41–43]. These hybrid networks are beyond the scope of discussion in this chapter but are covered in Chapters 7 and 8.

While the generation and analysis of networks were previously limited to biologists with programming experience, tools have recently been developed to simplify the process and broaden the availability of network analyses of molecular sequence data. This chapter introduces the different kinds of networks that are already available to biologists and a guide to how these networks can be constructed and analyzed for a large range of applications in molecular evolution. More precisely, this chapter will focus on three kinds

of network and the types of analyses that are possible using these networks: sequence similarity networks, gene-sharing networks, and multipartite graphs [23].

2 Sequence Similarity Networks (SSNs)

Sequence similarity networks are the bread and butter of network-based molecular sequence analyses, with a huge range of applications in molecular biology. The use of SSNs for molecular sequence analysis first came to the fore in the late 1990s and early 2000s, when SSNs were suggested as a way to analyze the rapid influx of new molecular sequence data due to advances in sequencing technology and reduced cost, as well as to predict gene functions and protein-protein interactions [39, 44–46]. One of the earliest formal and heuristic uses of SSNs was to define the COG groups of homologous families and facilitate prediction of the functions of large numbers of genes based on homology [39, 40]. The need for efficient computation and analyses for large biological databases still pervades; however, more recently SSNs have been increasingly appreciated as useful approaches to describe complex biological systems, including inferring the "social networks" of biological life forms [30], producing maps of genetic diversity [27], detecting distant homologues [47–49], and exploring gene and genome rearrangements [50, 51].

A SSN is a graph in which each node is a sequence and edges connect any two nodes that are similar at the sequence level above a certain threshold (e.g., coverage, percent identity, and *E*-value) as determined by their pairwise alignment (Box 1) (Fig. 1). While the principle behind SSN construction is simple, the expression of similarity data in this structure can enable the use of powerful

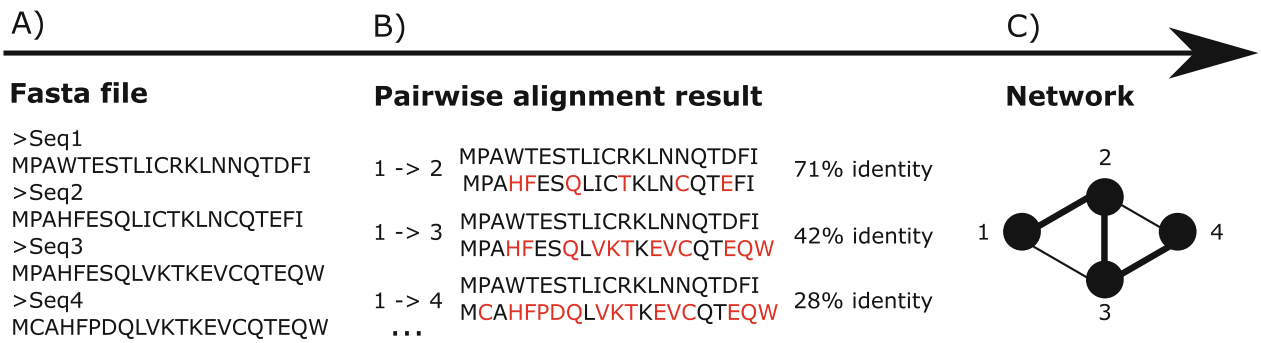


Fig. 1 Constructing a simple sequence similarity network. A set of sequences (protein or DNA) in fasta format (**a**) are aligned in pairs using alignment tools (such as BLAST). These alignments (**b**) are scored with metrics such as the percentage identity between two sequences (the number of identical nucleotides/amino acids displayed above) or the *E*-value of the alignment. In the resulting network (**c**), sequences are represented as nodes. Two sequence nodes are joined with an edge if they can be aligned above a define threshold, with the weight of the edge often based on percentage identity or *E*-value

algorithms for graph analyses to study complex biological phenomena. Construction of a SSN is also frequently the starting point in a diversity of further graph analyses. A SSN can be constructed directly from fasta formatted sequence files using pipelines, such as EGN [52], the updated and faster performing EGN2 (forthcoming), or PANADA [53]. Visualization of networks can be performed with programs such as Cytoscape [54] or Gephi [55], both of which also have a range of internal tools and external plugins for network analysis. While these programs are useful for the visualization and analysis of relatively small networks, it can be difficult to load large and complex networks with a lot of edges (e.g., ≥50,000 edges). In these cases the iGraph library offers an extremely powerful and well-supported implementation of a broad range of commonly used methods for both complex graph generation and analysis in R, Python, and C++ [56]. However, using iGraph requires knowledge of programming in at least one of these languages. An additional package for network analysis in Python is NetworkX [57]. It is our goal here to further generalize network approaches by explaining how evolutionary biologists with less programming knowledge could analyze their data. A list including many of the tools and programs available for SSN generation is available at https://omictools.com.

Box 1: How to Build Your Own Sequence Similarity Network

1. *Dataset assembly*: The first and most important step of SSN construction is the assembly of a dataset of sequences relevant to your biological question, usually in fasta format. This can be used as the initial input for wizards such as EGN or EGN2 [52], which can fully automate the process. The nature of the dataset is highly dependent on the research question, so here we focus on the practicalities of database assembly. To construct the similarity network, all sequences in the dataset are aligned against one another in a similarity search. This similarity search is often the time-limiting step in an analysis, and the total number of searches required is quadratic to the number of sequences in the dataset. For large datasets, it is useful to benchmark the alignment using a subset of the data to estimate the timescale for the alignment. Large datasets can generate huge outputs, not only due to the number of sequences but also the length of their identifier. One way to reduce the output size is to replace each sequence name in the fasta file with a unique integer. The use of integers will reduce disk space use and the memory consumption for any software used to analyze the sequence data.

(continued)

Box 1: (continued)

2. *Similarity search*: To generate a sequence similarity network, all sequences must be aligned against one another in an all-versus-all search, in which the dataset of sequences is searched against a database including the same sequences. For gene networks, the alignment is usually done with a fast pairwise aligner such as BLAST [58, 59] as implemented in EGN [52]. Filters are often used to remove low-complexity sequences from the search, as these can cause artefactual hits (BLAST options --seg yes, -soft-masking true). The BLAST method of alignment will be the focus of future discussion in this chapter; however, alternatives are available including BLAT [60] (also implemented in EGN), SWORD [61], USEARCH [62], and DIAMOND [63]. These alternatives generally include an option to produce a "BLAST" style tabulated output, making them compatible with programs commonly used in network analyses.

   Within alignment tools like BLAST, it is possible to assign thresholds, such as the maximum *E*-value of the alignment. It is not recommended to set minimal thresholds for some parameters (such as % sequence identity) unless required due to memory constraints so that you can generate networks from a single sequence alignment with different thresholds for comparison (e.g., comparison of a 30% similarity threshold to a 90% threshold, where edges will only be drawn between highly similar genes).

   Note: It may be intuitive to use additional CPUs to speed up the alignment process; however, in BLAST it can be more efficient to split the query file and launch multiple searches on separate cores instead of using the BLAST multithreading option. The pairwise alignment step is generally the most time-limiting part of generating a SSN, so benchmarking should be used to establish the optimal settings for the pairwise and/or determine the feasibility of a project given the size of the dataset and the available computational resources.

3. *Filtering similarity search results*: In an all-versus-all similarity search, any given query sequence will have a self-hit in the corresponding database. For example, with sequences A and B, a self-hit is query sequence A matching to sequence A in the database, cases of which must be removed prior to network construction (Fig. 2). When query sequence A in a similarity search is aligned with sequence B in the database, often the reciprocal result is also identified (an alignment between query sequence B and sequence A in the database). These are called reciprocal hits; while the sequences involved

(continued)

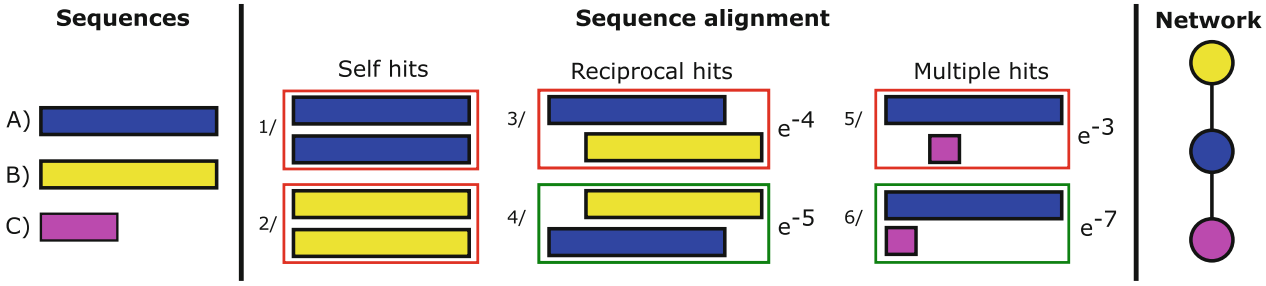


Fig. 2 Filtering sequence similarity results for network construction. In the output of an all-against-all sequence similarity search, there are a number of features that are often filtered out prior to network construction. Self-hits (1/ and 2/), where like sequences are paired in a sequence alignment, are not informative to network construction and are removed (highlighted by the red box surrounding the alignments). In cases where there are reciprocal hits (3/ and 4/) between two sequences, then only the alignment with the highest *E*-value is retained (highlighted with a green box around the retained alignment) to ensure only one edge representing the best possible alignment connects any two nodes in the network. The same is true for cases where a sequence has multiple hits against another sequence, such as when it aligns to another sequence in multiple positions (5/ and 6/)

Box 1: (continued)

are identical, the alignments and scores are not. Retaining both hits would generate two different edges between the same two nodes in a SSN, so generally only the best results from reciprocal hits are retained, based on a score such as the *E*-value (Fig. 2). Finally, a single query sequence may be significantly aligned multiple times in different positions of the same sequence in the database; however, for SSN construction only the best BLAST hit is generally retained (Fig. 2). The selection of the best BLAST hit is again generally often based on the *E*-value. Removing multiple hits against the same sequence allows the generation of an undirected network where a single edge connects two nodes, representing the best possible alignment between these nodes.

4. *Thresholding and network construction*: Constructing a SSN from a BLAST output is conceptually simple; an edge is created between two sequences (nodes) that have been aligned in the sequence similarity search. It is common to apply thresholding criteria such as minimal % ID and/or coverage and/or maximal *E*-value to determine whether an edge is drawn between two sequences in the network (Fig. 1). There are different ways to calculate the % coverage of an alignment. This could be based on the coverage of a single sequence in the alignment, selecting either the query or the database sequence in each alignment or the longest or shortest sequence in each alignment. Alternatively both (mutual coverage) can be used, retaining an alignment

(continued)

Box 1: (continued)
when both values are above a given threshold. Edges above the thresholding criteria can be assigned a weight based on these criteria, producing a weighted sequence similarity network that retains information of the properties of the alignment between two sequences (Fig. 1). It is often useful to construct and compare several SSNs with variable stringencies defining the edges between sequences, for example, to optimize gene family detection within the SSN (discussed below).

2.1 Scalability of Sequence Similarity Network Analysis

As with other computational approaches, the scale of network analysis is limited by the available computational resources. The limiting factor in terms of the size of network it is possible to construct is predominantly governed by the pairwise alignment. All sequences in the dataset need to be aligned against one another in a pairwise manner, meaning the number of alignments is quadratic to the size of the dataset. For example, computing an all-against-all comparison of 1,000,000 sequences requires computation of 10^{12} alignments. BLAST [64] is the standard tool for this step, with a relatively good speed and accuracy for sequence similarity searches; however, the use of BLAST can be a bottleneck for the analysis of large datasets. This is an especially important consideration given the growth in the number of gene and genome sequences available in public databases. Several rapid alignment tools such as BLAT [60], USEARCH [62], Rapsearch [65], and Diamond [63] have been proposed to overcome this issue. For example, Diamond benchmarks suggest that it is almost as accurate as BLAST but is at least three orders of magnitude faster.

A second point to consider from the perspective of scalability is the complexity and size of the graph and the complexity of the algorithms used in their analysis. Algorithms where the number of calculations is linear to the size of the graph can generally be run on huge graphs with sufficient computational resources, for example, finding connected components using the "deep search first" algorithm. Algorithms for community detection (e.g., PageRank [66], Louvain) are also linear and particularly suited for detecting groups of closely related sequences in huge graphs (discussed in Subheading 4). In contrast, computing graph statistics such as the betweenness centrality are not linear to the size of the graph, even using the relatively efficient Brande algorithm for calculation [67], and are therefore more difficult to calculate for huge graphs. This has led to the development of toolkits specifically designed for the analysis of huge graphs (e.g., NetworKit) [68]. A recent book summarizes the challenges of the analysis of huge networks and some of the algorithms that have been developed to face these challenges [69].

2.2 Exploiting Sequence Similarity Networks for Identification of Gene Families

A gene family is usually defined as a group of sequences that are similar at the sequence level, indicative of homology and potentially of shared functions; however, there is no uniform way to define this similarity [70, 71]. One of the early contributions of SSNs in molecular sequence analysis was the construction of the COG database of homologous protein sequences [39, 40]. This study attempted to define gene families based on similarity at the sequence level using the results of sequence similarity searches. Within the results of an all-versus-all BLAST search, groups of at least three proteins encoded by different genomes that were more similar to each other than they were to other proteins found in the same genomes were defined as a likely orthologous gene family. Orthologous gene families are group of genes in different genomes that show sequence similarity, likely as a result of their shared evolutionary history.

The idea of using graphs to identify gene families is now a core part of many graph-based analyses. Members of a gene family aggregate in a sub-network in a SSN. These sub-networks are called connected components (CCs) at these defined thresholds, i.e., clusters of nodes connected by edges either directly or indirectly (via intermediate nodes) (Fig. 3). The size (number of nodes and edges in a CC) and density (the proportion of potential connections between all nodes in a CC that are actually connected by edges in the graph) of CCs will depend on the thresholds used for

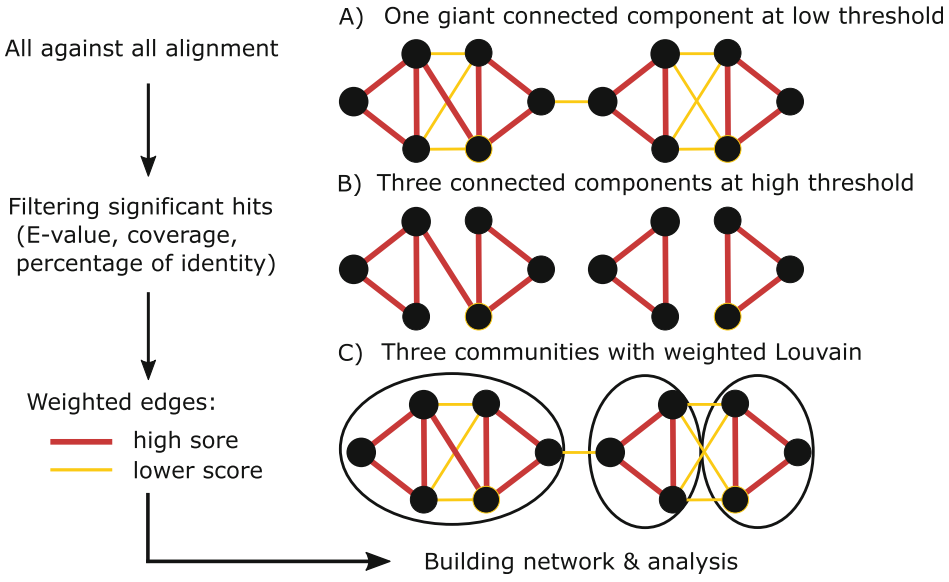


Fig. 3 Louvain community detection in a sequence similarity network. The network is assembled from the results of an all-versus-all alignment, as previously described. Edges can be weighted by *E*-value, percentage of identity, or bitscore. For the purpose of simplification, we consider strong or weak weights rather than actual values. (**a**) A giant connected component at relaxed threshold. (**b**) Three connected components at a more stringent threshold. (**c**) Three communities with Louvain clustering algorithm, taking into account edge weights

constructing the SSN as well as the relationships between sequences in the network. For example, for a given dataset at a given mutual coverage threshold, a threshold of 90% sequence identity will identify a large number of small connected components that only include highly similar genes, while at a threshold of 30% sequence identity, there will be fewer but larger connected components including genes with more variation in sequence similarity. Commonly used thresholds for detecting homologous gene families are an *E*-value $\leq e{-}5$, mutual coverage $\geq$80%, and a percentage of identity $\geq$30% [23].

CCs are often detected in a SSN using the Depth-First Search (DFS) algorithm; however, there are also other approaches for the detection of gene families based on the idea of detecting "communities" [72]. In some cases, a CC can be further separated into communities of sequences that share more similarity to one another than to other sequences in the CC and thus are more highly linked in the SSN (Fig. 3). Communities are commonly identified by using graph clustering algorithms such as Louvain [73], MCL [74], or OMA [75]; however, different clustering algorithms will result in different outputs. The Louvain weighted method is widely used because it is simple to implement and scales very well to large graphs (Figs. 3 and 4) [73]. MCL is a strong deterministic algorithm that has been implemented, for example, in tribeMCL [74] and orthoMCL [76]. A potential drawback of MCL is that it requires user specification of the "inflation index," a parameter which controls cluster granularity (or "tightness"). A high inflation

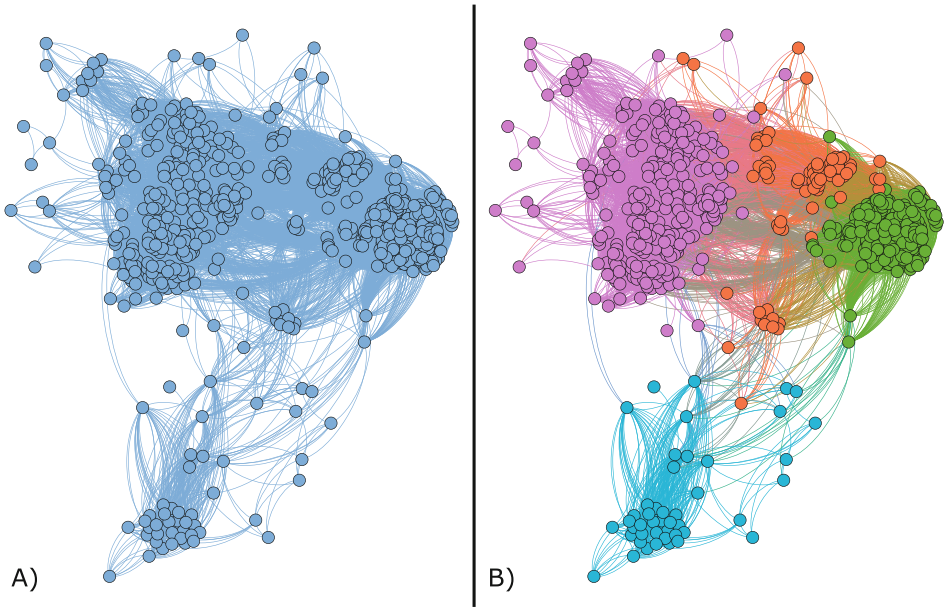


Fig. 4 Giant connected component before and after community detection. (**a**) A single giant connected component from a sequence similarity network. (**b**) The same giant connected component after application of a community detection algorithm. Node colors correspond to the newly assigned communities

index increases the tightness of clustering, producing a larger number of clusters that are smaller on average than those that would be obtained clustering the same dataset using a low inflation index. Selecting an appropriate inflation index is not trivial and requires optimization [74].

A number of the above approaches have been used to compile additional databases of orthology that can act as useful reference datasets. OMA is a program that uses graph-based algorithms and exact Smith-Waterman alignments to identify orthology between genes [77–80]. OMA is also available as a web browser [81] including a database of orthologues that, in 2015, included more than 2000 genomes and more than seven million proteins [75]. SILIX is a software package [82] that aims at building families of homologous sequences by using a transitive linkage algorithm, and HOGENOM [83] is a database that contains families inferred by SILIX for seven million proteins.

In addition to clustering genes into families, valuable information can be extracted from the connected components using network metrics. Highly conserved sequences tend to form CCs where most of the nodes are connected to each other by edges, while sequences from more divergent families will tend to form more sparsely interconnected CCs. This information can be easily assessed for each component using the clustering coefficient. Conserved families will have a clustering coefficient close to 1, even for stringent thresholds. Identifying such conserved families can be useful to produce multiple sequence alignments (MSA) needed for phylogenetic reconstruction, but SSNs have also been demonstrated to unravel relationships between distant homologues by linking distantly related sequences together [24, 29, 48]. In a SSN, two distant sequences A and C which do not share similarity according to BLAST can be linked together due to sequence B which shows similarity to both A and C.

The idea of distant homology has been particularly illuminating regarding chimeric organisms such as eukaryotes which carry homologous genes inherited from a bacterial ancestor and from an archaeal ancestor [29]. A common way to analyze sequence similarity networks is to identify certain "paths" of interest, for example, the shortest possible paths between two nodes. This notion describes the path between two nodes in a connected component that minimizes the sum of the edge weights. Alvarez-Ponce et al. used this approach to explore the topology of connected components in a SSN including the complete proteomes of 14 eukaryotes, 104 prokaryotes (including archaea and bacteria), 2389 viruses, and 1044 plasmids. Eight hundred and ninety-nine CCs contained sequences from all three domains, and of these 208 contained eukaryotic sequences that were not directly similar to one another but only linked to one another via a "eukaryote-archaea-bacteria-eukaryote" shortest path. These are putatively

distant homologues in eukaryotes that were present in both the archaeal host of the mitochondrial endosymbiont and in the alpha-proteobacterial endosymbiont, with both copies subsequently retained in eukaryotes and as such strong evidence for the chimeric origin of eukaryotes [29]. This demonstrates the utility of networks in the study of ancient evolutionary relationships including the origin of eukaryotes [28] or rooting the tree of life [84]. Simple path analysis for a network is possible using existing plug-ins within visualization tools such as Cytoscape [54] and Gephi [55].

2.3 Exploiting SSNs to Identify Signatures of "Tinkering" and Gene Fusion

When discussing identification of gene families, we have focused on networks where edges are drawn between protein sequences that show a high enough similarity across their entire length, defined by a high mutual coverage threshold (e.g., 80%). Sequence similarity can also be partial, for example, following gene remodeling or "tinkering" [85] producing new combinations of gene domains via gene fusion and fission events, or through the de novo sequence synthesis of gene extensions, adding to existing sequences. The term "Rosetta Stone sequence" was coined to define the formation of a new fusion protein in a species as the result of the fusion of two proteins that are found separate in another species, with authors originally predicting that these fusions could occur between proteins that physically interact in a common structural complex [86]. One of the earliest applications of sequence similarity searches to identify fusion proteins was an attempt to predict pairs of proteins that may physically interact in an organism based on whether they could be identified as a single "composite" fusion protein in another organism [44]. Beyond predicting protein-protein interactions, this kind of gene remodeling and recycling of existing gene parts has the potential to contribute to the expansion of functional diversity in genomes, creating new and unique combinations of domains and functions [51, 85, 87–91]. Similarity search-based screens have been implemented to identify composite genes and genome rearrangements in a range of prokaryotes [92–94], eukaryotes [87, 95–97], and viruses [98].

Early attempts to identify composite genes were based on the output of sequence similarity searches, but without formalizing the results of search methods into a graph structure. The first attempt to formalize the problem of identifying "composite" genes in networks was the "Neighborhood Correlation" approach, aiming to distinguish genuine multi-domain proteins sharing common ancestry (homologues) from novel multi-domain proteins that share domains due to insertions [99]. The later development of the FusedTriplets and MosaicFinder tools attempted to unify existing graph-based methods for detection of "composite" gene detection [50]. FusedTriplets is a graph-based implementation of the traditional gene-centered method for composite gene identification, originally introduced by Enright et al. [44], with additional cross-

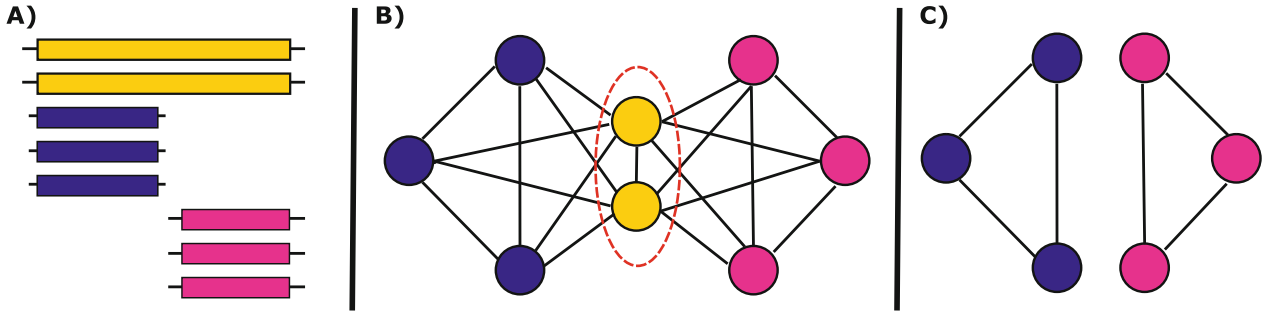


Fig. 5 Composite gene identification using "minimal clique separators." (**a**) A multiple sequence alignment of composite genes (yellow) with two components (blue and magenta). (**b**) The sequence similarity network corresponding to the multiple sequence alignment. The composite genes (yellow) are a minimal clique separator for the network. Their removal (shown in **c**) decomposes the network to the two separate component families

checks on the absence of similarity between the two component genes contributing to a composite gene based on varying thresholds [50, 100]. MosaicFinder is a gene family-centered approach which will only identify highly conserved composite gene families that form "minimal clique separators" (Fig. 5) [50]. This graph topology implies that MosaicFinder may fail to detect divergent (e.g., ancient or fast evolving) composite gene families which will tend to form "quasi-cliques" without perfect separation. CompositeSearch [101] (available at http://www.evol-net.fr/index.php/en/downloads) is a new program designed to overcome this limitation by identifying both conserved and divergent composite gene families (Box 2).

Box 2: How to Identify Composite Genes Using CompositeSearch

1. *BLAST search and filtering*: An all-versus-all BLAST search is carried out as described in Box 1. Filters can be applied on the *E*-value and sequence similarity but should not include a mutual query coverage threshold.
2. *CompositeSearch*: CompositeSearch takes a filtered BLAST output and a list of genes as the initial input. Two search algorithms are implemented: "fastcomposites" detects a list of potential composite genes and "composites" additionally detects potential composite gene families and component gene families. Additional options are included to filter the network based on a number of standard metrics (e.g., *E*-value, sequence similarity, mutual coverage) and set the maximum overlap allowed between different components aligned on the same potential composite gene. The definition of a maximum overlap allows adjustment for the

(continued)

Box 2: (continued)

tendency of BLAST to produce overhanging alignments [100]. The output includes a node, edge, and information file including information on number of nodes, edges, and family connectivity from family detection. Two outputs are included for composite gene detection, a "composites" file with detailed information on each predicted composite gene in fasta format and a "compositesinfo" file, summarizing the data. Similarly, two files provide detailed information on composite gene families and a summary of composite gene families.

3. *Filtering results*: By default, CompositeSearch outputs all possible composite genes in "fast" mode or composite gene families in the full mode. These are given alongside a number of different metrics designed to help to filter families for more confident predictions, including the gene family size, number of composites directly predicted within the gene family, the number of domains, the number of component families, the number of singleton component families (families including only one sequence), the connectivity of the family, and a score based on the overlap between different components mapped to the composite gene.

Recent studies have explored composite gene formation as a source of innovation by "tinkering" [85] during major evolutionary transitions. These can be especially interesting when exploring genome evolution following introgression, raising the possibility of formation of new composite genes using components with different evolutionary origins [20, 51, 102]. For example, the gain of a cyanobacterial endosymbiont at the origin of photosynthetic eukaryotes was accompanied by the transfer of whole cyanobacterial genes to its new host genome, with gene functions related to the role of the plastid [103–105]. Identification of composite genes related to the origin of photosynthetic eukaryotes unraveled novel symbiogenetic composite genes, and unique fusions of genes encoded in the nucleus of photosynthetic eukaryotes that included components derived from the plastid endosymbiont. As with whole genes transferred to the nucleus, several of these components had predicted functions related to the role of the plastid, including redox regulations and light response [51].

2.4 Exploiting SSNs for Ecological Studies

Ecological studies increasingly involve the assembly, analysis, and comparison of large metagenome datasets. In addition to identification of functions and organisms associated with a particular environment, these studies enable the investigation of important hypotheses in microbial ecology at the level of organism or

function, such as the often quoted hypothesis that "everything is everywhere, but the environment selects" from Bass Becking: the idea that microbial lineages are limitlessly dispersible in the environment, but the environmental conditions will select for certain lineages and control their distribution rather than any specific geographical separation [21].

Networks are useful for these kinds of ecological studies because existing graph algorithms can be used to investigate the structure of the network. When investigating gene (or gene-sharing networks), it is possible to distinguish nodes by labeling them based on their properties, such as categories for taxonomic or environmental origins (Fig. 6). A simple way to represent this visually is to color nodes based on these properties in Cytoscape or Gephi. A formal way to explore the relationships between node properties is to use network metrics such as conductance [106], modularity [73], and assortativity coefficient (normalized modularity) [107]. Assortativity and conductance are different metrics that attempt to answer the same type of question: do nodes labeled as belonging to a particular category, such as environmental origin, tend to be connected with other nodes labeled as belonging to the same category? More precisely, conductance quantifies whether a given category of nodes shares more edges between themselves than with nodes from different categories. A low conductance approaching zero indicates that nodes of a given category are highly connected to one another, with few connections to nodes from different categories. A higher conductance is indicative that nodes of this category tend to be more sparsely interconnected and share more connections with nodes from different categories. Assortativity is a measure of the preference for a category of nodes in a network to attach to other nodes

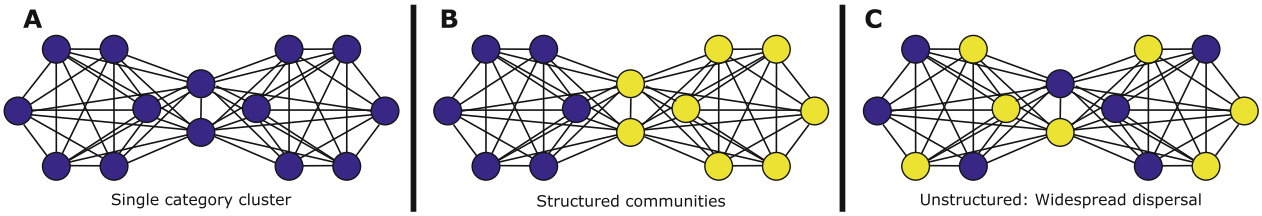


Fig. 6 Exploring distribution of annotations in sequence similarity networks. In this example, nodes within a single connected component are assigned two colors, blue and yellow, corresponding to their having a different categorical annotation (e.g., originating from a different environmental source). Using the example of environmental source, genes in cluster A would all have the same environmental source (blue), indicating an environment-specific cluster of genes. Genes in cluster B are found in two different environmental sources (blue and yellow); however, nodes of the same type are preferentially linked to each other in the network than to genes from different environmental sources. This would result in a positive assortativity coefficient approaching 1 for environment and a low conductance score, suggesting a strong environmental community structure. Genes in cluster C are also found in two different environmental sources; however, there is no clear pattern for the distribution of genes with regard to environment. This network would have an assortativity approaching 0 and a high conductance score

from the same category. Normalized assortativity values range between −1 and 1, where 0 indicates random distribution of categories within the network, 1 indicates that nodes from the same categories tend to be connected to one another in the network, and −1 indicates that nodes from different categories tend to be connected in the network. A detailed description of the algorithms used in these calculations can be found in [108].

2.4.1 Assortativity as a Tool to Study Geographical and Habitat Distributions of Microbes and Genes

Forster et al. used assortativity (among other network statistics, including the previously discussed shortest path analysis) to explore the geographical dispersion patterns of marine ciliates in a network generated from ciliate SSU-rDNA sequences [25]. Sequences were clustered into two different levels of gene family—CCs and Louvain communities (LCs) as previously described. Sequences were assigned categorical labels based on their geographical point of origin (eight locations) or habitat of origin (three habitats), and assortativity was calculated. If sequences, and thus species, are broadly distributed across geographical categories, then assortativity of SSU-rDNA sequences labeled with these geographical categories would be low because similar sequences would be found in different environments. Contrarily, if similar sequences tend to be from the same geographical category, indicative of endemism, then assortativity of sequence geographical origin will be high (Fig. 6). The majority of CCs and LCs showed a positive assortativity for geographical origin, higher than expected by chance, indicative of geographical community structure as opposed to global dispersal of ciliates. Similar approaches were used by Fondi et al. and applied to a collection of environmental metagenome samples to test the "everything is everywhere" hypothesis at the gene pool and functional level. Gene pools were more strongly associated with a particular ecological niche than with specific geographical location, supporting the idea that microbial genes are found everywhere but the environment selects for them [26].

2.4.2 Conductance in the Comparison of Lifestyles and Evolutionary Histories

Conductance is used to explore the clustering of pairs of different node categories in a connected component. In a study by Cheng et al., the proteomes of 84 prokaryote genomes were categorized into four broad redox groups based on their lifestyle, methanogens, obligate anaerobes, facultative anaerobes, and obligate aerobes [27]. For each CC in a pan-proteome sequence similarity network including all 84 genomes, the conductance was calculated for pairs of redox categories and compared to values obtained following random relabelling of the components. The distributions of conductance values for methanogens and for obligate anaerobes groups indicated that the sequences in these groups have features distinct from those in other groups, that anaerobes and aerobes tend to be dissimilar, and that their sequences are more isolated from one another in the SSN than expected by chance.

An additional example of the use of conductance is in exploring the propensity of a gene family to lateral gene transfer. Within a network of archaeal and bacterial genes, CCs showing a low conductance for both archaeal and bacterial sequences indicate that the bacterial and archaeal genes within the corresponding families are structured in two separate and conserved groups (Fig. 6). Structuring gene families into two groups would indicate that there was little or no evidence for lateral gene transfer between archaea and bacteria within this particular gene family. This kind of gene family is rare, with only 86 gene families from 40,584 (0.2%) meeting this criteria [24].

2.5 SSNs in Remote Homologue Identification: Shedding Light on the Microbial Dark Matter

Up to 99% of microbial species are not cultivable and thus have not been studied in isolated culture. Analysis of high-throughput sequencing and metagenomics datasets has shed light on these uncultivable organisms, often referred to as the "microbial dark matter" [109], and in some cases enabled the reconstruction of draft genomes [110–114]. A considerable portion of most metagenome studies have predicted ORFs showing no detectable similarity to any known proteins, termed metaORFans [115]. These can represent 25–85% of the total ORFs identified in metagenomes [22]. Identifying distant homologues of ORFans may help to predict their functions and begin to unravel the microbial dark matter. Recent work by Lopez et al. in 2015 probed the microbial diversity of metagenome datasets from a range of environments including the human gut microbiome, identifying homologues of genes from 86 ancient gene families that are distributed across archaea and bacteria. The majority of these gene families included environmental homologues that were highly divergent from any of their cultured homologues, and many branched deeply with the phylogenetic tree of life, highlighting our limited understanding of diverse elements of the microbial world and hinting at the existence of yet unknown major divisions of life [24] (Fig. 7).

2.6 Exploiting SSNs to Analyze Classifications

Metagenomic and genomic data are providing scientists with a tantalizing amount of sequence data, casting the analysis of the extent of biodiversity as a major research theme in biology [116–120]. In theory, existing organismal and viral classifications are invaluable tools to structure and analyze this biodiversity. However, the way taxonomical classifications are constructed raises questions about their naturalness and their actual application scope [38, 120–128], in particular regarding genetic diversity surveys. There are three major reasons for this. First, organismal and viral diversity is still largely undersampled, which means that existing classifications are incomplete [119, 120]. Therefore, taxonomically unassigned sequences cannot be readily used in class-based genetic diversity surveys, since this dark matter remains outside existing classes. Second, classifications are constructed

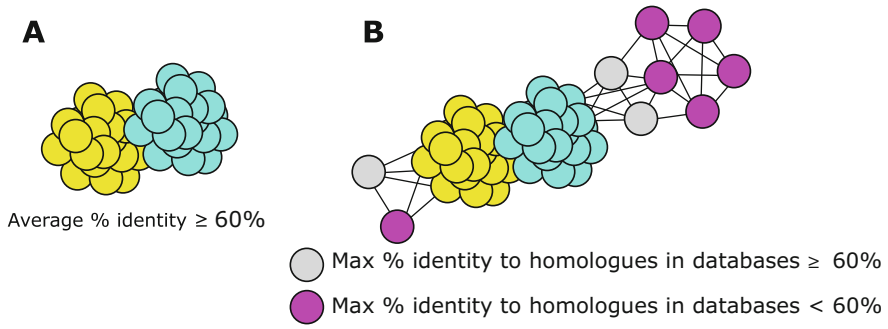


Fig. 7 Remote homologue detection to help characterize the microbial dark matter. (**a**) A hypothetical highly conserved cluster of genes from genomes present in sequence databases, where the average % of identity is high (≥60%). (**b**) The same cluster after addition of divergent environmental sequences to the network. Environmental sequences in gray are more similar to those already identified from genome surveys (≥60% max identity) so are connected directly to the conserved gene cluster in the network. More divergent sequences in pink have <60% maximum identity to their homologues in the database. Many of these are only identified as linked to the sequences from the conserved database via intermediate gray nodes. This is the notion of "transitive homology"

using different features (i.e., for viruses, a mix of phylogenetic, morphological, and structural criteria, such as replication properties in cell culture, virion morphology, serology, nucleic acid sequence, host range, pathogenicity, epidemiology, or epizootiology); therefore their classes do not necessarily offer immediate proxies for quantifying genetic diversity per se. Third, evolutionary processes responsible for both genetic and organismal diversity are diverse, and they operate at different tempos and modes in different lineages [49, 123, 129–141]. As a result, genetic diversity within classes and between classes can be heterogeneous, meaning that existing classifications may lack efficiency to discriminate, predict, or compare taxa on genetic bases, potentially hampering diversity studies, a profound practical issue at a time where the analysis of metagenomic sequences is becoming a priority in biology.

Addressing these challenges is notably crucial for viral studies. Recently, the executive committee of the ICTV [142] proposed that network analyses methods that create similarity metrics based on the detection of homologous genes and their genetic divergence constitute a valuable strategy to assist classification of viruses. Consistently, basic network properties and metrics (Table 1) can quantify (1) whether genetic diversity is consistent within and between the classes of existing classifications and (2) describe what classes are the most homogeneous and distinctive in terms of genetic diversity. Three criteria can be used to estimate intra-class genetic heterogeneity (Fig. 8a–c). First, the average edge weights (measured as % of identity, PID) between pairs of sequences from genomes of the

Table 1
Schematic properties of two extreme kinds of taxonomic classes with respect to their genetic diversity

"Ideal" classes	Not ideal classes
Low intra-class genetic diversity (high average PID)	High intra-class genetic diversity (low average PID)
High genetic cohesion (high average CCC)	Low genetic cohesion (low average CCC)
Core components (high maxCore%)	No core components (low maxCore%)
Obvious genetic distinctiveness (high conductance difference with random groups)	Limited genetic distinctiveness (conductance similar to random groups)
Exclusive pangenome (high % of exclusive CC)	No exclusive pangenome (low % of exclusive CC)

The three top properties inform about genetic diversity within classes (intra-class genetic diversity). The last two properties inform about the genetic distinctiveness (core and signature genes) of the classes. Interclass genetic heterogeneity identifies when genetic diversity of a class is not comparable with genetic diversity of another class in the classification. CCC, average proportion of genetic conservation between sequences from the same cluster and from the same taxonomic class; PID, average edge weights (% identity) between two sequences from genomes of the same class

same class provide a trivial measure of intra-class genetic diversity. Second, the average proportion of *C*onserved *C*anonical *C*onnections between sequences from the same connected component and from the same taxonomic class can be exploited (CCC, i.e., in each connected component of the SSN, the total number of edges connecting sequences of a given class i (intra-group edges, denoted E_{ii}) divided by the theoretical maximal number of possible edges between sequences of that class in the connected component (CCC $(i) = 2^{*}E_{ii}/(N_i \times (N_i - 1))$ where N_i is the number of sequences of class i present in the connected component). CCC ranges between 0 and 1. Within a connected component, if all pairs of sequences from the same class are directly connected, CCC equals 1, since all these sequences are more conserved than a given %ID threshold. By contrast, low CCC are observed when sequences from genomes from the same class lack cohesive evolution, for example, when some related sequences evolved so fast that they show less than the minimal similarity required to be directly connected to their homologues in the graph. Third, the genetic consistency of a class can be estimated by (1) identifying what cluster of sequences was present in the largest number of genomes of the class and then (2) by quantifying the proportion (in %) of the class members harboring that most ubiquitous cluster (maxCore%). When maxCore% of a class is <100%, it means that, for this dataset, there is no gene family shared by all members of that class (i.e., no core genes). The SSN structure can also serve to estimate the genetic distinctiveness of each class, i.e., whether sequences from a given class are

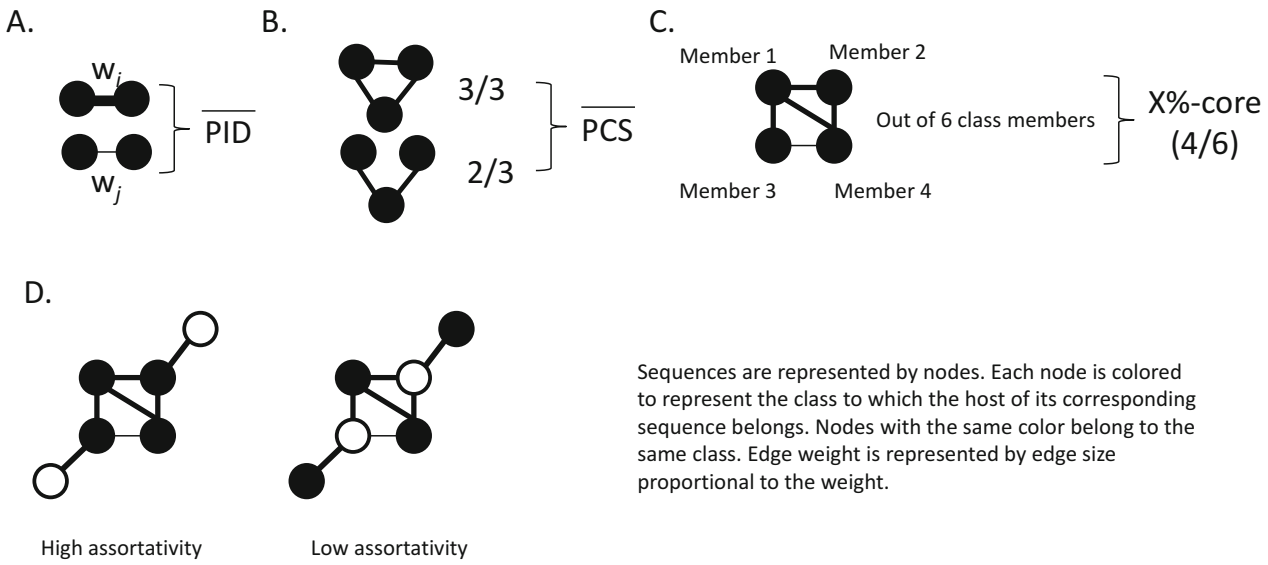


Fig. 8 Intra- and interclasses heterogeneity measurements in weighted similarity networks. Sequences are represented by nodes. Each node is colored to represent the taxonomic class to which its host belongs. Nodes with the same color belong to the same class. Edge weight is represented by edge size proportional to the weight. Subgraphs correspond to clusters of sequences. Direct neighbors have a greater similarity than the threshold set to allow such connections. PID, average edge weights (% identity) between two sequences from genomes of the same class; CCC, average proportion of genetic conservation between sequences from the same cluster and from the same taxonomic class; maxCore%, conductance; and %-exclusive components correspond to the estimates used to assess genetic consistency of classes

more similar to one another than they are to sequences from other classes (Fig. 8d, e). Such sequences could be used as classificatory features to assign members to the class. In a SSN, this property translates to a low ratio of interclass edges over intra-class edges and is measured by conductance (Fig. 8d). Likewise, the proportion of clusters comprised exclusively of sequences from one class, a diagnostic feature of the class, provides an estimate of the class genetic distinctiveness. Genetically highly distinct classes have a high % of such exclusive clusters. Based on these network measures, interclass genetic heterogeneity can simply be diagnosed by contrasting estimates of genetic consistency for all the above measures for each class. There is interclass heterogeneity within a classification when the mean PID, mean CCC, maxCore%, DRC, and % of exclusive components differ between classes.

Such network analyses show that virus classifications face a pragmatic issue: overall genetic distinctiveness allows relatively safe assignments of viral sequences to existing classes; however, genetic diversity of viral taxa of similar ranks differs among the tested classifications. Therefore, virus classifications (especially ICTV classification at the family level) should be used carefully to avoid inaccurate estimates in metagenomic diversity surveys. Classes with broader genetic diversity will tend to be more easily

detected in the environment than classes with reduced genetic diversity, since the former will necessarily be associated with more OTUs than the latter. Some alpha- and beta-diversity analyses of environmental data, which rely on counts and on contrasts of the abundance of taxonomic classes in different samples, will thus also be biased. A similar approach could be applied on different types of classified lineages, i.e., to identify what groups of bacteria, archaea, or eukaryotes with comparable taxonomical ranks are the most genetically heterogeneous and what ranks of their classification are the least genetically consistent.

3 Gene-Sharing Networks

Gene-sharing networks are often called "genome networks" as they are best suited for summarizing what genes are shared between different genomes, highlighting routes of gene sharing. The ability to explore gene sharing between all genomes in a network in a simple graph can have useful properties for reflecting microbial social life, inherently inclusive of gene sharing both as a consequence of vertical inheritance and lateral gene transfer (LGT). Bacteriophage and plasmid genomes are typically highly mosaic in nature due to a high level of horizontal gene transfer, making it difficult to classify their genomes [37, 143]. Lima-Mendez et al. proposed the use of gene-sharing networks as a new classification method that tackles this problem of mosaicism by classifying viruses based on their genome's content [37]. Constructing gene-sharing networks using subsets of genes from different functional categories of genes can also be useful in exploring what kinds of genes are being shared by different genomes.

In a gene-sharing network, each genome is represented by a node, and two nodes are connected by an edge when the two corresponding genomes share homologous genes or gene families (Fig. 9). These gene families can be identified from SSNs (of as CCs of LCs) or by alternative methods. In gene-sharing networks, edges can be weighted by the number of genes or gene families shared between the genomes. In this way, gene-sharing networks enable the study of microbial social life, quantitatively displaying the gene families shared between genomes both as a result of vertical transmission and lateral gene transfer.

Gene-sharing networks are useful tools for exploring overall patterns of gene sharing between genomes. Recently, Lord et al. developed BRIDES, a software package that specifically identifies different kinds of patterns in evolving gene-sharing networks after the addition of new genome nodes [144]. However, in gene-sharing networks the kind of gene families that are being shared is often overlooked. To explore how functions are shared between different genomes, gene-sharing networks can be built from genes

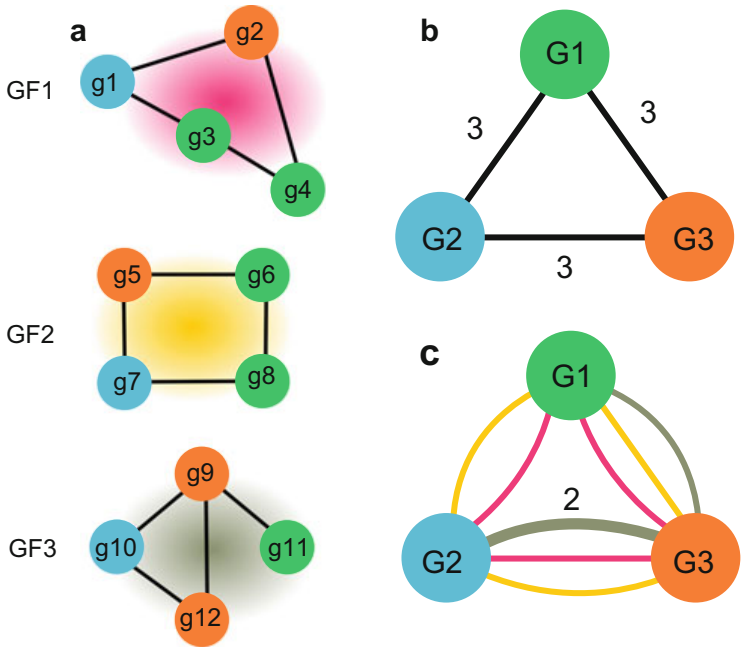


Fig. 9 Translating gene networks to gene-sharing networks. (**a**) Gene network for three gene families. Gene nodes are colored based on their genome of origin. The background color corresponds to the gene family color in part **c**. (**b**) The gene-sharing network corresponding to the gene network in **a**. Edges are weighted on the number of gene families shared by the genomes. (**c**) Multiplex gene-sharing network corresponding to the gene network in **a**. Genomes are connected by multiple edges with colors corresponding to different gene families. These edges are weighted based on the number of genes shared between two genomes for each family

using different subsets of functions (Fig. 10) [29]. An alternative form of the gene-sharing network is the multiplex network. In this network nodes can be linked by edges of different types, for example, each edge representing a different gene family or different functional groups of gene families, thus retaining additional information compared to a simpler gene-sharing network (Fig. 9) [23]. Multiplex networks can be useful for small-scale analyses; however, with large datasets they can rapidly become difficult to interpret and analyze. Importantly, multiplex networks are unimodal projections of bipartite graphs (discussed in the Subheading 14) which can provide greater clarity and have a number of attractive properties for the analysis of larger datasets.

3.1 Classification of Entities Using Gene-Sharing Networks

The possibility of summarizing gene sharing between sets of entities with complex evolutionary histories means that gene-sharing networks can be useful for classifying organisms based on their gene content. Lima-Mendez et al. analyzed bacteriophage genomes to generate two different phage gene-sharing networks that reflect their reticulate evolutionary history [37]. In the first gene-sharing network, phage genomes (nodes) were connected by edges when

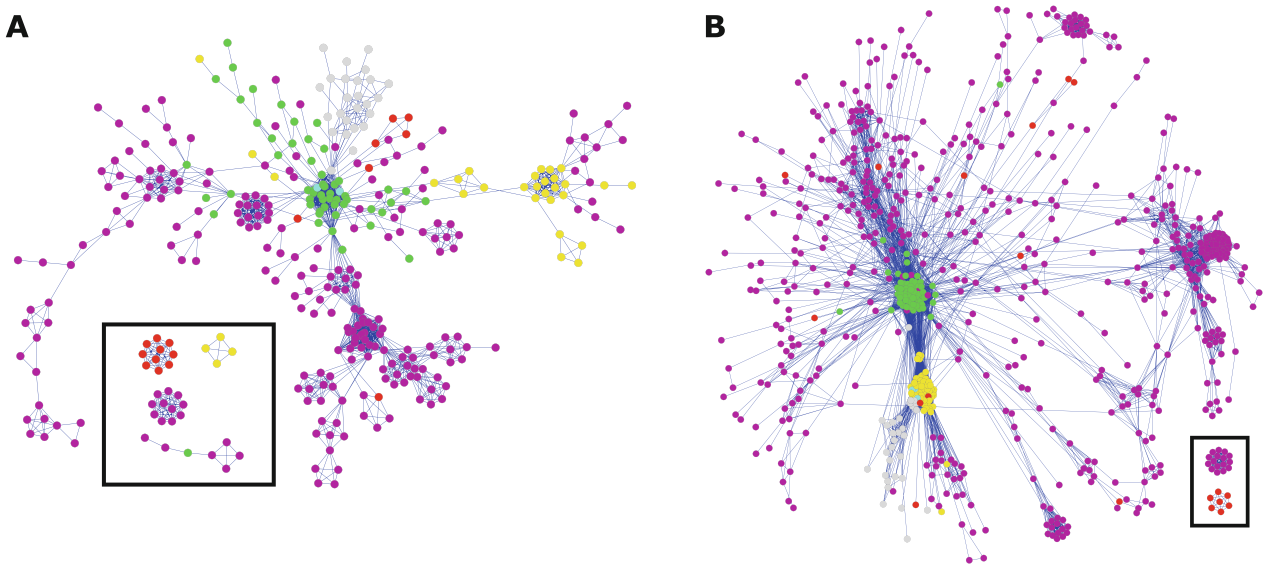


Fig. 10 Functional gene-sharing network reflecting the chimeric nature of eukaryotes. These gene-sharing networks describing how genes in different functional categories are shared between bacteria (green), archaea (yellow), eukaryotes (gray), plasmids (purple), and viruses (red) from a published dataset [29]. In both cases, a giant connected component is shown alongside examples of smaller connected components (**a**) Gene-sharing network for COG category D: cell division control. In this network, sequences of eukaryote origin (gray) cluster with bacterial sequences, reflecting their origin in the alphaproteobacterial endosymbiont that would become the mitochondrion. (**b**) Gene-sharing network for COG category K: transcription machinery. In this network, eukaryote sequence (gray) cluster with archaeal sequences, reflecting the origin of these genes in the archaeal host for the eukaryotic endosymbiont

they shared significant similarity at the sequence level. This gene-sharing network was clustered using the previously discussed MCL algorithm [145], identifying distinct groups of phages with sequence similarity. Following clustering, membership to a particular cluster was reassessed based on shared similarity with viruses in other clusters, reflecting their reticulate evolutionary history, allowing the generation of a matrix assigning a score describing the relative membership of any given viral genome to a particular classification group. In the second approach, Lima-Mendez et al. generated a "module"-based gene-sharing network, where edges are drawn between two phage genomes if they share a "module," in this case defined as a group of genes with similar phylogenetic profiles, enabling the exploration of what kinds of genes are shared between different groups of phages or are "signatures" for a particular group of phage genomes [37].

3.2 Exploring Routes of Gene Sharing in Gene-Sharing Networks

Two network metrics, also useful in the analysis of gene networks, can be used to attempt to identify "hubs" of gene sharing in the context of gene-sharing networks: node "degree" and "betweenness." Both metrics aim to determine the centrality of a node in a network. The degree of a node is simply the number of edges that it is connected to. The betweenness of a node is the frequency at

which it is found in all the possible shortest paths between any two nodes in the network. Halary et al. used gene-sharing networks based on DNA sequence similarity to explore gene sharing between prokaryotes and mobile genetic elements [30]. Plasmids were identified as hubs of gene sharing within this pool of genomes, suggesting that they are key vectors for genetic exchange between cellular genomes and a potential DNA reservoir shared by genomes. Phages were more peripheral in the network and mostly linked prokaryotes from the same lineage. Thus, gene-sharing networks provided insights on the evolutionary processes that shape the gene content of prokaryote genomes.

The importance of plasmids in genetic worlds was further highlighted by exploring plasmid gene-sharing networks without inclusion of prokaryote genomes [14, 36]. Connecting 2343 plasmid genomes based on shared gene content in a single graph demonstrated that plasmids tended to cluster based on the phylogenetic class of their corresponding host prokaryote rather than habitat but that more mobile plasmids tended to be more "central" in the graph, indicating that these were hubs of gene sharing. Specifically, routes of gene sharing for gene families including antibiotic resistance markers were identified between actinobacterial plasmids and gammaproteobacterial plasmids, suggesting that Actinobacteria may act as a reservoir for antibiotic resistance genes for Gammaproteobacteria [14].

The finding that plasmids are hubs of gene sharing for prokaryote genomes was supported by analysis of gene sharing in a proteobacterial phylogenomic network including 329 proteobacterial genomes [32]. A phylogenomic network is a type of phylogenetic network that has been constructed from fully sequenced genomes. In this example the phylogenomic network is an alternative to a gene-sharing network, in which genome nodes within a phylogeny are linked by edges if they share genes [34]. This study identified extensive evidence for lateral gene transfer among Proteobacteria, with at least one LGT event inferred in 75% of all gene families. Of these putative LGTs, more were related to plasmid-related genes than phage-related genes, suggesting plasmid conjugation was a more frequent source of gene transfer [32]. Directed graphs exploring directionality of LGT events between 657 prokaryote genomes allowed the polarization of 32,028 putative LGT events finding that frequency of recent events correlates with genome sequence similarity and most LGTs occurring between donor-recipient pairs with $<5\%$ difference in GC content, suggesting that there are some barriers to lateral gene transfer between prokaryotes but that these are not insurmountable [31]. Later reconstruction of transduction events linking phage donors and recipients in a phylogenomic network demonstrated that LGT by transduction was generally highest in similar genomes and between clusters of closely related species but that this constraint was occasionally broken, resulting in LGTs over long evolutionary distances [35].

4 Bipartite Graphs

Bipartite graphs are excellent at summarizing what genes are shared between sets of genomes, and as such are ideal for comparative genomics, including for the comparison of genomes reconstructed in metagenomic analyses. The potential to extend this approach to multilevel graphs, adding additional layers of information such as the environment in ecological studies, could provide a powerful summary of gene sharing in relatively complex datasets.

A multilevel network is a network in which edges exclusively connect nodes of different types, i.e., representing different levels of biological organization. Thus, a bipartite graph is a graph with two types of nodes (top and bottom nodes), where edges exclusively connect nodes of different types (Fig. 11) [146]. The types of nodes used can vary widely depending on the biological question, from linking diseases (top nodes) to their associated genes (bottom nodes) in order to explore the association between related disease phenotypes and their genetic causes [147, 148], to exploring the concept of flavor pairings in food based on a graph of ingredients (top nodes) and the flavor compounds they contain (bottom nodes) [149]. For applications in molecular biology, a typical example of a bipartite graph may describe the relationships between genomes (top nodes) and gene families (bottom nodes), with edges between nodes indicating that a genome encodes at least one member of the corresponding gene family (Fig. 11) [23, 33, 38, 150]. This kind of genome to gene family graph is particularly suited for the comparative analysis of the gene content of genomes in microbial communities and for exploring patterns of gene sharing, for example, between distantly related cellular genomes [33] or between cellular genomes and their mobile genetic elements (Corel et al. forthcoming). It is possible to represent all genes shared between a given set of genomes, as a result of both vertical inheritance and horizontal gene transfer, in a single bipartite graph [23].

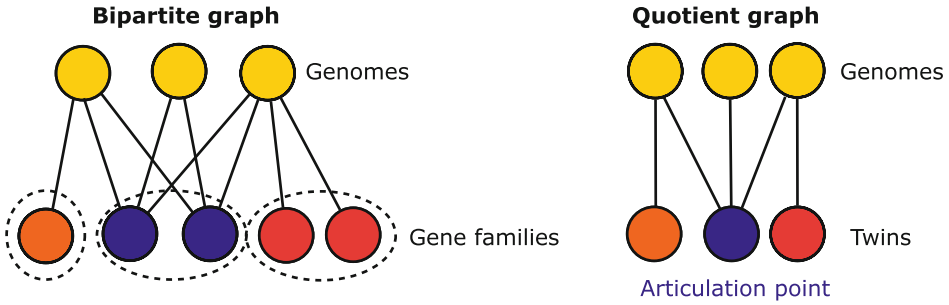


Fig. 11 A bipartite graph and its reduction to a quotient graph: (**a**) An example of a bipartite graph displaying how five gene families are shared between three genomes. (**b**) A reduced form of the bipartite graph in which gene families are combined to "twin" nodes if they share identical taxonomic distributions. A single "articulation point" connects all three genomes

This feature was utilized by Iranzo et al. to explore gene sharing among the entire dsDNA virosphere, a group of entities typified by high rates of molecular evolution and gene transfer [38]. In this case, bipartite modularity was identified in the graph to identify groups of related viral genomes and their shared genes, with the modularity of the graph optimized to Barber's bipartite modularity [151]. A number of additional methods have been developed for detection of module structures within a bipartite graph including for weighted graphs [152]. Two recently developed tools, AcCNET [150] and MultiTwin (forthcoming), have simplified the process of constructing and analyzing multilevel graphs without the need for custom programming (Boxes 3 and 4).

Box 3: Generating Gene-Sharing Networks and Bipartite Graphs

1. *Dataset assembly*: The same rules for dataset assembly as described in SSN generation apply to assembling the dataset for bipartite and gene-sharing graphs. It is especially important to maintain an annotation file that maps gene IDs to their genome of origin.
2. *Definition of gene families*: Gene family identification can be carried out following the construction of sequence similarity networks, as described in Subheading 2. There are a broad range of alternative approaches for construction of gene families that are beyond the scope of discussion in this chapter; however, all of these can also be applied to the generation of gene-sharing and bipartite graphs.
3. *Network construction*: From the definition of gene families, it is possible to construct both gene-sharing networks and bipartite graphs.
   (a) In a gene-sharing network, two genomes are connected by an edge when they encode genes belonging to the same gene family. Generating this kind of network can be automated from BLAST or fasta sequence data using EGN [52].
   (b) In a bipartite graph, there are two types of node, genome nodes and gene family nodes. An edge is drawn between a genome node and a gene family node if that genome encodes a member of the gene family. AcCNET [150] and MultiTwin (forthcoming) tools both include pipelines for generating bipartite graphs from sequence data. MultiTwin can also generate a bipartite graph from two files: a tab-delimited file mapping gene identifiers to their corresponding genome identifier and a tab-delimited file mapping gene identifiers to their corresponding gene family.

Two topological features of bipartite graphs can be used to facilitate studies of gene sharing by an exact decomposition of the bipartite graph: twins and articulation points [23, 153]. A bipartite graph can be reduced to a quotient graph, a reduced variant of the bipartite graph where nodes from the bipartite graph have been combined based on sharing similar properties without the loss of information. For twin nodes ("twins"), this reduction is based on the combination of bottom nodes that have identical neighbors into a single "twin" supernode in the quotient graph (Fig. 11). This is a useful way of reducing the size of large graphs without losing information, but twin nodes also have useful properties for graph interpretation. The genomes supporting a twin node (its neighbors) define a club of genomes that share genes, through common ancestry and/or horizontal transfer, and the number of gene families making up the twin gives a simple description of how many gene families are shared between this club. For example, in any given dataset, any "core" set of gene families encoded by all species in the analysis will be represented by a single twin node. The gene families combined in twin supernodes can be viewed as gene families that are likely to be transmitted together [23]. An articulation point is a node that, when removed, will split the graph into two or more connected components. Within a gene family-genome bipartite graph, articulation points are expected to help to identify "public genetic goods," gene families that are shared by distantly related entities that may confer an advantage independent of genealogy [23, 154], as well as selfish genetic elements such as transposases that also spread across multiple genomes.

Box 4: Considerations for the Construction and Analysis of Bipartite Graphs Using AcCNET and MultiTwin

The default workflow for both ACcNet and MultiTwin takes protein sequence data in fasta format as input and generates a bipartite graph alongside a number of graph summary statistics and outputs for visualization in standard tools (such as Gephi and Cytoscape) but with a number of important differences, including:

- *Graph levels*: Both AcCNET and MultiTwin can generate a bipartite graph using their default workflow; however, MultiTwin can also be used to explore additional graph levels by adding additional node types (e.g., a tripartite graph). Multipartite graphs mean that gene family level annotations can be associated with additional levels of biological information. This may be particularly useful for the comparison of samples in metagenomics studies or time course experiments, allowing gene families to be associated directly with features such as environmental origin or time point.

(continued)

Box 4: (continued)

- *Gene family identification*: AcCNET uses kClust [155] to assemble gene families, a kmer-based method for rapid assembly of clusters of homologous proteins from sequence data. By default, MultiTwin identifies gene families using an all-versus-all BLAST search, followed by identification of connected components at a given threshold, as previously discussed for gene family detection from SSNs. MultiTwin can also be used in a modular way allowing for additional customization, including the use of any custom gene family input in the form of a "community file": a tab-delimited file linking every gene/protein ID to a community identifier, with gene families defined using a clustering method of choice.
- *Edge weighting*: In AcCNET the edge weight is proportional to the inverse of the phylogenetic distance between proteins in a cluster from a given genome to other proteins within the same cluster. In MultiTwin, the default edge weight is based on the number of genes present in a gene family from any given genome.
- *Graph compression*: While both methods can be used to identify "twin" nodes, only MultiTwin generates a quotient graph from these twin nodes and identifies articulation points.

AcCNET is available at: https://sourceforge.net/projects/accnet

MultiTwin is available at: http://www.evol-net.fr/index.php/en/downloads

4.1 Using Bipartite Graphs to Explore Patterns of Gene Sharing Between Diverse Entities

The simplest application of a bipartite graph is the summary of all genes shared between genomes in a single parsable graph, and this feature has been used to explore gene sharing in the dsDNA virome [38], a range of *Escherichia coli* genomes to investigate *the E. coli* pangenome [150] and between a broad range of prokaryotes that include newly discovered organisms [33]. In their analysis of prokaryote genomes, Jaffe et al. used the notion of "twins" to explore patterns of gene sharing between prokaryotes, including Archaea and the recently discovered ultrasmall "Candidate Phyla Radiation" and TM6 bacteria with extremely unusual and reduced genomes. The group found evidence for lateral gene transfer between ultrasmall bacteria and other prokaryotes, consistent with the suggestion that the ultrasmall bacteria may be symbionts [33]. In their exploration of the dsDNA virome, Iranzo et al. used graph module detection, algorithms designed to identify groups of densely connected nodes in a graph, to identify sets of densely connected viral genes and genomes that included viruses with broad host ranges, as well as 14 hallmark viral genes that account for most of the gene sharing between all different viral modules [38].

5 Conclusions

This chapter has offered a brief introduction to the generation of commonly used sequence similarity networks in molecular biology and a guide to how they can be generated and applied to a broad range of studies (Fig. 12). Networks provide a highly scalable framework for the study of an increasingly broad range of applications in molecular biology and evolution and have already contributed to a number of important discoveries in the field. These include exploring patterns of introgression and horizontal transfer across all domains of life and mobile elements, the origin of eukaryotes, the contribution of new genes including novel fusion genes to major evolutionary transitions, shedding light on the "microbial dark matter" in metagenome sequencing datasets and in testing ecological hypotheses about organism and gene distribution and environmental selection. New methods and tools for network analysis are becoming increasingly user-friendly and accessible to biologists without extensive programming experience and enabling network analysis to become a more common part of a biologist toolkit in the analysis of molecular sequence data.

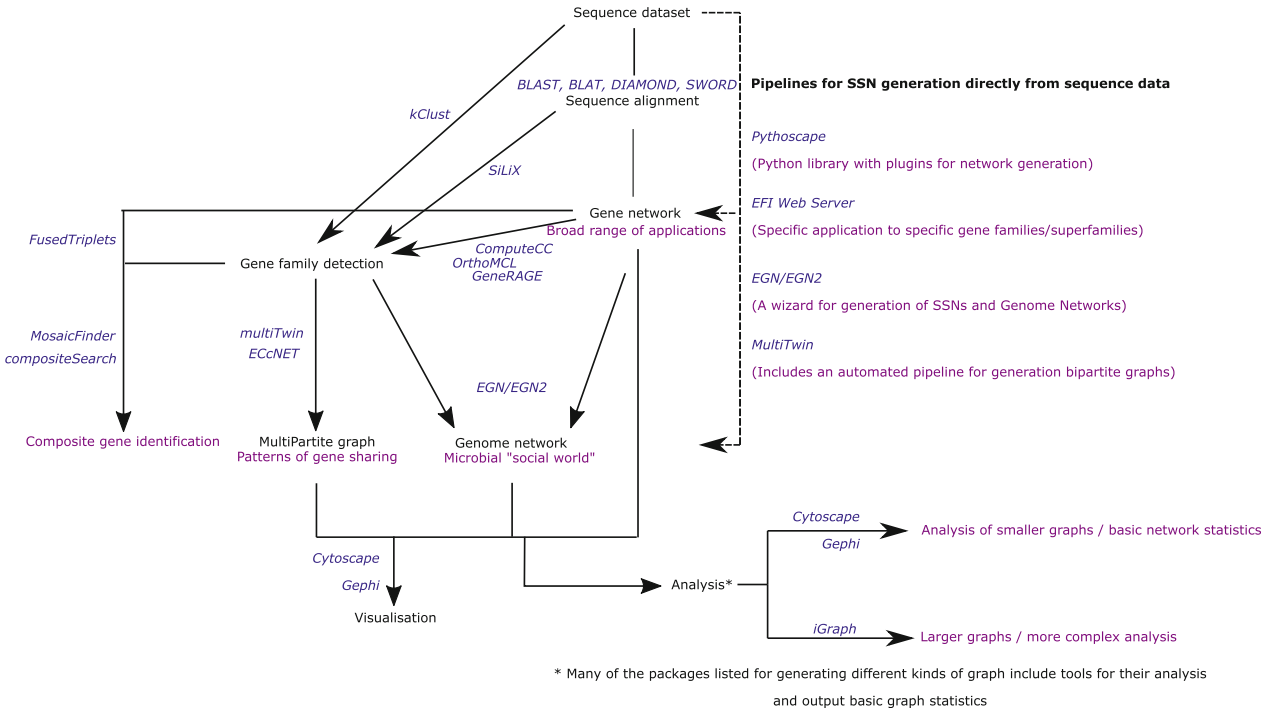


Fig. 12 A workflow highlighting some of the available routes for generation and analysis of SSNs, gene-sharing networks, and bipartite graphs. This workflow highlights just some of the many tools and routes for network construction and analysis

6 Exercises

The exercises use EGN [52] and require access to a local installation of BLAST+ [58] and Perl. The fasta sequence file "example.faa" provided with EGN includes a dataset of protein sequences from Archaea, Bacteria, Eukaryota, and mobile genetic elements, available at http://www.evol-net.fr/index.php/fr/downloads:

1. Perform a manual all-versus-all BLAST using search for a given protein sequence file from the unix terminal (requires local installation of BLAST). The output can be filtered to generate a network:
   (a) Make the blast database using the "*makeblastdb.*"
      - Command: "*makeblastdb -dbtype prot -in example.faa –out example*"
   (b) Performing the BLAST search using "blastp," remembering to output data in a tabular format for easy processing.
      - Command: "*blastp -query example.faa -db example -evalue 1e-5 -seg yes -soft_masking true - max_target_seqs 5000 -outfmt "6 qseqid sseqid evalue pident bitscore qstart qend qlen sstart send slen" -out protein.blastpout*"
2. Generate a SSN using EGN from example.faa (requires local installation of BLAST and download of EGN from http://www.evol-net.fr/index.php/fr/downloads):
   (a) Run EGN from the terminal using "perl egn.1.0.plus.pl" from the programs home directory.
   (b) Follow on-screen prompts sequentially to generate an alignment, filter the output, and generate a gene network with outputs compatible with both Cytoscape and Gephii.
3. Visualize SSN networks:
   (a) In Cytoscape: Import files named "*cc.*.txt*" as a network to visualize that set of connected components.
      - To associate nodes with their annotations, import "*cc*.atr*" as a table.
   (b) In Gephi: Open "*cc*.gxf*" files to import individual connected components from the network into Gephi. Use the "layout" menu to explore different kinds of layouts for the network.

Glossary

Articulation point	A node in a graph whose removal increases the number of connected components of the resulting graph.

Adjacency matrix	A numerical square matrix with row and columns labeled by network nodes, with 1 or 0 in the matrix indicating whether they are connected by an edge in the network.
Assortativity	A measure of the preference for labeled nodes in a network to attach to other nodes with identical labels. This is the Pearson correlation coefficient of the degrees of pairs of linked nodes. Assortativity $= \frac{\text{modularity}}{\text{modularity}_{\max}}$ with modularity defined below and modularity max as the modularity of a perfectly mixed network. $\text{modularity}_{\max} = \frac{1}{2m}\left(2m - \sum_{ij}\frac{k_i k_j}{2m}\delta(c_i - c_j)\right)$.
Betweenness	A centrality measure for a node in a graph. Precisely, this is the proportion of shortest paths between all possible pairs of nodes in a connected component that pass through this node. A betweenness close to 1 is indicative of a highly central gene, whereas close to 0 is more peripheral.
Bipartite graph	A graph with two types of nodes (top and bottom nodes), in which an edge only connects nodes of different types.
Club of genomes	A group of entities that replicated separately but exploit common genetic material that may not trace back to the last common ancestor.
Communities (also called modules)	In graph terminology, a community is defined as a group of nodes that are more connected between themselves than to nodes in the rest of the graph.
Composite gene	A gene that is made up of at least two component parts.
Component genes	Genetic fragments sharing partial similarity to a composite gene.
Conductance	A measure that quantifies whether a given category of nodes shares more edges between themselves than with the rest of the nodes in the graph. A low conductance approaching zero implies that there are few edges shared between this category of nodes and the rest of the graph, while a higher conductance implies more connectivity between that category of nodes and other nodes outside of the category. G a graph, G = {V, E}. With U & G a set of nodes that is assumed to not have more than half the total node. $\bar{U}$ = G\U. d(U)

	sum of degree of vertices in U. Conductance $= \frac{\sum_{i \in U, j \in \bar{U}} a_{i,j}}{\min\left(d(U), d\left(\bar{U}\right)\right)}$
Connected component	A subgraph in which any pair of nodes is connected, either directly or indirectly, and that is not connected to the rest of the graph.
Degree	The number of edges connected to a given node.
Endosymbiont	An organism that lives inside another to the mutual benefit of both organisms.
Edge	The link between two nodes in a network.
E-value	The number of alignments in a sequence similarity search expected to be seen by chance searching against a database of a certain size.
Introgression	Descent process through which the genetic material of an entity propagates into different host structures and is replicated within these new host structures.
Lateral gene transfer (LGT; or horizontal gene transfer, HGT)	Movement of genetic material between entities not mediated by vertical descent.
Louvain community	A graph community identified using the Louvain algorithm. Louvain algorithm is based on optimizing modularity.
Network (or graph)	A system of objects (nodes), some pairs of which are linked (edge).
Multipartite graph	Similar to a bipartite graph, but with any number of types of nodes exclusively connected to nodes of other types.
Multiplex graph	A graph where nodes can be connected by edges of different types.
Modularity	The fraction of edges falling within given groups (e.g., communities or functional categories) in a network, minus the fraction of edges that would be expected with a random distribution of edges. With m the total number of vertices, c_i the community of node i, $\delta()$ the Kronecker delta, and k_i the degree of modularity $= \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta\left(c_i - c_j\right)$.
Phylogenomic network	A phylogenetic network constructed from whole genome sequences where genomes are connected based on pairwise relationships including vertical and lateral gene transfer (LGT) events.

Public genetic goods	Common genetic materials shared by clubs of phylogenetically distinct genomes.
Quotient graph	A simplified graph whose nodes represent disjoint subsets of nodes of the original graph; an edge in this new graph connects two such new nodes whenever an edge in the original graph connects at least one element of a new node with at least one from the other.
Supporting genomes	The common set of neighbors that support a "twin" class in a multipartite graph.
Twins	Nodes in a multipartite graph that share identical sets of neighbors.

References

1. Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5:123–135. https://doi.org/10.1038/nrg1271
2. Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. Nature 440:623–630. https://doi.org/10.1038/nature04546
3. Williams TA, Foster PG, Cox CJ, Embley TM (2013) An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504:231–236. https://doi.org/10.1038/nature12779
4. Alsmark C, Foster PG, Sicheritz-Ponten T et al (2013) Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. Genome Biol 14:R19. https://doi.org/10.1186/gb-2013-14-2-r19
5. Hirt RP, Alsmark C, Embley TM (2015) Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. Curr Opin Microbiol 23:155–162. https://doi.org/10.1016/j.mib.2014.11.018
6. Nowack ECM, Price DC, Bhattacharya D et al (2016) Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of Paulinella chromatophora. Proc Natl Acad Sci U S A 113:12214–12219. https://doi.org/10.1073/pnas.1608016113
7. McCoy JM, Mi S, Lee X et al (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789. https://doi.org/10.1038/35001608
8. Kondo N, Nikoh N, Ijichi N et al (2002) Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. Proc Natl Acad Sci U S A 99:14280–14285. https://doi.org/10.1073/pnas.222228199
9. McInerney JO (2017) Horizontal gene transfer is less frequent in eukaryotes than prokaryotes but can be important (retrospective on DOI 10.1002/bies.201300095). BioEssays 39:1700002. https://doi.org/10.1002/bies.201700002
10. Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19:2226–2238
11. Dagan T, Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc Natl Acad Sci U S A 104:870–875. https://doi.org/10.1073/pnas.0606318104
12. Hooper SD, Mavromatis K, Kyrpides NC (2009) Microbial co-habitation and lateral gene transfer: what transposases can tell us. Genome Biol 10:R45. https://doi.org/10.1186/gb-2009-10-4-r45
13. Nelson-Sathi S, Sousa FL, Roettger M et al (2014) Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517:77–80. https://doi.org/10.1038/nature13805
14. Tamminen M, Virta M, Fani R, Fondi M (2012) Large-scale analysis of plasmid relationships through gene-sharing networks. Mol Biol Evol 29:1225–1240. https://doi.org/10.1093/molbev/msr292
15. Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. Trends

Genet 25:107–110. https://doi.org/10.1016/j.tig.2008.12.004
16. Vos M, Hesselman MC, te Beek TA et al (2015) Rates of lateral gene transfer in prokaryotes: high but why? Trends Microbiol 23:598–605. https://doi.org/10.1016/j.tim.2015.07.006
17. McInerney JO, McNally A, O'Connell MJ (2017) Why prokaryotes have pangenomes. Nat Microbiol 2:17040. https://doi.org/10.1038/nmicrobiol.2017.40
18. Niehus R, Mitri S, Fletcher AG, Foster KR (2015) Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nat Commun 6:8924. https://doi.org/10.1038/ncomms9924
19. Hotopp JCD, Clark ME, Oliveira DCSG et al (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317:1753–1756. https://doi.org/10.1126/science.1142490
20. Wolf YI, Kondrashov AS, Koonin EV (2000) Interkingdom gene fusions. Genome Biol 1: research0013.1. https://doi.org/10.1186/gb-2000-1-6-research0013
21. Becking LB (1934) Geobiologie of inleiding tot de milieukunde. W.P. Van Stockum & Zoon, Den Haag, The Hague, the Netherlands
22. Lobb B, Kurtz DA, Moreno-Hagelsieb G, Doxey AC (2015) Remote homology and the functions of metagenomic dark matter. Front Genet 6:234. https://doi.org/10.3389/fgene.2015.00234
23. Corel E, Lopez P, Méheust R, Bapteste E (2016) Network-thinking: graphs to analyze microbial complexity and evolution. Trends Microbiol 24:224–237. https://doi.org/10.1016/j.tim.2015.12.003
24. Lopez P, Halary S, Bapteste E (2015) Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life. Biol Direct 10:64. https://doi.org/10.1186/s13062-015-0092-3
25. Forster D, Bittner L, Karkar S et al (2015) Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. BMC Biol 13:16. https://doi.org/10.1186/s12915-015-0125-5
26. Fondi M, Karkman A, Tamminen MV et al (2016) "Every gene is everywhere but the environment selects": global geolocalization of gene sharing in environmental samples through network analysis. Genome Biol Evol 8:1388–1400. https://doi.org/10.1093/gbe/evw077
27. Cheng S, Karkar S, Bapteste E et al (2014) Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. Front Ecol Evol 2:72. https://doi.org/10.3389/fevo.2014.00072
28. Thiergart T, Landan G, Schenk M et al (2012) An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol 4:466–485. https://doi.org/10.1093/gbe/evs018
29. Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Natl Acad Sci U S A 110:E1594–E1603. https://doi.org/10.1073/pnas.1211371110
30. Halary S, Leigh JW, Cheaib B et al (2010) Network analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci U S A 107:127–132. https://doi.org/10.1073/pnas.0908978107
31. Popa O, Hazkani-Covo E, Landan G et al (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res 21:599–609. https://doi.org/10.1101/gr.115592.110
32. Kloesges T, Popa O, Martin W, Dagan T (2011) Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol Biol Evol 28:1057–1074. https://doi.org/10.1093/molbev/msq297
33. Jaffe AL, Corel E, Pathmanathan J et al (2016) Bipartite graph analyses reveal interdomain LGT involving ultrasmall prokaryotes and their divergent, membrane-related proteins. Environ Microbiol 18:5072–5081. https://doi.org/10.1111/1462-2920.13477
34. Dagan T (2011) Phylogenomic networks. Trends Microbiol 19:483–491. https://doi.org/10.1016/j.tim.2011.07.001
35. Popa O, Landan G, Dagan T (2017) Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. ISME J 11:543–554. https://doi.org/10.1038/ismej.2016.116
36. Fondi M, Fani R (2010) The horizontal flow of the plasmid resistome: clues from intergeneric similarity networks. Environ Microbiol 12:3228–3242. https://doi.org/10.1111/j.1462-2920.2010.02295.x

37. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R (2008) Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol 25:762–777. https://doi.org/10.1093/molbev/msn023
38. Iranzo J, Krupovic M, Koonin EV (2016) The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. MBio 7:e00978–e00916. https://doi.org/10.1128/mBio.00978-16
39. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637
40. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36. https://doi.org/10.1093/nar/28.1.33
41. Huson DH, Scornavacca C (2011) A survey of combinatorial methods for phylogenetic networks. Genome Biol Evol 3:23–35. https://doi.org/10.1093/gbe/evq077
42. Huson DH, Rupp R, Scornavacca C (2011) Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press, New York, NY
43. Nakhleh L (2011) Evolutionary phylogenetic networks: models and issues. In: Problem solving handbook in computational biology and bioinformatics. Springer, New York, pp 125–158
44. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90. https://doi.org/10.1038/47056
45. Pasternak G, Hochhaus A, Schultheis B, Hehlmann R (1998) Chronic myelogenous leukemia: molecular and cellular aspects. J Cancer Res Clin Oncol 124:643–660
46. Watanabe H, Otsuka J (1995) A comprehensive representation of extensive similarity linkage between large numbers of proteins. Bioinformatics 11:159–166. https://doi.org/10.1093/bioinformatics/11.2.159
47. Park J, Teichmann SA, Hubbard T, Chothia C (1997) Intermediate sequences increase the detection of homology between sequences. J Mol Biol 273:349–354. https://doi.org/10.1006/jmbi.1997.1288
48. Bolten E, Schliep A, Schneckener S et al (2001) Clustering protein sequences--structure prediction by transitive homology. Bioinformatics 17:935–941. https://doi.org/10.1093/bioinformatics/17.10.935
49. Bapteste E, Lopez P, Bouchard F et al (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. Proc Natl Acad Sci U S A 109:18266–18272. https://doi.org/10.1073/pnas.1206541109
50. Jachiet P-A, Pogorelcnik R, Berry A et al (2013) MosaicFinder: identification of fused gene families in sequence similarity networks. Bioinformatics 29:837–844. https://doi.org/10.1093/bioinformatics/btt049
51. Méheust R, Zelzion E, Bhattacharya D et al (2016) Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proc Natl Acad Sci U S A 113:3579–3584. https://doi.org/10.1073/pnas.1517551113
52. Halary S, McInerney JO, Lopez P, Bapteste E (2013) EGN: a wizard for construction of gene and genome similarity networks. BMC Evol Biol 13:146. https://doi.org/10.1186/1471-2148-13-146
53. Martin AJM, Walsh I, Di Domenico T et al (2013) PANADA: protein association network annotation, determination and analysis. PLoS One 8:e78383. https://doi.org/10.1371/journal.pone.0078383
54. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303
55. Bastian M, Heymann S, Jacomy M (2009) Gephi: an Open source software for exploring and manipulating networks. Third Int AAAI Conf Weblogs Soc Media. pp 361–362. https://doi.org/10.1136/qshc.2004.010033
56. Csárdi G, Nepusz T (2006) The igraph software package for complex network research. InterJ Complex Syst 1695
57. Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J (eds) Proc. 7th Python Sci. Conf, Pasadena, CA, pp 11–15
58. Camacho C, Coulouris G, Avagyan V et al (2009) BLAST+: architecture and applications. BMC Bioinform 10:421. https://doi.org/10.1186/1471-2105-10-421
59. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool.pdf. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
60. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12:656–664.

https://doi.org/10.1101/gr.229202. Article published online before March 2002
61. Vaser R, Pavlović D, Šikić M (2016) SWORD—a highly efficient protein database search. Bioinformatics 32:i680–i684. https://doi.org/10.1093/bioinformatics/btw445
62. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461
63. Buchfink B, Xie C, Huson DH (2014) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176
64. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
65. Ye Y, Choi J-H, Tang H (2011) RAPSearch: a fast protein similarity search tool for short reads. BMC Bioinform 12:159. https://doi.org/10.1186/1471-2105-12-159
66. Page L, Brin S, Motwani R, Winograd T (1998) The PageRank citation ranking: bringing order to the web. Technical Report. Stanford InfoLab
67. Brandes U (2001) A faster algorithm for betweenness centrality*. J Math Sociol 25:163–177. https://doi.org/10.1080/0022250X.2001.9990249
68. Staudt CL, Sazonovs A, Meyerhenke H (2016) NetworKit: a tool suite for large-scale complex network analysis. Network Science 4 (4):508–530. https://doi.org/10.1017/nws.2016.20
69. Teng S-H (2016) Scalable algorithms for data and network analysis. Now Publishers Inc, Hanover, MA
70. Dayhoff MO (1976) The origin and evolution of protein superfamilies. Fed Proc 35:2132–2138
71. Heger A, Holm L (2000) Towards a covering set of protein family profiles. Prog Biophys Mol Biol 73:321–337. https://doi.org/10.1016/S0079-6107(00)00013-4
72. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci U S A 99:7821–7826. https://doi.org/10.1073/pnas.122653799
73. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp. https://doi.org/10.1088/1742-5468/2008/10/P10008
74. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584. https://doi.org/10.1093/nar/30.7.1575
75. Altenhoff AM, Kunca N, Glover N et al (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res 43:D240–D249. https://doi.org/10.1093/nar/gku1158
76. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503
77. Dessimoz C, Cannarozzi G, Gil M et al (2005) OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. Springer, Berlin, pp 61–72
78. Dessimoz C, Boeckmann B, Roth ACJ, Gonnet GH (2006) Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. Nucleic Acids Res 34:3309–3316. https://doi.org/10.1093/nar/gkl433
79. Roth ACJ, Gonnet GH, Dessimoz C (2008) Algorithm of OMA for large-scale orthology inference. BMC Bioinform 9:518. https://doi.org/10.1186/1471-2105-9-518
80. Altenhoff AM, Gil M, Gonnet GH et al (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS One 8:e53786. https://doi.org/10.1371/journal.pone.0053786
81. Schneider A, Dessimoz C, Gonnet GH (2007) OMA browser exploring orthologous relations across 352 complete genomes. Bioinformatics 23:2180–2182. https://doi.org/10.1093/bioinformatics/btm295
82. Miele V, Penel S, Duret L (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinform 12:116. https://doi.org/10.1186/1471-2105-12-116
83. Penel S, Arigon A-M, Dufayard J-F et al (2009) Databases of homologous gene families for comparative genomics. BMC Bioinform 10:S3. https://doi.org/10.1186/1471-2105-10-S6-S3
84. Dagan T, Roettger M, Bryant D, Martin W (2010) Genome networks root the tree of life between prokaryotic domains. Genome Biol Evol 2:379–392. https://doi.org/10.1093/gbe/evq025
85. Jacob F (1977) Evolution and tinkering. Science 196:1161–1166
86. Marcotte EM, Pellegrini M, Ng HL et al (1999) Detecting protein function and

protein-protein interactions from genome sequences. Science 285:751–753

87. Kawai H, Kanegae T, Christensen S et al (2003) Responses of ferns to red light are mediated by an unconventional photoreceptor. Nature 421:287–290. https://doi.org/10.1038/nature01310
88. Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20:1313–1326. https://doi.org/10.1101/gr.101386.109
89. Marsh JA, Teichmann SA (2010) How do proteins gain new domains? Genome Biol 11:126. https://doi.org/10.1186/gb-2010-11-7-126
90. Promponas VJ, Ouzounis CA, Iliopoulos I (2014) Experimental evidence validating the computational inference of functional associations from gene fusion events: a critical survey. Brief Bioinform 15:443–454. https://doi.org/10.1093/bib/bbs072
91. McLysaght A, Guerzoni D (2015) New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc B Biol Sci 370:20140332. https://doi.org/10.1098/rstb.2014.0332
92. Enright AJ, Ouzounis CA (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 16:451–457. https://doi.org/10.1093/bioinformatics/16.5.451
93. Snel B, Bork P, Huynen M (2000) Genome evolution. Gene fusion versus gene fission. Trends Genet 16:9–11
94. Enright AJ, Ouzounis CA (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol 2:RESEARCH0034
95. Patthy L (2003) Modular assembly of genes and the evolution of new functions. Genetica 118:217–231
96. Nakamura Y, Itoh T, Martin W (2007) Rate and polarity of gene fusion and fission in Oryza sativa and Arabidopsis thaliana. Mol Biol Evol 24:110–121. https://doi.org/10.1093/molbev/msl138
97. Ekman D, Björklund ÅK, Elofsson A (2007) Quantification of the elevated rate of domain rearrangements in metazoa. J Mol Biol 372:1337–1348. https://doi.org/10.1016/j.jmb.2007.06.022
98. Jachiet P-AA, Colson P, Lopez P, Bapteste E (2014) Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. Genome Biol Evol 6:2195–2205. https://doi.org/10.1093/gbe/evu168
99. Song N, Joseph JM, Davis GB et al (2008) Sequence similarity network reveals common ancestry of multidomain proteins. PLoS Comput Biol 4:e1000063. https://doi.org/10.1371/journal.pcbi.1000063
100. Yanai I, Derti A, DeLisi C (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc Natl Acad Sci U S A. https://doi.org/10.1073/pnas.141236298
101. Pathmanathan JS, Lopez P, Lapointe F-J, Bapteste E (2018) CompositeSearch: a generalized network approach for composite gene families detection. Mol Biol Evol 35:252–255. https://doi.org/10.1093/molbev/msx283
102. Dorrell RG, Gile G, McCallum G et al (2017) Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. elife. https://doi.org/10.7554/eLife.23717
103. Martin W, Stoebe B, Goremykin V et al (1998) Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393:162–165. https://doi.org/10.1038/30234
104. Martin W, Rujan T, Richly E et al (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A 99:12246–12251. https://doi.org/10.1073/pnas.182432999
105. Reyes-Prieto A, Hackett JD, Soares MB et al (2006) Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. Curr Biol. https://doi.org/10.1016/j.cub.2006.09.063
106. Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2008) Statistical properties of community structure in large social and information networks. In: Proceeding 17th Int. Conf. World Wide Web - WWW '08. ACM Press, New York, p 695
107. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67:26126. https://doi.org/10.1103/PhysRevE.67.026126
108. Newman M (2010) Networks. An introduction. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001
109. Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol

57:369–394. https://doi.org/10.1146/annurev.micro.57.030502.090759
110. Williams TA, Embley TM (2014) Archaeal? Dark matter? And the origin of eukaryotes. Genome Biol Evol 6:474–481. https://doi.org/10.1093/gbe/evu031
111. Castelle CJJ, Wrighton KCC, Thomas BCC et al (2015) Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr Biol 25:690–701. https://doi.org/10.1016/j.cub.2015.01.014
112. Brown CT, Hug LA, Thomas BC et al (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523:208–211. https://doi.org/10.1038/nature14486
113. Spang A, Saw JH, Jørgensen SL et al (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173–179. https://doi.org/10.1038/nature14447
114. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH et al (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358. https://doi.org/10.1038/nature21031
115. Prakash T, Taylor TD (2012) Functional assignment of metagenomic data: challenges and applications. Brief Bioinform 13:711–727. https://doi.org/10.1093/bib/bbs033
116. Hingamp P, Grimsley N, Acinas SG et al (2013) Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. ISME J 7:1678–1695. https://doi.org/10.1038/ismej.2013.59
117. de Vargas C, Audic S, Henry N et al (2015) Eukaryotic plankton diversity in the sunlit ocean. Science 348:1261605–1261605. https://doi.org/10.1126/science.1261605
118. Sunagawa S, Coelho LP, Chaffron S et al (2015) Structure and function of the global ocean microbiome. Science 348:1261359–1261359. https://doi.org/10.1126/science.1261359
119. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA et al (2016) Uncovering earth's virome. Nature 536:425–430. https://doi.org/10.1038/nature19094
120. Shi M, Lin XD, Tian JH et al (2016) Redefining the invertebrate RNA virosphere. Nature. https://doi.org/10.1038/nature20167
121. van Regenmortel MH, Mayo MA, Fauquet CM, Maniloff J (2000) Virus nomenclature: consensus versus chaos. Arch Virol 145:2227–2232
122. Gibbs AJ (2000) Virus nomenclature descending into chaos. Arch Virol 145:1505–1507
123. Lawrence JG, Hatfull GF, Hendrix RW (2002) Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol 184:4891–4905
124. Franklin LR (2007) Bacteria, sex, and systematics. Philos Sci 74:69–95. https://doi.org/10.1086/519476
125. Bapteste E, Boucher Y (2008) Lateral gene transfer challenges principles of microbial systematics. Trends Microbiol 16:200–207. https://doi.org/10.1016/j.tim.2008.02.005
126. Bapteste E, O'Malley MA, Beiko RG et al (2009) Prokaryotic evolution and the tree of life are two different things. Biol Direct 4:34. https://doi.org/10.1186/1745-6150-4-34
127. Andam CP, Williams D, Gogarten JP (2010) Natural taxonomy in light of horizontal gene transfer. Biol Philos 25:589–602. https://doi.org/10.1007/s10539-010-9212-8
128. Koonin EV, Dolja VV (2014) Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol Mol Biol Rev 78:278–303. https://doi.org/10.1128/MMBR.00049-13
129. Lederberg J, Tatum EL (1946) Gene recombination in Escherichia coli. Nature 158:558
130. Zinder ND, Lederberg J (1952) Genetic exchange in Salmonella. J Bacteriol 64:679–699
131. Levin BR (1988) Frequency-dependent selection in bacterial populations. Philos Trans R Soc Lond B Biol Sci 319:459–472
132. Rodriguez-Valera F (2004) Environmental genomics, the big picture? FEMS Microbiol Lett 231:153–158
133. Chen I, Christie PJ, Dubnau D (2005) The ins and outs of DNA transfer in bacteria. Science 310:1456–1460. https://doi.org/10.1126/science.1114021
134. Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3:504–510. https://doi.org/10.1038/nrmicro1163
135. Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 3:722–732. https://doi.org/10.1038/nrmicro1235
136. Dagan T, Martin W (2009) Getting a better picture of microbial evolution en route to a network of genomes. Philos Trans R Soc Lond B Biol Sci 364:2187–2196. https://doi.org/10.1098/rstb.2009.0040

137. Kulp A, Kuehn MJ (2010) Biological functions and biogenesis of secreted bacterial outer membrane vesicles. Annu Rev Microbiol 64:163–184. https://doi.org/10.1146/annurev.micro.091208.073413
138. McDaniel LD, Young E, Delaney J et al (2010) High frequency of horizontal gene transfer in the oceans. Science 330:50. https://doi.org/10.1126/science.1192243
139. Dubey GP, Ben-Yehuda S (2011) Intercellular nanotubes mediate bacterial communication. Cell 144:590–600. https://doi.org/10.1016/j.cell.2011.01.015
140. Desnues C, La Scola B, Yutin N et al (2012) Provirophages and transpovirons as the diverse mobilome of giant viruses. Proc Natl Acad Sci U S A 109:18078–18083. https://doi.org/10.1073/pnas.1208835109
141. Kutschera VE, Bidon T, Hailer F et al (2014) Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. Mol Biol Evol 31:2004–2017. https://doi.org/10.1093/molbev/msu186
142. Simmonds P (2014) Methods for virus classification and the challenge of incorporating metagenomic sequence data. J Gen Virol. https://doi.org/10.1099/jgv.0.000016
143. Iranzo J, Koonin EV, Prangishvili D, Krupovic M (2016) Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsid-less mobile elements. J Virol 90:11043–11055. https://doi.org/10.1128/JVI.01622-16
144. Lord E, Le Cam M, Bapteste É et al (2016) BRIDES: a new fast algorithm and software for characterizing evolving similarity networks using breakthroughs, roadblocks, impasses, detours, equals and shortcuts. PLoS One 11:e0161474. https://doi.org/10.1371/journal.pone.0161474
145. van Dongen SM (2001) Graph clustering by flow simulation. PhD thesis, University of Utrecht
146. Borgatti SP, Everett MG (1997) Network analysis of 2-mode data. Soc Netw 19:243–269. https://doi.org/10.1016/S0378-8733(96)00301-2
147. Goh K-I, Cusick ME, Valle D et al (2007) The human disease network. Proc Natl Acad Sci U S A 104:8685–8690. https://doi.org/10.1073/pnas.0701361104
148. Himmelstein DS, Baranzini SE, Rand V et al (2015) Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. PLoS Comput Biol 11:e1004259. https://doi.org/10.1371/journal.pcbi.1004259
149. Ahn Y-Y, Ahnert SE, Bagrow JP et al (2011) Flavor network and the principles of food pairing. Sci Rep 1:196. https://doi.org/10.1038/srep00196
150. Lanza VF, Baquero F, de la Cruz F, Coque TM (2017) AcCNET (Accessory Genome Constellation Network): comparative genomics software for accessory genome analysis using bipartite networks. Bioinformatics 33:283–285. https://doi.org/10.1093/bioinformatics/btw601
151. Barber MJ (2007) Modularity and community detection in bipartite networks. Phys Rev E 76:66102. https://doi.org/10.1103/PhysRevE.76.066102
152. Beckett SJ (2016) Improved community detection in weighted bipartite networks. R Soc Open Sci 3:140536. https://doi.org/10.1098/rsos.140536
153. Diestel R (2010) Graph theory. Springer, New York
154. McInerney JO, Pisani D, Bapteste E, O'Connell MJ (2011) The public goods hypothesis for the evolution of life on Earth. Biol Direct 6:41. https://doi.org/10.1186/1745-6150-6-41
155. Hauser M, Mayer CE, Söding J (2013) kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinform 14:248. https://doi.org/10.1186/1471-2105-14-248

3.5 Des graphes pour étudier la diversité d'eucaryotes marins unicellulaires phylogénétiquement proches des animaux.

En 2015, Forster et al. ont utilisé une approche basée sur les réseaux pour décrire la diversité et la répartition géographique des ciliés (Forster et al. 2015).

Avec le Dr. Alicià Arroyo-Sanchez du laboratoire d'Iñaki Ruiz-Trillo (Université de Barcelone, Espagne), nous avons développé cette approche pour étudier les Holozoa unicellulaires marins, en utilisant un jeu de données de métabarcoding (région V9 du 18S rRNA) produit par le consortium TARA (Sunagawa et al. 2015). Le clade des Holozoa est un groupe qui inclut les animaux et comprend plusieurs lignées d'eucaryotes unicellulaires (Fig. 18).

Fig. 18. Exemple de la diversité des Holozoa
A gauche *Salpingoeca sp* un choanoflagellé, à droite une *Ciona roulii*, un Tunicier. [1]

L'étude des Holozoa est donc importante pour comprendre l'origine des animaux et de la multicellularité. Notre objectif était de décrire la diversité des Holozoa, et de trouver des nouvelles lignées appartenant ou non aux groupes d'Holozoa déjà connus. Une des

[1] Kichigin / Shutterstock.com - https://doris.ffessm.fr/

difficultés de cette étude tenait dans le peu d'information phylogénétique disponible pour la description de la diversité des Holozoa, en raison de la petite taille du V9 amplifié par le consortium TARA. Nous nous sommes servis des réseaux de similarité de séquences pour extraire un maximum d'information malgré la taille réduite de ce marqueur. Nous avons notamment utilisé des métriques associées aux réseaux et particulièrement des mesures de centralité et d'association. Cela nous a permis d'identifier les séquences les plus pertinentes pour décrire une nouvelle diversité chez les Holozoa : des séquences excentriques dans le réseau de similarité et éloignées des séquences de référence. Nos résultats mettent en lumière une diversité d'Holozoa marins qui n'est pas encore décrite et dont on ne connaît pour l'instant que la séquence V9 du 18S RNA. De plus, nous prédisons l'existence de relations symbiotiques ou parasitaires non décrites impliquant ces taxa. Enfin, ce travail a aussi permis de détecter une nouvelle lignée d'Holozoa que le Dr. Alicià Arroyo-Sanchez a eu l'honneur de nommer MASHOL.

3.5.1 Article 5, "Gene similarity networks from the Tara Oceans expedition unveil geographical distribution, ecological interactions, and novel diversity among unicellular relatives of animals", in prep.

Gene similarity networks from the *Tara* Oceans expedition unveil geographical distribution, ecological interactions, and novel diversity among unicellular relatives of animals

Alicia S. Arroyo[1*], Romain Lannes[2*], Colomban de Vargas,
Eric Bapteste[2♦] & Iñaki Ruiz-Trillo[1,3,4♦]

[1] Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta, 37-49, 08003 Barcelona, Spain

[2] Institut de Biologie Paris-Seine (IBPS), UPMC Université Paris 06, Sorbonne Universités, Paris, France

[3] Departament de Genètica, Microbiologia I Estadística, Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Avinguda Diagonal 643, 08028 Barcelona, Spain

[4] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

ABSTRACT

The Holozoa clade is comprised of animals and several unicellular lineages. Thus, understanding the full diversity of unicellular holozoans is essential to address the origins of animals and other evolutionary questions. However, the full diversity of these lineages is poorly known. In this study, we analysed 18S rDNA metabarcoding data from the global *Tara* Oceans expedition with the objective of finding new diversity within or between unicellular Holozoa lineages. We used similarity networks to overcome the low phylogenetic information contained in the metabarcoding dataset (composed of sequences from the short V9 region of the gene). We constructed similarity networks by combining two datasets: unknown environmental sequences from *Tara* Oceans and known reference sequences from GenBank, and blasting them all against all. We calculated network metrics to compare environmental to reference sequences. These metrics reflected the divergence between both types of sequences in a mathematical way and provided an effective way to mine the *Tara* Oceans dataset to search for evolutionary relevant new diversity, further validated by phylogenetic placements. Our results showed that unicellular holozoans from *Tara* Oceans were not similar to the extant references, expanding the known diversity of these lineages. Novelties were mainly found in Acanthoecida choanoflagellates, branching off several already described subgroups. We also found 21 OTUs that did not cluster to any other existing lineage and thus, could be a new holozoan group. Moreover, we also explored for the first time the geographical distribution of the extant holozoan lineages around the globe, and the ecological interactions they may have with animals. Results showed that, although ubiquitous, each lineage exhibited a different distribution pattern. We also checked for potential associations between unicellular Holozoa and animals, and identified a positive correlation between the abundance of new animal hosts and the ichthyosporean *Creolimax frangrantissima*, as well as for other holozoans that were previously reported as free-living. Overall our analyses provide a fresh perspective into the diversity and ecology of unicellular holozoans, highlighting the amount of undescribed diversity in this important clade of the tree of life.

Keywords

networks, metabarcoding, 18S, molecular diversity, unicellular Holozoa, novelty

INTRODUCTION

An important evolutionary question that feeds on our current knowledge of diversity is the origin of animals. To understand animal origins, we first need to have a good phylogenetic framework (Ruiz-Trillo *et al.*, 2007). We now know that animals are closely related to several unicellular lineages, namely Choanoflagellatea, Filasterea, and Ichthyosporea, all together forming the Holozoa clade (Lang *et al.*, 2002; Ruiz-Trillo *et al.*, 2008; Shalchian-Tabrizi *et al.*, 2008; Torruella *et al.*, 2015; Grau-Bové *et al.*, 2017). Therefore, to understand how the unicellular ancestor of animals looked like and how animals evolved from that ancestor, one must study unicellular holozoans. However, the real diversity of Holozoa is still mostly unknown (del Campo *et al.*, 2015; Arroyo *et al.*, 2018). Therefore, improving our knowledge about Holozoa diversity may change current interpretations on the evolutionary transition towards animal multicellularity (Ruiz-Trillo *et al.*, 2007; del Campo *et al.*, 2014).

To fill this gap and provide a more accurate perspective on the unicellular Holozoa diversity, we analysed the longest and largest metabarcoding marine dataset: the 18S ribosomal RNA gene (hereafter 18S or 18S rDNA) from the *Tara* Oceans expedition (de Vargas *et al.*, 2015; Pesant *et al.*, 2015), of which a third did not match any reference in databases (de Vargas *et al.*, 2015). A drawback of this dataset is the absence of full-length 18S sequences, being composed by the relatively small V9 region (around 130 bp long), located at the end of the 18S (Hugerth *et al.*, 2014). Because these short amplicons contain too little phylogenetic information to resolve phylogenies, we followed a different strategy to unravel new molecular diversity within Holozoa (Amaral-Zettler *et al.*, 2009).

To overcome the issue of the limited phylogenetic signal, we decided to analyse this dataset using gene similarity networks. Networks have been preferentially applied to study ecological interactions, such as predator-prey, parasite-host or mutualism (Logares *et al.*, 2014; Krabberød *et al.*, 2017; Layeghifard *et al.*, 2017; Pilosof *et al.*, 2017; Valverde *et al.*, 2018). They are now becoming widely adopted to explain complex evolutionary processes, such as horizontal gene transfer, gene domain fusion, and gene or genome introgression (Corel *et al.*, 2016; Pathmanathan *et al.*, 2018; Ocaña-Pallarès *et al.*, 2019). To our knowledge, there are very few metabarcoding studies that used networks to describe

novelty in metabarcoding datasets (Forster *et al.*, 2015), even though this methodology offers a structure to test evolutionary questions in massive high-throughput data and to mine large datasets for sequences of interest.

The main objective of our analysis was to find novelty along unicellular holozoans. Moreover, we also analysed the geographical distribution of unicellular Holozoa and looked for co-occurrence patterns between some unicellular Holozoa parasitic lineages and their corresponding animal hosts.

We detected novel unicellular Holozoa diversity, in particular within Choanoflagellatea and Ichthyosporea. Specifically, we found unicellular Holozoa Operational Taxonomic Units (OTUs) branching off several acanthoecid subgroups (for example Choanoflagellate H), *Syssomonas multiformis* and *Creolimax fragrantissima*. We also retrieved 15 Filasterea-related OTUs, detecting this clade for very first time in an environmental survey. Interestingly, we also identified a putative novel unicellular Holozoa group that could not be located within any other known lineage. This clade, that we tentatively named as MASHOL (for MArine Small HOLozoa clade), was composed of 21 OTUs (6,244 reads in total). We also observed that the freshwater environmental group FRESCHO3 could have diverged from a marine clade, showing another marine-to-freshwater transition in choanoflagellates. Finally, our results suggested novel associations between animals and ichthyosporeans. For example, the ichthyosporean *C. fragrantissima* could be associated with a wider range of animal hosts than previously described. Other associations were identified between the environmental clades marine ichthyosporeans 1 (MAIP1) and marine opisthokonts 2 (MAOP2), and different animal phyla, adding other ecological dynamics to the unicellular relatives of animals.

RESULTS AND DISCUSSION

Initial datasets & network construction

The main objective of this study was to look for potential new diversity of unicellular Holozoa and to address, for the first time, the geographical distribution of the clade around the globe. We used metabarcoding data from the V9 region of the 18S rRNA gene. We combined two datasets: an environmental dataset of OTUs (Operational Taxonomic Units) and a

reference dataset with known holozoan sequences. The environmental dataset came from the worldwide *Tara* Oceans expedition (de Vargas 2015), which included a total of 1,086 samples from 210 oceanic stations, 3 water column layers and 10 size fractions (further details about sampling procedures can be found in Pesant *et al.*, 2015). The reference dataset was built by collecting sequences from both GenBank Nucleotide and PR2 databases (see Materials and Methods).

The initial unicellular Holozoa network was built from 2,426 sequences (2,197 from *Tara* Oceans, 229 from the reference dataset). In the network, each node represented either an environmental OTU from *Tara* Oceans (hereafter ENV) or a sequence from the reference database (hereafter REF) (**Figure 1**). The basic structure of the network consisted of Connected Components (CCs): subgraphs of the network in which there is always a path between all nodes (**Figure 1**). The initial network was subsequently partitioned using increasing sequence similarity thresholds (≥85%, ≥87%, ≥90%, ≥95% and ≥97%), resulting in a more fragmented network (**Figure 1**). CCs could be classified in three types: CCs in which all nodes were environmental (CC_{ENV}), CC in which all nodes were reference (CC_{REF}) and CC in which there were both types of nodes (CC_{MIX}).

The topology of the network was constant in all thresholds, meaning that the number of CC_{ENV} was always the largest, followed by a CC_{MIX} and CC_{REF} (**Supplementary Figure 1**), which indicated the presence of abundant divergent groups of environmental sequences.

Definition of novelty

We explored the network structure to search for molecular diversity. To do so, we calculated different metrics that are grouped in four categories:

I. ***Closeness centrality*** (**Figure 1** and Supplementary Material 1): It defines to which extent a node (sequence) is central in the network. Typically, a peripheral sequence is more divergent than the rest of the nodes in the network because it shares less similarity. Therefore, we tested whether and which environmental sequences (ENV) were significantly more peripheral than reference sequences (REF), since this suggests that those ENV sequences extends the current known diversity of Holozoa.

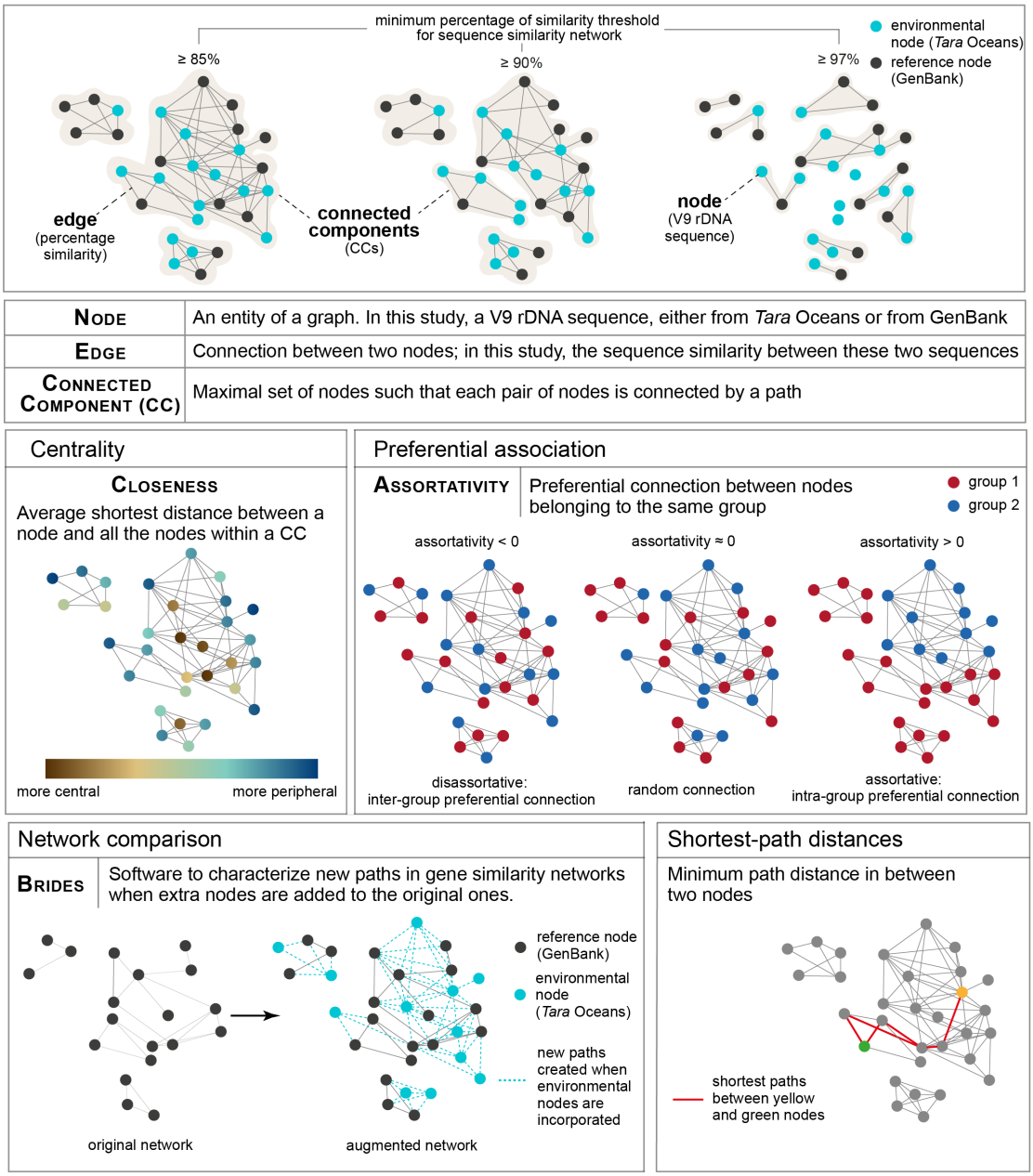


Figure 1. Network metrics. Upper panel: once the unicellular Holozoa network was constructed, different similarity thresholds were applied to gain a more detailed structure of their diversity. Lower panels: network metrics computed in this study to address molecular novel diversity in unicellular Holozoa. A more technical explanation of closeness and assortativity can be found in Supplementary Material 1, and of BRIDES in Supplementary Figure 2.

II. ***Preferential association*** (***Assortativity,*** **Figure 1** and Supplementary Material 1): Assortativity quantifies whether nodes that belong to the same category are more connected with each other rather than with nodes from other categories. For example, a significant preferential association between ENV nodes would indicate the existence of groups of similar environmental sequences, distinct from sequences from already described Holozoa.

III. ***Network comparison (path analyses by BRIDES,*** (**Figure 1** and **Supplementary Figure 2*)***: It quantifies the new paths created in an augmented network when new sequences (e.g. ENV) are added to an original network (with only REF), as in Lord *et al.*, 2016. In particular, this allows the evaluation of whether newly added sequences fill in some gaps between the original sequences (B and S paths indicating that added sequences are intermediate diversity with respect to known sequences) or fail to do so (the I path indicating that added sequences do not present such an intermediate diversity).

IV. ***Shortest-path distance*** (**Figure 1**): Shortest paths describe the minimal number of edges between any pairs of nodes in a network. We used these metrics to quantify the distance between ENV and REF nodes in the graph. By definition, increasingly divergent ENV sequences will be located increasingly far from REF sequences. If ENV and REF sequences are located in distinct CCs, there is no path between them, thus the shortest path distance for such pairs of nodes is infinite.

All these steps of graph-mining pointed towards evolutionary relevant ENV sequence candidates, for which phylogenetic placement could be finally computed (see Materials and Methods).

The structure of the unicellular Holozoa network shows potential new diversity

The general structure of the network provided an overview of the unicellular Holozoa diversity and the potential new diversity (**Figure 2**).

First, we computed the closeness of all nodes (**Figure 1**, **Figure 3** and Supplementary Material 1) to test

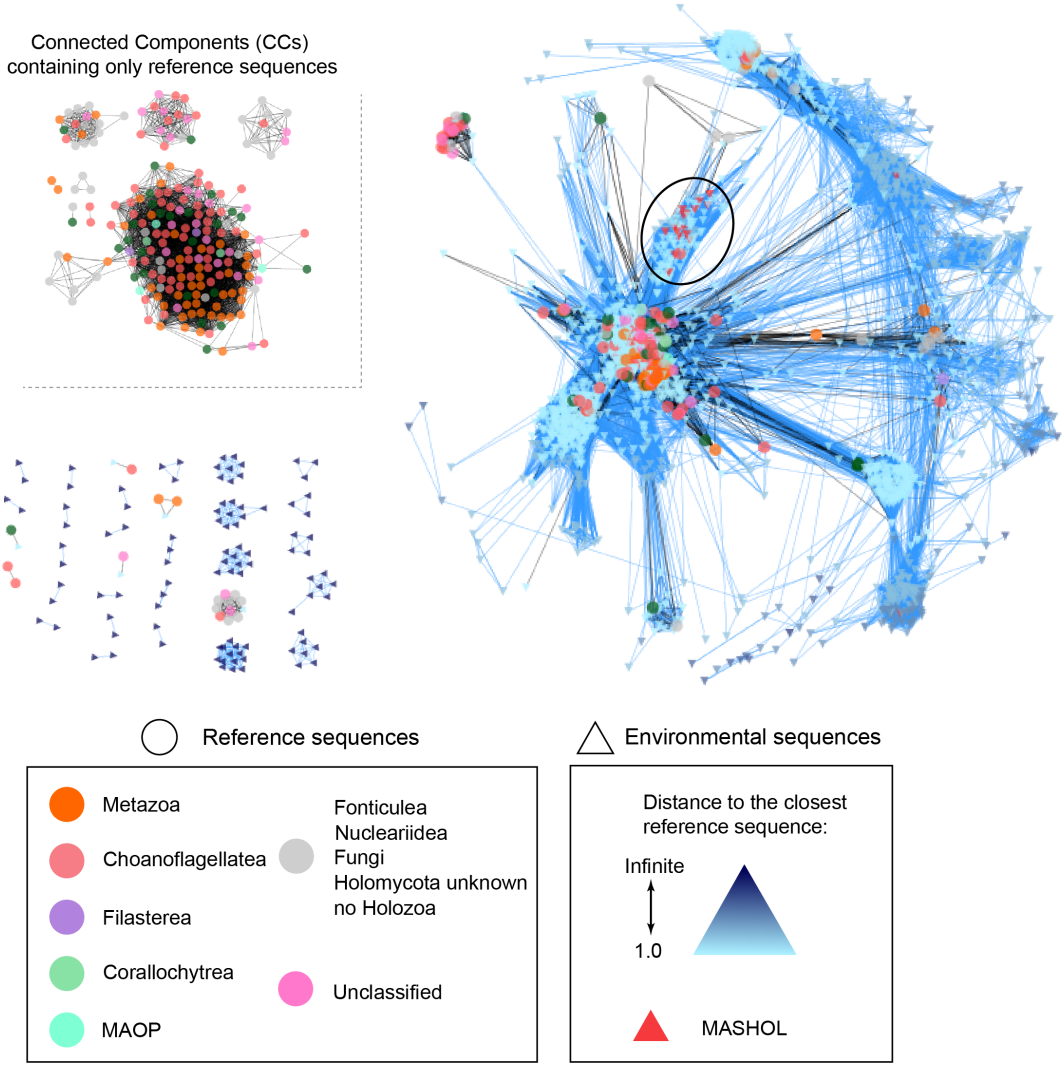


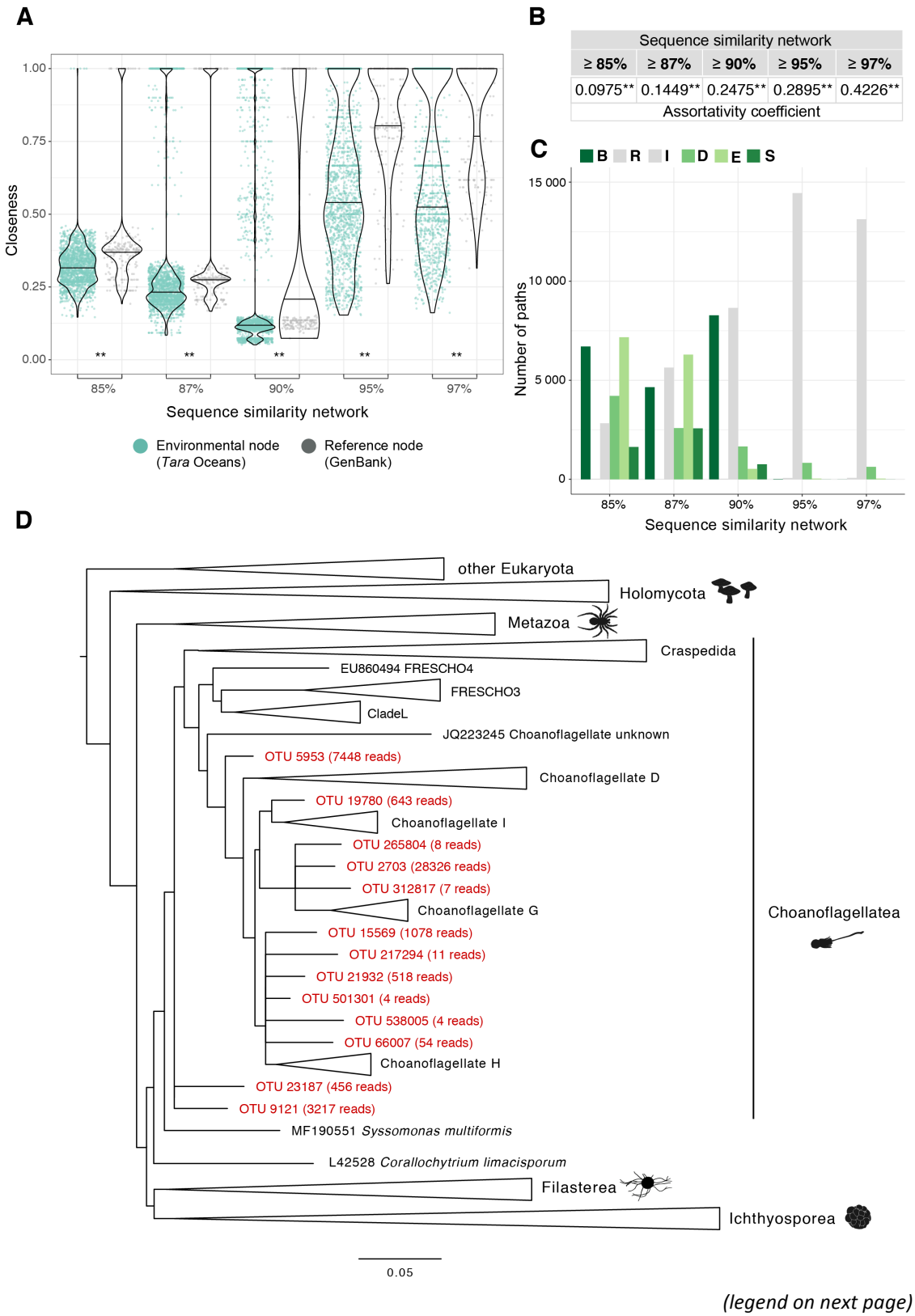
Figure 2. Unicellular Holozoa network at ≥85% similarity threshold. Environmental nodes from *Tara* Oceans are depicted with triangles that are coloured according to the distance to their shortest reference sequence (right panel). Reference nodes from GenBank dataset are depicted with circles that are coloured according to the taxaonomy (left panel). Connected Components composed of only reference nodes are located in the top right corner. The novel Holozoa group described in this paper, MASHOL (for MArine Small HOLozoa), is shown in red triangles and pointed in the network with a black circle. Raw network data can be found in Supplementary Material 4.

whether the distribution of closeness values for REF nodes was (i) significantly different and (ii) significantly higher than the distribution of closeness values for ENV nodes, using Wilcoxon signed-rank test. The results showed that ENV nodes were significantly more peripheral than REF nodes (Wilcoxon signed-rank test, p-value<0.01**) (**Figure 3A**) in all networks. This result indicates a high amount of potential new diversity in our unicellular Holozoa dataset from *Tara* Oceans. Not only the closeness distributions for REF nodes

were significantly higher than that for ENV nodes, but also their shapes were different. At ≥85, ≥87 and ≥90% similarity thresholds, most closeness values of both ENV and REF distributions were low (95% confident interval between 0.2-0.4, approximately), and only few nodes presented a closeness value of 1. On the other hand, at ≥95 and ≥97% thresholds, when the network was more disconnected, the distributions of closeness values for ENV nodes were scattered along a wider range of higher closeness values (~0.2-1). This change reflected the fragmentation of the network into more, but smaller CCs.

Next, we analysed the preferential connection between ENV nodes, which showed greater similarity between ENV sequences than between ENV and REF sequences. For every network, we computed (i) a distribution of null assortativity values by randomly shuffling the ENV and REF node labels and we contrasted these values with (ii) the assortativity values of all our real networks (see Materials and Methods). All networks were significantly assortative (one sample t-test, p-value<0.01**) (**Figure 3B**). This tendency for intra-group preferential linkage suggests a lack of representation of oceanic Holozoa in the reference dataset before the *Tara* Ocean expedition, and thus stresses the high level of potential new diversity in *Tara* Oceans.

Moreover, we checked is these results were not trivially explained by the unequal amount of ENV sequences (2,197) compared to REF sequences (229) in our initial dataset. Basically, we repeated the same analysis using networks constructed from the same number of ENV and REF nodes (see the "Control test" section in Materials and Methods and Supplementary Table). Regarding closeness, at ≥90%, ≥95%, and ≥97% similarity thresholds we obtained the same results than for real networks: the distributions of closeness values for REF nodes were significantly higher than that for ENV nodes. At lower



(legend on next page)

Figure 3. Network approach to the analysis of novel diversity of unicellular Holozoa. **(A) Closeness** distribution of reference nodes was significantly higher than that of environmental nodes. This showed that environmental nodes were located at the periphery of the connected components because they were more divergent. Two asterisks mark the significance of the Wilcoxon signed-rank test when p-value<0.01. **(B) Assortativity** values were significantly positive in all networks, meaning that environmental nodes tended to connect preferentially together rather than with reference nodes. **(C) BRIDES analysis**. Environmental OTUs from unicellular Holozoa created new paths with respect to the original reference network, as green bars show (see Supplementary Figure 2 for details about each type of path). **(D) New molecular groups in Choanoflagellata**. Phylogenetic placement of the OTUs that created breakthroughs and shortcuts at ≥85% similarity threshold in (C; in red) against a curated reference tree of unicellular Holozoa. We computed the placement using the RAxML-EPA algorithm with the GTR+CAT+I evolutionary model (Berger *et al.,* 2011). Several OTUs branched off some acanthoecid clades, such as Choanoflagellate I, G and H, showing a different diversity from the extant known species. This novel molecular diversity is well supported by the high abundance of some OTUs (shown as the number in the brackets) and the good quality of their placement (Supplementary Figure 3A,B). Alignments and the full phylogenetic tree can be found in Supplementary Material 2.

thresholds, however, these differences were usually non-significant, but the closeness values of REF nodes were never lower to that of ENV nodes in the evenly sampled networks. Assortativity values were also positive and significantly different for all control networks, as in the actual networks.

Overall, these metrics (closeness and assortativity) indicated that our environmental dataset of unicellular holozoans from *Tara* Oceans expanded the current known diversity of this group. Moreover, we proved that these results were not an artefact from the unequal number of ENV sequences compared to REF. We then mined this molecular diversity to uncover evolutionary relevant Holozoa groups.

Identification of a potential novel Holozoa group. New molecular diversity found in Acanthoecida (Choanoflagellatea)

To identify new groups of interest, we first performed network comparisons using BRIDES software (**Figure 1**, **Supplementary Figure 2**) (see Materials and Methods and Lord *et al.*, 2016). This allowed us to contrast the topologies of networks built exclusively from REF nodes (original networks) with that in which ENV nodes had been included (augmented networks).

BRIDES analysis showed that ENV sequences of unicellular Holozoa from *Tara* Oceans created numerous new paths in the augmented similarity networks (**Figure 3C**), guiding the discovery of evolutionary relevant novel

sequences. First, despite the enhanced molecular diversity provided by the *Tara* Oceans dataset, some REF nodes remained disconnected, indicating that the diversity of most ENV sequences was not intermediate with respect to some REF nodes. This was especially noticeable for networks built at high similarity thresholds. At ≥97% ID, the vast majority of paths were *impasses* (I), meaning that ENV sequences did not create bridges between REF sequences in the augmented network (**Supplementary Figure 2**). This is logical because, given the high level of stringency, only sequences from the closest related holozoan lineages would connect in the CC, confirming the general divergent nature of ENV sequences. Interestingly, when lowering the similarity threshold required to connect sequences in the networks, the proportion of *impasses* decreased, showing that some of these divergent ENV sequences started to connect some REF sequences. Still, at ≥85%, some Holozoa REF sequences remained disconnected, indicating that the *Tara* Oceans dataset did not provide evidence for intermediate groups for all known Holozoa (i.e., in terms of diversity, there remained persistent gaps within the Holozoa tree). Possible explanations to this enormous amount of *impasses* may be: (i) a lack of sufficient sampling effort, (ii) the presence of intermediate ENV sequences in other habitats but not in the marine water column, or (iii) the nature of the Holozoa clade, which may be comprised of some significantly divergent lineages without intermediate diversity between them.

On the other hand, *breakthroughs* (B) and *shortcuts* (S) were increasingly observed in networks at lower thresholds (**Figure 3C**). These two types of paths correspond to sequences that introduce either new connections in the known diversity (B) or new intermediate sequences within known groups (S). Thus, B paths indicated which ENV sequences could possibly branch in between two groups in a phylogenetic tree, whereas S paths indicated which ENV sequences could possibly branch within a group (**Supplementary Figure 2**). Overall, the presence of a high proportion of B and S paths (≥85% = 36.93%, ≥87% = 33.22%, ≥90% = 45.42%) suggested that *Tara* Oceans data hinted at the existence of oceanic clades that could help to better resolve the Holozoa phylogeny.

We corroborated this putative new diversity performing a phylogenetic placement analysis (see Materials and Methods). We selected the OTUs that created *breakthroughs* and *shortcuts* in

the network at 85% similarity threshold (**Figure 3D**). These OTUs unravelled novelty within Acanthoecida, one of the two subgroups of Choanoflagellatea. A group of 6 sequences (with a total of 1,675 reads) branched off Choanoflagellate H, suggesting a potential novel environmental group of acanthoecids. Another group of 3 sequences (including one of the most abundant OTUs in the whole *Tara* Oceans dataset: OTU 2703, with more than 28,000 reads) appeared to be the sister group of Choanoflagellate G. The importance of this result lies in the fact that these OTUs do not cluster together with the already morphologically described Choanoflagellate G species (i.e., *Acanthocorbis unguiculata, Acanthoeca spectabilis*, *Savillea micropora*, *Helgoeca nana*), but branch at an internal node, showing the divergent nature of these OTUs. We also recovered the second earliest diverging acanthoecid (OTU 5953, with 7,448 reads), splitting differently from the reference sequence JQ223245, which was already identified as a divergent choanoflagellate (del Campo *et al.*, 2015). Finally, several OTUs were clustered in freshwater environmental choanoflagellate groups, such as FRESCHO3 or FRESCHO1, which shows a wider ecosystem range in which these species can inhabit. We confirmed the good quality of these placements gauging the likelihood and distance between placements (**Supplementary Figure 3A,B**). Alignments and the full phylogenetic tree of Figure 3D can be found in Supplementary Material 2.

Our second approach to examine in detail the novelty in unicellular Holozoa was performing shortest-path distance analysis between every ENV node and its closest REF node (**Figure 1**). The longer the distance, the more divergent the ENV sequence is, because many steps are required to reach the nearest REF sequence. The most extreme case is the infinite distance, shown by ENV nodes belonging to exclusively environmental CCs. Our results showed that indirect connections to REF (when there is more than 1 step from ENV to REF) were the most abundant, ranging from 92.5% of all ENV nodes at ≥85% similarity network to 69.83% at ≥97% (**Figure 4A**). In addition, networks at higher similarity thresholds (≥95% and ≥97%) exhibited a high proportion of infinite distances (15.39% of ENV nodes at ≥95% similarity threshold; 30.56% at ≥97% similarity threshold) (**Figure 4A**). We then extracted those OTUs to perform phylogenetic placement against a curated reference Holozoa tree (see

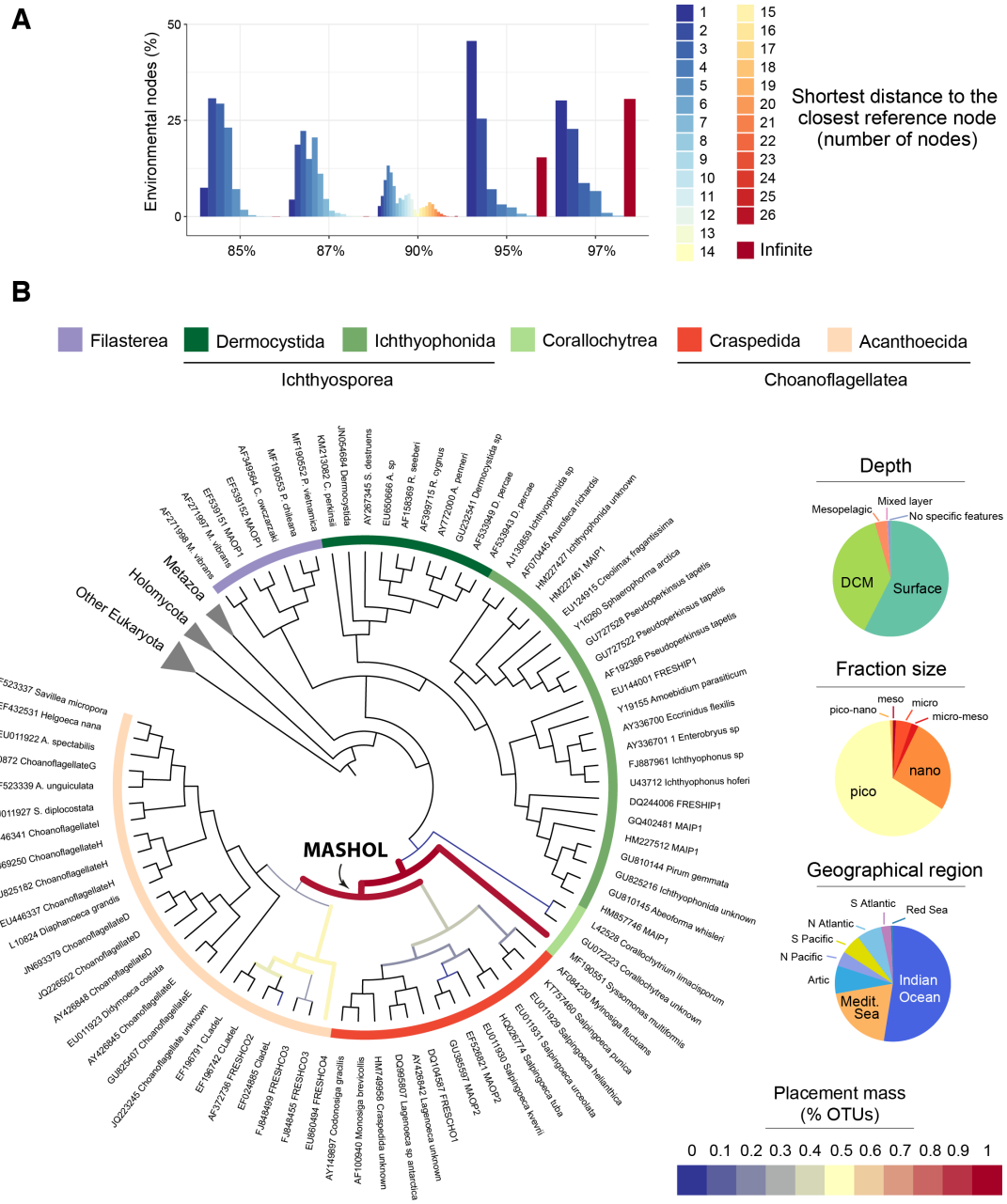


Figure 4. Potential new group of unicellular Holozoa (MASHOL) found branching off Choanoflagellatea. (A) Shortest path analysis showed that a considerable proportion of environmental nodes have infinite distance with their closest reference node (15.39% in the network at ≥95% similarity threshold; 30.56% in the network at ≥97%). These ENV nodes were not connected to any reference node whatsoever, suggesting a substantial amount of divergent diversity. **(B)** Phylogenetic placement of the 21 OTUs that exhibited infinite distance in the networks at ≥95% and ≥97% similarity thresholds in (A). All OTUs were allocated in internal branches, outside Choanoflagellatea and *Syssomonas multiformis*, depicted as a thick magenta line. The lack of high support (measured as Likelihood Weight Ratio or LWR) in the placements suggests a deep uncertainty about the exact placement of these sequences in the Holozoa tree of life (Supplementary Figure 3D). However, their narrow scattering over the tree and their clear position in internal rather than external branches open up the possibility for these OTUs to be a potential new Holozoa group that we tentatively named as MASHOL (for MArine Small HOLozoa). Phylogenetic placement was carried out using RAxML-EPA algorithm (Berger *et al.,* 2011) under the GTR+CAT+I evolutionary model. Alignments and the full phylogenetic tree can be found in Supplementary Material 3.

Materials and Methods). The deepest novelty (understood as the diversity that lays in internal nodes in the tree) was observed in the networks at ≥95% and ≥97% thresholds. We performed a specific phylogenetic placement of this deep novelty, shown in **Figure 4B**. A group of 21 OTUs with a total abundance of 6,244 reads was located in the most internal branch outside Choanoflagellata, specifically scattered across the internal branches of choanoflagellates and *Syssomonas multiformis*. These OTUs were mainly recovered in the pico (0.8-3/5 µm) and nano (3/5-20 µm) size fractions from the Indian Ocean and Mediterranean Sea. Inspired by its uncertain phylogenetic position and the small size, we tentatively named this group as MASHOL (standing for MArine Small HOLozoa). The quality of the placement test revealed that the placements had very low Likelihood Weight Ratio (**Supplementary Figure 3D**), although all of them were located around the same internal branches in the tree. As Mahé *et al.*, 2017 pointed out, these low-probability placements do not necessarily mean that they are incorrect, but they hold a high molecular distance with the reference sequences in the tree. This result indicates that these OTUs do not really belong to any of the already known unicellular holozoan lineages, although its exact position is deeply uncertain.

Unicellular holozoans are globally distributed, with some lineages showing specific geographical patterns

We next evaluated the geographical distribution of unicellular Holozoa across oceans, layers of the water column, and sizes.

In general, all lineages of unicellular Holozoa were widely distributed across the world's oceans (**Figure 5A**). Ichthyosporeans were the most homogeneously dispersed group across all oceans. There were, however, some exceptions. Within Choanoflagellata, for example, Acanthoecida OTUs were more abundant in the Arctic samples (60.29% of total abundance) compared to Craspedida (4.5%) (**Figure 5A**). These results are consistent with previous morphological studies of choanoflagellates in sea ice (Thomsen *et al.*, 1997), although these choanoflagellates were more extensively found in the Antarctic rather than in the Arctic oceans, as we observed. OTUs assigned to Filasterea were widely distributed, but their abundance was higher in the samples coming from the South Pacific Ocean (43.37%), Red Sea (24.7%) and Indian Ocean (16.97%)

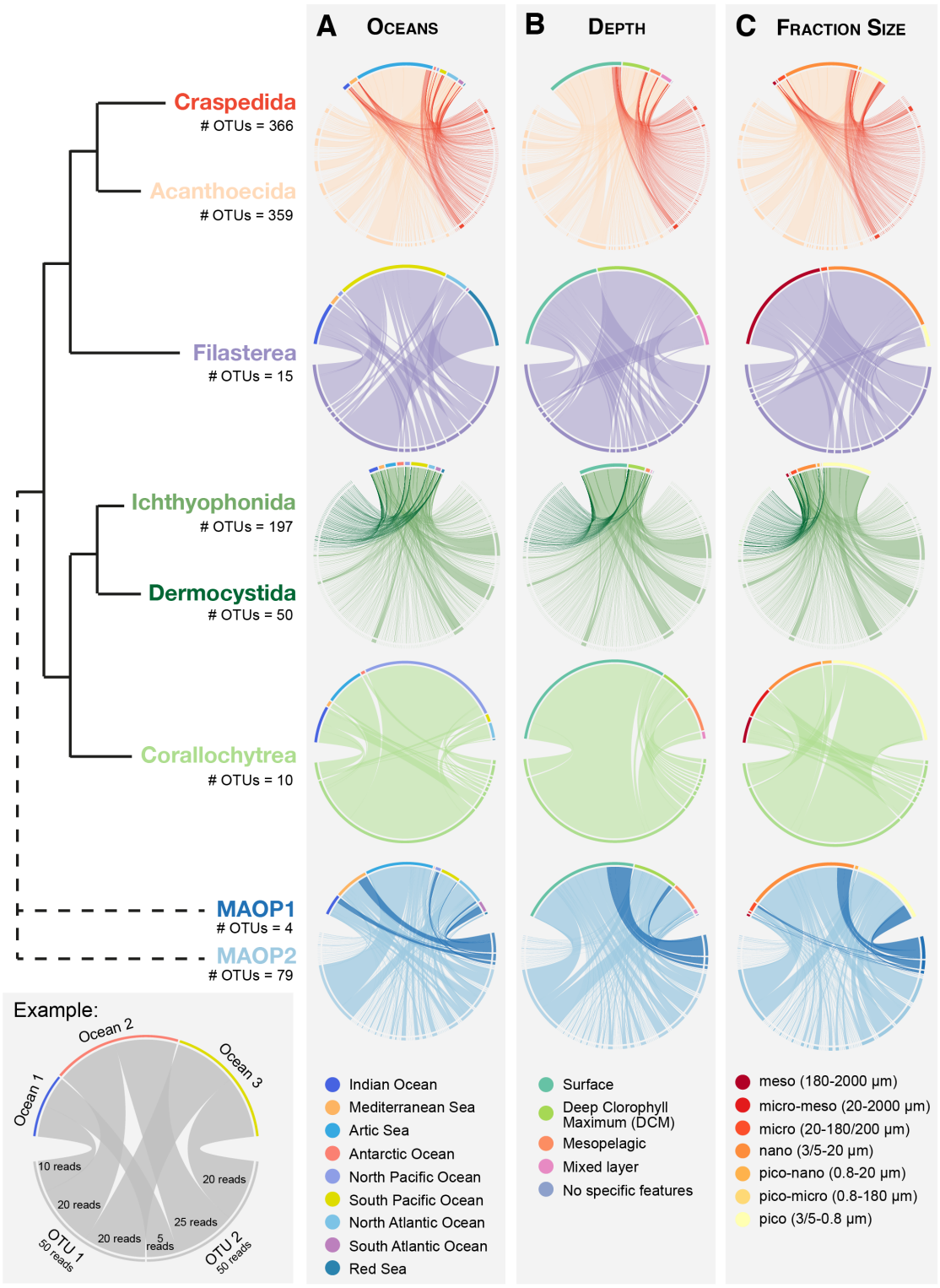


Figure 5. Geographical distribution of unicellular Holozoa OTUs from the *Tara* Oceans expedition. As depicted in the example (bottom left panel), chord diagrams show OTUs on the bottom half of the circle, and oceanic regions, depths, and fraction sizes on the upper half. Each OTU is represented by a line, whose thickness depicts the OTU's abundance in that particular place. In general, all unicellular holozoans were widespread and located in surface or DCM layers of the water column. However, some had different preferential geographical locations (i.e., MAOP1 *vs* MAOP2, or Craspedida *vs* Acanthoecida), or fraction sizes (i.e., Ichthyophonida *vs* Dermocystida, or Craspedida *vs* Acanthoecida). Note that the thickness of each OTU is relative to the amount of OTUs in each group, so comparisons between lineages are not possible. Numbers below group names indicate the number of OTUs.

Corallochytrea group were widely distributed, although the OTU with the highest abundance (OTU 30781, 248 reads) was mainly located in the North Pacific Ocean (**Figure 5A**). Both the Indian Ocean and the Arctic Ocean held 30% of the reads of corallochytreans (**Figure 5A**). On the contrary, the presence of corallochytreans in the Atlantic Ocean seemed to be insignificant. Regarding the environmental group of marine opisthokonts 1 and 2 (MAOP1 and MAOP2, respectively), they showed a pattern of distribution similar to Choanoflagellata. MAOP2 appeared to be most abundant and with more OTUs than MAOP1, in contrast to what had been found in coastal European waters (del Campo *et al.*, 2015). Moreover, while MAOP1 was not found in the Arctic or the Antarctic Oceans, MAOP2 exhibited 36% of its abundance in the Arctic, expanding to the maximum the range of geographical locations in which this environmental group has been found up to now (**Figure 5A**) (Romari and Vaulot, 2004; Amacher *et al.*, 2009; Edgcomb *et al.*, 2011; Marshall and Berbee, 2011). Assortativity coefficients of geographical distribution across oceans and oceanic provinces showed positive values in all networks (Supplementary Table). Even though these values were not very high (a range from 0.016 in the network at ≥85% similarity threshold to 0.046 in the network at ≥97%), it shows a tendency of OTUs from the same geographical region to be more associated between them, hence genetically more similar, than with OTUs from other regions.

Regarding the depth in the water column, the majority of the unicellular Holozoans were preferentially located in the surface or Deep Chlorophyll Maximum (DCM) layers (**Figure 5B**). This tendency to be present in the upper layers of the water column was supported by the positive assortativity coefficient (Supplementary Table). Even though these are low positive numbers, they were significantly different from the random shuffled distribution (one sample t-test, p-value<0.01**), which supported the tendency for a shallower preference location.

Finally, unicellular holozoans were recovered from a wide range of sizes (**Figure 5C**). For example, within Choanoflagellata, the majority of Acanthoecida abundance (69.37%) was present in the nano fraction (3/5-20 μm), followed by 19.4% in the pico fraction (0.8-3/5 μm). Filasterean reads were mainly found in meso (43.18%) and nano (46.21%) fractions. Ichthyosporeans had a different pattern of sizes according to

subgroup (**Figure 5C**). The distribution of Dermocystida reads was shifted towards the largest fractions (10.96%, 19.98% and 57.73% in meso, micro and nano fractions, respectively). On the contrary, the distribution of Ichthyophonida reads was shifted towards the smallest fractions (24.46% in nano and 61.97% in pico fractions). OTUs associated with Corallochytrea were preferentially found in the pico, nano and pico-nano fractions (0.8-20 µm). Finally, both MAOP groups were more present in the smallest fractions: nano (54.94%) and pico (37.81%), which differs from previous findings that showed MAOP dominating the micro fraction (del Campo and Ruiz-Trillo, 2013). Yet, these results are consistent with these authors, who already suggested that MAOP group might be composed by species with different sizes. The group might also undergo a life cycle with several stages that include different cell sizes. The preferential location of different lineages in different size fractions can be seen in the assortativity values (Supplementary Table). In all networks, assortativity coefficients of fraction sizes were the highest among all elements considered (depths, oceanic provinces, oceans and size). These values were also significant compared to the distribution of random shuffled labels (one sample t-test, p-value<0.01**), indicating a tendency for unicellular Holozoa lineages to be retrieved from specific size fractions.

Co-occurrence of the ichthyosporean *Creolimax fragrantissima* and its putative animal hosts, some of them detected for the first time

Some of these unicellular species, specially the Ichthyosporea, have been previously described as animal parasites or symbionts (Mendoza *et al.*, 2002; Glockling *et al.*, 2013). To see whether our data could illuminate us on this aspect, we checked if there was any association between the presence of unicellular Holozoa and animals.

Our results showed that there were indeed significant positive and negative correlations between unicellular Holozoa and animals (**Figure 6A**). The strongest correlation (Spearman's rank correlation coefficient, ρ_S=0.6-0.8, p<0.01**) was shown between OTUs associated with *Creolimax fragrantissima* and several animal phyla: Entoprocta (Barentsiidae), Mollusca (Polyplacophora), Tardigrada, and Porifera (Homoscleromorpha, Calcarea and Demospongiae). To see if we could detect other associations but monotonic and linear (as Spearman and

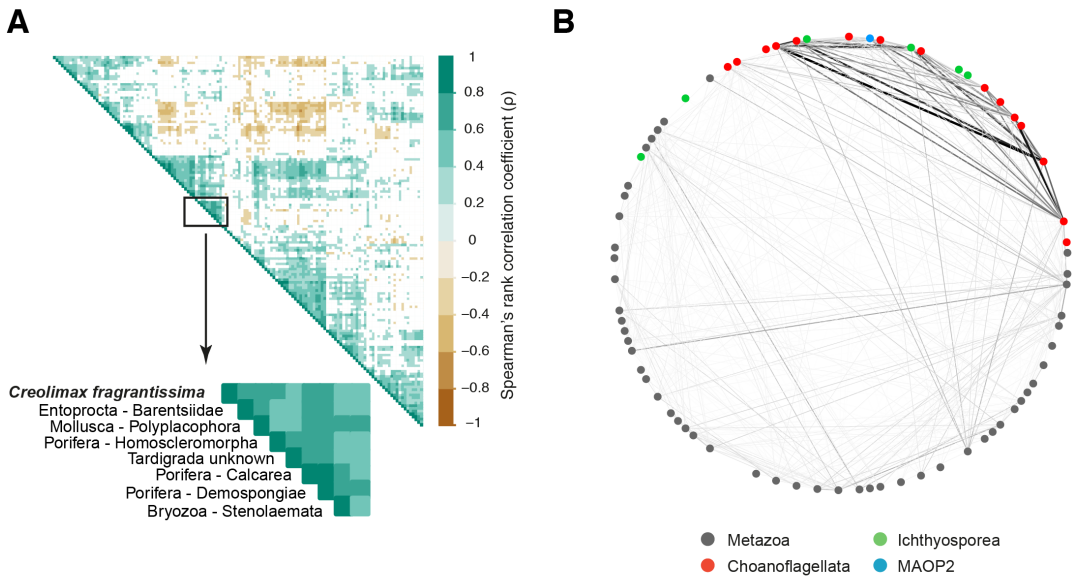


Figure 6. Co-occurrence analysis between unicellular Holozoa OTUs and animal classes from *Tara* Oceans. (A) Heatmap representing the Spearman's rank correlation coefficient (ρ). The ichthyosporean symbiont *Creolimax fragrantissima* had the strongest correlation coefficient (ρ_S=0.6-0.8, p<0.01**) with several animal phyla, suggesting a wider diversity of animal hosts in which this organism can dwell. Full heatmap can be found in Supplementary Figure 4A. (B) Network depicting other possible associations, besides monotonic and linear. The environmental clades marine ichthyosporeans 1 (MAIP1) and marine opisthokonts 2 (MAOP2) were connected with several animal phyla, suggesting non-exclusive free-living lifestyles, or coincidence due to the use of same ecological resources. Full network can be found in Supplementary Figure 4B.

Pearson describe, respectively), we used a bipartite network (**Figure 6B**). We corroborated the previous finding of *Creolimax fragrantissima* with several animal phyla, specifically with Polyplacophora (ρ_S=0.465), Calcarea (ρ_S=0.352) and Demospongiage (ρ_S=0.311). *C. fragrantissima* was isolated 27 times from invertebrate guts, mostly from a sipunculid species, but also one tunicate, sea cucumber and chiton (Marshall *et al.*, 2008). Thus, our results corroborated some symbiotic relationships (with Polyplacophora, commonly known as chiton) and suggested some other putative hosts (Entoprocta, Tardigrada, and Porifera).

We also found that the environmental group marine ichthyosporeans 1 (MAIP1) was connected to Acoelomorpha, Arthropoda (Hexapoda, Crustacea), Bryozoa, Cnidaria, Nematoda (Enoplea) and Chordata (Tunicata, Craniata). This result suggests that the environmental group MAIP1 may be associated with animal phyla and not being exclusively free-living. Another interesting result was the interaction between MAOP2 and Ctenophora (ρ_S=0.409) or Mollusca

(Cephalopoda) (ρ_S=0.317), which could imply that these taxa use the same resources or have some ecological interaction, as it was found for other environmental groups (Lima-Mendez *et al.*, 2015; Lambert *et al.*, 2018).

Overall, these results suggest more complex ecological interactions between parasitic/symbiotic unicellular holozoans and animals. These biotic effects (grazing, pathogenicity, and parasitism) have been reported to explain 82% of the variability in the *Tara* Oceans interactome, giving a greater importance to these interspecific connections (Lima-Mendez *et al.*, 2015). However, we refuse to claim that correlation implies causation. What is certain though is that metabarcoding has a great power to assess diversity in its multiple forms, from pure ecological and evolutionary studies to applied conservationism, which is of vital importance in a world of threat to biodiversity.

MATERIALS & METHODS

Datasets

The initial environmental dataset was provided by the *Tara* Oceans consortium, which contained a total of 474,303 Operational Taxonomic Units (OTUs) from all eukaryotic clades. The reference database was obtained by merging three different databases: GenBank, PR2-Opistho and PR2_V9. First, we downloaded two databases from GenBank: nucleotide (nt) and environmental nucleotide (env_nt) by January 25th 2018. We retrieved 18S rDNA sequences from these databases by searching them using the human 18S sequence as a query (AC139250, positions 551,257 to 553,055). This sequence had been previously confirmed to contain the *Tara* Oceans V9 primer sequences. BLASTn parameters were: e-value <1E-10, percentage of identity $\geq$60% and maximum target sequences of $9{,}9 \cdot 10^7$ (for nt) and $9{,}9 \cdot 10^8$ (for env_nt). From the BLASTn output, we implemented two filtering processes. In the first one, we retrieved the sequences that contained both *Tara* Oceans V9 primer sequences. We then trimmed the sequences just to have the V9 region. In the second step, we kept those sequences whose length was comprised between 80 and 120 base pairs to keep the most frequent length range of this region (Amaral-Zettler *et al.*, 2009).

The second database, PR2-Opistho, was a well-curated and updated version of the original PR2 database for Opisthokonta clade. This database (PR2-Opistho) was also trimmed with the *Tara* Oceans primer sequences just to keep the V9 region.

The third database, PR2_V9, was generated by the *Tara* Oceans consortium (de Vargas *et al.*, 2015). Because both PR2-Opistho and PR2_V9 were originally generated from PR2 database, we eliminated redundancies and kept the taxonomical annotation from the PR2-Opistho database. Finally, we combined all databases, producing a final reference database of 49,379 eukaryotic sequences.

To retrieve the unicellular Holozoa sequences, we performed a phylogenetic placement of both environmental and reference datasets against an eukaryotic reference tree, and took those that branched within Holozoa and outside animals. A phylogenetic placement consists of mapping short amplicons (in this case, *Tara* Oceans OTUs) into a fixed reference tree made from full-length 18S rDNA sequences. This reference was constructed using 130 full 18S sequences that covered all eukaryotic groups. We performed the phylogenetic placement using the RAxML-EPA algorithm (Berger *et al.*, 2011), and we selected the sequences that were placed into unicellular Holozoa using the C++ script extract_clade_placements from Genesis software v0.18.1 (Czech and Stamatakis, 2016). Therefore, the starting dataset of unicellular Holozoa contained 2,426 sequences (2,197 were environmental from *Tara* Oceans while 229 were reference sequences). This dataset can be found in Supplementary Material 4.

Network construction

We built the initial similarity network based on a blast all-against-all of the unicellular Holozoa dataset. We used BLASTn v2.7.1+ (Camacho *et al.*, 2009), with the following options: e-value <1E-10, percentage of identity ≥85%, maximum number of HSPs 1 and maximum target sequences 3,000.

We used the cleanblastp script from CompositeSearch software to filter the output in order to remove auto-loops and reciprocal connections (A-B would be the same as B-A) (Pathmanathan *et al.*, 2018). Final networks were obtained by setting up a mutual cover threshold of ≥95% and increasing sequence similarity thresholds: ≥85%, ≥87%, ≥90%, ≥95%, and ≥97%. These networks can be found in Supplementary Material 4.

Network node annotation

To annotate taxonomically every node in the network, we performed a BLAST of the initial 2,426 holozoan sequences against the PR2-Opistho database, using the following parameters: e-value <1E-50 and ≥97% percentage of identity. Under these conditions, only 438 sequences

could be annotated. Thus, we decided to use a phylogenetic method to taxonomically assign the rest of the unannotated OTUs: tax2tree algorithm (McDonald *et al.*, 2012). This software requires the structure of the phylogenetic tree of both reference and unannotated sequences. Then, it assigns the taxonomy to the unannotated tips, given a file with the taxonomical information of the annotated tips. We could successfully annotate 1,503 additional sequences. Thus, a total of 1,941 sequences (78.8% of the initial dataset) could be taxonomically annotated.

Network analysis

To address the molecular diversity and novelty of unicellular Holozoa, we analysed all network metrics using NetworkX v2.1 library on python 3.5.1 (Hagberg *et al.*, 2008).

Novelty assessment: preferential connection

Assortativity is a property of the network that measures the preferential connection between nodes belonging to the same group (Newman, 2003; Forster *et al.*, 2015) (**Figure 1**). To compute its significance, we first calculated a distribution of null assortativity values for each network, because it may be different than 0, which is associated with a random distribution of the nodes in the graph (**Figure 1** and Supplementary Material 1). Namely, we randomly shuffled the labels of the nodes 100 times while keeping the same network topology. For example, one ENV node (i.e., a node composed of an environmental sequence) could turn out to be ENV or REF (i.e., a node composed of a reference sequence) after the shuffling. For all these 100 random networks, we computed the assortativity, generating the distribution of assortativity values for random networks. We next computed the actual value of assortativity in the networks (**Figure 3A** and Supplementary Table), for each tested pair of categories (ENV *vs* REF; IND *vs* MEDIT *vs* ARCTIC *vs* ANTAR *vs* NPAC *vs* SPAC *vs* NATL *vs* SATL *vs* REDS; SURF *vs* DCM *vs* MES *vs* MIX *vs* ZZZ; MESO *vs* MICRO_MESO *vs* MICRO *vs* NANO *vs* PICO_NANO *vs* PICO_MICRO *vs* PICO).

Control test

We performed a control test to check whether our results could be explained by the large difference in the amount of ENV sequences (2,197) compared to REF sequences (229). We subsampled randomly 10% of the original ENV sequences 100 times and combined them with the same REF dataset and

then, performed all the analyses in the same way as for the real networks.

Novelty assessment: BRIDES

BRIDES software characterizes new paths that are created when extra nodes are added to an original network (Lord *et al.*, 2016). For every sequence similarity network, we first kept only the REF nodes (original network) and then, we added the ENV nodes of unicellular Holozoa (augmented networks) to compute BRIDES using the default parameters.

Novelty assessment: phylogenetic placement

To validate the putative novel diversity previously obtained with BRIDES and shortest-path analysis, we performed a phylogenetic placement of the OTUs into our curated reference Holozoa tree, which can be found in Supplementary Material 5. We aligned the sequences using PaPaRa with default parameters (Berger and Stamatakis, 2011) and manually examined the alignment and corrected wrong positions in Geneious v9.0.5 (Kearse *et al.*, 2012). We then trimmed the non-homologous positions with trimAl 1.4.rev15, setting the gap threshold option at 0.2 for the alignment coming from BRIDES analysis (Capella-Gutiérrez *et al.*, 2009). Regarding the alignment from the shortest-path analysis the trimming was done manually, removing those positions with a mean pairwise identity over all pairs below 30%. We performed the phylogenetic placement using the RAxML-EPA algorithm (Berger *et al.*, 2011). The final tree of Figure 4B was enhanced using iTOL (Letunic and Bork, 2016).

We validated the quality of the phylogenetic placement using the placement_histograms script from Genesis package v0.18.1 (Czech and Stamatakis, 2016). The first parameter computed was the EDPL (Expected Distance between Placement Locations). For every OTU, it calculates the weighted distance between all placement positions. In other words, EDPL quantifies to which extent all placements from an OTU are scattered over the tree. In both groups, EDPL values were extremely small (<0.05) (**Supplementary Figure 3A,C**). Considering that most branches in the tree had less than 0.05 nucleotide substitutions per site, it meant that the majority of the OTUs were located within the same branch. However, the quality of these placements was not high, measured as the distribution and frequency of Likelihood Weight Ratio values (LWR). This was especially drastic in the placements of MASHOL OTUs (**Supplementary Figure 3D**),

which shows the uncertainty in the location of the group.

Geographical distribution

We described the geographical distribution of unicellular Holozoa lineages, as well as the distribution along the water column and size fractions, through circular layouts using "circlize" package in Rstudio (Gu *et al.*, 2014; RStudio, 2017)

Co-occurrence patterns

To test the association between unicellular Holozoa and animal OTUs, we carried out a co-occurrence analysis. First, we filtered the dataset to keep those OTUs that were present in at least 3 samples (out of 1,086 total samples in *Tara* Oceans). Then, we summed up OTU abundances if these OTUs belonged to the same class in animals or the same genus/species in unicellular Holozoa. We used "corrplot" and "Hmisc" libraries in Rstudio v.1.1.383 to perform the analyses (RStudio, 2017; Wei *et al.*, 2017; Harrell, 2019). These consist of building a correlation matrix among all pairwise comparisons and then, extract the significant relationships (Spearman's significance<0.01**), which finally were plotted in a heatmap.

There was a possibility, however, that some associations could be neither monotonic nor linear. In that case, we would not be able to detect them using Spearman's or Pearson's correlation coefficients. We used instead MICtools package (Albanese *et al.*, 2018), which is able to identify a wider range of relationships in large datasets and assess their statistical significance. Final networks were created using Cytoscape 3.3.0 (Shannon *et al.*, 2003).

SUPPLEMENTARY MATERIAL

The Supplementary Material of this article is available at FigShare: https://doi.org/10.6084/m9.figshare.8020427.v2

AUTHOR CONTRIBUTIONS

ASA, IRT, RL and EB designed the study. CDV provided the raw data. ASA and RL performed all analyses. ASA designed the figures and wrote the manuscript. EB and IRT supervised the project and reviewed the manuscript.

FUNDING

This work was supported by an European Research Council Consolidator Grant (ERC- 2012-Co-616960), and grants (BFU2014- 57779-P and BFU2017-90114-P) from Ministerio de Economía y Competitividad (MINECO), Agencia Estatal de Investigación (AEI), and Fondo Europeo de Desarrollo Regional (FEDER) to IRT. EB was founded by the European

Research Council (FP7/2017- 2013 Grant Agreement #615274).

CONFLICT OF INTEREST

None declared

REFERENCES

Albanese, D., Riccadonna, S., Donati, C., and Franceschi, P. (2018) A practical tool for maximal information coefficient analysis. *Gigascience* **7**: 1–8.

Amacher, J., Neuer, S., Anderson, I., and Massana, R. (2009) Molecular approach to determine contributions of the protist community to particle flux. *Deep Sea Res. Part I* **56**: 2206–2215.

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., and Huse, S.M. (2009) A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. *PLoS One* **4**: e6372.

Arroyo, A.S., López-Escardó, D., Kim, E., Ruiz-Trillo, I., and Najle, S.R. (2018) Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina. *Front. Ecol. Evol.* **6**: 99.

Berger, S.A., Krompass, D., and Stamatakis, A. (2011) Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Syst. Biol.* **60**: 291–302.

Berger, S.A. and Stamatakis, A. (2011) Aligning short reads to reference alignments and trees. *Bioinformatics* **27**: 2068–2075.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

de Vargas, C., Stephane, A., Nicolas, H., Johan, D., Frederic, M., Ramiro, L., et al. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1–12.

del Campo, J., Mallo, D., Massana, R., de Vargas, C., Richards, T. a., and Ruiz-Trillo, I. (2015) Diversity and distribution of unicellular opisthokonts along the European coast analysed using high-throughput sequencing. *Environ. Microbiol.* n/a-n/a.

del Campo, J. and Ruiz-Trillo, I. (2013) Environmental Survey Meta-analysis Reveals Hidden Diversity among Unicellular Opisthokonts. *Mol. Biol. Evol.* **30**: 802–805.

del Campo, J., Sieracki, M.E., Molestina,

R., Keeling, P., Massana, R., and Ruiz-Trillo, I. (2014) The others: our biased perspective of eukaryotic genomes. *Trends Ecol. Evol.* **29**: 252–259.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Corel, E., Lopez, P., Méheust, R., and Bapteste, E. (2016) Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends Microbiol.* **24**: 224–237.

Czech, L. and Stamatakis, A. (2016) Genesis. A Toolkit for Working with Phylogenetic Data. *https://github.com/lczech/genesis*.

Edgcomb, V., Orsi, W., Bunge, J., Jeon, S., Christen, R., Leslin, C., et al. (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J.* **5**: 1344–1356.

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S., et al. (2015) Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biol.* **13**: 1–16.

Glockling, S.L., Marshall, W.L., and Gleason, F.H. (2013) Phylogenetic interpretations and ecological potentials of the Mesomycetozoea (Ichthyosporea). *Fungal Ecol.* **6**: 237–247.

Grau-Bové, X., Torruella, G., Donachie, S., Suga, H., Leonard, G., Richards, T.A., and Ruiz-Trillo, I. (2017) Dynamics of genomic innovation in the unicellular ancestry of animals. *eLife* **6**: e26036.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014) circlize implements and enhances circular visualization in R. *Bioinformatics* **30**: 2811–2812.

Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008) Exploring network structure, dynamics, and function using NetworkX. In, Varoquaux,G., Vaught,T., and Millman,J. (eds), *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA USA, pp. 11–15.

Harrell, F.E. (2019) Hmisc: Harrell Miscellaneous. *https://github.com/harrelfe/Hmisc*.

Hugerth, L.W., Muller, E.E.L., Hu, Y.O.O., Lebrun, L.A.M., Roume, H., Lundin, D., et al. (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One* **9**:.

Kearse, M., Moir, R., Wilson, A., Stones-

Havas, S., Cheung, M., Sturrock, S., et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Krabberød, A.K., Bjorbækmo, M.F.M., Shalchian-Tabrizi, K., and Logares, R. (2017) Exploring the oceanic microeukaryotic interactome with metaomics approaches. *Aquat. Microb. Ecol.* **79**: 1–12.

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., and Galand, P.E. (2019) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *ISME J.* **13**: 388-401.

Lang, B.F., O'Kelly, C., Nerad, T., Gray, M.W., and Burger, G. (2002) The closest unicellular relatives of animals. *Curr. Biol.* **12**: 1773–1778.

Layeghifard, M., Hwang, D.M., and Guttman, D.S. (2017) Disentangling Interactions in the Microbiome: A Network Perspective. *Trends Microbiol.* **25**: 217–228.

Letunic, I. and Bork, P. (2016) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* **44**: 127–128.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., et al. (2015) Determinants of community structure in the global plankton interactome. *Science* **348**: 1262073–1262073.

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., et al. (2014) Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* **24**: 813–821.

Lord, E., Le Cam, M., Bapteste, É., Méheust, R., Makarenkov, V., and Lapointe, F.J. (2016) BRIDES: A new fast algorithm and software for characterizing evolving similarity networks using breakthroughs, roadblocks, impasses, detours, equals and shortcuts. *PLoS One* **11**: 5–7.

Mahé, F., de Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., et al. (2017) Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nat. Ecol. Evol.* **1**: 0091.

Marshall, W.L. and Berbee, M.L. (2011) Facing Unknowns: Living Cultures (Pirum gemmata gen. nov., sp. nov., and Abeoforma whisleri, gen. nov., sp. nov.) from Invertebrate Digestive Tracts Represent an Undescribed Clade within the Unicellular Opisthokont Lineage Ichthyosporea (Mesomycetozoea).

Protist **162**: 33–57.

Marshall, W.L., Celio, G., McLaughlin, D.J., and Berbee, M.L. (2008) Multiple Isolations of a Culturable, Motile Ichthyosporean (Mesomycetozoa, Opisthokonta), Creolimax fragrantissima n. gen., n. sp., from Marine Invertebrate Digestive Tracts. *Protist* **159**: 415–433.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., Desantis, T.Z., Probst, A., et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**: 610–618.

Mendoza, L., Taylor, J.W., and Ajello, L. (2002) The Class Mesomycetozoea: A Heterogeneous Group of Microorganisms at the Animal-Fungal Boundary. *Annu. Rev. Microbiol.* **56**: 315–344.

Newman, M.E.J. (2003) Mixing patterns in networks. *Phys. Rev. E* **67**: 1–13.

Ocaña-Pallarès, E., Najle, S.R., Scazzocchio, C., and Ruiz-Trillo, I. (2019) Reticulate evolution in eukaryotes: Origin and evolution of the nitrate assimilation pathway. *PLOS Genet.* **15**: e1007986.

Pathmanathan, J.S., Lopez, P., Lapointe, F.-J., and Bapteste, E. (2018) CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection. *Mol. Biol. Evol.* **35**: 252–255.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., et al. (2015) Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**: 150023.

Pilosof, S., Porter, M.A., Pascual, M., and Kéfi, S. (2017) The multilayer nature of ecological networks. *Nat. Ecol. Evol.* **1**: 0101.

Romari, K. and Vaulot, D. (2004) Composition and temporal variability of picoeukaryote communities at a coastal site of the English Channel from 18S rDNA sequences. *Limnol. Oceanogr.* **49**: 784–798.

RStudio, T. (2017) Rstudio: Integrated Development for R. *http://www.rstudio.com*.

Ruiz-Trillo, I., Burger, G., Holland, P.W.H., King, N., Lang, B.F., Roger, A.J., and Gray, M.W. (2007) The origins of multicellularity: a multi-taxon genome initiative. *Trends Genet.* **23**: 113–118.

Ruiz-Trillo, I., Roger, A.J., Burger, G., Gray, M.W., and Lang, B.F. (2008) A Phylogenomic Investigation into the Origin of Metazoa. *Mol. Biol. Evol.* **25**: 664–672.

Shalchian-Tabrizi, K., Minge, M.A., Espelund, M., Orr, R., Ruden, T.,

Jakobsen, K.S., and Cavalier-Smith, T. (2008) Multigene Phylogeny of Choanozoa and the Origin of Animals. *PLoS One* **3**: e2098.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**: 2498–2504.

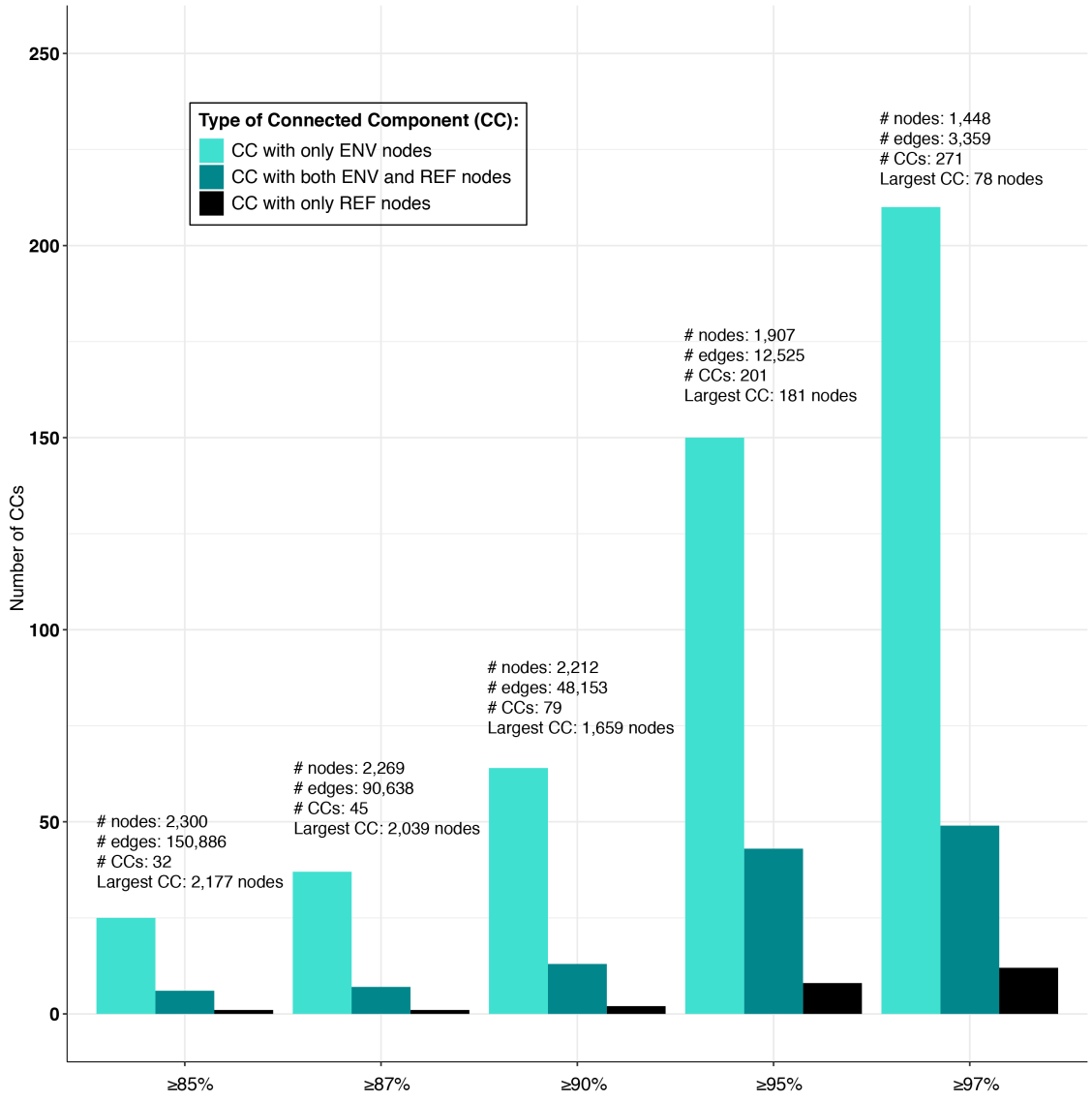
Thomsen, H.A., Garrison, D.L., and Kosman, C. (1997) Choanoflagellates (Acanthoecidae, Choanoflagellida) from the Weddell sea, Antarctica, taxonomy and community structure with particular emphasis on the ice biota; with preliminary remarks on Choanoflagellates from Arctic sea ice (Northeast Water Polynya, G. *Arch. fur Protistenkd.* **148**: 77–114.

Torruella, G., de Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M.A., del Campo, J., et al. (2015) Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Curr. Biol.* **25**: 1–7.

Valverde, S., Piñero, J., Corominas-Murtra, B., Montoya, J., Joppa, L., and Solé, R. (2018) The architecture of mutualistic networks as an evolutionary spandrel. *Nat. Ecol. Evol.* **2**: 94–99.

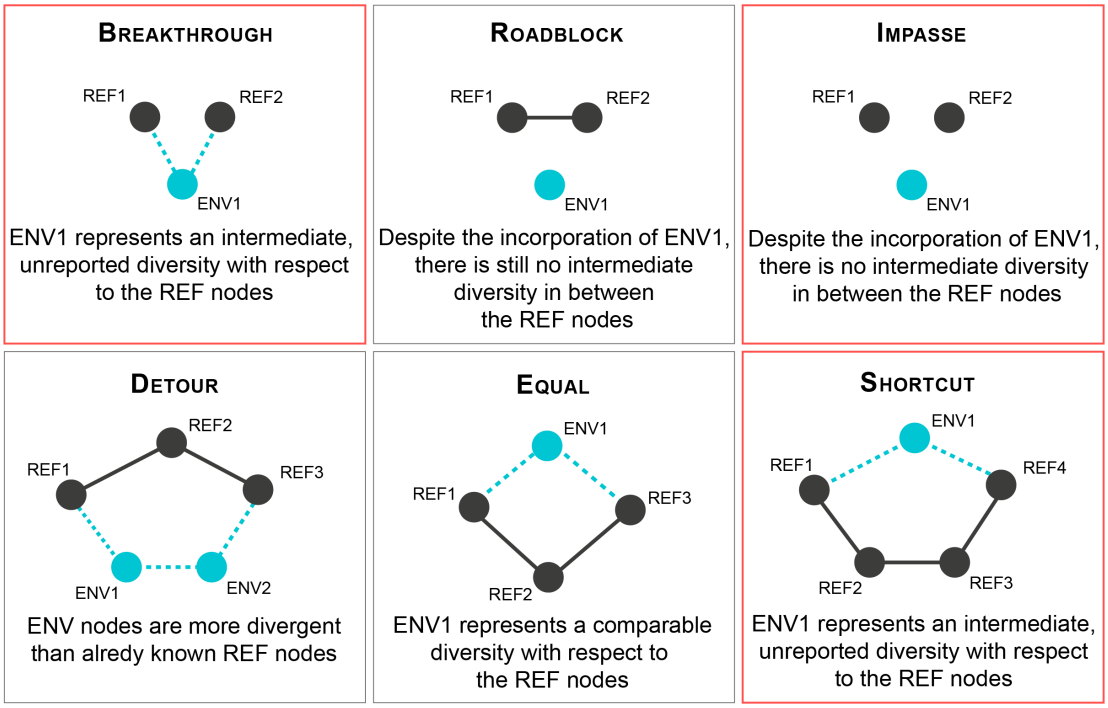
Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., and Zemla, J. (2017) corrplot: Visualization of a Correlation Matrix. *https://github.com/taiyun/corrplot*.

SUPPLEMENTARY FIGURES

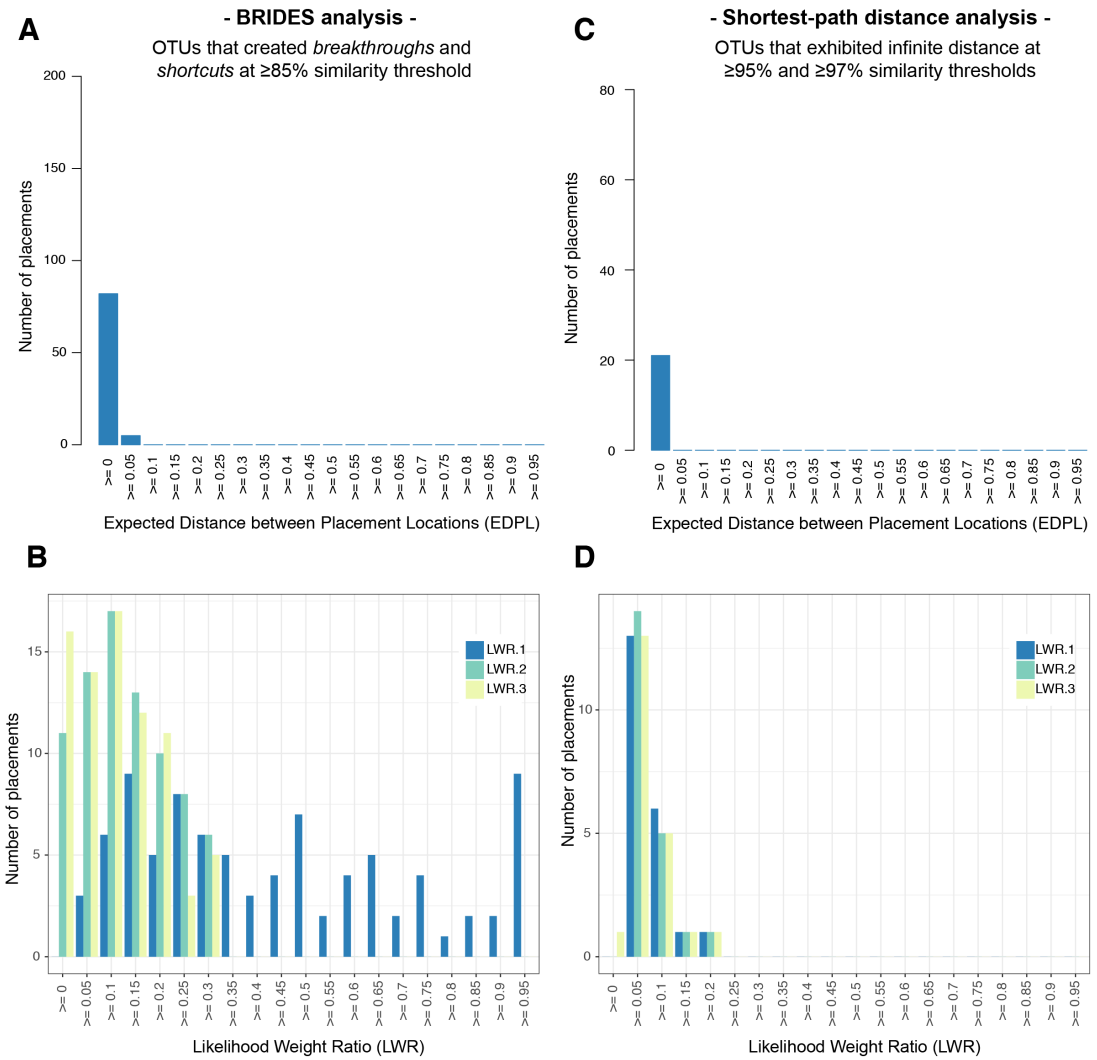


Supplementary Figure 1. Topological metrics of each network. Connected Components (CCs) with only environmental nodes exceeds the rest of CCs because of the unequal amount of environmental sequences compared to reference sequences in the original database (2,197 environmental sequences; 230 reference sequences). Number of nodes reflects only the nodes that are connected, not singletons. This is the reason why the number of nodes decreases as the similarity threshold increases.

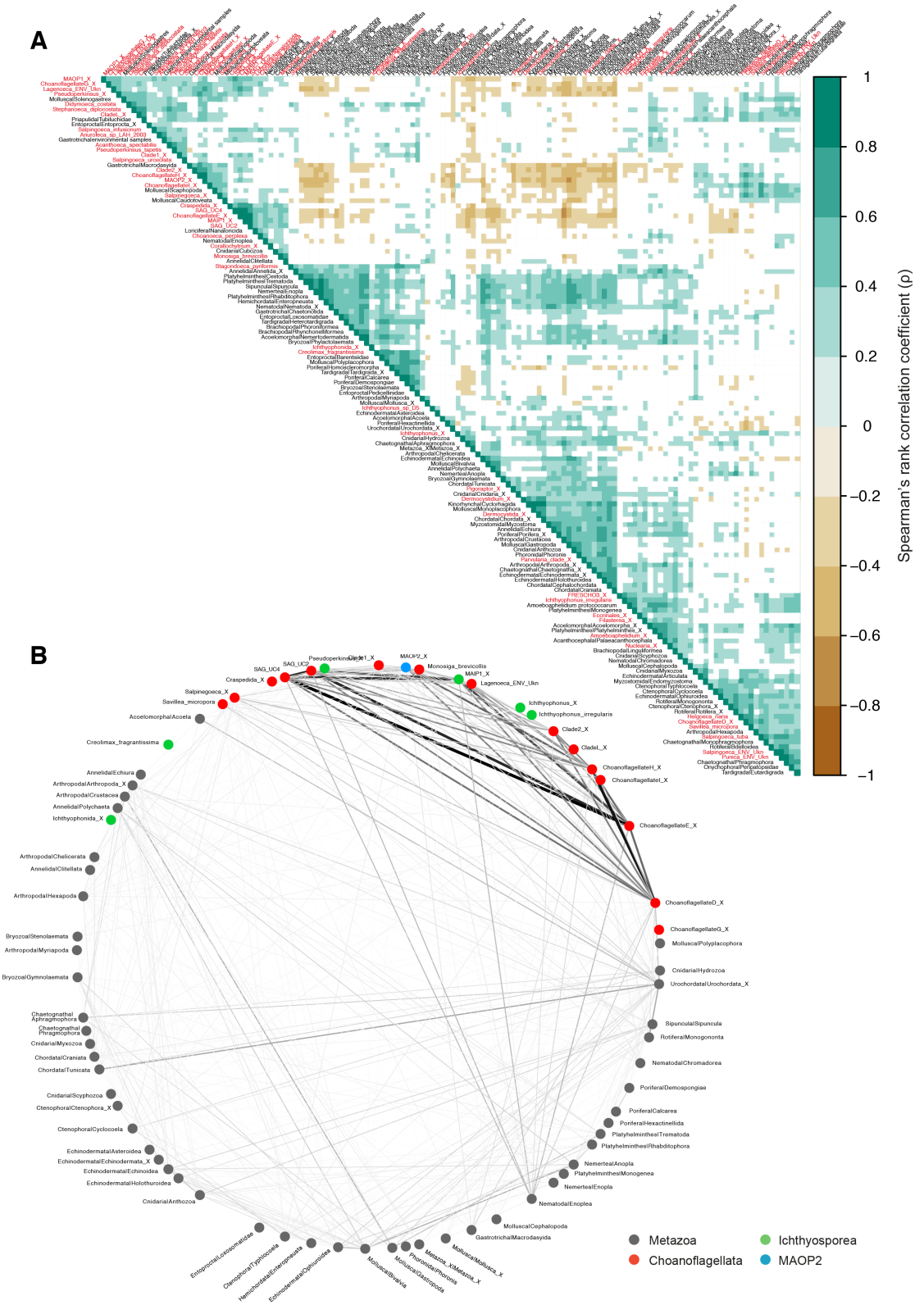
BRIDES statistic	path in the original network	path in the augmented network
Breakthrough	impossible	possible
Roadblock	possible	impossible
Impasse	impossible	impossible
Detour	shorter distance	longer distance
Equal	equal distance	equal distance
Shortcut	longer distance	shorter distance



Supplementary Figure 2. BRIDES paths. An illustration of all BRIDES paths, together with the possible biological interpretation. Blue nodes and edges are generated by the environmental sequences (ENV), which are added to the original network only made from reference sequences (REF), depicted by black nodes and edges. We focused on the paths highlighted with a red box because they were the simplest to interpret from a biological standpoint.



Supplementary Figure 3. Phylogenetic placement validation. **(A,C)** The Expected Distance between Placement Locations (EDPL) indicates whether one OTU is scattered over the tree or not. The smaller the EDPL, the better is the placement because it is located in a specific area of the tree. **(B,D)** Barplot represents the first three most probable Likelihood Weight Rations (LWR) of each OTU. In (D) the distribution of the placements was left-tailed, showing the uncertainty of the placement.



Supplementary Figure 4. Co-occurrence analysis of unicellular Holozoa OTUs and animal classes from *Tara* Oceans. (A) Significant correlations (Spearman's significance<0.01**) range from negative values (brown) to positive ones (blue). "_X" sign after a taxa means "unknown". Unicellular Holozoa are depicted in red. **(B)** Significant correlations (Maximal Information Coefficient, MICe, between 0.08-0.638) displayed among unicellular Holozoa.

3.6 Recherche d'homologues par itération

Dans le but d'annoter fonctionnellement et taxonomiquement des séquences d'intérêt, j'ai recherché des relations d'homologie avec des séquences de référence. Deux séquences sont homologues si elles descendent de la même séquence ancestrale (1.3.2). Afin d'inférer des relations d'homologie, j'ai utilisé des méthodes d'alignement de séquences par paires comme BLAST (1.3.3). Deux séquences sont considérées homologues si elles s'alignent avec au moins un certain pourcentage d'identité et sur une longueur suffisante, avec un score d'E-value significatif. Cette méthode permet de retrouver des séquences dont la relation d'homologie est directement détectable. Cependant, deux séquences peuvent être homologues mais avoir divergeés suffisamment pour que cette homologie ne soit plus détectable en comparant directement les séquences. Néanmoins, il peut exister une troisième séquence homologue dont la relation d'homologie est encore détectable avec les deux séquences précédentes. On a alors un exemple de non transitivité, où on peut suggérer l'homologie de deux séquences par similarité d'alignement en trouvant une séquence intermédiaire (Fig. 19).

En réalité, une telle approche est équivalente à construire une composante connexe dans

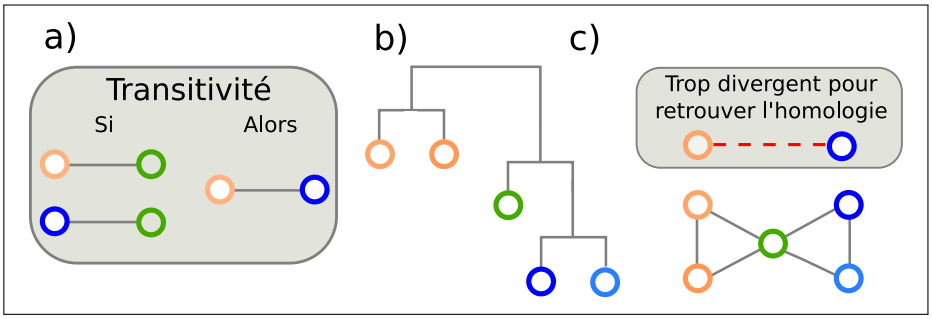


Fig. 19. Transitivité, composante connexe et homologie.
a) Définition de la transitivité. b) Arbre phylogénétique représentant l'histoire évolutive d'un gène. c) Une composante connexe dans un réseau de similarité permet de proposer des relations d'homologie même quand la transitivité est perdue.

un réseau de similarité de séquences. En effet, si les critères pour établir un lien sont suffisamment contraignants (>30 % d'identité, >80 % de couverture mutuelle, $<5.10^{-5}$ E-value), on peut raisonnablement faire l'hypothèse qu'une composante connexe est un ensemble de séquences homologues. Cependant, l'histoire évolutive des séquences peut inclure des insertions et des délétions. Par ailleurs, l'insertion d'un domaine protéique va créer une homologie partielle entre les séquences donneuse et receveuse, parce que les deux séquences homologues ne le sont pas forcément sur toute leurs longueurs. Dans la recherche d'homologues fonctionnels (i.e. des séquences homologues qui partagent la même fonction), cette propriété peut être source de faux positifs; l'ajout d'un domaine à une protéine peut modifier sa localisation cellulaire ou ses partenaires. J'ai donc développé un programme qui permet de rechercher par itération dans un jeu de données des homologues proches et distants à une famille de gènes de départ. A chaque itération, les séquences précédemment identifiées comme homologues sont utilisées comme nouvelles séquences requêtes pour interroger le jeu de données. Cependant, il est possible que les séquences identifiées possèdent des homologies partielles, une succession d'homologies partielles peut conduire à des séquences qui ne s'alignent plus sur les séquences d'origine. Pour éviter d'être induit en erreur par ce biais, si l'utilisateur le souhaite, il peut utiliser une option qui permet de vérifier que la position de l'alignement entre séquences se conserve au fil des itérations. J'ai de plus rajouté des analyses qui permettent à l'utilisateur de sélectionner les séquences les plus pertinentes, par exemple les séquences les plus divergentes, ou au contraire les plus proches des séquences de référence. Ces analyses consistent en une première annotation taxonomique par comparaison au meilleur hit dans une base de donnée de référence, mais aussi en une analyse basée sur les réseaux qui permet d'identifier des groupes de séquences et leurs positions par rapport aux séquence de références. Ce programme est le premier que j'ai écrit pour être distribué.

L'article que je présente ci-dessous est une "Application Note" décrivant le fonctionnement du programme ISHF. Cependant, avec mes encadrants, nous avons décidé de le publier avec une analyse biologique pour montrer ses capacités. Cette analyse n'est pas encore incluse dans l'article et n'est pas à un stade assez avancé pour justifier une partie

spécifique dans ce manuscrit. J'ai donc inclus ici le plan de cette analyse en cours. Son objectif est de trouver des séquences basales dans l'arbre du vivant, si possible qui branchent entre les archées et les bactéries. De par leur nature basale, de telles séquences auraient le pouvoir de nous apporter des informations sur l'ancêtre des archées et/ou des bactéries. De plus des séquences qui brancheraient de manière robuste entre archée et bactérie seraient le signe potentiel d'un nouveau domaine du vivant. Pour ce faire, nous avons d'abord identifié des familles de gènes conservées et potentiellement présentes chez l'ancêtre commun des archées et des bactéries.

Nous avons construit un jeu de données représentatif de la diversité microbienne de plus de 2 millions de séquences. Nous avons aligné ce jeu de données contre lui-même avec DIAMOND sur MESU (le supercalculateur de l'université). Nous avons obtenu un alignement de plus d'un milliard d'arêtes. Avec Python et IIgraph, il était impossible de charger en mémoire un tel graphe, j'ai donc écrit un code en Rust (Matsakis et al. 2014) qui utilise la librairie "petgraph" (https://docs.rs/petgraph/0.4.13/petgraph/) pour en extraire les composantes connexes. Pour chaque composante connexe, nous avons vérifié qu'elle ne se scindait pas si on ne considérait uniquement les séquences d'origines bactérienne et archée. Ensuite nous avons calculé l'assortativité des séquences d'origine bactérienne et archée de ces composantes connexes (Fig. 20) pour valider l'ancienneté de ces familles de gènes.

Une assortativité proche de 1 indique que les groupes archées et bactéries ne se mélangent que très faiblement dans les réseaux. Cette approche permet d'une part d'identifier des familles de séquences présentes à la fois chez les archées et les bactéries, d'autre part le peu de liens entre séquences d'origines différentes nous permet de sélectionner des familles de gènes où les transfert horizontaux de gènes entre domanines du vivant sont rares. Une étape de validation visuelle a permis de sélectionner 20 familles de gènes qui semblent précéder la séparation entre archées et bactéries (Fig. 21).

Nous avons ensuite utilisé ISHF pour rechercher des homologues distants de ces familles dans le jeu de données de TARA Océans, dans l'espoir de trouver des séquences environnementales qui branchent à la base de l'arbre du vivant. Les analyses sont en cours au moment où j'écris ces lignes.

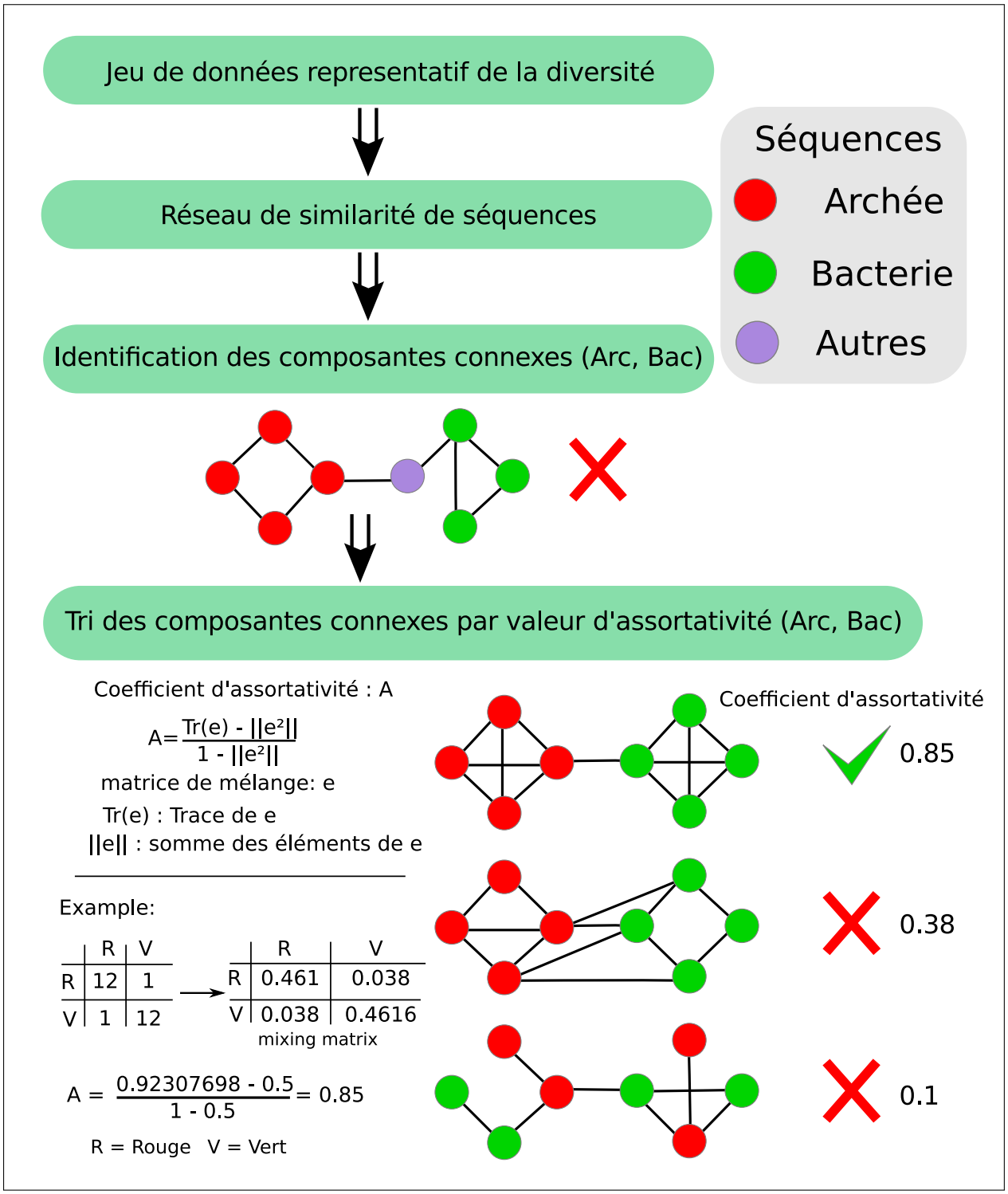


Fig. 20. Identification de familles de gènes ancestrales

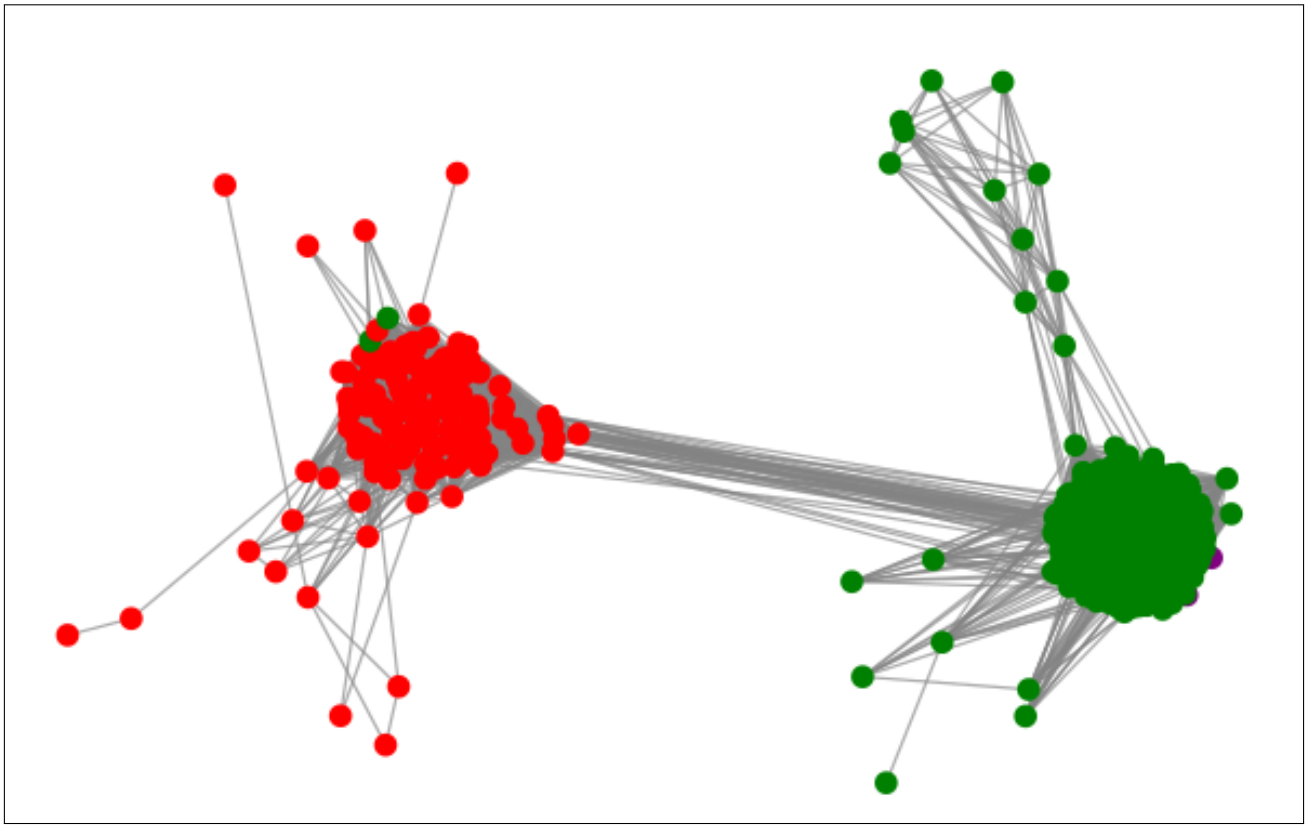


Fig. 21. Example d'une famille de gène ancestrale identifiée: une métallopeptidase
Les nœuds représentent des séquences nucléiques: rouge d'origine bactérienne, verte d'origine archée. Un lien représente un pourcentage d'identité supérieur ou égal à 30% avec 80% de couverture mutuelle et une E-value inférieure à 1.10^{-5}

3.6.1 Article 6, "Iterative Safe Homologs Finder", in prep

Iterative Safe Homologs Finder

Romain LANNES 1, Philippe LOPEZ1 and Eric Bapteste1
1 Université, Institut de Systématique, Evolution, Biodiversité (ISYEB),
Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, EPHE,
Université des Antilles, 7, quai Saint Bernard, 75005 Paris, France
Corresponding Author: romain.lannes@gmail.com

Abstract

Homology detection, i.e. detection of common ancestry, is a standard method for automatic sequence annotation. Homology between sequences that have diverged a long time ago or are evolving fast is not always detectable by direct alignment. Here, we present Iterative Safe Homologs Finder (ISHF), a Python pipeline, that uses an iterative alignment procedure, using previously detected sequences to recover remote homologs of a gene family from a large data set. We investigated the presence of deep branching prokaryotes in the tree of life. We identified putative ancient genes families using sequence similarity networks. We found remote homologs of these ancient gene families using ISHF in large metagenomic data sets, hinting at hidden deep branching microbial diversity in environmental data sets.

Keywords

Homology, Metagenomics, Microbiology

Introduction

Homology detection, i.e. detection of common ancestry, is a standard method for automatic sequence annotation. However, homology between sequences that have diverged a long time ago or are evolving fast is not always detectable by direct alignment. In principle, detection of remote homology in environmental data sets may lead to the discovery of deep branching organisms in the tree of life. The identification of

such divergent new lineages could deeply transform our knowledge about microbiology, by identifying deeply branching lineages. Here, we present Iterative Safe Homologs Finder (ISHF) a Python pipeline, that uses an iterative alignment procedure, using previously detected sequences to recover remote homologs of a gene family from a large data set. ISHF takes as input two fasta files, one containing the seed gene family and the other the data set in which remote homologs of that gene family will be searched for. ISHF starts with a pre-iteration step: after renumbering the seed and target sequences, ISHF filters out target sequences that do not pass the mutual coverage threshold criterion defined by the option (default 80%). This way, ISHF limits the size of the target database, reducing the search space and limiting artifactual matches to partial sequences. ISHF formats the reduced data set as a database for alignment according to the selected aligner, i.e. BLASTP [1], DIAMOND [2] or MMseqs 2 [3]. The first iteration step begins by aligning sequences of the seed family to the target data set. Following iteration steps align previously recovered target sequences to the target data set. Alignment is filtered by threshold values (i.e. the percentage of identity, mutual coverage and E-value) between a query and a target . When several alignments exist for a pair of sequences, the best alignment is selected by the E-value. Such iterative alignment may be prone to false positives. In particular, due to partial homology the alignment position of target sequences retrieved after multiple iterations from the original seed sequences might shift. To circumvent this problem, ISHF asserts the conservation of the alignment position with respect to seed sequences through iterations. By default, ISHF only checks the conservation of alignment position from the previous round of iteration, but this behavior can be changed to assert the conservation of alignment position from the original seed sequences. This option can also be removed. Iteration stops when no more sequences are retrieved or when a user defined maximal number of iterations is reached. Sequence identifiers are changed back from numbers to the original identifiers. Then, ISHF outputs a fasta file with all the sequences found and performs an all-against-all alignment of these sequences. ISHF is then able to perform post analyses, on the extracted gene families. For example, ISHF

can taxonomically annotate retrieved sequences by comparing them to NCBI nr reference databases. It can perform graph analyses, short path analyses and graph based clustering of the homologs whithin a gene family using the Louvain algorithm [4]. ISHF provides annotations at the sequence and at the cluster level, allowing for the selection of the more relevant sequences or clusters (e.g. group of very divergent homologs with respect to the seed sequences).

Results

Benchmark on simulated data

Simulated data set creation

We created a simulated data set to investigate ISHF precision and recall. We made a root sequences data set, and we evolved randomly selected sequences from the root data set along various trees. The root sequences data set was composed of 762 sequences: 381 randomly generated sequences and 381 merged sequences, produced by merging randomly selected first or last half of previously generated sequences. We merged sequences to investigate if domain sharing could lead to false positive detection despite ISHF check option for alignment conservation. We constructed 378 trees from a perfectly symmetrical tree with 64 leaves. For each node under the root node in the symmetrical tree we built a new tree by multiplying all branch lengths from this node to the root by a factor f. We used three factors (f=2,f=4,f= 8) and built one tree for each. In total, we generated 126 trees by factor. We used pyvolve [5] to evolve a randomly selected sequence for each of the 378 trees, producing 378 sequences families with 64 sequences each for a total of 24,192 sequences.

ISHF on simulated data set

Each simulated sequence family was elvolced along a tree with a fast branch and many normal branches. For each sequence family, we selected 8 sequences from the normal branches. We launched ISHF using these selected sequences as seeds against, including merged sequences. Extensive results are shown in supplementary data. Logically, ISHF performance is greatly influenced by the speed of the fast evolving part of the tree. With a factor f=2, ISHF achieved a 100% of true positive detection of homologs and, on average, needs 2 rounds to find all sequences from a sequence family. With an evolving factor of f=4, ISF retrieved 83% of the fast evolving sequences and as expected uses more iterations to achieve its performance (4 iterations on average). Indeed, ISHF uses intermediate sequences to recover diverging sequences. With an evolving factor of f=8, ISHF performs in average 1 iteration and failed to recover fast evolving sequences the evolutionary distance being too far to recover (non existing) intermediate sequences.

Factors	T.P.	F.P.	N.S.E.	F.E.	N.R.
2	1.0	0.0	1.0	1.0	2.325
4	0.9187	0.0	1.0	0.8373	4.1032
8	0.5079	0.0	1.0	0.0159	1.0635
merged	0.8089	0.0	1.0	0.6177	2.497

Table 1 : Average results of ISHF on simulated data. T.P.: True Positive, F.P.: False Positive, N.S.E.: Average proportion for “Normal Speed Evolving”, F.E.: Average proportion for “Fast Evolving”, N.R.: Average Number of rounds by ISHF.

Case study on real data

We used ISHF to investigate the presence of divergent microbial species. Microorganisms that diverged billions of years ago may differ from currently known organisms. We expect to retrieve the most divergent organisms by identifying

homologous sequences from lineages that diverged the earliest, e.g. branching deep in archaea, bacteria or even between. After identifying ancient gene families, we used ISF to look for remote homologous sequences of those genes families in the TARA Oceans dataset.

Identification of ancient gene families

Using sequence similarity networks (SSN) [6], we identified conserved gene families that may have been present in the ancestors of archaea and bacteria. SSN is a way to represent relations between sequences, where a node represents a sequence and an edge is drawn between two nodes if sequences align over a threshold (30% of identity, 80% mutual cover and 10-5 evalue). Such ancient families, that may pre-date archaea and bacteria divergence, present a clear pattern if represented with SSN: two groups, one archaeal the other bacterial, sharing few similarities between them (Fig2 left networks). Using a homemade protocol, we detected 20 such ancient gene families in metagenomes. (methionine aminopeptidase 1, iron-sulfur cluster binding protein-like, RecA-like recombinase protein, 50S ribosomal protein L15, 30s ribosomal protein S5, 30S ribosomal protein S17, 50S ribosomal protein L17, 30S ribosomal protein S19, 50S ribosomal protein L23/L25, 60s ribosomal protein L1, 50S ribosomal subunit protein L3 50S ribosomal protein L16, fibronectin binding protein A, transcription termination antitermination factor, nicotinamide-nucleotide adenylyltransferase, Alanyl-tRNA synthetase, 50S ribosomal protein L13, M50 family peptidase, tryptophan-tRNA synthetase, 40S ribosomal subunit protein S9, glycoside hydrolase family protein, protein translocase subunit SecF, protein translocase membrane subunit SecD, RNA polymerase beta subunit I, Acetoacetyl-CoA reductase, putative 30S ribosomal protein S2)

Mining TARA Ocean

Using ISHF, we investigated the presence of remote homologs of the selected genes families in TARA OCEAN metagenomic data set [7]. Our results show a diversity of environmental sequences that have few to no detectable similarities to reference databases and would have been missed by direct homology search (Table 2, Iteration column). We are currently investigating these groups of divergent sequences in order to determine their taxonomy and diversity.

function	iteration	number of sequences	starting sequences	sequences found
methionine aminopeptidase 1	10	53068	1136	51932
iron-sulfur cluster binding protein-like	3	786	133	653
RecA-like recombinase protein	10	43398	1264	42134
50S ribosomal protein L15	8	18182	919	17263
30s ribosomal protein S5	10	15639	864	14775
30S ribosomal protein S17	6	14958	923	14035
50S ribosomal protein L17	9	14855	893	13962
30S ribosomal protein S19	8	14031	936	13095
50S ribosomal protein L23/L25	7	15431	857	14574
60s ribosomal protein L1	5	14174	910	13264
50S ribosomal subunit protein L3	9	17743	941	16802
50S ribosomal protein L16	9	15675	947	14728
fibronectin binding protein A	7	1491	236	1255
transcription termination antitermination factor	10	121843	835	121008
nicotinamide-nucleotide adenylyltransferase	10	238695	431	238264
Alanyl-tRNA synthetase	7	12959	859	12100
50S ribosomal protein L13	6	15092	942	14150
M50 family peptidase	10	26074	658	25416
tryptophan-tRNA synthetase	10	51348	1152	50196
40S ribosomal subunit protein S9	10	57807	993	56814
glycoside hydrolase family protein	10	34896	1059	33837
protein translocase subunit SecF	10	57013	634	56379
protein translocase membrane subunit SecD	10	36018	634	35384
RNA polymerase beta subunit I	10	19477	713	18764
Acetoacetyl-CoA reductase	5	1999	163	1836
putative 30S ribosomal protein S2	10	15874	842	15032

Table 2 Summary of ISHF search on TARA OCEANS for the selected families.
We stopped ISHF after 10 iterations

Discussion

ISHF, is designed to find remote homologues of a gene family whithin a large data set. It can be used to mine environmental data set, but also to extend gene family using large reference databases. ISHF is modular and supports three majors aligner software. Hidden Markov Model (HMM) will also be integrated in the next version of ISHF. HMM built from Position specific scoring matrices are very powerful tools to detect functional domains using conserved positions in a multiple sequence alignment. To be used effectively they require a good alignment without contamination for their construction, i.e. are alignment constructed with proteins having the same domains. Thus running iterative procedure with HMM is challenging in terms of false positive rates.

References

[1] Altschul, Stephen F. et al. (1990). “Basic local alignment search tool”. In: J. Mol. Biol. 215.3, pp. 403–410. DOI : 10.1016/S0022-2836(05)80360-2.

[2] Buchfink, Benjamin, Chao Xie, and Daniel Huson (2015). “Fast and sensitive protein alignment using DIAMOND”. In: Nat. Methods 12.1, pp. 59–60. DOI : 10 . 1038 /nmeth.3176.

[3] Steinegger, Martin and Johannes Söding (2017). “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets”. In: Nat. Biotechnol. 35.11, pp. 1026–1028. DOI : 10.1038/nbt.3988.

[4] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10). https://doi.org/10.1088/1742-5468/2008/10/P10008

[5] Spielman, SJ and Wilke, CO. 2015. Pyvolve: A flexible Python module for simulating sequences along phylogenies. PLOS ONE. 10(9): e0139047.

[6] Watson, Andrew K. et al. (2019). “The Methodology Behind Network Thinking: Graphs to Analyze Microbial Complexity and Evolution”. In: Methods Mol. Biol. Vol. 1910, pp. 271–308. DOI : 10.1007/978-1-4939-9074-0_9.

[7] Sunagawa, Shinichi et al. (2015). “Ocean plankton. Structure and function of the global ocean microbiome.” In: Science 348.6237, p. 1261359. DOI : 10 . 1126 / science.1261359.

Discussion

Durant ma thèse, je me suis intéressé à la matière noire microbienne. J'ai eu la chance de la réaliser durant une période de découvertes qui remettent en question certaines de nos connaissances en microbiologie. Certaines des découvertes récentes devraient encore être validées par des méthodes alternatives. En effet, l'isolement et la culture d'organismes restent le plus haut niveau de preuve de leur existence et facilitent l'étude des organismes ainsi isolés. Actuellement, les techniques de culture des micro-organismes sont en train d'évoluer pour prendre en compte les relations obligatoires, syntrophiques, parasitaires et symbiotiques qui peuvent exister (Overmann et al. 2017, figure 4). Cette évolution en cours a déjà donné des résultats avec la coculture de souches proches de Lokiarchaeota (Imachi et al. 2019). De plus, un travail important doit être mené en vue de l'intégration de ces découvertes en un modèle descriptif et prédictif des communautés microbiennes. C'est une approche en spirale (Papale F. et Bapteste E., in prep) où les découvertes font évoluer les modèles actuels, qui à leur tour permettent de nouvelles avancées. Pour autant, de nombreuses limites et interrogations demeurent, en particulier des limites techniques et technologiques sur le stockage et le traitement de grands volumes de données mais également sur la validité et l'interprétation des résultats de certaines expériences.

1 Techniques d'analyses d'organismes non cultivables

L'analyse de micro-organismes non cultivables reste un défi. Le séquençage de cellules uniques ne permet pas à l'heure actuelle d'obtenir des génomes complets. La validité des GAMs est sujette à caution pour une grande part de la communauté scientifique (Garg et al. 2019), quand elle n'est pas tout simplement impossible faute d'une profondeur de séquençage insuffisante. En effet, les contigs sont regroupés par similarité d'abondance et / ou de composition en tétranucléotides. Il est difficile d'estimer les possibles contaminations et la complétude des génomes ainsi assemblés. L'outil existant (Parks et al. 2015)(ChekM) utilise des connaissances reposant *a priori* sur ce à quoi devrait ressembler un génome. Légitimes, ces interrogations ont amené à comparer les résultats obtenus avec du séquençage à cellule unique (Alneberg et al. 2018). Cette analyse montre que les résultats produits par ces deux méthodes sont en adéquation, nonobstant des particularités et limitations

propres à chaque méthode. Le regroupement en génomes est moins sensible. Ce sont des techniques complémentaires et il est possible de les utiliser ensemble. Le regroupement des contigs en GAMS est une technologie récente, en évolution et qui de par son succès va sans doute évoluer rapidement dans les prochaines années. Récemment Bowers et al. a proposé un ensemble d'informations minimum à fournir pour publier des GAMs ou des génomes obtenues par séquençage de cellules uniques (Bowers et al. 2017). Néanmoins, cette technologie a permis de véritables percées dans l'étude de la matière noire microbienne. Au moment où j'écris ces lignes, deux études viennent d'être publiées: la première annonce la culture d'une archée proche de Lokiarchaeota (Imachi et al. 2019), le deuxième prétend que les GAMs sont des constructions aléatoires (Garg et al. 2019). Leur synchronicité est révélatrice de la période que traverse la microbiologie environnementale. On y voit un effort de standardisation, de mise en place de contrôle de qualité et de validation expérimentale des résultats obtenus.

2 Problèmes liés à la classification des micro-organismes

Une autre problématique concerne la standardisation de la classification phylogénétique des micro-organismes. Premièrement, la classification actuelle n'est pas standardisée au sens où existent différentes ressources (GreenGenes, SILVA, NCBI Taxonomy) (McDonald et al. 2012; Yilmaz et al. 2014) qui diffèrent sur certains points (Yarza et al. 2014). De plus, la classification actuelle n'est pas respectueuse des relations évolutives: elle possède des groupes polyphylétiques, un héritage de la classification morphologique et métabolique des organismes. Il existe également un biais dans la répartition des rangs taxonomiques. Les groupes étudiés de façon intensive ont tendance à posséder plus de groupes phylogénétiques à diversité égale que des groupes moins étudiés. Il existe ainsi une grande disparité dans la diversité de séquences observées pour le 16 rRNA entre des groupes définis dans SILVA (Yilmaz et al. 2014). Par exemple, la famille Enterobacteriaceae qui contient des douzaines de genres, est équivalente, en terme de diversité, au genre Bacillus. Il est également important de rappeler que la phylogénie basée sur le 16S rRNA est limitée (1.4.1.1). En effet, le 16S rRNA est souvent présent en plusieurs copies dans les génomes, et on ne sait pas en séquencer

la totalité à partir d'amorce nucléotidique et les amorces nucléotidiques que l'on utilise pour le séquençage dirigé sont construites sur des connaissances *a priori*. Cette problématique de la classification des micro-organismes prend de l'ampleur avec l'apparition des GAMs et la découverte de nouveaux phyla. Une classification robuste, se basant sur les relations évolutives et standardisée en terme de diversité, est essentielle à l'interprétation des résultats de microbiologie environnementale et de la description de la diversité. Une telle classification est sans doute une chimère, les relations évolutives et la diversité sont souvent décrites sur des marqueurs présélectionnés. De plus, une phylogénie en arbre ne permet pas de décrire les relations horizontales qui peuvent exister entre espèces. Récemment, Parks et al. ont proposé une base de données (GTDB) phylogénétique basée sur 127 marqueurs phylogénétiques où la distance à la racine est normalisée pour corriger les variation de vitesse évolutive entre les groupes (Parks et al. 2018). Cette classification n'est sans doute pas parfaite mais elle a le mérite de posséder des niveaux de classifications qui décrivent une même diversité moléculaire.

3 Limitations technologiques à la recherche en microbiologie environnementale

L'analyse de jeux de données toujours plus grands nécessite une croissance en capacité de stockage, mémoire vive et puissance de calcul. C'est une limitation que j'ai dû affronter tout au long de ma thèse, remplissant plusieurs fois le disque dur de mon ordinateur mais aussi celui du supercalculateur de l'université. J'ai également dû apprendre à programmer en langage bas niveau pour pouvoir charger un réseau avec plus d'un milliard et demi d'arêtes en mémoire. Certaines librairies informatiques comme "Bio Python" n'ont pas été conçues pour faire face à des jeux de données de l'ordre de la centaine de Go. Dans beaucoup de cas, le bioinformaticien aujourd'hui doit réimplémenter des solutions à son problème en prenant en compte ces limitations de la mémoire. Il est probable que ces librairies vont évoluer avec le temps mais pour l'instant cela peut être perçu comme un frein à la reproductibilité. *In fine*, cette réalité va forcer une adaptation par l'acquisition de matériel dédié toujours plus performant à titre individuel ou en commun, au développement

de nouveaux formats, de nouvelles structures de données, algorithmes et librairies mais aussi à l'acquisition de nouvelles méthodes et organisations de travail. Le budget du parc de stockage d'analyse informatique va sans aucun doute augmenter fortement. Actuellement, la solution adoptée en France passe par la mutualisation des moyens et la création d'une entité responsable de l'entretien et de l'accessibilité des ressources de calcul: l'Institut Français de Bio-informatique (IFB). Les ressources déployées par l'IFB et rendues accessibles au plus grand nombre par l'utilisation d'interfaces graphiques, sont adaptées à de nombreuses applications. Néanmoins, ces ressources se révèlent pour l'instant (trop) limitées pour la génomique environnementale. Les supercalculateurs sont souvent optimisés pour le calcul et non pour des tâches lourdes en lecture/écriture sur disque. Cette limitation relative dans nos capacités à traiter les données est aujourd'hui un frein à la recherche en microbiologie environnementale.

4 Recours aux heuristiques pour l'alignement de séquences

J'ai utilisé tout au long ma thèse l'alignement de séquences. Cependant, je me suis rendu compte que BLAST souffre de problèmes majeurs lorsqu'on l'utilise sur de très grandes bases de données. En effet, BLAST renvoie les X (par défaut 500) premiers alignements qu'il trouve et non pas les X meilleurs. Ce comportement, implémenté pour obtenir un gain de vitesse peut avoir pour conséquence que BLAST ne retrouve pas une séquence (identique à la séquence de départ) pourtant présente dans la base de donnée, si la base de donnée est de grande taille. Les alternatives à BLAST, comme DIAMOND et MMSEQS2, sont moins sensibles et consomment beaucoup plus de mémoire physique et vive. Après plusieurs tests et discussions avec des pairs, il semble aujourd'hui que MMSEQS2 soit la meilleure alternative à BLAST. MMSEQS2 retourne les X meilleurs alignements et non les X premiers. S'il consomme des quantités importantes de mémoire vive et physique, MMSEQS2 est beaucoup plus rapide que BLAST. Finalement, en terme de sensibilité, MMSEQS2 est proche de BLAST. Cet exemple illustre (de mon point de vue) comment les limites techniques conduisent au développement de nouveaux algorithmes et de nouvelles heuristiques pour les contourner. BLAST a été publié en 1990, DIAMOND en 2016 et

MMSEQS2 en 2017. En 2018, j'ai eu la chance de voir une présentation du Dr Dessimoz de l'Université de Lausanne. Son approche a pour objectif de faire une base de données de référence de famille d'homologues (OMA) et consiste à aligner avec Smith et Waterman toutes les séquences d'un jeu de données (Altenhoff et al. 2018). En effet, le Dr Dessimoz et son équipe souhaitent obtenir l'alignement optimal (bien que ce soit relatif aux paramètres utilisés pour l'alignement) et non une heuristique. Avec l'utilisation de ressources de calculs dédiées, en 14 ans, ils ont réussi à intégrer 2 500 génomes dans leur analyse. Bien que remarquable, ce chiffre, comparé au nombre de génomes complet présents dans nos bases de données (43 665), représente au mieux 5,7% de la diversité. Le recours aux heuristiques est donc toujours indispensable pour traiter les données de génomique environnementale avec des ressources limitées en un temps raisonnable.

5 Proposition du concept "d'autotrophie communautaire"

Nous avons montré la présence des gènes nécessaires à la réalisation de métabolismes autotrophes, notamment la fixation du carbone, dans la fraction de taille ultra-petite de TARA Océans. Nous ne pouvons dire si un organisme possède l'ensemble des séquences nécessaires à la réalisation d'une voie métabolique. Il est envisageable qu'un organisme parasite ne réalise qu'une partie bénéficiaire, pour lui, d'une voie métabolique. La présence de voies métaboliques incomplètes dans un organisme peut également être la signature d'un processus de réduction de génome en cours. Une corrélation entre la taille d'un génome et le volume cellulaire a d'ailleurs été décrite (Baker et al. 2010). Si les organismes ultra-petits ne contiennent qu'une partie des gènes permettant la fixation du carbone, on peut alors imaginer des relations syntrophiques au sein de communautés de procaryotes ultra-petits, en accord avec la théorie de la reine noire (Morris et al. 2012). Des organismes en interactions récurrentes peuvent développer des relations de syntrophie où une voie métabolique est répartie entre les différents organismes, ce qui serait compatible avec une réduction du génome mais aussi avec une répartition du coût de biosynthèse de cette voie métabolique. Cette hypothèse donne lieu à un problème de définition: si une voie métabolique dite autotrophe est réalisée par plusieurs organismes, elle ne peut

plus être qualifiée d'autotrophe, au niveau d'organisation des cellules individuelles. En effet, si la réalisation d'un métabolisme dépend de la présence d'organismes différents, cela contredit la définition d'autotrophie. Pour la transformation de matière inorganique en matière organique par différents organismes on pourrait alors parler "d'autotrophie communautaire", la communauté devenant l'unité pertinente de sélection de référence, à la place des individus (Doolittle et al. 2017). L'étude et la compréhension des relations syntrophiques et d'interdépendances au sein des communautés de micro-organismes sont donc désormais un champ de recherche dont les résultats pourraient modifier profondément notre compréhension de la vie microbienne (Pande et al. 2017; Libby et al. 2019; Embree et al. 2015). Il est possible que l'interdépendance et/ou la coopération métabolique soient la norme et non l'exception comme les études exclusives sur des organismes pouvant être cultivés en culture pure a pu le laisser penser.

6 Utilisation des graphes en microbiologie environnementale

Finalement, nous avons utilisé une approche originale basée sur la théorie des graphes pour étudier la diversité microbienne dans un jeu de donnée de métabarcoding. Cela nous a permis de comparer les séquences connues aux séquences environnementales par rapport à leurs positions dans le réseau et leurs associations préférentielles. Cette approche nous a permis d'identifier un nouveau clade d'Holozoa extérieur aux choanoflagellés, eucaryotes unicellulaires groupe frère des animaux. De plus, les réseaux peuvent permettre d'isoler les séquences qui résument le mieux la diversité (Forster et al. 2019). J'ai également, utilisé des propriétés des réseaux pour identifier des familles de gènes probablement antérieurs à la séparation archées bactéries. Ces familles de gènes permettront, je l'espère, d'identifier une diversité basale dans l'arbre du vivant.

Les réseaux sont une représentation des données adaptée pour décrire des relations entre des objets. Ils permettent d'extraire de l'information de ces relations, soit directement déduite de leurs structures soit calculée en utilisant des propriétés des réseaux. Leur application dans de nombreux domaines a pour conséquence qu'il existe de nombreux développements théoriques et pratiques les concernant. De plus, ils ont été appliqués avec succès à des

systèmes titanesques comme les réseaux sociaux et semblent donc être un outil de choix face au déluge de données provoqué par le séquençage haut débit. Une des limitations à l'emploi des réseaux, que j'ai pu observer durant ma thèse est le manque de solution pour la visualisation de très grand réseaux (plus de quelques millions d'objets). Il donc parfois impossible d'avoir une image du réseau et l'on doit ce contenter des descripteurs numériques. Les réseaux sont néanmoins particulièrement adaptés à l'étude des données de microbiologie environnementale où il existe différentes relations entre les différents objets biologiques. C'est la raison pour laquelle ils sont de plus en plus utilisés par la communauté scientifique pour étudier : les similarités de séquences, les relations d'homologie, les relations d'introgression, les co-occurrences dans l'environnement ou les communautés bactériennes, les voies métaboliques et les interactions (que ce soit une relation trophique, de parasitisme, de syntrophie ou de symbiose).

Conclusion et perspectives

Durant cette thèse, je me suis intéressé à la matière noire microbienne. Mon travail a permis de mettre en évidence le rôle écologique de micro-organismes ultra-petits dans certaines voies métaboliques autotrophiques des océans. Il a également permis de proposer que les CPR et DPANN jouent un rôle à ce jour inconnu, dans la dynamique des communautés microbiennes grâce à la présence de gènes impliqués dans des systèmes de quorum sensing. Il a permis de décrire une diversité jusque là inconnue d'Holozoa unicellulaires marins et de développer une méthode adaptée à la recherche d'homologues distants dans de grands jeux de données. Mon travail s'intègre ainsi dans l'évolution de la microbiologie moderne qui essaye d'une part de faire face aux jeux de données titanesques qu'elle produit mais aussi également de formuler des hypothèses et *in fine* d'améliorer nos modèles et notre compréhension des micro-organismes et de leurs interactions.

Le prolongement du travail de cette thèse est multiple:

- Notre analyse (3.2.1) révèle une potentielle diversité d'archées inconnues qui semblent posséder de nombreux gènes (certains très divergents) impliqués dans le métabolisme du carbone. L'isolation de ces archées, si elles existent, permettrait de décrire un nouveau groupe phylogénétique d'archées mais aussi possiblement de découvrir de nouveaux métabolismes de fixation du carbone.

- La confirmation *in vitro* ou *in vivo* des voies métaboliques "autotrophiques communautaires" mise en évidence dans les bactéries ultra-petites est souhaitable. S'il est avéré que les micro-organismes sont capables de réaliser des voies métaboliques dites autotrophes de manière communautaire, cela serait une découverte majeure de la microbiologie environnementale qui remettrait en question notre connaissance des interactions dans les communautés microbiennes.

- Un effort de séquençage de la diversité des Holozoa est nécessaire. En effet, avec la Dr Arroyo-Sanchez (3.5.1), nous avons montré qu'il existe une diversité d'Holozoa caractérisée uniquement par leurs fragments V9 du 18s RNA. Etudier la diversité des Holozoa peut nous permettre de mieux comprendre comment s'est effectuée la transition vers la multicellularité pour les animaux, nous informer sur leurs histoires

évolutives et inférer des caractéristiques de l'ancêtre commun aux animaux.

- Le développement de méthodes basées sur les réseaux pour l'étude de la microbiologie environnementale doit continuer. Notamment, la modélisation par des réseaux des relations syntrophiques et des capacités métaboliques d'un métagénome ou de GAMs, pourrait permettre d'aider à la prédiction de l'évolution des communautés bactériennes.

Références bibliographiques

Albertsen, Mads et al. (2013). "Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes". In: *Nat. Biotechnol.* 31.6, pp. 533–538. DOI: 10.1038/nbt.2579.

Allen, Rosalind J and Bartlomiej Waclaw (2019). *Bacterial growth: A statistical physicist's guide*. DOI: 10.1088/1361-6633/aae546.

Alneberg, Johannes et al. (2014). "Binning metagenomic contigs by coverage and composition". In: *Nat. Methods* 11.11, pp. 1144–1146. DOI: 10.1038/nmeth.3103.

Alneberg, Johannes et al. (2018). "Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes". In: *Microbiome* 6.1, p. 173. DOI: 10.1186/s40168-018-0550-0.

Altenhoff, Adrian M et al. (2018). "The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces." In: *Nucleic Acids Res.* 46.D1, pp. D477–D485. DOI: 10.1093/nar/gkx1019.

Altschul, Stephen F. et al. (1990). "Basic local alignment search tool". In: *J. Mol. Biol.* 215.3, pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Ambardar, Sheetal et al. (2016). "High Throughput Sequencing: An Overview of Sequencing Chemistry". In: *Indian J. Microbiol.* 56.4, pp. 394–404. DOI: 10.1007/s12088-016-0606-4.

Amit Roy, Samit Ray, Samit Ray, and Amit Roy (2014). "Molecular Markers in Phylogenetic Studies-A Review". In: *J. Phylogenetics Evol. Biol.* 02.02, pp. 1–9. DOI: 10.4172/2329-9002.1000131.

Anantharaman, Karthik et al. (2016a). "Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum". In: *PeerJ* 4, e1607. DOI: 10.7717/peerj.1607.

Anantharaman, Karthik et al. (2016b). "Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system". In: *Nat. Commun.* 7.1, p. 13219. DOI: 10.1038/ncomms13219.

Arber, W and S Linn (1969). "DNA Modification and Restriction". In: *Annu. Rev. Biochem.* 38.1, pp. 467–500. DOI: 10.1146/annurev.bi.38.070169.002343.

Azua-Bustos, Armando et al. (2019). "Aeolian transport of viable microbial life across the Atacama Desert, Chile: Implications for Mars". In: *Sci. Rep.* 9.1, p. 11024. DOI: 10.1038/s41598-019-47394-z.

Bahram, Mohammad et al. (2019). "Newly designed 16S rRNA metabarcoding primers amplify diverse and novel archaeal taxa from the environment". In: *Environ. Microbiol. Rep.* 11.4, p. 487. DOI: 10.1111/1758-2229.12684.

Baker, Brett J et al. (2010). "Enigmatic, ultrasmall, uncultivated Archaea." In: *Proc. Natl. Acad. Sci. U. S. A.* 107.19, pp. 8806–11. DOI: 10.1073/pnas.0914470107.

Bębenek, Anna and Izabela Ziuzia-Graczyk (2018). "Fidelity of DNA replication—a matter of proofreading". In: *Curr. Genet.* 64.5, pp. 985–996. DOI: 10.1007/s00294-018-0820-1.

Bedree, Joseph K et al. (2018). "Quorum Sensing Modulates the Epibiotic-Parasitic Relationship Between Actinomyces odontolyticus and Its Saccharibacteria epibiont, a Nanosynbacter lyticus Strain, TM7x." In: *Front. Microbiol.* 9, p. 2049. DOI: 10.3389/fmicb.2018.02049.

Belilla, Jodie et al. (2019). "Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area". In: *bioRxiv*, p. 658211. DOI: 10.1101/658211.

Benjak, Andrej, Claudia Sala, and Ruben C. Hartkoorn (2015). "Whole-Genome Sequencing for Comparative Genomics and De Novo Genome Assembly". In: *Methods Mol. Biol.* Vol. 1285, pp. 1–16. DOI: 10.1007/978-1-4939-2450-9_1.

Berg, Ivan A (2011). "Ecological aspects of the distribution of different autotrophic CO2 fixation pathways." In: *Appl. Environ. Microbiol.* 77.6, pp. 1925–36. DOI: 10.1128/AEM.02473-10.

Bernard, Guillaume et al. (2018). "Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery". In: *Genome Biol. Evol.* 10.3, pp. 707–715. DOI: 10.1093/gbe/evy031.

Blöchl, E et al. (1997). "Pyrolobus fumarii, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C." In: *Extremophiles* 1.1, pp. 14–21.

Boisvert, Sébastien et al. (2012). "Ray Meta: scalable de novo metagenome assembly and profiling". In: *Genome Biol.* 13.12, R122. DOI: 10.1186/gb-2012-13-12-r122.

Bowers, Robert M et al. (2017). "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea". In: *Nat. Biotechnol.* 35.8, pp. 725–731. DOI: 10.1038/nbt.3893.

Braman, Jeff. (2002). *In vitro mutagenesis protocols*. Humana Press, p. 287.

Brimacombe, R and W Stiege (1985). "Structure and function of ribosomal RNA." In: *Biochem. J.* 229.1, pp. 1–17. DOI: 10.1042/bj2290001.

Brock, T D and H Freeze (1969). "Thermus aquaticus gen. n. and sp. n., a nonsporulating extreme thermophile." In: *J. Bacteriol.* 98.1, pp. 289–97.

Brown, Christopher T. et al. (2015). "Unusual biology across a group comprising more than 15% of domain Bacteria". In: *Nature* 523.7559, pp. 208–211. DOI: 10.1038/nature14486.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson (2015). "Fast and sensitive protein alignment using DIAMOND". In: *Nat. Methods* 12.1, pp. 59–60. DOI: 10.1038/nmeth.3176.

Cano, R J and M K Borucki (1995). "Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber." In: *Science* 268.5213, pp. 1060–4. DOI: 10.1126/science.7538699.

Castelle, Cindy J. et al. (2018a). "Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations". In: *Nat. Rev. Microbiol.* 16.10, pp. 629–645. DOI: 10.1038/s41579-018-0076-2.

Castelle, Cindy J. and Jillian F. Banfield (2018b). "Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life". In: *Cell* 172.6, pp. 1181–1197. DOI: 10.1016/j.cell.2018.02.016.

Castelle, Cindy J. et al. (2015). "Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling". In: *Curr. Biol.* 25.6, pp. 690–701. DOI: 10.1016/j.cub.2015.01.014.

Chien, A, D B Edgar, and J M Trela (1976). "Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus." In: *J. Bacteriol.* 127.3, pp. 1550–7.

Chisholm, Sallie W. et al. (1992). "Prochlorococcus marinus nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll a and b". In: *Arch. Microbiol.* 157.3, pp. 297–300. DOI: 10.1007/BF00245165.

Delcher, Arthur L. et al. (2007). "Identifying bacterial genes and endosymbiont DNA with Glimmer". In: *Bioinformatics* 23.6, pp. 673–679. DOI: 10.1093/bioinformatics/btm009.

Dewhirst, Floyd E et al. (2010). "The human oral microbiome." In: *J. Bacteriol.* 192.19, pp. 5002–17. DOI: 10.1128/JB.00542-10.

Donoghue, Philip C J and Ziheng Yang (2016). "The evolution of methods for establishing evolutionary timescales." In: *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 371.1699. DOI: 10.1098/rstb.2016.0020.

Doolittle, W. Ford and Austin Booth (2017). "It's the song, not the singer: an exploration of holobiosis and evolutionary theory". In: *Biol. Philos.* 32.1, pp. 5–24. DOI: 10.1007/s10539-016-9542-2.

Dyall, S. D., Mark T Brown, and Patricia J Johnson (2004). "Ancient Invasions: From Endosymbionts to Organelles". In: *Science (80-.).* 304.5668, pp. 253–257. DOI: 10.1126/science.1094884.

Embree, Mallory et al. (2015). "Networks of energetic and metabolic interactions define dynamics in microbial communities." In: *Proc. Natl. Acad. Sci. U. S. A.* 112.50, pp. 15450–5. DOI: 10.1073/pnas.1506034112.

Errington, Jeff (2003). "Regulation of endospore formation in Bacillus subtilis". In: *Nat. Rev. Microbiol.* 1.2, pp. 117–126. DOI: 10.1038/nrmicro750.

Ferreira, A. C. et al. (1997). "Deinococcus geothermalis sp. nov. and Deinococcus murrayi sp. nov., Two Extremely Radiation-Resistant and Slightly Thermophilic Species from Hot Springs". In: *Int. J. Syst. Bacteriol.* 47.4, pp. 939–947. DOI: 10.1099/00207713-47-4-939.

Field, Christopher B. et al. (1998). "Primary production of the biosphere: integrating terrestrial and oceanic components". In: *Science* 281.5374, pp. 237–40. DOI: 10.1126/science.281.5374.237.

Fijalkowska, Iwona J., Roel M. Schaaper, and Piotr Jonczyk (2012). "DNA replication fidelity in Escherichia coli: a multi-DNA polymerase affair". In: *FEMS Microbiol. Rev.* 36.6, p. 1105. DOI: 10.1111/J.1574-6976.2012.00338.X.

Finn, Robert D et al. (2014). "Pfam: the protein families database." In: *Nucleic Acids Res.* 42.Database issue, pp. D222–30. DOI: 10.1093/nar/gkt1223.

Forster, Dominik et al. (2015). "Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms". In: *BMC Biol.* 13.1, p. 16. DOI: 10.1186/s12915-015-0125-5.

Forster, Dominik et al. (2019). "Improving eDNAbased protist diversity assessments using networks of amplicon sequence variants". In: *Environ. Microbiol.* Pp. 1462–2920.14764. DOI: 10.1111/1462-2920.14764.

Foster, Kevin R. and Thomas Bell (2012). "Competition, Not Cooperation, Dominates Interactions among Culturable Microbial Species". In: *Curr. Biol.* 22.19, pp. 1845–1850. DOI: 10.1016/J.CUB.2012.08.005.

Fowler, David et al. (2013). "The global nitrogen cycle in the Twentyfirst century". In: *Philos. Trans. R. Soc. B Biol. Sci.* 368.1621. DOI: 10.1098/rstb.2013.0164.

Fox, G E et al. (1977). "Classification of methanogenic bacteria by 16S ribosomal RNA characterization." In: *Proc. Natl. Acad. Sci. U. S. A.* 74.10, pp. 4537–41. DOI: 10.1073/pnas.74.10.4537.

Friedberg, Errol C. (2003). "DNA damage and repair". In: *Nature* 421.6921, pp. 436–440. DOI: 10.1038/nature01408.

Ganai, Rais A. and Erik Johansson (2016). "DNA Replication—A Matter of Fidelity". In: *Mol. Cell* 62.5, pp. 745–755. DOI: 10.1016/j.molcel.2016.05.003.

Gardner, Richard C. et al. (1981). "The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing". In: *Nucleic Acids Res.* 9.12, pp. 2871–2888. DOI: 10.1093/nar/9.12.2871.

Garg, Sriram G. et al. (2019). "Anomalous phylogenetic behavior of ribosomal proteins in metagenome assembled genomes". In: *bioRxiv*, p. 731091. DOI: 10.1101/731091.

Gómez, Felipe et al. (2019). "Ultra-small microorganisms in the polyextreme conditions of the Dallol volcano, Northern Afar, Ethiopia". In: *Sci. Rep.* 9.1, p. 7907. DOI: 10.1038/s41598-019-44440-8.

Gong, Xue et al. (2019). "Alterations in the human gut microbiome in anti N methylDaspartate receptor encephalitis". In: *Ann. Clin. Transl. Neurol.* acn3.50874. DOI: 10.1002/acn3.50874.

Gray, Michael W. (1988). "Organelle origins and ribosomal RNA". In: *Biochem. Cell Biol.* 66.5, pp. 325–348. DOI: 10.1139/o88-042.

Gray, Michael W, Gertraud Burger, and Franz Lang (1999). "Mitochondrial Evolution Michael". In: *Science (80-.).* 283.5407, pp. 1476–1481. DOI: 10.1101/cshperspect.a011403.

Gumsley, Ashley P et al. (2017). "Timing and tempo of the Great Oxidation Event." In: *Proc. Natl. Acad. Sci. U. S. A.* 114.8,

pp. 1811–1816. DOI: 10.1073/pnas.1608824114.

Hadziavdic, Kenan et al. (2014). "Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers". In: *PLoS One* 9.2. Ed. by Christian R. Voolstra, e87624. DOI: 10.1371/journal.pone.0087624.

Harris, J Kirk, Scott T Kelley, and Norman R Pace (2004). "New perspective on uncultured bacterial phylogenetic division OP11." In: *Appl. Environ. Microbiol.* 70.2, pp. 845–9. DOI: 10.1128/aem.70.2.845-849.2004.

Henkin, Tina M (2016). "Classic Spotlight: Bacterial Endospore Resistance, Structure, and Genetics." In: *J. Bacteriol.* 198.14, p. 1904. DOI: 10.1128/JB.00312-16.

Hug, Laura A. et al. (2016). "A new view of the tree of life". In: *Nat. Microbiol.* 1.5, p. 16048. DOI: 10.1038/nmicrobiol.2016.48.

Hugenholtz, P et al. (1998). "Novel division level bacterial diversity in a Yellowstone hot spring." In: *J. Bacteriol.* 180.2, pp. 366–76.

Hyatt, Doug et al. (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification". In: *BMC Bioinformatics* 11, p. 119. DOI: 10.1186/1471-2105-11-119.

Hyatt, Doug et al. (2012). "Gene and translation initiation site prediction in metagenomic sequences". In: *Bioinformatics* 28.17, pp. 2223–2230. DOI: 10.1093/bioinformatics/bts429.

Imachi, Hiroyuki et al. (2019). "Isolation of an archaeon at the prokaryote-eukaryote interface". In: *bioRxiv*, p. 726976. DOI: 10.1101/726976.

Jaffe, Alexander L et al. (2019). "Lateral Gene Transfer Shapes the Distribution of RuBisCO among Candidate Phyla Radiation Bacteria and DPANN Archaea". In: *Mol. Biol. Evol.* 36.3. Ed. by Daniel Falush, pp. 435–446. DOI: 10.1093/molbev/msy234.

Jahn, Ulrike et al. (2008). "Nanoarchaeum equitans and Ignicoccus hospitalis: new insights into a unique, intimate association of two archaea." In: *J. Bacteriol.* 190.5, pp. 1743–50. DOI: 10.1128/JB.01731-07.

Jain, Miten et al. (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads". In: *Nat. Biotechnol.* 36.4, pp. 338–345. DOI: 10.1038/nbt.4060.

Johnson, Paul W. and John McN. Sieburth (1979). "Chroococcoid cyanobacteria in the sea: A ubiquitous and diverse phototrophic biomass1". In: *Limnol. Oceanogr.* 24.5, pp. 928–935. DOI: 10.4319/lo.1979.24.5.0928.

Johnson, R E et al. (2000). "Fidelity of human DNA polymerase eta." In: *J. Biol. Chem.* 275.11, pp. 7447–50. DOI: 10.1074/jbc.275.11.7447.

Kang, Dongwan D. et al. (2015). "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities". In: *PeerJ* 3, e1165. DOI: 10.7717/peerj.1165.

Karnkowska, Anna et al. (2016). "A eukaryote without a mitochondrial organelle". In: *Curr. Biol.* 26.10, pp. 1274–1284. DOI: 10.1016/j.cub.2016.03.053.

Kim, Jongsun and Douglas C. Rees (1994). "Nitrogenase and Biological Nitrogen Fixation". In: *Biochemistry* 33.2, pp. 389–397. DOI: 10.1021/bi00168a001.

KIMURA, MOTOO (1968). "Evolutionary Rate at the Molecular Level". In: *Nature* 217.5129, pp. 624–626. DOI: 10.1038/217624a0.

Kimura, Motoo (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences". In: *J. Mol. Evol.* 16.2, pp. 111–120. DOI: 10.1007/BF01731581.

Kuehbacher, T. et al. (2008). "Intestinal TM7 bacterial phylogenies in active inflammatory bowel disease". In: *J. Med. Microbiol.* 57.12, pp. 1569–1576. DOI: 10.1099/jmm.0.47719-0.

Kumar, Sudhir (2005). "Molecular clocks: four decades of evolution". In: *Nat. Rev. Genet.* 6.8, pp. 654–662. DOI: 10.1038/nrg1659.

Lafontaine, Denis L.J. and David Tollervey (2001). "The function and synthesis of ribosomes". In: *Nat. Rev. Mol. Cell Biol.* 2.7, pp. 514–520. DOI: 10.1038/35080045.

Lake, James A. (1988). "Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences". In: *Nature* 331.6152, pp. 184–186. DOI: 10.1038/331184a0.

Landmark, Donald G. and Milkos Müller (1973). "a Cytoplasmic Organelle Flagellate Tritrichomonas foetus , and Its Role in

Pvruvate Metabolism *". In: *J. Biol. Chem.* 248.22, pp. 7724–7729.

Lane, D J et al. (1985). "Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses." In: *Proc. Natl. Acad. Sci. U. S. A.* 82.20, pp. 6955–9. DOI: 10.1073/pnas.82.20.6955.

Lannes, Romain et al. (2019). "Carbon Fixation by Marine Ultrasmall Prokaryotes". In: *Genome Biol. Evol.* 11.4. Ed. by Laura A Katz, pp. 1166–1177. DOI: 10.1093/gbe/evz050.

Lebowitz, Jacob, Marc S. Lewis, and Peter Schuck (2009). "Modern analytical ultracentrifugation in protein science: A tutorial review". In: *Protein Sci.* 11.9, pp. 2067–2079. DOI: 10.1110/ps.0207702.

LEHMAN, I R et al. (1958). "Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli." In: *J. Biol. Chem.* 233.1, pp. 163–70.

Li, Dinghua et al. (2015). "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph". In: *Bioinformatics* 31.10, pp. 1674–1676. DOI: 10.1093/bioinformatics/btv033.

Libby, Eric et al. (2019). "Syntrophy emerges spontaneously in complex metabolic systems". In: *PLOS Comput. Biol.* 15.7. Ed. by James O'Dwyer, e1007169. DOI: 10.1371/journal.pcbi.1007169.

Lin, Hsin-Hung and Yu-Chieh Liao (2016). "Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes". In: *Sci. Rep.* 6.1, p. 24175. DOI: 10.1038/srep24175.

Ling, Zongxin et al. (2014). "Altered fecal microbiota composition associated with food allergy in infants." In: *Appl. Environ. Microbiol.* 80.8, pp. 2546–54. DOI: 10.1128/AEM.00003-14.

Luef, Birgit et al. (2015). "Diverse uncultivated ultra-small bacterial cells in groundwater". In: *Nat. Commun.* 6.1, p. 6372. DOI: 10.1038/ncomms7372.

Lukashin, A. (1998). "GeneMark.hmm: new solutions for gene finding". In: *Nucleic Acids Res.* 26.4, pp. 1107–1115. DOI: 10.1093/nar/26.4.1107.

Martin, William F. et al. (2017). "The Physiology of Phagocytosis in the Context of Mitochondrial Origin". In: *Microbiol. Mol. Biol. Rev.* 81.3. DOI: 10.1128/MMBR.00008-17.

Matsakis, Nicholas D. et al. (2014). "The rust language". In: *Proc. 2014 ACM SIGAda Annu. Conf. High Integr. Lang. Technol. - HILT '14*. Vol. 34. 3. New York, New York, USA: ACM Press, pp. 103–104. DOI: 10.1145/2663171.2663188.

Maxam, A. M. and W. Gilbert (1977). "A new method for sequencing DNA." In: *Proc. Natl. Acad. Sci.* 74.2, pp. 560–564. DOI: 10.1073/pnas.74.2.560.

McDonald, Daniel et al. (2012). "An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea." In: *ISME J.* 6.3, pp. 610–8. DOI: 10.1038/ismej.2011.139.

McGrath, Annette (2007). "Chapter 11 Genome Sequencing and Assembly". In: *Perspect. Bioanal.* 2, pp. 327–355. DOI: 10.1016/S1871-0069(06)02011-8.

McNally, Luke and Sam P Brown (2015). "Building the microbiome in health and disease: niche construction and social conflict in bacteria." In: *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 370.1675. DOI: 10.1098/rstb.2014.0298.

Medvedev, Paul et al. (2011). "Paired de Bruijn Graphs: A Novel Approach for Incorporating Mate Pair Information into Genome Assemblers". In: *J. Comput. Biol.* 18.11, pp. 1625–1634. DOI: 10.1089/cmb.2011.0151.

Merlino, Giuseppe et al. (2018). *Microbial ecology of deep-sea hypersaline anoxic basins*. DOI: 10.1093/femsec/fiy085.

Miller, Melissa B. and Bonnie L. Bassler (2001). "Quorum Sensing in Bacteria". In: *Annu. Rev. Microbiol.* 55.1, pp. 165–199. DOI: 10.1146/annurev.micro.55.1.165.

Mindell, David P. (1992). "Phylogenetic consequences of symbioses: Eukarya and Eubacteria are not monophyletic taxa". In: *Biosystems* 27.1, pp. 53–62. DOI: 10.1016/0303-2647(92)90046-2.

Moran, Nancy A. and Daniel B. Sloan (2015). "The Hologenome Concept: Helpful or Hollow?" In: *PLOS Biol.* 13.12, e1002311. DOI: 10.1371/journal.pbio.1002311.

Morris, J Jeffrey, Richard E Lenski, and Erik R Zinser (2012). "The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss." In: *MBio* 3.2, e00036–12. DOI: 10.1128/mBio.00036-12.

Mullis, K.B. et al. (1989). *Process for amplifying, detecting, and/or cloning nucleic acid sequences*.

Namiki, Toshiaki et al. (2012). "MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads". In: *Nucleic Acids Res.* 40.20, e155–e155. DOI: 10.1093/nar/gks678.

Nap, Jan Peter and Ton Bisseling (1990). *Developmental biology of a plant-prokaryote symbiosis: The legume root nodule*. DOI: 10.1126/science.250.4983.948.

Nurk, Sergey et al. (2017). "metaSPAdes: a new versatile metagenomic assembler". In: *Genome Res.* 27.5, pp. 824–834. DOI: 10.1101/gr.213959.116.

Ohno, Susumu (1970). *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-86659-3.

Olivier, Valérie et al. (2018). "Micro-inflammation et translocation bactérienne d'origine digestive dans la maladie rénale chronique". In: *Néphrologie & Thérapeutique* 14.3, pp. 135–141. DOI: 10.1016/j.nephro.2017.10.005.

Olsen, G J and C R Woese (1993). "Ribosomal RNA: a key to phylogeny." In: *FASEB J.* 7.1, pp. 113–123. DOI: 10.1096/fasebj.7.1.8422957.

O'Malley, Maureen A. (2014). *Philosophy of microbiology*. Cambridge: Cambridge University Press, pp. 1–269. DOI: 10.1017/CBO9781139162524.

Overmann, Jörg, Birte Abt, and Johannes Sikorski (2017). "Present and Future of Culturing Bacteria". In: *Annu. Rev. Microbiol.* 71.1, pp. 711–730. DOI: 10.1146/annurev-micro-090816-093449.

Pace, Norman R. (2006). "Time for a change". In: *Nature* 441.7091, pp. 289–289. DOI: 10.1038/441289a.

Pande, Samay and Christian Kost (2017). "Bacterial Unculturability and the Formation of Intercellular Metabolic Networks". In: *Trends Microbiol.* 25.5, pp. 349–361. DOI: 10.1016/J.TIM.2017.02.015.

Parada, Alma E., David M. Needham, and Jed A. Fuhrman (2016). "Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples". In: *Environ. Microbiol.* 18.5, pp. 1403–1414. DOI: 10.1111/1462-2920.13023.

Parks, Donovan H. et al. (2015). "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes". In: *Genome Res.* 25.7, p. 1043. DOI: 10.1101/GR.186072.114.

Parks, Donovan H et al. (2018). "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life". In: *Nat. Biotechnol.* 36.10, pp. 996–1004. DOI: 10.1038/nbt.4229.

Partensky, F., W. R. Hess, and D. Vaulot (1999). "Prochlorococcus, a Marine Photosynthetic Prokaryote of Global Significance". In: *Microbiol. Mol. Biol. Rev.* 63.1, p. 106.

Petsch, S. T. (2013). "The Global Oxygen Cycle". In: *Treatise Geochemistry Second Ed.* Vol. 10. Elsevier, pp. 437–473. DOI: 10.1016/B978-0-08-095975-7.00811-1.

Pikuta, Elena V., Richard B. Hoover, and Jane Tang (2007). "Microbial Extremophiles at the Limits of Life". In: *Crit. Rev. Microbiol.* 33.3, pp. 183–209. DOI: 10.1080/10408410701451948.

Ponomarova, Olga and Kiran Raosaheb Patil (2015). "Metabolic interactions in microbial communities: untangling the Gordian knot". In: *Curr. Opin. Microbiol.* 27, pp. 37–44. DOI: 10.1016/J.MIB.2015.06.014.

Pontefract, Alexandra et al. (2017). "Microbial diversity in a hypersaline sulfate lake: A terrestrial analog of ancient mars". In: *Front. Microbiol.* 8.SEP, p. 1819. DOI: 10.3389/fmicb.2017.01819.

Postler, Thomas Siegmund and Sankar Ghosh (2017). "Understanding the Holobiont: How

Microbial Metabolites Affect Human Health and Shape the Immune System". In: *Cell Metab.* 26.1, pp. 110–130. DOI: 10.1016/J.CMET.2017.05.008.

Probst, Alexander J. et al. (2017). "Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO ₂ concentrations". In: *Environ. Microbiol.* 19.2, pp. 459–474. DOI: 10.1111/1462-2920.13362.

Probst, Alexander J. et al. (2018). "Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface". In: *Nat. Microbiol.* 3.3, pp. 328–336. DOI: 10.1038/s41564-017-0098-y.

Quince, Christopher et al. (2017). "Shotgun metagenomics, from sampling to analysis". In: *Nat. Biotechnol.* 35.9, pp. 833–844. DOI: 10.1038/nbt.3935.

Reuter, Jason A, Damek V Spacek, and Michael P Snyder (2015). "High-throughput sequencing technologies." In: *Mol. Cell* 58.4, pp. 586–97. DOI: 10.1016/j.molcel.2015.05.004.

Rinke, Christian et al. (2013). "Insights into the phylogeny and coding potential of microbial dark matter". In: *Nature* 499.7459, pp. 431–437. DOI: 10.1038/nature12352.

Rosano, Germán L and Eduardo A Ceccarelli (2014). "Recombinant protein expression in Escherichia coli: advances and challenges." In: *Front. Microbiol.* 5, p. 172. DOI: 10.3389/fmicb.2014.00172.

Round, June L. and Sarkis K. Mazmanian (2009). "The gut microbiota shapes intestinal immune responses during health and disease". In: *Nat. Rev. Immunol.* 9.5, pp. 313–323. DOI: 10.1038/nri2515.

Ruppert, Krista M., Richard J. Kline, and Md Saydur Rahman (2019). "Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA". In: *Glob. Ecol. Conserv.* 17, e00547. DOI: 10.1016/J.GECCO.2019.E00547.

Sagan, Lynn (1967). "On the origin of mitosing cells". In: *J. Theor. Biol.* 14.3, 225–IN6. DOI: 10.1016/0022-5193(67)90079-3.

Saitou, N and M Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." In: *Mol. Biol. Evol.* 4.4, pp. 406–425. DOI: 10.1093/oxfordjournals.molbev.a040454.

SANGER, F and H TUPPY (1951a). "The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates." In: *Biochem. J.* 49.4, pp. 481–90. DOI: 10.1042/bj0490481.

— (1951b). "The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates." In: *Biochem. J.* 49.4, pp. 463–81. DOI: 10.1042/bj0490463.

Sanger, F. and A.R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". In: *J. Mol. Biol.* 94.3, pp. 441–448. DOI: 10.1016/0022-2836(75)90213-2.

Sanger, F, S Nicklen, and A R Coulson (1977). "DNA sequencing with chain-terminating inhibitors." In: *Proc. Natl. Acad. Sci. U. S. A.* 74.12, pp. 5463–7. DOI: 10.1073/pnas.74.12.5463.

Schirrmeister, Bettina E, Muriel Gugger, and Philip C J Donoghue (2015). "Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils." In: *Palaeontology* 58.5, pp. 769–785. DOI: 10.1111/pala.12178.

Sharon, Itai et al. (2013). "Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization". In: *Genome Res.* 23.1, pp. 111–120. DOI: 10.1101/GR.142315.112.

Shendure, Jay A. et al. (2011). "Overview of DNA Sequencing Strategies". In: *Curr. Protoc. Mol. Biol.* 96.1, pp. 7.1.1–7.1.23. DOI: 10.1002/0471142727.mb0701s96.

Shuman, Stewart (2009). "DNA ligases: progress and prospects." In: *J. Biol. Chem.* 284.26, pp. 17365–9. DOI: 10.1074/jbc.R900017200.

Siekevitz, Philip (1957). "Powerhouse of the Cell". In: *Sci. Am.* 197.1, pp. 131–144. DOI: 10.1038/scientificamerican0757-131.

Sims, David et al. (2014). “Sequencing depth and coverage: key considerations in genomic analyses”. In: *Nat. Rev. Genet.* 15.2, pp. 121–132. DOI: 10.1038/nrg3642.

Smith, T.F. and M.S. Waterman (1981). “Identification of common molecular subsequences”. In: *J. Mol. Biol.* 147.1, pp. 195–197. DOI: 10.1016/0022-2836(81)90087-5.

Sohn, Jang-il and Jin-Wu Nam (2016). “The present and future of de novo whole-genome assembly”. In: *Brief. Bioinform.* 19.1, bbw096. DOI: 10.1093/bib/bbw096.

Soucy, Shannon M., Jinling Huang, and Johann Peter Gogarten (2015). “Horizontal gene transfer: building the web of life”. In: *Nat. Rev. Genet.* 16.8, pp. 472–482. DOI: 10.1038/nrg3962.

Sousa, Filipa L. et al. (2016). “Lokiarchaeon is hydrogen dependent”. In: *Nat. Microbiol.* 1.5, p. 16034. DOI: 10.1038/nmicrobiol.2016.34.

Spang, Anja et al. (2015). “Complex archaea that bridge the gap between prokaryotes and eukaryotes”. In: *Nature* 521.7551, pp. 173–179. DOI: 10.1038/nature14447.

Staden, R. (1979). “A strategy of DNA sequencing employing computer programs”. In: *Nucleic Acids Res.* 6.7, pp. 2601–2610. DOI: 10.1093/nar/6.7.2601.

Staley, J T and A Konopka (1985). “Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats”. In: *Annu. Rev. Microbiol.* 39.1, pp. 321–346. DOI: 10.1146/annurev.mi.39.100185.001541.

Steinegger, Martin and Johannes Söding (2017). “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets”. In: *Nat. Biotechnol.* 35.11, pp. 1026–1028. DOI: 10.1038/nbt.3988.

Sun, Dong-Lei et al. (2013). “Intragenomic Heterogeneity of 16S rRNA Genes Causes Overestimation of Prokaryotic Diversity”. In: *Appl. Environ. Microbiol.* 79.19, p. 5962. DOI: 10.1128/AEM.01282-13.

Sunagawa, Shinichi et al. (2015). “Ocean plankton. Structure and function of the global ocean microbiome.” In: *Science* 348.6237, p. 1261359. DOI: 10.1126/science.1261359.

Theis, Kevin R et al. (2016). “Getting the Hologenome Concept Right: an Eco-Evolutionary Framework for Hosts and Their Microbiomes.” In: *mSystems* 1.2, e00028–16. DOI: 10.1128/mSystems.00028-16.

Timasheff, Serge N. et al. (1958). “The molecular weight of ribonucleic acid prepared from ascites-tumor cells”. In: *Biochim. Biophys. Acta* 27.3, pp. 662–663. DOI: 10.1016/0006-3002(58)90412-8.

Uritskiy, Gherman et al. (2019). “Halophilic microbial community compositional shift after a rare rainfall in the Atacama Desert”. In: *ISME J.* DOI: 10.1038/s41396-019-0468-y.

Van Der Giezen, Mark (2009). “Hydrogenosomes and mitosomes: Conservation and evolution of functions1”. In: *J. Eukaryot. Microbiol.* Vol. 56. 3, pp. 221–231. DOI: 10.1111/j.1550-7408.2009.00407.x.

Venter, J C et al. (2001). “The sequence of the human genome.” In: *Science* 291.5507, pp. 1304–51. DOI: 10.1126/science.1058040.

Watson, Andrew K. et al. (2019). “The Methodology Behind Network Thinking: Graphs to Analyze Microbial Complexity and Evolution”. In: *Methods Mol. Biol.* Vol. 1910, pp. 271–308. DOI: 10.1007/978-1-4939-9074-0_9.

Wheeler, David A. et al. (2008). “The complete genome of an individual by massively parallel DNA sequencing”. In: *Nature* 452.7189, pp. 872–876. DOI: 10.1038/nature06884.

Whitfield, John (2002). “Portrait of a serial killer”. In: *Nature*. DOI: 10.1038/news021001-6.

Whitman, W B, D C Coleman, and W J Wiebe (1998). “Prokaryotes: the unseen majority.” In: *Proc. Natl. Acad. Sci. U. S. A.* 95.12, pp. 6578–83. DOI: 10.1073/pnas.95.12.6578.

Woese, C R (1987). “Bacterial evolution.” In: *Microbiol. Rev.* 51.2, pp. 221–71.

Woese, C. R. and G. E. Fox (1977). “Phylogenetic structure of the prokaryotic domain: The primary kingdoms”. In: *Proc. Natl. Acad. Sci.*

74.11, pp. 5088–5090. DOI: 10.1073/pnas.74.11.5088.

Woese, C R, O Kandler, and M L Wheelis (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." In: *Proc. Natl. Acad. Sci. U. S. A.* 87.12, p. 4576. DOI: 10.1073/PNAS.87.12.4576.

Wrighton, Kelly C et al. (2012). "Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla." In: *Science* 337.6102, pp. 1661–5. DOI: 10.1126/science.1224041.

Wrighton, Kelly C et al. (2014). "Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer". In: *ISME J.* 8.7, pp. 1452–1463. DOI: 10.1038/ismej.2013.249.

Wrighton, Kelly C et al. (2016). "RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria". In: *ISME J.* 10.11, pp. 2702–2714. DOI: 10.1038/ismej.2016.53.

Wu, Yu-Wei, Blake A. Simmons, and Steven W. Singer (2016). "MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets". In: *Bioinformatics* 32.4, pp. 605–607. DOI: 10.1093/bioinformatics/btv638.

Yang, D et al. (1985). "Mitochondrial origins." In: *Proc. Natl. Acad. Sci. U. S. A.* 82.13, pp. 4443–7. DOI: 10.1073/pnas.82.13.4443.

Yarza, Pablo et al. (2014). "Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences". In: *Nat. Rev. Microbiol.* 12.9, pp. 635–645. DOI: 10.1038/nrmicro3330.

Yilmaz, Pelin et al. (2014). "The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks". In: *Nucleic Acids Res.* 42.D1, pp. D643–D648. DOI: 10.1093/nar/gkt1209.

Zhang, F., Y. Wen, and X. Guo (2014). "CRISPR/Cas9 for genome editing: progress, implications and challenges". In: *Hum. Mol. Genet.* 23.R1, R40–R46. DOI: 10.1093/hmg/ddu125.

Zhang, Haowen, Chirag Jain, and Srinivas Aluru (2019). "A comprehensive evaluation of long read error correction methods". In: *bioRxiv*, p. 519330. DOI: 10.1101/519330.

Zuckerkandl, Emile and Linus Pauling (1965). "Molecules as documents of evolutionary history". In: *J. Theor. Biol.* 8.2, pp. 357–366. DOI: 10.1016/0022-5193(65)90083-4.

Résumé

L'objectif de cette thèse a été d'identifier des micro-organismes encore inconnus présents dans divers environnements et de caractériser certains de leurs métabolismes. Cette diversité non identifiée, à la fois taxonomique et fonctionnelle, est communément appelée matière noire microbienne. J'ai utilisé et développé de nouvelles méthodes de réseaux, et notamment des réseaux de similarité de séquences, afin d'exploiter de très grands jeux de données de séquences, issus de projets de métagénomique. En particulier, mon travail a mis en évidence le rôle écologique de micro-organismes ultra-petits dans certaines voies métaboliques autotrophes des océans. Il montre également que les CPR et DPANN, bactéries et archées ultra-petites récemment découvertes, participent à la dynamique des communautés microbiennes via des systèmes de quorum sensing homologues à ceux d'organismes mieux caractérisés. Une application des réseaux de similarité de séquences à des données de métabarcoding a également révélé une diversité jusque là inconnue d'Holozoa, qui pourrait nous permettre de mieux comprendre la transition vers la multicellularité des Metazoa. Enfin, j'ai développé une méthode et un logiciel destiné à la recherche d'homologues distants de protéines d'intérêt dans de très grands jeux de données, tels que ceux issus de la métagénomique. Cette méthode, maintenant validée, devrait permettre de rechercher des séquences appartenant à des organismes encore inconnus et très divergents, dans l'espoir de découvrir de nouveaux phylums profonds, voire même de nouveaux domaines du vivant.

Summary

The objective of this thesis was to identify as yet unknown microorganisms present in various environments and to characterize some of their metabolisms. This unidentified diversity, both taxonomic and functional, is commonly referred to as microbial dark matter. I have used and developed new network methods, including sequence similarity networks, to exploit very large sequence datasets from metagenomic projects. In particular, my work has highlighted the ecological role of ultra-small micro-organisms in some autotrophic metabolic pathways in the oceans. It also shows that CPR and DPANN, recently discovered ultra-small bacteria and archaea, participate in the dynamics of microbial communities through quorum sensing systems similar to those of better characterized organisms. An application of sequence similarity networks to meta-barcoding data also revealed a previously unknown diversity of Holozoans, which could allow us to better understand the transition to multicellularity of Metazoans. Finally, I have developed a method and software for searching for remote homologs of proteins of interest in very large datasets, such as those from metagenomics. This method, now validated, should make it possible to search for sequences belonging to still unknown and very divergent organisms, in the hope of discovering new deep branching phyla, or even new domains of life.