Biodiversité microbienne environnementale : des gènes et des génomes

Gisèle Bronner

Université Clermont Auvergne

Biodiversité microbienne environnementale : des gènes et des génomes

Gisèle BRONNER

MCU (65) Université Clermont Auvergne

UMR – CNRS 6023 : Laboratoire microorganismes : Génome et Environnement

École Doctorale : Science de la Vie, Santé, Agronomie, Environnement

en vue de l'obtention de l'Habilitation à diriger des recherches

13 Juillet 2021

Rapporteurs

GARCZAREK Laurence, DR CNRS – CNRS / Université Paris Sorbonne

HILL David, PR – Université Clermont Auvergne

PERRIERE Guy, DR CNRS – CNRS / Université de Lyon

Examinateurs

BITTNER Lucie, MCU MNHN / Université Paris Sorbonne

DEBROAS Didier, PR – Université Clermont Auvergne

Laboratoire Microorganismes : Génome et Environnement

Sommaire

Présentation

1 - Curriculum Vitae **3**

2 - Synthèse de la carrière **5**

3 - Activité scientifique **6**

4 - Responsabilités collectives **11**

5 - Activité pédagogique **12**

Mémoire d'habilitaiton

1 – Biodiversité **17**

2 - Biodiversité microbienne **18**

3– Contexte de génomique environnementale **19**

4– Positionnement du travail présenté dans ce rapport **19**

PARTIE I : APPROCHE *METABARCODING* ou LA DECRIPTION DE LA DIVERSITE SPECIFIQUE

1 – Notion d'espèce et définition opérationnelle de l'espèce en microbiologie **23**

1.1 – Les OTUs, mesure opérationnelle de la richesse spécifique 26

1.2– Limites de la notion d'OTU dans un contexte de séquençage de seconde génération

2– Traitement des données de *metabarconding* issues de NGS : PANAM **32**

2.1 – PANAM pour une affiliation phylogénétique des séquences d'amplicons 33

2.2 - Performance des affiliations phylogénétiques sur amplicons NGS 36

2.3 – Quelques illustrations 38

3 – Une meilleure estimation de la diversité ? **45**

3.1 – Les variants exacts de séquences 45

3.2 – Apport des unités phylogénétiques de diversité 51

4 – Perspectives **53**

PARTIE II : APPROCHE CELLULE UNIQUE ; DYNAMIQUE EVOLUTIVE DU PANGENOME A L'ECHELLE D'UNE POPULATION NATURELLE

Préambule **57**

1 - Le concept de pangénome **59**

1.1 - De l'espèce phénotypique au pangénome 59

1.2 – Pangénome, transfert de gènes et unité de diversité 60

1.3 – Dimension écologique du pangénome 6**1**

2 – Controverse sur l'origine du pangénome **63**

2.1 – Pourquoi un pangénome ? 63

2.2 – Une organisation génomique en îlots 64

2.3 – Étudier le pangénome à partir de populations naturelles 65

3 - Analyse pangénomique d'une population bactérienne environnementale **66**

3.1 – Quelle représentativité d'un pangénome populationnel ? 67

3.2 – Organisation du pangénome de sous-populations de *Prochlorococcus marinus* HLII

3.3 – Caractérisation des COGs flexibles spécifiques des sous-populations

4– Dynamique évolutive des compartiments génomiques **74**

4.1 - Taux de substitution et signatures des compartiments génomiques 75

4.2 - Signature évolutive et distribution des COGs dans les sous-populations 77

5 – Bilan et perspectives **78**

PARTIE III : « LA MATIERE NOIRE EUCARYOTE » ET AUTRES TRAVAUX

1 – Le projet MICROSTORE **85**

1.1 – Mise en place d'une chaîne de traitement pour l'annotation des transcriptomes

1.2 – Perspectives : Phylogénomique des eucaryotes unicellulaires lacustres 91

2 – Représentation des connaissances en cartographie comparée des génomes de mammifères

3 - Biodiversité et génomique végétale **94**

4 - Big data, grille et complexité biologique **95**

Références Bibliographiques **97**

Mémoire d’habilitation

1 – Biodiversité

La diversité de la nature n'est pas continue mais consiste en des entités discrètes composées d'individus et séparées les unes des autres par des discontinuités. Celles-ci, désignées sous le terme d'*espèce,* sont considérées comme les unités de base de la diversité et sont les unités fondamentales considérées par le biologiste. Les espèces sont cependant composées de nombreux individus caractérisés par des phénotypes *(phena)* variés et, lorsque ces phénotypes sont très hétérogènes, il peut arriver que des individus appartenant effectivement à une même espèce soient affiliés à tort à des espèces différentes (Figure 1.1). La notion de biodiversité inclut non seulement l'ensemble des espèces et leur histoire évolutive, mais aussi la variabilité génétique au sein et entre populations d'espèces (une population est un groupe d'individus se reproduisant entre eux plus fréquemment qu'avec des individus extérieurs à la population Freudenstein *et al.* 2017), ainsi que la répartition de celles-ci dans les habitats locaux, les écosystèmes et les paysages (National Research Council (US) Committee on Noneconomic and Economic Value of Biodiversity, 1999). Les associations d'espèces dans un environnement (les communautés) reflètent à la fois l'histoire des "stocks" présents à un endroit à un moment donné et les réponses différentes des communautés à des différences physico-chimiques de l'environnement et des interactions entre les membres de la communauté.

Figure 1.1 : Dimorphisme sexuel ayant conduit à la définition de deux espèces différentes. Le mâle bleu irisé du -satyr céruléen *Caeruleuptychia helios* (à gauche) et la femelle (à droite) ont été affiliées à la même espèce sur la base d'une analyse de leur l'ADN (d'après Nakahara *et al.* 2018).

La manière la plus immédiate pour appréhender la biodiversité au sein d'un écosystème consiste à dénombrer les espèces (plus généralement les taxa, c'est-à-dire des entités conceptuelles regroupant des êtres vivants sur la base de caractères partagés) présentes dans un écosystème, en les pondérant par leur abondance, ou en les replaçant dans un contexte phylogénétique. On estime alors une diversité alpha. Lorsque l'on compare différents écosystèmes, ceux-ci peuvent partager un nombre d'espèces

semblables bien que leurs compositions taxonomiques soient différentes. On peut alors évaluer la différence « compositionnelle » dans « l'environnement » par le calcul de la diversité beta. Les diversités alpha et beta permettent de caractériser les unités de biodiversité, mais pas d'évaluer les interactions spécifiques, ni le rôle que peuvent jouer les espèces - individuellement ou collectivement - dans le fonctionnement des écosystèmes (National Research Council (US) Committee on Noneconomic and Economic Value of Biodiversity, 1999). Ces derniers aspects sont appréhendés à travers l'étude de la biodiversité fonctionnelle, définie comme « la variation des traits biologiques dans l'espace fonctionnel occupé par une unité écologique » (Escalas *et al.* 2019). Les traits fonctionnels correspondent aux caractères biologiques des organismes (respiration, nutrition, croissance, reproduction...) qui impactent leur valeur sélective (c'est-à- dire la capacité des individus à produire une descendance viable, également appelée *fitness*) *via* ses effets sur leur croissance, leur reproduction ou leur survie. Ils déterminent les interactions de ces organismes avec les conditions abiotiques du milieu et les interactions avec les autres espèces. En ce sens, ils sont une des clés du passage de la réponse fonctionnelle des individus au fonctionnement de l'écosystème (Violle *et al.* 2007). Pour appréhender cette diversité fonctionnelle, il est nécessaire de considérer différents niveaux d'organisation biologique depuis les gènes, les espèces, les communautés, jusqu'à la planète dans son ensemble.

2 - Biodiversité microbienne

Parmi les entités qui concourent au fonctionnement des écosystèmes, les communautés microbiennes sont connues depuis longtemps pour jouer un rôle clé dans le fonctionnement général de la biosphère (Falkowski *et al.* 2008). Elles interviennent en effet dans de nombreux processus biogéochimiques, sont les médiatrices de processus vitaux des écosystèmes comme la production primaire, le cycle des nutriments, la propagation des maladies et la transformation de polluants (Ducklow, 2008, Giller *et al.* 2004). De façon surprenante, leur diversité spécifique et fonctionnelle et les mécanismes régissant leur dispersion et leur histoire évolutive demeurent encore mal compris.

La compréhension de la diversité fonctionnelle d'une communauté dépend de la mesure de traits fonctionnels qui, pour les micro-organismes, sont difficiles à évaluer à l'échelle du phénotype et qui nécessitent souvent leur mise en culture. Or la grande majorité des micro-organismes restent encore de nos jours difficiles à mettre en culture. Par contre, la relative simplicité de la physiologie microbienne et des modalités de la régulation génétique de ces traits (dont l'induction dépend de la taille des populations, de l'activité cellulaire et des conditions de l'environnement) facilite l'association entre gènes et fonctions et permet d'appréhender l'écologie fonctionnelle des communautés microbiennes à travers l'étude de leurs génomes, de leurs transcriptomes ou de leurs protéomes (Esacalas et al 2019). L'essor des approches moléculaires ces dernières décennies (comme la PCR, le séquençage, les empreintes génétiques), le développement des techniques "omiques" et les avancées en matière de puissance de calcul informatique, permettent maintenant d'accéder à une fraction de micro-organismes jusqu'alors inaccessibles par les techniques culturales et d'approfondir ces questions.

3– Contexte de génomique environnementale

Le développement des approches de génomique environnementale a permis de mettre en lumière un ensemble de nouveaux éléments remettant en cause notre vision de la diversité et notre compréhension du monde microbien. En premier lieu, les approches de métagénomique ont révélé une diversité microbienne largement sous-estimée, incluant la découverte de nouveaux phyla (Rinke *et al.* 2013 Castelle *et al.* 2015, Castelle et Banfield 2018), la redéfinition de certains groupes taxonomiques (Parks *et al.* 2018, Keeling et Burki 2019), ou la réévaluation des hypothèses précisant l'origine phylogénétique des eucaryotes (Spang et al 2015, Eme *et al.* 2017). Les études métagénomiques ont également révélé que la plupart des espèces bactériennes ne sont pas clonales (Venter *et al.* 2004, Vergin *et al.* 2007, Rosen *et al.* 2015). Ces éléments remettent dès lors en cause la définition de l'espèce chez les bactéries et les archées. Enfin, il a été mis en évidence une grande diversité de profils génomiques en termes de contenu en gènes et de fonctions portés au sein d'une même « espèce microbienne », associée à un taux de renouvellement important de ce contenu (Coleman *et al.* 2006, Bhaya *et al.* 2007, Biller *et al.* 2014). Ces observations ont donné lieu au développement du concept de pangénome, sous-tendant l'existence d'un pool de gènes communs à l'ensemble des individus d'une espèce et une constellation de gènes accessoires qui peuvent constituer autant de profils fonctionnels au sein même des espèces (Medini *et al.* 2020). Ces «constats» remettent aussi en cause notre vision de la notion de génome au sein d'une *espèce*, de l'organisation de l'information génétique dans ces génomes, ainsi que la nature des processus qui gouvernent leur composition génique et fonctionnelle. Ceci a également un impact important sur la manière dont on doit concevoir les interactions microbiennes dans le cadre des études d'écologie des communautés notamment.

L'écologie des communautés vise à comprendre les interactions entre les différents acteurs (populations / espèces) au sein des communautés, la caractérisation de propriétés émergentes associées à ces assemblages, ainsi que celle de leur impact sur le fonctionnement de l'écosystème. Le flou dans la définition de l'espèce bactérienne ou archéenne, associé à la faible caractérisation taxonomique des communautés (découverte de beaucoup de nouvelles unités taxonomiques sans référence proche dans les phylogénies) rend la résolution de la composition spécifique des communautés microbiennes procaryotiques complexe. Il en résulte un glissement récent des questions d'écologie des communautés depuis l'interrogation du *qui* vers le *quoi*, à savoir, identifier les fonctions qui sont réalisées indépendamment de la question de qui les porte (Koskella *et al.* 2017). Cependant une telle approche laisse en suspend la question du *comment*, c'est à dire l'identification des facteurs biologiques, évolutifs ou environnementaux qui gouvernent la formation et le maintien ou non des assemblages microbiens.

4– Positionnement du travail présenté dans ce rapport

A l'échelle des micro-organismes, les fréquences alléliques peuvent changer au cours d'une génération par le fait de transferts horizontaux de gènes (Koonin et Wolf, 2009) de sorte que ces changements peuvent se produire suffisamment rapidement pour affecter des interactions écologiques (Messer *et al.* 2016, Good *et al.* 2017). Le fait que les processus écologiques (changement de l'abondance des individus dans le temps) dans les communautés microbiennes se superposent avec les processus

évolutifs (changement de la fréquence des gènes dans le temps) chez les micro-organismes (Shapiro 2018) a conduit de nombreux auteurs à argumenter que la génomique des populations microbiennes ne peut être séparée de l'écologie. C'est dans ce contexte que je souhaite placer les travaux que je présente ici.

Ceux-ci relèvent de l'étude de la biodiversité microbienne de l'environnement à l'échelle du gène et du génome, également appelée génomique environnementale. Ces travaux ont dans un premier temps porté sur le développement de méthodes bio-informatiques pour la caractérisation de la biodiversité microbienne des communautés aquatiques naturelles de l'environnement par des approches de métagénomique (Roux *et al.* 2011) et de *metabarcoding* en séquençage haut débit (Taib *et al.* 2013). Ils ont,dans un second temps, porté sur la compréhension des mécanismes évolutifs à même d'expliquer la diversité génétique des populations microbiennes aquatiques, libres issues de l'environnement. Les travaux présentés portent sur des modèles procaryotes et eucaryotes.

Ce manuscrit est organisé en trois parties.

Dans la première partie, je présente les travaux en lien avec l'analyse de la diversité microbienne par des approches de *metabarcoding*. Après une introduction présentant le problème de la définition de l'unité de mesure de la diversité microbienne, je présente la notion d'unité taxonomique opérationnelle (OTU) et les différentes approches développées pour les inférer. Je décris ensuite les contraintes induites par les nouvelles technologies de séquençage (NGS) pour l'estimation de la diversité microbienne et présente l'approche retenue dans l'équipe à travers le développement de la chaîne de traitement PANAM (travaux de thèse de Najwa Taib). J'illustre celle-ci à travers la présentation de quelques travaux en collaboration avec des écologues microbiens. Dans un second temps, je reviens sur le débat actuel entre OTU et ESV pour la caractérisation de la diversité microbienne et présente les arguments en faveur de l'utilisation d'unités phylogénétiques de diversité.

Dans la seconde partie, je présente les travaux relatifs à la caractérisation des unités de diversité microbiennes dans les populations naturelles et l'étude des forces évolutives qui gouvernent la dynamique de leur pangénome. Après une brève comparaison des approches d'analyse de la diversité taxonomique basée sur des gènes marqueurs (*metabarcoding*) par rapport à l'analyse de génomes complets, j'introduis les éléments essentiels à la notion de pangénome. Les travaux concernant l'analyse de la dynamique évolutive du génome accessoire d'une population environnementale~~s~~ de bactéries appartenant au genre *Prochlorococcus*, écotype HLII sont ensuite présentés (travaux de thèse de Hélène Gardon).

Pour terminer, la troisième partie est dédiée à la présentation de travaux en collaboration. J'introduirai le projet MICROSTORE, porté par l'équipe (C. Lepère) auquel je suis associée et qui constitue un développement complémentaire à mes travaux de recherche. Cette partie me permet par ailleurs de présenter brièvement les travaux que j'ai effectués pour l'essentiel avant mon intégration au sein du LMGE .

PARTIE I

APPROCHE *METABARCODING*

ou

LA DESCRIPTION DE LA DIVERSITE SPECIFIQUE

Le développement des approches de métagénomique / génomique environnementale a permis de mettre en lumière un ensemble de nouveaux éléments remettant en cause notre compréhension du monde microbien. En premier lieu, les approches de métagénomique ont révélé une diversité microbienne largement sous-estimée, incluant la découverte de nouveaux phyla (Rinke *et al.* 2013, Castelle *et al.* 2015, Castelle et Banfield 2018). Comme dit plus haut, la diversité peut, en premier lieu, être appréhendée par le dénombrement des espèces qui occupent un espace (écosystème).

1 – Notion d'espèce et définition opérationnelle de l'espèce en microbiologie

La littérature scientifique traitant de la nature des espèces est riche et n'est pas le propos ici (se reporter à Freudenstein *et al.* 2017 et ses références). Il semble intéressant cependant de revenir sur la distinction entre le concept de l'espèce d'une part et l'espèce considérée d'un point de vue opérationnel d'autre part. Le concept d'espèce est une notion principalement ontologique au sens où elle fait référence à « une idée du type d'entité désignée par le terme espèce » et doit servir de référence pour la classification correcte des « espèces ». L'espèce *opérationnelle* doit, quant-à-elle, permettre de restituer de manière pragmatique l'ensemble des propriétés (phénotypiques) propres aux individus qui constituent cette espèce. Il s'agit alors d'une notion épistémologique traduisant la définition de critères opérationnels par lesquels une espèce peut être reconnue dans la nature (Feudenstein *et al.* 2017))

Le concept d'espèce est « multiforme » et dépend, pour beaucoup, du point de vue adopté pour caractériser cette unité de diversité. Dans les travaux de Darwin (1859), les espèces sont vues comme des lignées, c'est-à-dire un continuum d'individus, constituées dans le temps. Ce concept a toutefois longtemps été éclipsé par le concept d'espèce biologique proposé par Mayr selon lequel « Les espèces sont des groupes de populations naturelles qui se reproduisent effectivement ou potentiellement entre elles et qui sont isolées des autres groupes sur le plan de la reproduction » (Mayr 1942). Une population implique donc une unité de lieu et de temps et constitue « un groupe dans lequel des individus adjacents

échangent au moins occasionnellement des gènes entre eux de manière reproductive, et dans lequel des individus adjacents se reproduisent plus fréquemment qu'avec des individus extérieurs à la population » (Freudenstein 2017).

La vision de l'espèce comme lignée évolutive, est revenue en force à partir des années 70, alimentée par la pensée phylogénétique ainsi que de nombreuses études empiriques dans la continuité des travaux de Woese (Woese 1977, Doolitlle 1999). Cette vision de « l'espèce comme lignée » est devenue commune dans la pratique et la systématique actuelle (systématique phylogénétique) est maintenant largement basée sur cette idée (Lecointre et Le Guyader 2017). L'espèce taxonomique est alors décrite comme un groupe monophylétique et se caractérise par une coalescence exclusive d'allèles (De Queiroz 2007).

La définition de l'espèce, tant biologique que phylogénétique, trouve ses limites dans le monde microbien en particulier chez les procaryotes du fait de l'absence de reproduction sexuée (*stricto sensu*) et de l'existence d'échanges de matériel génétique (transferts horizontaux de gènes), y compris entre individus distants, c'est à dire n'appartenant pas à la même lignée phylogénétique (Médigue *et al.* 1991, Doolittle 1999, Ochman *et al.* 2000). Cependant, les bactéries forment clairement des groupes génétiquement et phénotypiquement distincts. Des modèles explicatifs, autres que celui de l'espèce biologique ont été proposés pour décrire la structuration de la diversité microbienne comme par exemple le modèle d'écotype (Cohan 2001 – Figure 1.2). Dans ce modèle, les groupes génotypiques correspondent à des niches écologiques et des événements périodiques de sélection purgent la variation génétique dans chaque niche séparément, induisant une cohérence génétique au sein des écotypes et une différenciation génotypique entre écotypes (Kumar *et al.* 2015).

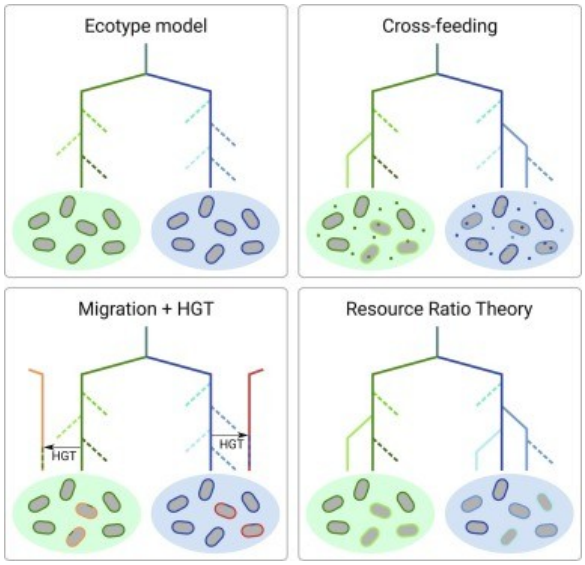
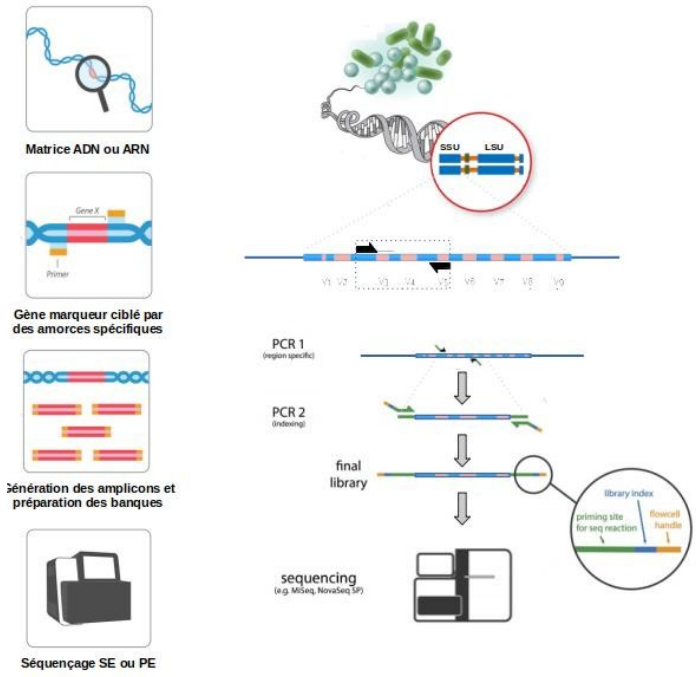


Figure 1.2 : Quatre modèles de diversité microbienne. Les ovales colorés représentent différentes niches. Le génotype des bactéries est représenté par la couleur des parois, les lignées phylogénétiques sont représentées au dessus. Le modèle de l'écotype propose que chaque niche soit occupée par des génotypes uniques. Le modèle de l'alimentation croisée propose que les génotypes peuvent survivre grâce aux métabolites sécrétés par les autres. Le modèle de migration et transfert de gènes suppose que les génotypes migrent et acquièrent du matériel génétique leur permettant de survivre dans la communauté dans une nouvelle niche. Enfin, le modèle du rapport de ressources suggère que différents génotypes s'approprient les ressources de la niche qui ne limitent pas la croissance des génotypes coexistants (d'après Mitri 2019).

BOX 1 - Gène marqueur, amplicon et lectures



Dans un étude de diversité microbienne, l'ADN d'un ensemble d'individus est extrait à partir d'un échantillon issu de l' environnement. Un gène marqueur de référence (en général l'ARNr 16S ou 18S) est spécifiquement ciblé par des amorces et amplifié par PCR. Le produit de cette amplification est enrichi de la portion d'ADN spécifique du gène ciblé, appelé amplicon. Les séquences produites lors du séquençage de ces amplicons sont appelées lectures (*reads*). Celles-ci peuvent couvrir une extrémité de l'amplicon (séquençage *single-end*) ou les deux (séquençage *paired-end*).

D'un point de vue opérationnel, les espèces procaryotes sont définies à partir de caractères phénotypiques et génotypiques (similarité de séquences sur un gène marqueur, similarité nucléotidique moyenne ou nombre de gènes partagés à l'échelle des génomes (Richter et Rossello-Mora 2009, Achtman et Wagner 2008). Lorsque les traits phénotypiques ne peuvent pas être décrits, une désignation provisoire de l'espèce candidate peut être proposée sur la base d'une discrimination exclusivement génétique. Celle-ci sera alors décrite sous la dénomination de *Candidatus sp*. Cette situation est souvent retrouvée dans les études de microbiologie de l'environnement dont une grande partie de la diversité ne peut pas être maintenue en culture, par manque de connaissances des condition de croissance de ces organismes. Ces limites empêchent de fait l'étude et la caractérisation de leur physiologie. Ainsi, la diversité taxonomique des communautés naturelles est dans les faits évaluée de manière pragmatique, essentiellement à travers l'analyse de gènes marqueurs conservés tels que la petite sous-unité du ribosome (ARNr 16S et 18S) et les définitions d'unités taxonomiques opérationnelles ou OTU (*Operational Taxonomic Unit*). Plus récemment, des approches profitant des avancées technologiques liées au séquençage et reposant sur la caractérisation du pourcentage d'identité nucléotidique moyen à l'échelle des génomes (ANI) ont été proposées. Dans ce cas, un seuil de similarité est fixé pour assigner deux génomes à une même espèce (Goris *et al.* 2007, Konstantidinis et Tiedje 2005, Richter et Rossello-Mora 2009). Bien que les approches présentées ici ne soient pas exhaustives, on peut noter que toutes utilisent un seuil universel d'identité entre les séquences comparées pour définir une unité opérationnelle de diversité / espèce.

1.1 – Les OTUs, mesure opérationnelle de la richesse spécifique

1.1.1 – OTUs basés sur une clusterisation selon un seuil de similarité

Les OTUs sont définis à partir de l'analyse de la séquence de la petite sous-unité du ribosome (SSU, ou ARNr 16S chez les procaryotes et 18S chez les eucaryotes). Celle-ci est ciblée et amplifiée expérimentalement par PCR pour produire des amplicons (BOX 1). Les séquences des amplicons présentant au moins 97% d'identité sont ensuite agrégées entre elles pour constituer les OTUs. Ce seuil de 97% a été défini par Stackebrandt et Goebel (1994) comme équivalent au seuil d'hybridation ADN-ADN de 70% observé dans des expériences de ré-association réalisées entre les membres d'espèces bactériennes préétablies, issus d'organismes mis en culture. Les OTUs sont le plus souvent traités comme l'observation d'une espèce (à noter que pour les approches par ANI, il est admis qu'un seuil de clusterisation à 95% d'identité produit des résultats semblables à ceux obtenus au seuil d'hybridation ADN-ADN de 70% cité ci-avant).

De nombreuses méthodes de clusterisation ont été proposées pour définir des OTUs moléculaires. La description la plus concurremment faite de ces approches distingue la clusterisation d'OTUs à partir du jeu de séquences uniquement (*de novo*) et les approches par clusterisation sur des séquences références (*closed-reference* ou *open-reference*). Il existe par ailleurs des distinctions algorithmiques, qui ont un impact non moins r igeable sur la qualité des OTUs générés. Ainsi on peut distinguer les algorithmes de classification hiérarchique, les méthodes gloutonnes et les méthodes non basées sur la définition d'un seuil.

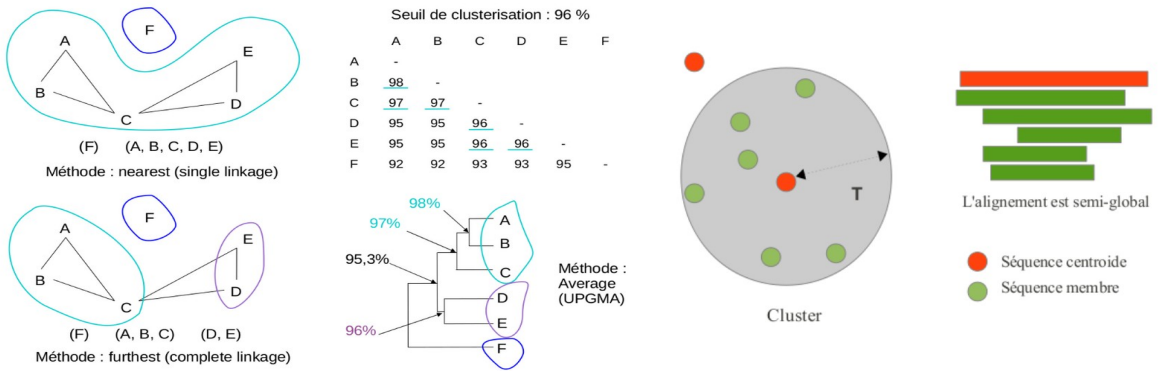


Figure 1.3 : Représentation schématique de quelques algorithmes de clusterisation (A) trois stratégies de la clusterisation hiérarchique au seuil de 96% d'identité. Les lettres représentent des séquences, les lignes entre les lettres précisent un lien entre les séquences au seuil défini. La clusterisation par liaison simple produit un groupe ABCDE et un singleton F. L'approche par liaison complète, génère deux groupes, ABC et CDE et un singleton F. Ces deux groupes sont ici distincts car les distances entre (A,B) et (D,E) sont supérieures au seuil défini. (B) clusterisation gourmande implémentée dans UCLUST. Le premier cluster testé dont le centroïde partage avec la séquence requête un pourcentage d'identité inférieur ou égale au seuil fixé, intégrera ce cluster (d'après http://www.drive5.com)

Les méthodes de classification hiérarchique de type liaison complète, liaison simple ou liaison moyenne (*complete linkage, single linkage, average linkage*), sont utilisées dès les premières études de diversité microbienne par séquençage Sanger. La liaison complète, qui associe au sein d'un même OTU toutes les séquences partageant entre elles un similarité supérieure au seuil fixé est souvent préférée, car elle

seule garantit que toutes séquences groupées dans un même OTU sont distantes au maximum du seuil défini (Figure 1.3). Contrairement aux approches de classification hiérarchique par liaison complète, les classifications par liaison simple sont basées sur la transitivité des relations de similarité. Ainsi, si une séquence A est similaire à une séquence B au seuil défini et que B est similaire à C, alors on considère que A est similaire à C. Ces méthodes sont implémentées dans des outils tels que Dothur (Schloss et Handelsman 2005) et Mothur (Schloss *et al.* 2009). Du fait de l'accroissement de la taille des jeux de données liés aux technologies NGS, ces approches ont été progressivement abandonnées, en particulier la liaison complète, car trop coûteuses en ressources informatiques (temps et mémoire).

Les algorithmes de clustering gloutons (CD-HIT (Li et Godzik 2006), UCLUST, USEARCH (Edgar 2010), VSEARCH (Rogne *et al.* 2016) et approches dérivées) sont des heuristiques des classifications hiérarchiques ; elles fournissent une solution rapide, raisonnable mais pas nécessairement optimale. Le principe général de ces algorithmes consiste à traiter progressivement un jeu de séquences en commençant par une. L'idée est qu'à chaque ajout d'une nouvelle séquence, on choisit la solution qui semble la meilleure pour le problème posé en l'état des données considérées (on parle d'optimum local), en espérant que ces choix amèneront à une solution globale optimale. Pour réduire le temps de calcul associé à la construction des OTUs, les méthodes gloutonnes définissent pour chaque cluster (OTU) une séquence de référence, appelée *seed* ou *centroïde* contre laquelle les séquences non encore classées seront comparées. Ainsi, toutes les séquences d'un OTU définies à un seuil de similarité donné ne seront pas comparées entre elles et il est possible que certains couples de séquences au sein de l'OTU ne satisfassent pas le seuil de similarité requis. Par ailleurs, selon les algorithmes considérés, les séquences peuvent être traitées par ordre de taille décroissante, par ordre d'abondance décroissante ou de manière non ordonnée. Un inconvénient majeur des approches gloutonnes est que le résultat dépend largement de l'ordre dans lequel les séquences sont analysées. Ainsi, il n'est pas garanti que différents ordonnancements des séquences produisent les mêmes OTUs, ni les mêmes séquences représentatives d'un OTU.

1.1.2 - OTUs basés sur une clusterisation indépendante d'un seuil de similarité

Il est peu à peu apparu que l'équivalence entre la définition d'un OTU au seuil de 97 % d'identité de la séquence de l'ARNr et l'espèce microbienne telle que proposée par Stackebrandt and Goebel (1994) n'était pas universelle à l'échelle des microorganismes (Stachenbrandt et Ebers 2006). Au seuil de 97 % d'identité, près de 46 % des souches bactériennes types (c'est à dire d'espèces de référence maintenues en laboratoire) analysées sur la base de leur séquence d'ARNr 16S quasi complète sont assemblées dans des OTUs « mixtes » (Mysara *et al.* 2017). De même, les OTUs générés selon ce même seuil peuvent parfois agréger des séquences appartenant à des lignées phylogénétiques distinctes, et conduire à une hétérogénéité écologique des OTUs inférés (Koeppel and Wu 2013). Ceci peut s'expliquer notamment par des fluctuations de la vitesse d'évolution de lARNr 16S (ou 18S) entre lignées ou espèces, qui induit une divergence plus ou moins marquée entre les séquences (Caron *et al.* 2009, Brown *et al* 2015).

Si différentes lignées évoluent à différentes vitesses, il n'existe pas un seuil unique à l'ensemble de l'arbre de la vie et il conviendrait alors de s'affranchir de l'utilisation d'un seuil de clusterisation universel

à l'ensemble des taxa microbiens. Ainsi il a été proposé d'appliquer des seuils de clusterisation différentiels, calculés dynamiquement lors de l'analyse des amplicons (Mysara *et al* 2017). Ceux-ci sont d'abord assignés taxonomiquement au niveau de la famille, puis clusterisés sur la base d'un seuil pré-défini spécifique à chaque famille taxonomique.

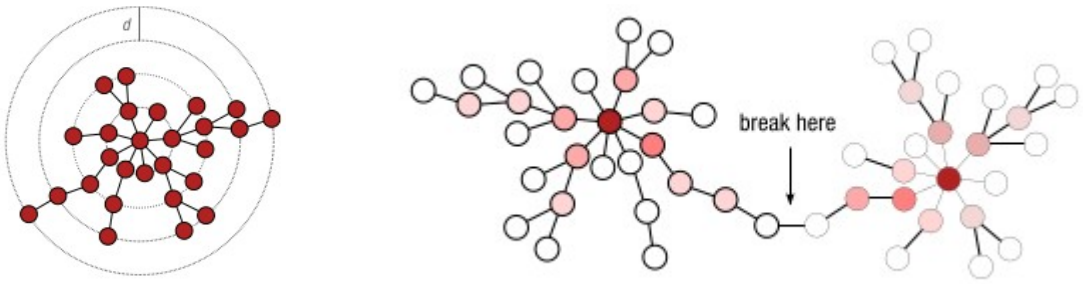


Figure 1.4 : Approche de clusterisation par SWARM. SWARM regroupe les amplicons itérativement en utilisant une distance local de différentiation d, permettant aux OTUs de croître jusqu'à leurs limites naturelles (*i.e.*, aucun autre amplicon ne peut être ajouté). Afin d'éviter la sur-agrégation des amplicons, l'abondance de chaque amplicon est considérée (dégradé de rouge, plus la couleur est intense, plus l'amplicon est abondant). Les liaisons rares caractérisées par des abondances faibles sont alors rompues. Une option *fastidious* permet de postuler l'existence d'amplicons virtuels de liaison pour greffer de petits OTUs sur les plus grands (d'après Mahé *et al.* 2015).

D'autres approches n'utilisent pas de seuil de clusterisation. SWARM par exemple (Mahé *et al* 2014, 2015) analyse des graphes en « plaçant les amplicons dans un espace multidimensionnel discrétisé » et considère que chaque variant d'une séquence peut être représenté dans un espace voisin distant d'une valeur équivalente à la distance d'édition qui sépare les deux séquences (Figure 1.4). Dans cet espace discrétisé, une séquence différant d'un nucléotide par rapport à une séquence de référence lui sera voisine à une distance d=1 ; l'espace occupé par cette séquence sera dit "plein". Les clusters générés par l'outil SWARM (appelés essaims et non des OTUs) sont donc des continuum d'espaces non vides séparés d'une distante *d* avec une séquence voisine depuis la séquence graine de l'amplicon considéré.

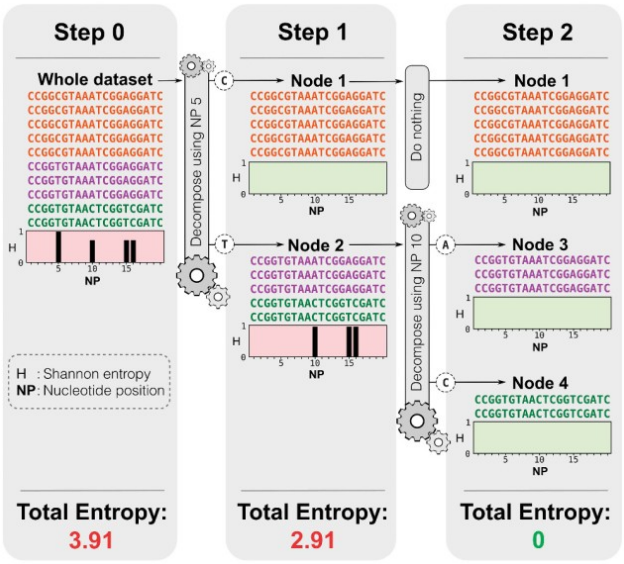


Figure 1.5 : Décomposition itérative d'un jeu de séquences fictives selon le critère d'entropie minimale (entropie de Shannon). A chaque étape, les lectures sont groupées dans des nœuds caractérisés par un profil d'entropie (indiqués par des barres noires à chaque position) et une valeur d'entropie totale pour le nœud. A chaque itération, les nœuds avec une valeur d'entropie supérieure à 0 sont décomposés selon la distinction établie par la position du nucléotide d'entropie maximale. L'algorithme se termine lorsque l'entropie des nœuds est inférieure à un seuil minimal m' calculé dynamiquement en fonction des données (d'après Eren *et al.* 2015).

La méthode MED (*Mimimal Entropy Decomposition*) est, quant-à-elle ,basée sur la décomposition d'entropie minimale pour partitionner les données environnementales en unités phylogénétiques homogènes (Eren *et al.* 2015). En effet calculer l'entropie de Shannon pour caractériser des signatures nucléotidiques, permet de distinguer des signaux biologiques pertinents du bruit de fond sans avoir à calculer des similitudes entre paires de séquences (Eren *et al.* 2015). A partir de l'ensemble des données considérées (de même taille), MED calcule l'entropie à chaque position nucléotidique le long des séquences cibles, puis partitionne le jeu de séquences en les distribuant dans différents sous-ensembles définis à partir des positions d'entropie maximale (Figure 1.5). L'algorithme s'arrête quand plus aucun pic d'entropie n'est discernable pour un jeu de données.

1.2– Limites de la notion d'OTU dans un contexte de séquençage de seconde génération

1.2.1 – Un saut technologique

L'émergence des technologies de séquençage à haut débit au milieu des années 2000 (NGS) a permis aux biologistes d'étudier la diversité microbienne à partir de millions de séquences (au lieu de quelques dizaines). Ce saut technologique rend inopérantes les approches d'analyse « traditionnelles » du fait de la quantité des données produites, mais aussi en raison de leur qualité. Cette quantité de données, jamais atteinte dans le domaine de la microbiologie environnementale, a nécessité un changement d'échelle dans les ressources informatiques mobilisées, tant pour son stockage que pour son traitement. A titre d'exemple, la technique Illumina HiSeq2500 peut produire jusqu'à 600 Gigaoctets (Go) de données ce qui correspond à $6x10^9$ séquences (Scholz *et al.* 2012) et nécessite environ 0.6 Téraoctets (To) d'espace disque. De nos jours, les séquenceurs Illunima NovaSeq produisent jusqu'à 6 Tb de données par *run*.

Du point de vue de leur qualité, les séquences produites par les technologies NGS se caractérisent par des biais dans la production des séquences (Figure 1.6) et des taux d'erreur caractéristiques (tableau 1.1).

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile
454 GS FLX Titanium XLR70	Up to 600; 450 mode (SE, PE)*	450 Mb*	~1 M*	10 h*	1%, indel‡
Illumina MiSeq v3	75 (PE)	3.3–3.8 Gb	44–50 M (PE)*	21–56 h*	0.1%, substitution‡
	300 (PE)*	13.2–15 Gb*			
Illumina HiSeq2500 v4	100 (PE)	360–400 Gb		5 d	0.1%, substitution‡
	125 (PE)*	450–500 Gb*		6 d*	

Tableau 1.1 : Caractéristiques de quelques plateformes de séquençage NGS. SE : *single end*, PE : *paired-end*. * Données constructeur. (adapté de Goodwin *et al.* 2016).

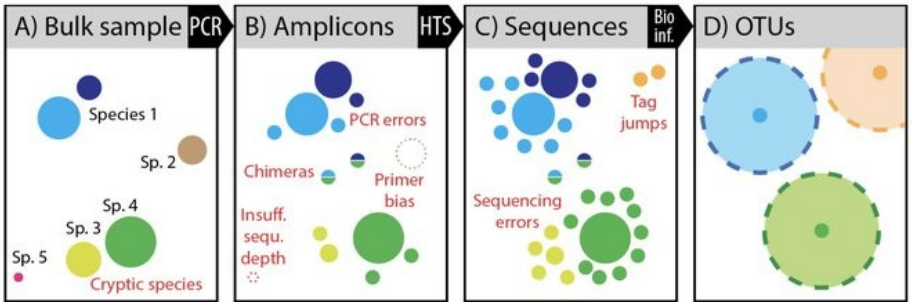


Figure 1.6 : Aperçu des biais qui peuvent apparaître aux différentes étapes du *metabarcoding* (A) Différentes espèces (différentes couleurs) avec différentes abondances (taille des cercles) sont présentes dans l'échantillon. (B) Après extraction de l'ADN et amplification de marqueurs ciblés (ARNr par exemple), certaines espèces peuvent être perdues du fait des biais dans les *primers* qui échouent à amplifier certains taxons ou d'une quantité de matériel insuffisante pour être « captées » par les *primers* dans le cas d'espèces rares. Des erreurs dans les séquences initiales peuvent également être introduites lors de l'étape d'amplification (substitutions, insertions, délétions, chimérisme) conduisant à la génération de séquences dérivées de séquences originales (petits cercles en périphérie des cercles initiaux, de même couleur que ceux-ci). (C) Des erreurs de séquences peuvent également être introduites lors de l'étape de séquençage, incluant également de possibles « sauts de tag » dans le cas d'échantillons multiplexés. (D) Regroupement des lectures issues d'amplicons pour former des OTUs et définition d'une séquence de référence appelée *seed* ou centroïde (d'après Elbrecht *et al.* 2018).

Ces biais et erreurs peuvent se produire à l'étape d'amplification PCR des séquences initiales (c'est-à dire-en amont du séquençage) et induire la perte ou la sur-représentation de certains types de séquences présentes dans l'échantillon, la production de séquences artéfactuelles porteuses de substitutions, d'insertions ou de délétions nucléotidiques, ou correspondant à la fusion de deux séquences différentes (chimères). Des erreurs peuvent également se produire lors du séquençage à proprement parler. Il est admis que les erreurs de « lecture » des séquences-type lors de l'amplification se produisent aléatoirement, sont indépendantes entre elles et entre séquences. Comme les séquences « vraies » présentes dans l'échantillon original, les séquences porteuses d'erreurs vont être amplifiées de manière exponentielle dès lors qu'elles existent. Ainsi, bien que rares à leur apparition (une erreur apparaîtra sur un brin d'ADN), elles peuvent constituer un nombre significatif de séquences à l'issue du séquençage. Si on admet que les séquences « vraies » sont en plusieurs exemplaires dans l'échantillon original, l'abondance des séquences porteuses d'erreurs devrait toutefois être plus faible que celles des séquences « vraies ». Contrairement aux erreurs d'amplification, qu'aucune information technique ne permet de détecter, les erreurs de séquençage peuvent dans une certaine mesure être tracées par l'analyse d'informations qualitatives associées à la production des séquences (Figure 1.7). Il a pu être établi que ces erreurs ne sont pas indépendantes et pour certaines technologies, elles ne sont pas aléatoires (Shirmer *et al.* 2015, Callahan *et al.* 2019). Dans ce contexte, l'usage des OTUs répond à d'autres objectifs que la seule définition d'une unité de base de la diversité microbienne. Ils permettent en effet de réduire la dimension du jeu de données issu du séquençage massif, de supprimer des variants artéfactuels ou d'agréger des variants de séquences légitimes (vraies) qui peuvent réellement exister au sein d'un génome bactérien (c'est le cas des espèces avec plusieurs opérons ARNr dans leur génome (Acinas *et al.* 2004), et eucaryotes (Stoddard *et al.* 2015, Nearing *et al.* 2018, Callahan *et al.* 2019) ou entre taxa proches. Ainsi, bien que répondant à deux objectifs difficilement conciliables (conceptuels et méthodologiques), les OTUs sont devenus l'unité de mesure de la richesse spécifique en écologie microbienne, la notion d'espèce n'étant cependant jamais totalement absente dans l'esprit des microbiologistes.

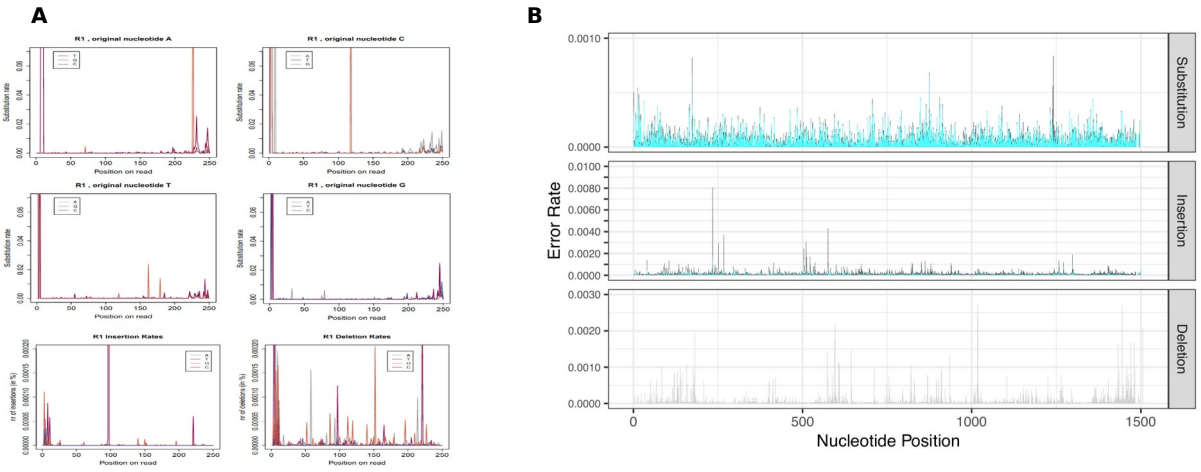


Figure 1.7 : Distribution des erreurs de séquençage le long d'un amplicon en fonction dde la position du nucléotide et du score de qualité. Sont présentés les taux de substitutions, d'insertions et de délétions (A) Miseq ; (B) PacBio CCS sur les lectures non chimériques et non contaminantes. Les bases de faible qualité qualité sont représentées par des couleurs plus foncées (d'après Shirmer *et al.* 2015 et Callahan *et al.* 2019).

1.2.2 – Une réduction de l'information exploitable

Le gène de l'ARNr se caractérise par la présence de régions hypervariables (v1 à v9) (Goebel and Stackebrandt, 1994) présentant chacune des taux de variabilité différents (Figure 1.8). Cette caractéristique a participé à sa popularité car ces régions donnent accès à des niveaux de résolution taxonomique différents (Liu *et al.* 2007, 2008, Wang *et al.* 2007). Cependant, cela introduit une difficulté majeure dans un contexte de NGS. En effet, là où le séquençage Sanger permettait de générer des séquences de près de 1 kb, les NGS produisent des séquences courtes (entre 50 et 450 pb), associées à des taux d'erreur relativement élevés. Ces contraintes techniques ont eu pour conséquence de contraindre les biologistes à cibler des zones réduites de l'ARNr 16S ou 18S, induisant une réduction de l'information exploitable.

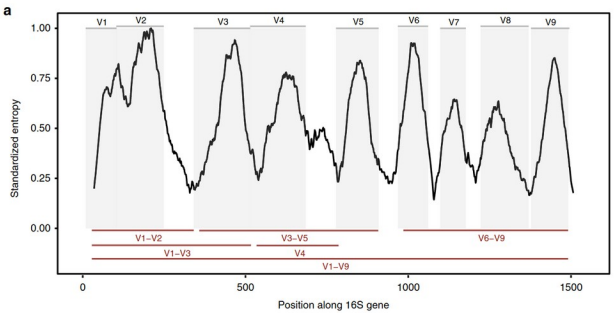


Figure 1.8 : Représentation de la variabilité nucléotidique le long de la séquence du l'ARN 16S de *Escherichia coli K-12 MG1655* (NCBI Gene ID 947777), La variabilité est exprimée par l'entropie de Shannon (d'après Johnson *et al.* 2019).

La région choisie peut alors avoir un impact important sur la diversité microbienne observée (Schmalenberger *et al.* 2001, Taib *et al.* 2013, Mysara *et al.* 2017) et il devient nécessaire d'adapter les seuils de clusterisation des OTUs à la région considérée. Ainsi, pour l'ARNr 16S, le seuil de clusterisation varie de 96% à 98% selon la région étudiée (Kim *et al.* 2011) ; ces auteurs préconisent par ailleurs l'utilisation des régions v1-v3 et v1-v4 pour une meilleure affiliation des bactéries, et de la région v1-v3 pour les archées. Pour les eucaryotes, Caron *et al.* (2009) préconise un seuil de 95% pour clusteriser les séquences de protistes au niveau de l'espèce. Chez ces derniers, les régions v4 et v9 sont les plus

utilisées pour des études de diversité (Amaral-Zettler *et al.* 2009, Behnke *et al.* 2011, Dunthorn *et al.* 2012, Debroas *et al.* 2017). Dans une étude plus récente, Mysara et collaborateurs montrent cependant chez les bactéries qu'au niveau taxonomique de la famille les profils de conservations de l'ARNr 16S varient en fonction des régions variables ciblées et que ces variations ne sont pas cohérentes entre les familles. Par exemple, la famille des *Methanomicrobiaceae* présente une conservation de séquence de la région V9 plus forte que celle de la famille des *Methanosarcinaceae*, alors même que ces deux familles appartiennent à la même classe d'archées (Mysara *et al.* 2017).

2- Traitement des données de *metabarconding* issues de NGS : PANAM

Comme nous venons de le voir, le développement des technologies de séquençage de seconde génération (NGS) ont profondément modifié la manière dont la biodiversité microbienne pouvait être appréhendée. Les études de diversité ont dès lors nécessité la mise en place de nouvelles stratégies d'analyse et la définition d'outils adaptés pour prendre en compte le changement d'échelle lié à la masse de données produites, mais également les contraintes techniques associées à ces nouvelles technologies.

Dans les chaînes de traitement pour l'analyse de la diversité microbienne adaptée aux amplicons de nouvelle génération (tel que Mothur (Schloss *et al.* 2009), QIIME (Caporaso *et al.* 2010, Bolyen *et al.* 2019), PyroTagger (Kunin and Hugenholtz, 2010), FROGs (Escudié *et al.* 2018)), le regroupement des lectures issues du séquençage en OTUs permet d'énumérer les unités cohésives de diversité génétique présentes dans l'échantillon. Le nombre et la taille des OTUs sont alors utilisés pour qualifier la diversité de l'échantillon indépendamment de toute référence extérieure (approche *de novo*). Il est également possible d'analyser ces OTUs, dans un contexte taxonomique en comparant les séquences représentatives de ceux-ci à des séquences de référence (approche *closed-reference)*. Ceci permet de relier les OTUs à ce qui est connu dans les bases de données de référence et de les associer à des groupes taxonomiques dont les métabolismes et la fonction dans les écosystèmes sont peut être connus et décrits. Ces outils implémentent par ailleurs des approches probabilistes (classifieur bayessien RDP) ou de recherche de similitude (BLAST) pour la caractérisation taxonomique des séquences environnementales.
Les approches phylogénétiques sont une alternative pour l'affiliation taxonomique (Liu *et al.* 2008). Elles présentent l'avantage de prendre en compte les relations de parenté entre différents organismes et d'analyser la séquence dans un contexte évolutif. Ces approches permettent ainsi d'attribuer une taxonomie aux séquences qui ne possèdent pas de séquences de référence proches dans les bases de données, mais aussi de décrire de nouveaux clades (Figure 1.11). Largement utilisées dans les analyses de diversité microbienne antérieures à l'avènement des NGS, ces approches ont été délaissées dans le contexte du séquençage de seconde génération. Outre la complexité algorithmique des méthodes phylogénétiques, la taille des séquences générées par les techniques de séquençage semblait constituer une limitation à leur utilisation. Ainsi, les topologies des phylogénies construites à partir de fragments courts (~200 pb) ciblant différentes régions hypervariables du gène de l'ARNr 16S sont significativement différentes de celles obtenues à partir des séquences complètes du gène, indiquant une perte du signal phylogénétique. Les topologies des phylogénies varient par ailleurs en fonction des régions

hypervariables considérées du fait des variations de vitesses d'évolution le long de ce gène (Liu *et al.* 2007, Liu *et al.* 2008 ; Huse *et al.* 2008 ; Youssef *et al.* 2009). Jeraldo et collaborateurs (Jeraldo *et al.* 2011) soulignent cependant qu'à partir de fragments de 400 pb on peut obtenir des phylogénies semblables aux phylogénies des séquences complètes. C'est dans ce contexte qu'a partir des années 2008 nous avons développé PANAM, une chaîne de traitement des amplicons de seconde génération (Phylogenetic Analysis of Next Generation Amplicons – Taib *et al.* 2013) proposant une approche phylogénétique pour l'affiliation taxonomique des séquences issues de NGS.

2.1 – PANAM pour une affiliation phylogénétique des séquences d'amplicons

L'originalité de PANAM est de proposer une chaîne de traitement permettant l'affiliation taxonomique des séquences par une approche phylogénétique. Comme d'autres chaînes de traitement populaires (QIIME MOTHUR), PANAM est organisé en modules dédiés i) au pré-traitement des séquences d'amplicons et à la génération des OTUs, ii) à l'affiliation taxonomique des OTUs (PANAM à proprement parler) et iii) au calcul d'indices de diversité (Figure 1.9). Cet outil permet de traiter des jeux de données NGS sur un ordinateur personnel, les différentes étapes étant réalisées de manière séquentielle. Le temps de traitement varie de 20 minutes pour 1 000 OTUs à un peu moins de 7 jours pour 1 million d'OTUs.

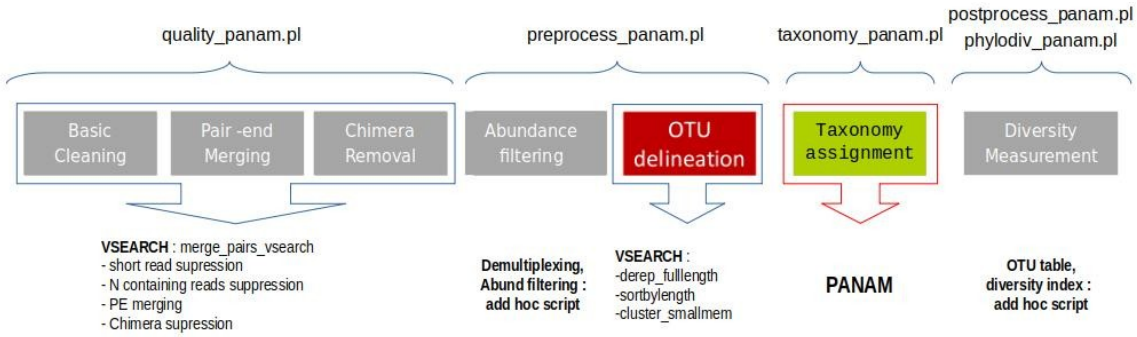


Figure 1.9 : Chaîne de traitement des amplicons par PANAM2. Après une évaluation de la qualité et un nettoyage des séquences (suppression des séquences courtes ou contenant des nucléotides non identifiés (N)), les lectures par paires sont assemblées et démultiplexées (dans le cas de séquençage simultané de plusieurs échantillons), c'est à dire que le jeu de données est décomposé en fonction des échantillons. Les séquences chimériques sont détectées et supprimées, les jeux de données par échantillon sont normalisés et les séquences faible abondance sont supprimées. Les amplicons sont ensuite regroupés en OTUs selon l'algorithme glouton cluster_smallmem de la suite logiciel VSEARCH avant d'être affiliés taxonomiquement sur la base d'une approche phylogénétique. Une table d'OTUs décrivant la taxonomie des OTUs et leur abondance dans les échantillons analysés est produite et divers indices de diversité sont calculés.

PANAM comprend une base de données de séquences de référence, un fichier de taxonomie et des profils d'alignements de référence. Cette chaîne de traitement écrite en perl est décrite dans sa version initiale dans l'article de Taib *et al.* 2013. Dans sa version la plus récente, PANAM utilise les outils VSEARCH (Rogne *et al.* 2013 pour la recherche de similarité entre séquences et la formation des OTUs, HMMER (Eddy 1998) pour l'alignement des séquences représentatives des OTUs contre les séquences de référence et FASTTREE (Price *et al.* 2010) pour la phylogénie.

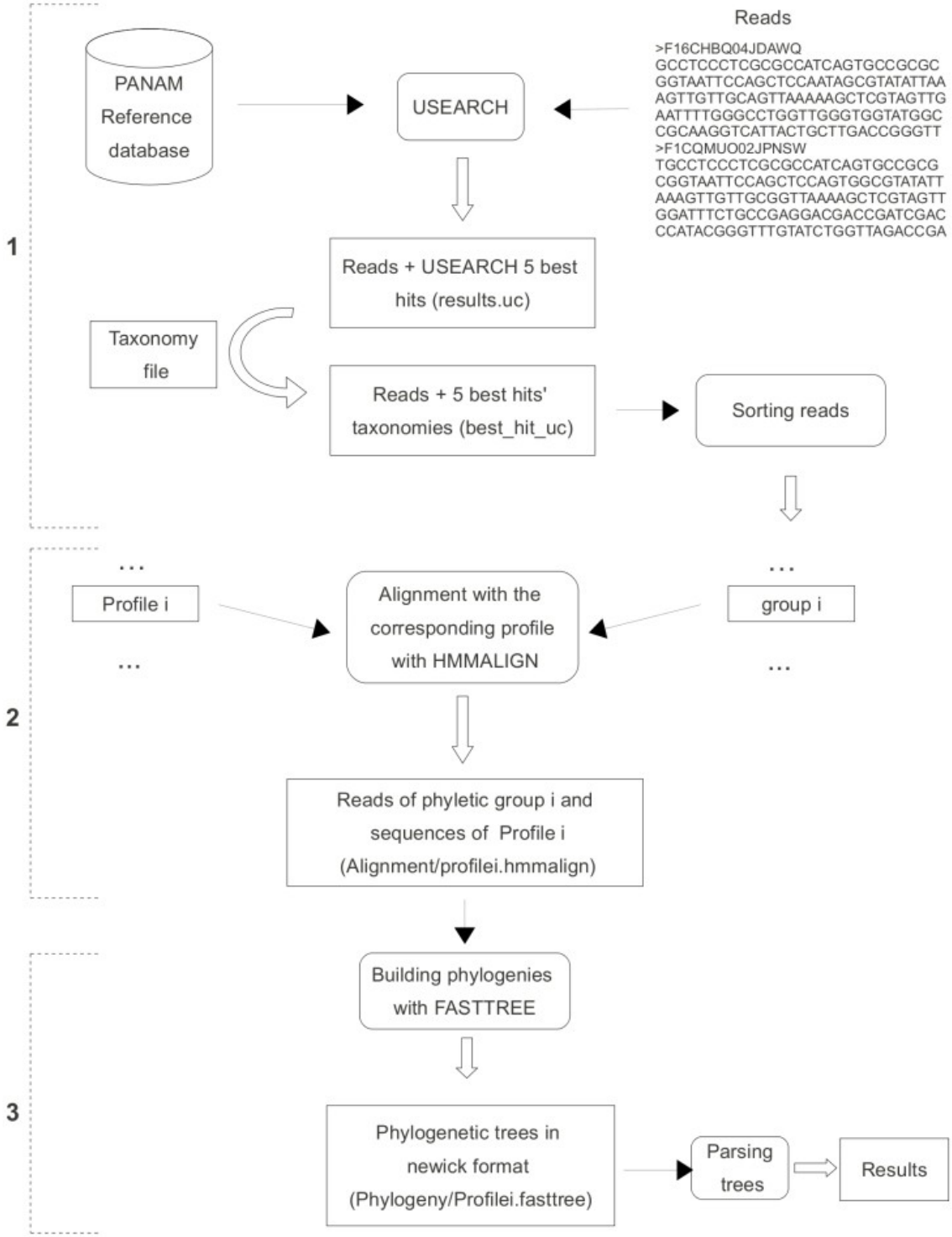
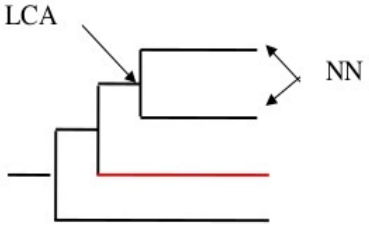


Figure 1.10 : Affiliation phylogénétique par PANAM. Une classification primaire trie et répartit les lectures en groupes selon la taxonomie de leur meilleur résultat USEARCH/VSEARCH (1). Un fichier contenant les lectures alignées avec les les séquences de référence du groupe correspondant est généré par une approche de type alignement de profil par HMMER (2). Ce fichier est utilisé par FASTTREE pour construire un arbre phylogénétique, qui est ensuite analysé pour attribuer une taxonomie à chaque lecture et pour signaler les clades putatifs (3) (d'après Taib *et al.* 2013).

Les lectures issues des NGS analysées par PANAM sont comparées à une base de données dédiée de séquences de référence extraites de la base de données SSURef du projet de base de données SILVA (Pruesse *et al.* 2007, Yilmaz *et al.* 2014). Celle-ci fournit des informations taxonomiques, une évaluation de la qualité et un alignement expertisé des séquences d'ARNr 16S et 18S. La base de données PANAM inclut les séquences de plus de 1200 pb, avec un score de qualité > 75%, et une valeur pintail > 50 tel que défini dans SILVA. La base de données complète contient 302 832 séquences soit 11 028 séquences d'archées, 258 573 de bactéries et 33 231 d'eucaryotes. Pour accélérer le traitement phylogénétique, les alignements des séquences de références sont divisés en 37 groupes phylétiques (dont 24 groupes pour les eucaryotes, 10 pour les bactéries et 3 pour les archées). Les groupes de plus de 1000 séquences sont regroupés au seuil de 97% d'identité.
Pour chaque groupe phylétique, un groupe externe contenant une séquence de chacun des autres groupes phylétiques plus deux séquences de métazoaires est ajouté à l'alignement pour permettre l'enracinement de l'arbre phylétique produit. Il permet également de préciser la parenté des séquences divergeant à proximité de la racine du groupe phylétique auquel elles sont affiliées. Les alignements des séquences de référence de chaque groupe phylétique sont résumées dans un profil HMM. Un fichier contenant la taxonomie EMBL et SILVA de chaque séquence de la base de données de référence est également généré

Pour réaliser l'affiliation phylogénétique dans PANAM (Figure 1.10), les séquences représentatives des OTUs préalablement calculés sont d'abord comparées à la base de données de référence avec VSEARCH. Chaque amplicon est ensuite ajouté au profil phylétique correspondant à la taxonomie de la séquence de référence de meilleur score, ainsi que les séquences de référence des cinq meilleurs résultats de la comparaison. Une fois assignées à leur groupe phylétique, les séquences représentatives des OTUs sont alignées sur le profil de l'alignement de référence correspondant de manière à produire des alignements incluant les séquences de référence et les séquences représentatives des OTUs à affilier.
Pour chaque groupe phylétique de référence, un arbre phylogénétique est calculé en utilisant FASTTREE (modèle Jukes-Cantor + Cat et 100 bootstrap). Si le nombre de séquences représentatives des OTUs appartenant à un groupe phylétique dépasse 30.000, ce groupe est divisé en fichiers contenant moins de 30 000 séquences. Les arbres sont ensuite analysés pour générer des fichiers contenant la taxonomie des séquences insérées et signaler les clades environnementaux qui pourraient être mis en évidence à partir de séquences représentatives des OTUs (Figure 1.11). Une évaluation de la taxonomie est proposée selon le principe de l'ancêtre commun le plus proche ou du plus proche voisin (BOX 2).

BOX 2 – NN et LCA



- Le plus proche voisin (NN) : la séquence d'intérêt (en rouge) va acquérir la taxonomie complète de la première séquence référencée à partir du nœud fils qu'elle partage avec son ancêtre commun.

- L'ancêtre commun le plus récent (LCA) : chaque nœud référence la taxonomie commune à tous ses descendants. La séquence d'intérêt hérite de la taxonomie du premier nœud fils qu'elle partage avec son ancêtre commun. Dans ce cas, la taxonomie sera plus robuste mais pourra être incomplète.

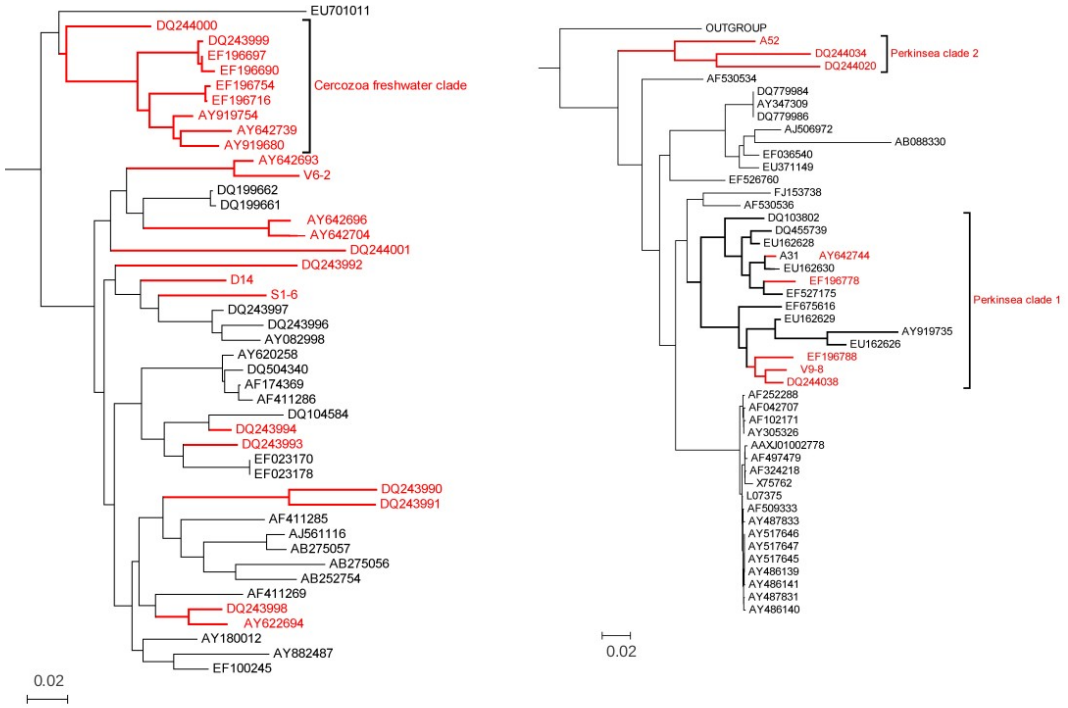


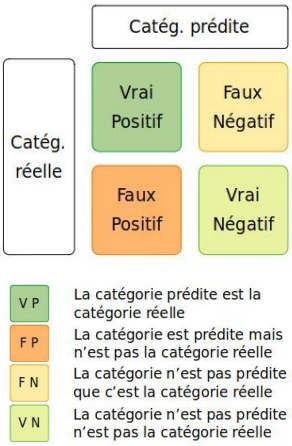
Figure 1.11 : Phylogénies des *Cercozoa* et *Perkinsea* générées par PANAM après l'insertion de séquences environnementales. Les séquences insérées (issues de la technologie NGS Roche 454) sont représentées en rouge. Les clades *freshwate* tels que décrits à partir de séquences d'ARNr 18S quasi complètes obtenues à partir de séquençage Sanger sont correctement restitués (d'après Taib *et al.* 2013).

2.2 - Performance des affiliations phylogénétiques sur amplicons NGS

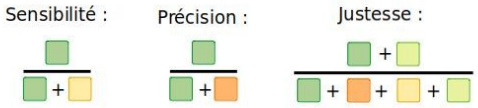
Dans le cadre de la thèse de Najwa Taib, nous avons montré que les affiliations par le biais d'une analyse phylogénétique sont précises et surpassent la classification par apprentissage ou par comparaison de séquences, en particulier pour la classification des eucaryotes unicellulaires pour les amplicons de taille modérée (>350 pb) (Taib *et al.* 2013). Pour ce qui concerne les eucaryotes, différentes régions du gène de l'ARNr 18S ont été testées par la simulation d'amplicons d'une longueur de 400 bp à partir d'une région conservée ciblée par des amorces directes sur les régions V1-V2, V3-V4, V4, V5-V6, V6,V7, V8-V9 (amorces directes :NSF4, NSF370, NSF573 NSF963, NSF1179 et NSF1419). Les résultats d'affiliation au niveau du genre différaient en fonction de la méthode d'affiliation (similitude, classification bayessienne, phylogénie NN ou LCA), de la région du gène de l'ARNr ciblée et des taxons considérés. En considérant les méthodes d'affiliation, la plus grande précision a été obtenue par l'approche PANAM-LCA (comprise entre 64,2% (V5-V6) et 79,2% (V8-V9) au niveau du genre) à l'exception de l'amplicon V5-V6 pour lequel les affiliations par BLAST sont plus précises. La précision de PANAM-LCA pour la région V4, la plus couramment ciblée dans les études de *metabarcoding*, est de 76,8 %. La « très bonne performance » de la région V8-V9 pour les affiliations des eucaryotes a largement été soulignée (cette région est utilisée dans le projet TARA océan). Il mérite cependant d'être noté que cette région est souvent absente des bases de données publiques, les résultats obtenus à partir de cette région ont été basés uniquement sur 300 séquences incluses dans la base de données de référence. Pour les bactéries, la restitution

BOX 3 – Sensibilité, précision, justesse

Pour évaluer les performances des outils de classification, on considère en général un ensemble d'entités dont on connaît *a priori* la catégorie, mais que l'on cherche à retrouver *via* les outils de classification. A partir des résultats de la classification de ces entités et du dénombrement des résultats corrects (vrais positifs / vrais négatifs) et des résultats incorrects (faux positifs / faux négatifs), on évalue les performances des outils de classification en termes de :



- sensibilité / rappel : part des entités correctement attribuées à une catégorie rapportée à l'ensemble des entités qui appartiennent effectivement à cette catégorie ;
- précision : part des entités attribuées correctement à une catégorie rapportée à l'ensemble des entités qui ont été attribuées à cette catégorie ;
- justesse : part de prédictions correctes rapportée à l'ensemble des prédictions.



taxonomique inférée par PANAM est équivalente aux résultats obtenus avec les méthodes d'affiliation implémentées dans QIIME : au niveau du genre, la précision varie de 82,7% pour RDP à 90,3% pour PANAM-NN et 91,4 % pour BLAST (Figure 1.12).

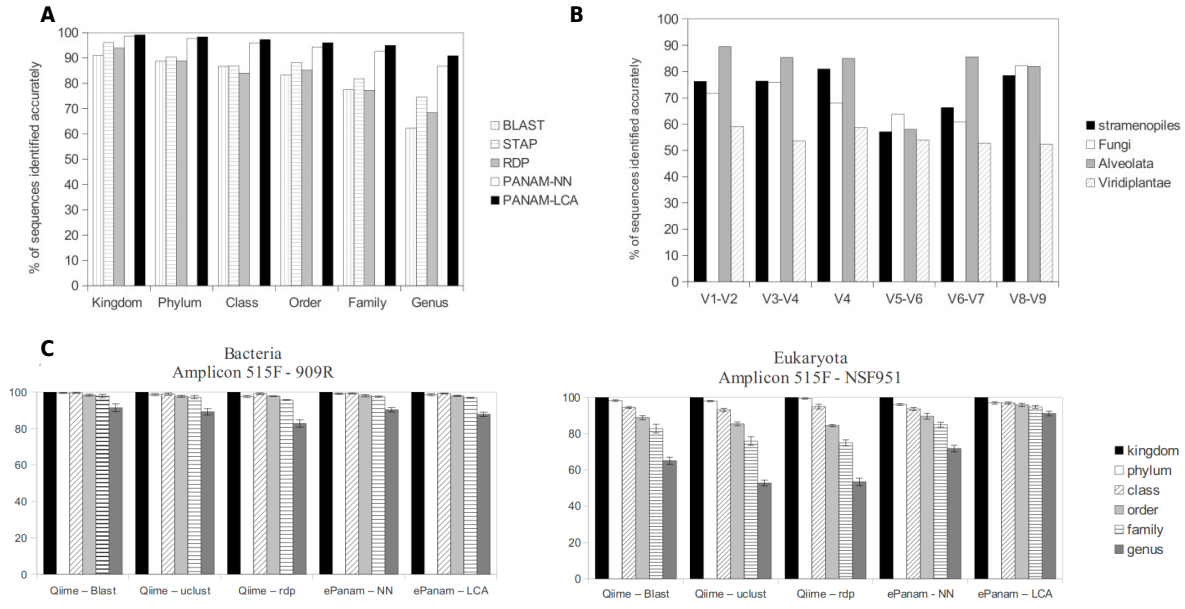


Figure 1.12 : Précision de l'affiliation phylogénétique de PANAM. (A) Précision de l'affiliation phylogénétique de PANAM-LCA, PANAM-NN, STAP, BLAST et le classifieur RDP sur les séquences ccomplètes d'ARNr 18S. (B) Précision de l'affiliation phylogénétique en relation avec la région variable ciblée. La spécificité a été testée avec PANAM-LCA et une longueur de séquence égale à 400 pb. (C) Spécificité des affiliations par rang taxonomique sur la région V4 telle que définies par les amorces de référence pour différents outils. Dans tous les cas, 1 000 séquences de longueur quasi complètes ont été choisies au hasard dans la base de données de référence d'où elles ont été retirées et retirées de celle-ci pour les simulations. Pour PANAM, les simulations ont été répétées 5 fois (d'après Taib *et al.* 2013).

La publication en 2018 de l'outil HmmUFOtu pour l'affiliation taxonomique de lectures assemblées ou non, confirme l'efficacité de l'approche phylogénétique sur l'analyse de similarité telles que celles proposée dans QIIME pour produire des affiliation précise sur de rangs taxonomiques fins (Zheng *et al.* 2018).

Les différences enregistrées entre la qualité des annotations entre les bactéries et les eucaryotes reposent principalement sur le fait que la diversité eucaryote est moins richement décrite dans les bases de données que ne peuvent l'être les clades bactériens ou archéens (Bik *et al.* 2012 , Taib *et al.* 2013). Ces résultats illustrent la supériorité de l'approche phylogénétique pour caractériser les communautés peu étudiées ou les micro-organismes pour lesquels nous ne connaissons pas toute l'étendue de la diversité.

2.3 – Quelques illustrations

Le développement de PANAM a permis à l'équipe d'analyser la biodiversité microbienne sur des modèles bactériens, archéens ou eucaryotes unicellulaires dans le cadre de projets propres ou en collaboration.

A travers la présentation de points particuliers de quatre études en écologie microbienne, j'illustre dans cette partie l'apport du *metabarcoding* et de son traitement par PANAM pour la caractérisation de la diversité microbienne dans les environnements aquatiques. Les deux premières études portent sur la caractérisation de communautés picooeucaryotes dans différents lacs (Taib *et al.* 2013, Li *et al.* 2017 - partie 2.3.1) et me permettent d'illustrer l'apport du *metabarconding* pour la découverte et d'unités taxonomiques minoritaires dans les écosystèmes. La troisième étude (Hugoni *et al.* 2013 - Partie 2.3.2) traite de la dynamique temporelle et de l'activité de la communauté archéenne d'une zone côtière en méditerranée. Elle permet d'illustrer l'apport de la phylogénie dans la discrimination de groupes archés caractérisés par des dynamiques saisonnières et des activités contrastées. La quatrième étude (Hugoni *et al.* 2015 - partie 2.3.3) est centrée sur l'étude de la dynamique de la communauté microbienne nitrifiante dans l'estuaire de la Charente. Cette étude me permet d'illustrer comment l'association de données environnementales avec les informations phylogénétiques restituées par PANAM peut amener à définir des écotypes, c'est-à-dire des groupes d'OTUs spécifiques, caractéristiques de conditions environnementales particulières.

Collaborations

- Isabelle MARY - Isabelle DOMAIZON – Hélène AGUOGE – Pierre GALAND (EC2CO 2010-12)
- Shengnan LI (doctorante) - Xiaoli SHI
- Kabilan Mani – Judith Bragança (Bourse d'étude du gouvernement indien)

2.3.1 – Des *Mamiellales* lacustres

La plupart des espèces eucaryotes sont définies par des différences morphologiques, mais comme la majorité des micro-organismes issus de l'environnement, leurs caractéristiques phénotypiques peuvent difficilement être décrites dans la mesure où on ne sait pas encore les mettre en culture. Aussi, la composition des communautés de protistes dans les écosystèmes lacustres est essentiellement décrite à travers des approches moléculaires.

Une des difficultés liée à la description des communautés de protistes réside dans la disparité des observations faites selon les techniques moléculaires utilisées pour caractériser ces communautés. Ainsi certains taxa (*Chlorophyta* et les *Haptophyta),* bien que constituant une proportion significative des communautés sur la base de comptages par la méthode FISH (*Hybridation In Situ en Fluorescence*) (Lepère *et al.* 2010), se trouvent en faible proportion voire absent avec des approches de clonage-séquençage (Lepère *et al.* 2006, Tarbe *et al.* 2013, Morris *et al.* 2002, Lefranc *et al.* 2005). Dans ce contexte, nous avons appliqué une approche de *metabarcoding* NGS (Roche 454) sur huit lacs et réservoirs de la région Auvergne Rhône-Alpes pour caractériser la composition des communautés de protistes des eaux de surface durant la période de stratification thermique (Taib *et al.* 2013). Les amplicons générés ciblant la région V4-V5 du gène de l'ARNr 18S. Après nettoyage, tri et normalisation des données, les lectures ont été clusterisées à un seuil de 95 % d'identité (Caron *et al.* 2009, Mangot *et al.* 2013). En plus de décrire les communautés de ces lacs (ce dont je m'abstiendrai ici), le traitement des données par PANAM a permis de montrer que peu d'OTUs eucaryotes lacustres avaient été décrits jusqu'alors. Ces OTUs correspondent pour l'essentiel à des taxa retrouvés dans différents environnements par les approches de clonage-séquençage, mais en relativement faible abondance et pour lesquels il existe peu d'informations disponibles. De nouveaux OTUs appartenant à des taxa encore jamais décrits dans les écosystèmes lacustres ont également été trouvés. C'est le cadres Foraminifères et des *Mamiellales* qui, bien que rares, forment des clades lacustres soutenus par des valeurs de bootstrap élevées (Figure 1.13).

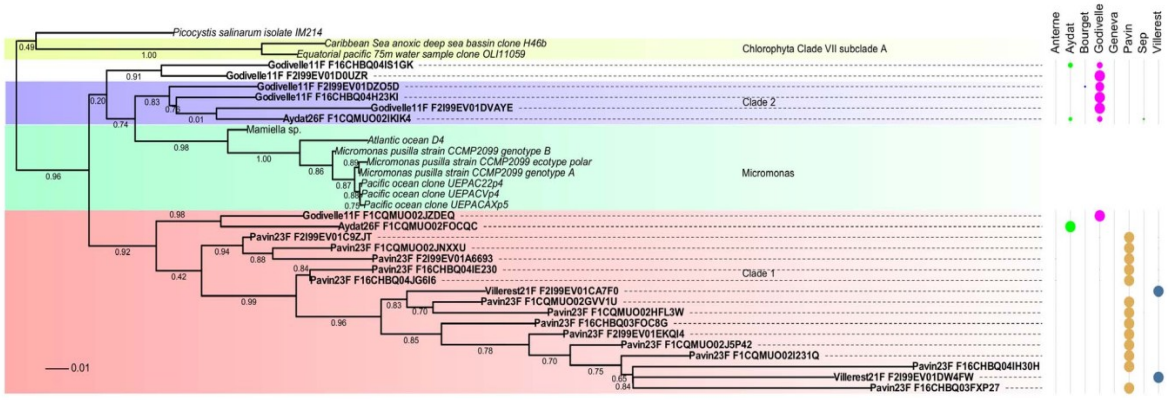


Figure 1.13 : Principaux clades détectés au sein des *Mamiellophyceae* à partir d'amplicons ARNr 18S (425pb). Les OTUs ont été générés au seuil de 95 % d'identité. La distribution des OTUs au sein des différents lacs étudiés montrent une présence du clade 1 principalement dans le lac Pavin tandis que le clade 2 est retrouvé dans le lac de la Godivelle (d'après Taib *et al.* 2013).

C’est à notre connaissance la première fois qu'un clade étroitement associé aux *Mamiellales* est détecté dans des lacs. Ceux-ci constituent, par contre, le groupe de microalgues photosynthétiques dominant du picoplancton marin, en particulier dans les eaux côtières (Tardin et Vaulot 2019). Ces résultats suggèrent que de nouvelles approches ciblant spécifiquement les organismes photosynthétiques dans en milieu lacustre sont nécessaires afin d’avoir une meilleure description de ces communautés (Marin et Melkonian 2010). Une telle approche a été mise en œuvre dans les travaux de Shengnan Li, portant sur la caractérisation de la communauté picoplanctonique eucaryote photosynthétique (PPE) de deux lac chinois (Li *et al.* 2017). Dans cette étude, la communauté PPE a été spécifiquement ciblée par une approche de tri cellulaire par cytométrie en flux, avant d’être analysée par *metabarcoding* NGS (Illumina Miseq). L’analyse de ces données avec PANAM a permis non seulement de confirmer l’efficacité de l’approche d’enrichissement de la communauté PPE ciblée dans les échantillons, puisque plus de 70 % des lectures traitées appartenaient au groupe des eucaryotes photosynthétiques, mais également de révéler un OTU lacustre partageant 99 % d’identité avec une séquence affiliée au genre *Ostreococcus* sp. (groupe des *Mamiellales*), confirmant l’existence d’un taxon lacustre affilié aux *Mamiellales.*

2.3.2 - Dynamique des communautés archéennes en mer Méditerranée

Les études sur la croissance microbienne océanique ont révélé une corrélation positive entre l'abondance et l'activité des communautés bactériennes dans les eaux côtières de surface (Campbell *et al.* 2011, Gaidos *et al.* 2011). Les microbes les plus abondants contribuent largement au fonctionnement de l’écosystème, du fait même de leur abondance, toutefois ceux-ci ne sont pas toujours les plus actifs au sein de la communauté (Campbell et Kirchman 2012, Lennon et Jones 2011). Il a ainsi été montré que la fraction rare (les organismes peu abondants) des communautés peut être active (Jones et Lennon 2010), mais que le taux de croissance des OTUs qui la composent diminuait lorsque leur abondance augmentait (Campbell *et al.* 2011). ainsi, contrairement à l’idée avancée selon laquelle la biosphère rare constituerait un fraction « dormante » de la communauté, c’est à dire une faction inactive, mais en capacité de se réactiver si les conditions venaient à changer (Pedrós-Alió 2006), ces résultats suggèrent que la biosphère rare joue un rôle dans le fonctionnement de la communauté.
Les archées marines sont des acteurs du plancton microbien contribuant significativement aux cycles biogéochimiques (Hugoni 2013). Dans l’étude rapportée ici, une approche *metabarcoding* a permis d’étudier la structure, l’activité et la dynamique saisonnière à long terme de la communauté archéenne dans les eaux de surface du nord-ouest de la Méditerranée (Hugoni *et al.* 2013). Le séquençage NGS (Roche 454) de l'ADNr et de l'ARNr 16S de 40 échantillons prélevés mensuellement sur 3 ans a permis de calculer un rapport d’abondance ARNr/ADNr et d’en déduire un indice de la croissance, indicateur de l’activité des organismes étudiés (Campbell *et al.* 2011, Gaidos *et al.* 2011).

Dans cette étude, sont considérés comme rares les OTUs constitués de moins de 0.2 % des séquences et présents au plus une fois dans les jeux de séquences normalisés à l’échelle des échantillons (soit 488 séquences).

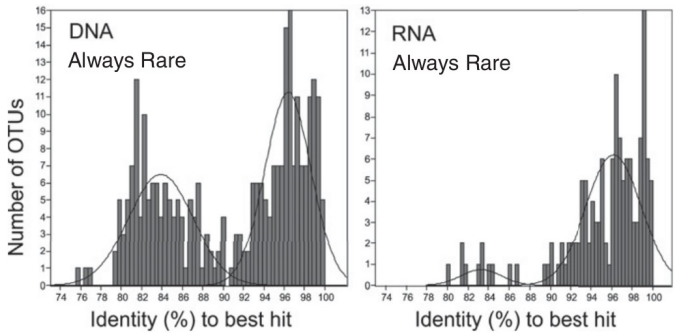


Figure 1.14 : Distribution du pourcentage d'identité entre les séquences de la base de données SILVA et les séquences d'ADNr et d'ARNr 16S des OTUs toujours rares. Un modèle de distributions normales (lignes noires) identifie les groupes d'OTUs communs (c'est-à-dire un pourcentage d'identité élevé) ou peu communs (c'est-à-dire un pourcentage d'identité faible) dans SILVA (d'après Hugoni *et al.* 2013).

Sur l'ensemble des OTUs abondants détectés, 15 présentaient une différence d'abondance entre les saisons hiver et été. Certains OTUs peuvent donc être abondants à certaines saisons et devenir rares à d'autres. Quand ils sont détectés, les OTUs abondants sont actifs. Près de la moitié des OTUs rares présentaient des similitudes faibles avec les séquences de la base de données SILVA, ce qui suggère que ces OTUs appartiennent à une fraction sous-échantillonnée des écosystèmes. Ces derniers sont retrouvés dans le jeu de séquences extrait à partir des ADN génomiques, mais n'apparaissent pas dans la fraction ARN (Figure 1.14). Ce déficit en ARNr des OTUs toujours rares et présentant une similitude faible avec les séquences de références issues de SILVA suggère que ces OTUs ne sont pas actifs.

Parmi les OTUs actifs, ceux affiliés au groupe d'archées marines MGI constituent une communauté affichant une dynamique saisonnière, incluant des OTUs abondants et d'autres toujours rares. La plupart des OTUs sont affiliés au clades MGI.B et MGI.A, étroitement lié à *Nitrosopumilus maritimus*, capable d'oxyder l'ammonium (Könneke *et al.* 2005). Outre les clades MGI.A et MGI.B déjà décrits et comportant à la fois des OTUs abondants et rares, l'approche phylogénétique de PANAM a permis de mettre en évidence deux nouveaux clades (MGI.C, MGI.D) constitués exclusivement d'OTUs rares (Figure 1.15). Ces clades, semblent par ailleurs répondre aux changements des conditions saisonnières puisqu'ils sont plutôt actifs en hiver et ont des dynamiques d'activité différentes de celles des clades abondants (Hugoni *et al.* 2013). Les indices d'activité mesurés sur ces différents clades suggèrent que les OTUs du clade MGI.B sont plus actifs que ceux du clade MGI.A, même s'ils ne sont pas les plus abondants dans l'écosystème. Le clade MGI.C semble actif lorsqu'il est présent, alors que le clade MGI.D ne l'est pas toujours.

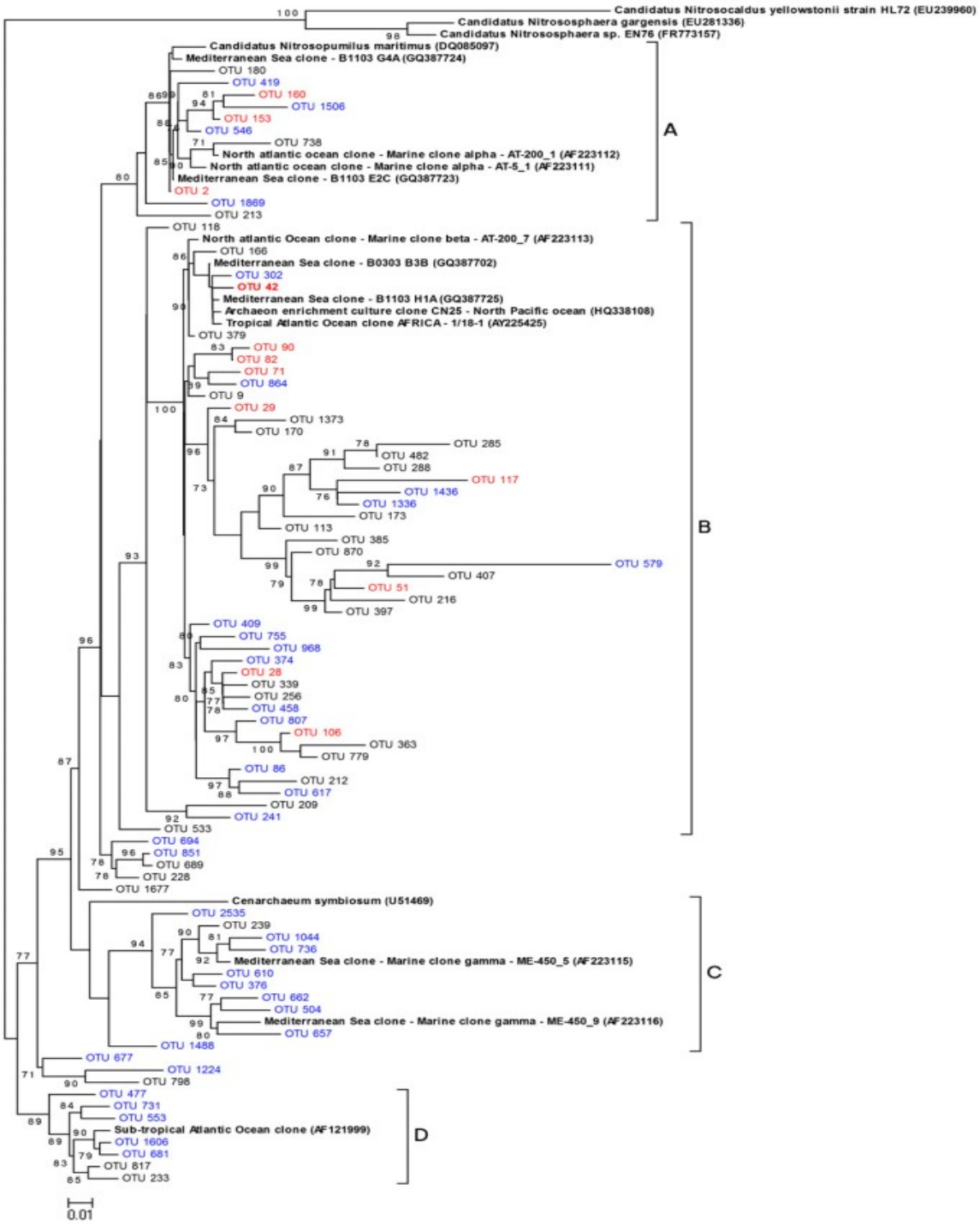


Figure 1.15 : Arbre phylogénétique incluant les séquences représentatives des OTUs associés au groupe d'archées marines MGI issues d'un échantillon prélevé sur le site de l'Observatoire microbien de Banyuls-sur-Mer. Les séquences de référence extraites de GenBank sont en gras. Les OTUs abondants sont représentés en rouge, les rares sont en bleu. Les valeurs de bootstrap >70 sont exprimées en pourcentage de 100 répliques (barre d'échelle : 10% de divergence de séquence) (d'après Hugoni *et al.* 2013).

2.3.3 – Spécialisation de niches des communautés nitrifiantes le long d'un gradient de salinité

L'eau, l'oxygène, le carbone, l'azote et l'hydrogène constituent 95% de la composition atomique des organismes vivants qui puisent ces ressources de façon sélective dans leur environnement. La circulation et le recyclage de ces éléments sont assurés *via* des flux de matière entre les organismes autotrophes, hétérotrophes et l'environnement (on parle de cycles biogéochimiques). Ainsi, le bon déroulement de ces cycles est assuré par une complémentarité écophysiologique des différents intervenants, et la pérennité des écosystèmes est assurée par le maintien d'une autorégulation, ou homéostasie, entre ces différents compartiments du cycle. Les micro-organismes y jouent un rôle primordial par leur implication dans des fonctions très spécifiques, notamment dans le cycle de l'azote (Hugoni 2013).

Dans le cadre du cycle de l'azote, la distribution des archées oxydant l'ammonium (AOA) dans les eaux océaniques et les écosystèmes d'eau douce est bien établie (Mincer *et al.* 2007, Galland *et al.* 2010, Hugoni *et al.* 2013, Vissers *et al.* 2013), mais leur contribution relative par rapport aux bactéries oxydant l'ammonium (AOB) reste floue en particulier dans les systèmes estuariens. En effet, alors que certaines études rapportent que les AOB dominent dans des conditions estuariennes salines (Mincer *et al.* 2007), d'autres montrent que les archées nitrifiantes (AOA) y sont souvent plus nombreuses (Beman *et al.* 2008) voire dominantes (Bernhard et Bollmann 2010). Dans ce contexte, la structure des communautés nitrifiantes (part relative des AOA et AOB) dans les eaux de surface de l'estuaire de la Charente a été suivie pendant 1 an (Hugoni *et al.* 2015). La structure des communautés archéennes potentiellement actives a été étudiée sur gradient de salinité par une approche de *metabarconding* ciblant la région V3-V5 du gène de l'ARNr 16S.

La quantification de l'abondance du transcrit (ARN) du gène codant pour l'enzyme AmoA, impliqué dans la réaction d'oxydation de l'ammonium a montré une transition dans les populations actives de l'AOB dans les eaux douces à l'AOA dans les eaux marines (Hugoni *et al.* 2015). Ceci est en accord avec des études précédentes qui suggèrent que les l'activité nitrifiante des AOB est inhibée dans un environnement à haute salinité (Bernhard *et al.* 2007). Ces auteurs ont toutefois montré que les AOB présentaient une large gamme de tolérance à la salinité, laissant à penser que d'autres paramètres environnementaux doivent être pris en compte pour comprendre la dynamique de leur activité.

La phylogénie réalisée sur les OTUs archéens affiliés au clade MGI a permis de caractériser six sous-clades majeurs (Figure 1.16), reflétant l'existence de communautés distinctes le long de l'estuaire. Les OTUs retrouvés en eau douce appartenaient principalement au clade A spécifique de ce compartiment et au clade propre aux sédiments, tandis que ceux isolés dans les eaux mésohalines et marines sont associés au clade A marin. Les OTUs communs aux stations d'eau douce et mésohaline appartiennent au clade affilié aux sédiments, tandis que les OTUs spécifiques des stations mésohalines ou marines sont principalement retrouvées dans le clade A marin. Des groupes monophylétiques associés à une variabilité saisonnière de l'activité d'oxydation de l'ammonium ont été identifiés au sein du clade A marin, suggérant une spécialisation écologique de ces clades qui pourraient constituer de possibles

écotypes. Bien que les niches écologiques ne puissent pas être définies dans cette étude (il n'y a pas de lien statistique clair avec les paramètres environnementaux mesurés), ce travail confirme l'idée qu'au sein des archées le clade MGI est composé de différents écotypes, comme précédemment proposé en milieu marin (Hugoni *et al.* 2013, Sintes *et al.* 2013) et lacustre (Auguet *et al.* 2012, Auguet et Casamayor 2013, Restrepo-Ortiz *et al.* 2013).

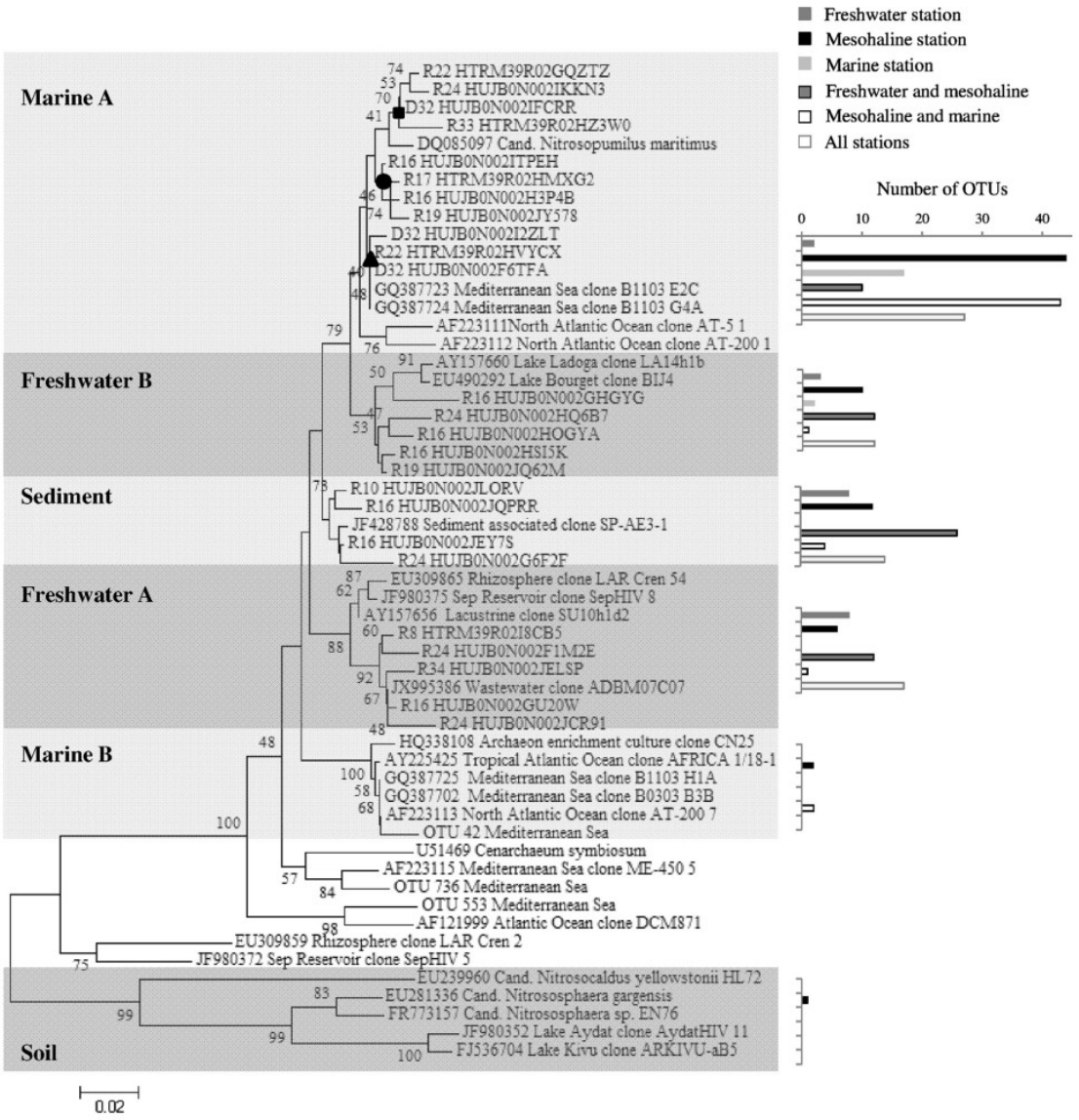


Figure 1.16: Arbre phylogénétique des OTUs archéens potentiellement actifs affiliés au clade MGI issus de trois stations le long de l'estuaire de la Charente. Les valeurs de bootstrap >40 sont indiquées. Les histogrammes représentent le nombre d'OTUs dans chaque station, dans les stations d'eau douce et mésohalines, dans les stations mésohalines et marine, et dans les trois stations. Les changements saisonniers d'activité pour les OTUs de la station mésohaline au sein du cluster A marin sont illustrés par une iconographie (■ : septembre à novembre ; ● : avril à août ; ▲ : décembre à mars, d'après Hugoni *et al.* 2015).

3 – Une meilleure estimation de la diversité ?

Un point majeur dans les études des communautés microbiennes est la caractérisation des micro-organismes présents dans l'environnement et donc dans l'échantillon. L'un des principaux constat associé à l'étude de la diversité microbienne par des approches de séquençage NGS réside dans l'augmentation importante du nombre d'OTUs caractérisés et donc de la richesse spécifique du monde microbien (BOX 4). Cette accroissement de la richesse spécifique résulte notamment d'une plus grande profondeur d'échantillonnage des milieux étudiés, ce qui permet d'accéder à des espèces microbiennes minoritaires dans les écosystèmes et non encore décrites. Cette diversité insoupçonnée est souvent appelée *microbial dark matter* en référence à la matière noire des physiciens.
Certains auteurs remettent cependant en cause une partie de cette diversité nouvelle et rare, et argumentent que celle-ci est la conséquence du traitement inapproprié des séquences ayant accumulé de multiples erreurs lors des étapes d'amplification et de séquençage inhérentes au *metabarcoding* (Sogin *et al.* 2006, Quince *et al.* 2009,). Ces erreurs peuvent en effet fausser l'image des communautés, tant sur l'estimation du nombre que de la taille ou des contours des OTUs générés. Récemment, les OTUs ont été remis en cause au profit de ESV (*Exact Sequence Variant*), en particulier dans le contexte de l'estimation de la richesse spécifique dans les écosystèmes.

3.1 – Les variants exacts de séquences

3.1.1 – Algorithmes de *débruitage*

Les procédures de *débruitage* des données de séquençages pour limiter l'inflation de la richesse spécifique consiste à appliquer une correction sur les séquences d'amplicons identifiées comme suspectes. Il s'agit d'un traitement bio-informatique des données issues du séquençage visant à inférer l'ensemble des séquences-type présentes dans l'échantillon au départ, pour pouvoir ensuite identifier les variants de séquences (*Exact Sequence Variant* - ESV) et les réaffecter à leur séquence type originale. Les premières méthodes de correction d'erreur ont été développées pour traiter les séquences artéfactuelles issues d séquençage Roche 454 (Quince *et al. 2009,* 2011, Reeder et Knight, 2010, Rosen *et al.* 2012). Elles ont par la ensuite été adaptées aux plateformes Illumina et sont en cours d'adaptation pour celle PacBio (Callahan *et al.* 2019).

Trois approches de *débruitage* ont été développées pour traiter des lectures issues des technologies Illumina. Toutes s'accordent à considérer l'abondance des lectures comme un marqueur de leur inexactitude et modélisent, soit les erreurs de séquençage directement, soit les effets des erreurs de séquençage sur la distribution des lectures erronées dans le jeu de séquences pour inférer les séquences *vraies*, c'est-à-dire celles présentes dans l'échantillon au départ. Par ordre chronologique, les trois outils les plus couramment cités sont UNOISE2 (Edgar 2016), DADA2 (Callahan *et al.* 2016) et DEBLUR (Amir *et al.* 2017).

UNOISE2 (Edgar 2016 – Figure 1.17) utilise une stratégie de clusterisation gloutonne des lectures triées par ordre d'abondance décroissant. Cet outil est basé sur l'idée que les erreurs issues de l'amplification par PCR, puis du séquençage, vont générer un nuage de séquences dérivées d'une séquence d'origine telles que les séquences ayant accumulé des erreurs tôt dans le processus seront plus abondantes et auront accumulé plus d'erreurs. Celles-ci seront donc plus distantes (moins semblables) que les séquences ayant accumulé des erreurs plus tardivement. A partir d'un type de séquence, il est alors possible de définir une représentation grossière du phénomène d'accumulation des mutations par un modèle phénoménologique, c'est-à-dire estimé à partir des observations représentant le biais d'abondance B(C,M) maximal autorisé pour une séquence M par rapport à une séquence C. Il s'agit en fait de comparer l'abondance d'une séquence M à la valeur du biais d'abondance prédit B(d) à un seuil de divergence donné d. Si l'abondance de M est supérieure au biais d'abondance prédit B(C,M) entre les séquences M et C, alors la séquences M est trop abondante pour que sa présence dans l'échantillon puisse être expliquée par l'amplification d'une séquence erronée issue du C et cette séquence formera un zOTU (*zero-radius* OTUs). Dans le cas contraire, la séquence M est inclue dans C dont l'abondance est corrigée en conséquence.

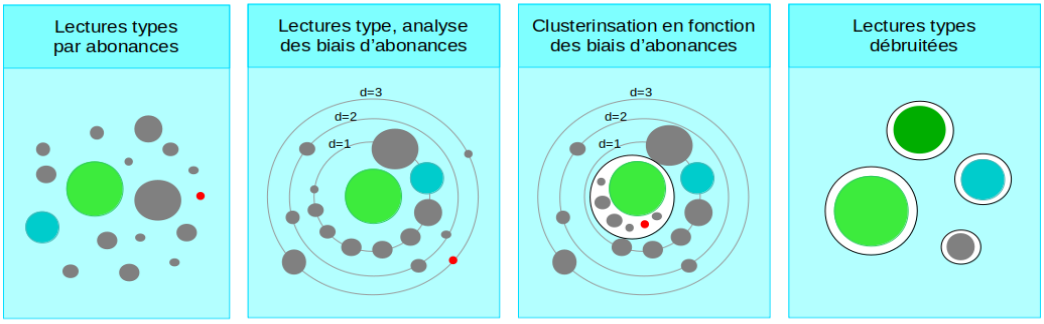


Figure 1.17 : *Débruitage* par UNOISE2 : les séquences *dérépliquées* constituent des séquences types traitées par ordre d'abondance décroissante. La distance entre chaque séquence type et le type le plus abondant est calculée, ainsi que son biais d'abondance maximal. Si l'abondance de la séquence type évaluée, rapporté à sa distance au type le plus abondant est inférieure au biais d'abondance maximal théorique, la séquence type sera incluse dans le cluster du type de la séquence la plus abondante (d'après Edgar 2016).

DEBLUR (Amir *et al.* 2017 – Figure 1.18) est basé sur une stratégie de ré-estimation des abondances des différents types d'amplicons à partir de l'estimation de l'abondance des lectures erronées issues du séquençage. On part ici du principe que les erreurs de séquençage sont aléatoires mais que leur nature précise ne peut être connue. Il s'agit alors de modéliser une approximation de ces erreurs en estimant i) la probabilité moyenne d'obtenir une séquence erronée et ii) la limite supérieure de la probabilité B(K) pour un amplicon avec une abondance T dans l'échantillon, de générer une lecture avec K erreur.
Partant du principe que les types de séquences les plus abondants correspondent à des séquences *vraies* de l'échantillon, on peut à partir d'une séquence de type i, et pour chaque autre séquence de type j distante de (différant de) K nucléotides, estimer la proportion de séquences de type j correspondant à une lecture issue de i mais ayant accumulée K erreurs. Il devient alors possible de proposer des abondances corrigées pour chaque type de lecture dans le jeu de données. L'approche est répétée pour les différents types de séquences considérées par ordre d'abondance décroissante. Si l'abondance d'un type vient à devenir nulle, le type est supprimé. Les types maintenus à l'issue de cette analyse sont appelés sub-OTU.

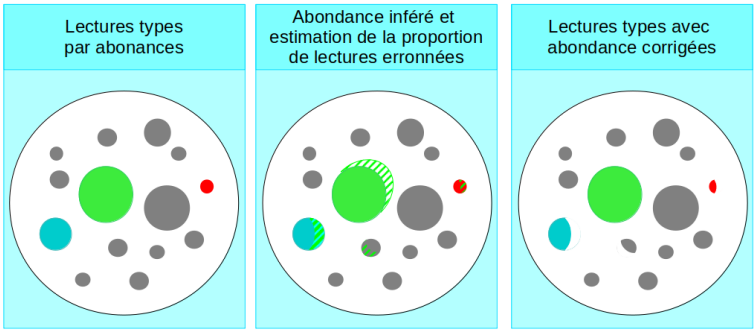


Figure 1.18 : *Débruitage* par DEBLUR (Amir *et al.* 2017). Pour un type de séquence avec un abondance donnée, connaissant la probabilité qu'une séquence produite à partir d'une séquence type soit erronée, on peut estimer le nombre total de séquences (justes ou contenant des erreurs) correspondant à cette séquence type dans le jeu de données avant séquençage et donc corriger l'abondance attendue pour cette séquence type. 1 – Soit *ci* le nombre de séquences de type ri ; *c'i* le nombre de séquences de type *ri* dans l'échantillon d'origine. On cherche à corriger *ci* et inférer *c'i* ; 2 – On calcule une approximation de *c'i* ; 3 - *c'i = ci *(1 / 1 - a)* où a est la probabilité moyenne d'obtenir un *misread* (0.5% par défaut).

DADA2 (Callahan *et al.* 2016 – Figure 1.19) est l'adaptation pour Illumina d'une méthode initialement développée pour la technologie Roche 454 (Rosen *et al.* 2012). L'approche consiste à i) modéliser les erreurs de séquençage et ii) estimer une *p*-value d'abondance *pA* associée au fait qu'une séquence de type i soit trop abondante pour être expliquée par des erreurs de séquençage. Cette méthode utilise les données qualitatives associées aux lectures pour estimer un modèle d'erreur propre à chaque jeu de données.

A partir d'une partition initiale unique, contenant l'ensemble des types de séquences, et considérant le type le plus abondant comme séquence de référence de cette partition, la probabilité d'abondance pour chaque autre type dans la partition est calculée. Si les probabilités d'abondance minimale pour les types minoritaires sont inférieures au seuil *WA* – paramètre du logiciel, ces types minoritaires sont extraits de la partition initiale pour former le type de référence d'une nouvelle partition (une *pA* faible indique qu'il y a plus de séquences de ce type qu'attendu par le seul fait des erreurs de séquençage). Chaque type de séquence restant est ensuite comparé aux différents types de référence et assignés à la partition à laquelle il ressemble le plus. Ce processus est répété tant que de nouvelles partitions sont inférées. Enfin, chaque partition est résumée par sa séquence de référence pour constituer un ASV (*Amplified Sequenced Variant*) dont l'abondance sera égale à la somme des abondances de tous les types inclus dans la partition.

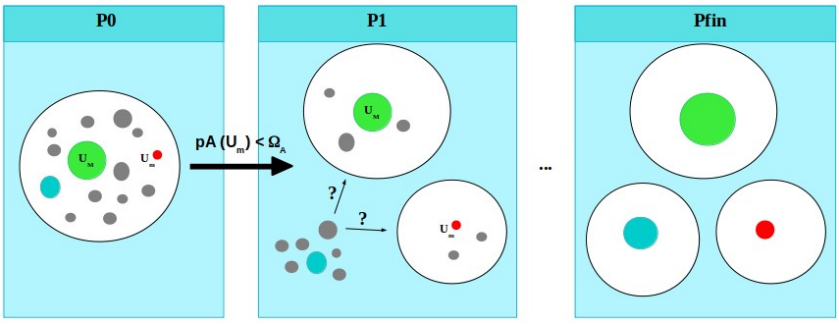


Figure 1.19 : *Débruitage* par DADA2 (Callahan *et al.* 2016). On peut modéliser le nombre de séquences de type *i* qui sont produites à partir d'un amplicon de type *j*, sachant qu'on a effectivement observé une lecture de type i. Ceci définit la probabilité d'abondance *pA* de la séquence de type *i* si celle-ci est la conséquence d'un séquençage erroné d'une séquence originale de type j. La probabilité d'abondance *pA* permet de quantifier l'idée qu'une séquence de type *i* est trop abondante pour être expliquée seulement par des erreurs de PCR et de séquençage. Elle est d'autant plus faible que la probabilité des erreurs de substitutions est faible.

DADA2 a récemment été adapté pour pouvoir traiter également des lectures CCS PacBio (Callahan *et al.* 2019). Ceci implique une modification de la structure de données utilisées pour gérer des séquences longues et une optimisation de la procédure de comparaison des séquences par paires. Enfin, le modèle d'erreur a été modifié de manière à coller au plus près au schéma d'erreurs caractérisant le séquençage PacBio.

Sur des aspects techniques, DADA2 est un package R qui nécessite d'avoir les données brutes au format fastq. UNOISE3 et DEBLUR peuvent traiter des lectures *paired-end* assemblées. DADA2 et UNOISE3 peuvent traiter des échantillons réunis (échantillons *poolés*), ce qui permet de prendre en compte les erreurs de type *batch* (Nearing *et al.* 2018), alors que DEBLUR utilise une stratégie de parallélisation sur les échantillons pour réduire le temps de traitement des données (Amir *et al.* 2017). Tous ces outils possèdent un étape de détection des chimères.

Différentes études comparant ces outils sur des données issues de communautés contrôlées (ou *mock*) ont montré que le nombre d'ESV restitués est dans tous les cas supérieur au nombre d'organismes présents dans les communauté initiales, mais inférieur au nombre d'OTUs obtenus par les approches par clusterisation (Nearing *et al.* 2018, Prodan *et al.* 2020). DADA2 est l'outil le plus sensible, mais produit plus d'ESV douteux, tandis que UNOISE et DEBLUR présente une meilleure spécificité (Prodan *et al.* 2020). UNOISE3 tend à trouver plutôt les ESV les plus abondants. DADA2 est plus sensible pour la détection des taxa rares, mais reste moins sensible que les approches par OTU. Enfin, bien que retrouvant le moins d'ESV en général, DEBLUR présente une bonne stabilité, c'est-à-dire qu'il permet de retrouver les mêmes unités de diversité lorsque les échantillons sont traités indépendamment (ceci peut être intéressant quand on souhaite exploiter des échantillons existants à partir de ressources telles que le Earth Microbiome Project – Amir *et al.* 2017). Bien que ces approches soient efficaces pour détecter des taxa abondants, leurs résultats ne sont que partiellement chevauchants (Figure 1.20). Ceci s'explique en partie par leur particularité dans la manière de capter les séquences de faible abondance.

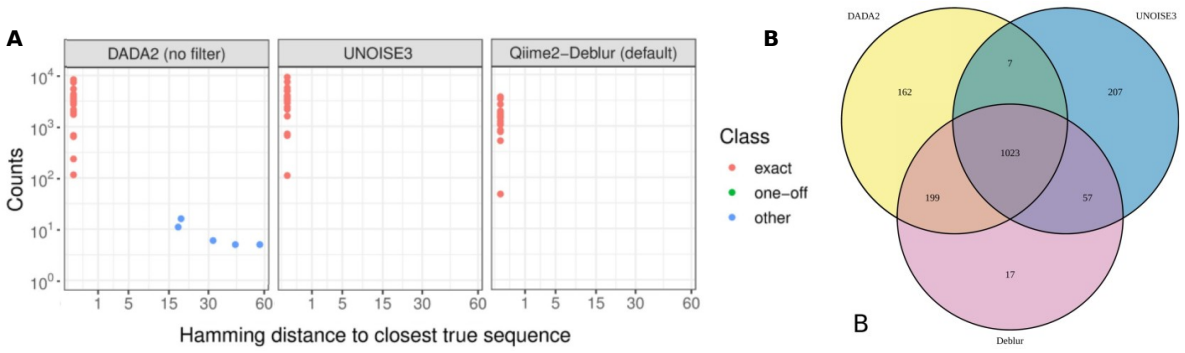
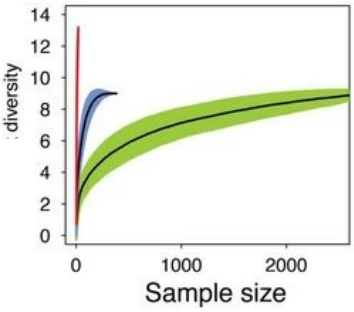


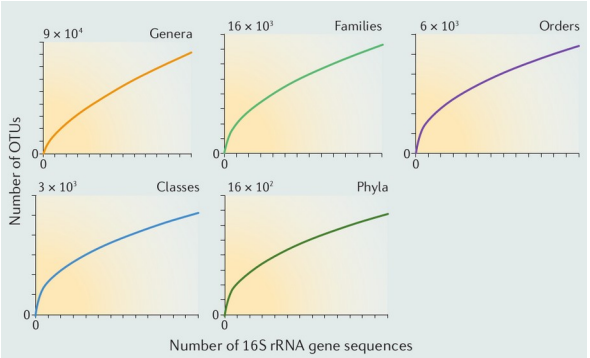
Figure 1.20 : Comparaison des approches UNOISE, DEBLUR et DADA2. A – Abondance des ESV prédits en fonction de la distance (Hamming) à la référence (sur communauté *mock* – DADA2 retrouve l'ensemble des ESV attendus mais prédit aussi quelques ESV faux, UNOISE et DEBLUR ne produisent aucun ESV suspect mais ne retrouvent pas les ESV peu abondants, effectivement présents dans le jeu de données. B – Recouvrement des ESV détectés par les trois méthodes de *débruitage* sur une communauté réelle, après filtrage des lectures rares (< 0.002% du jeu de données) (adapté de Prodan *et al.* 2020).

BOX 4 – Estimation de la diversité spécifique

La diversité alpha permet de décrire la biodiversité d'un milieu en termes du nombre d'espèces présentes, pondérées par leur abondance ou replacées dans un contexte phylogénétique. Cette diversité alpha se décline à travers :

- L'estimation de la richesse spécifique, correspondant au dénombrement des différentes espèces qui constituent cette diversité. En écologie microbienne, la richesse est approximée par le comptage des OTUs ou ESV identifiés dans l'échantillon analysé.
- L'estimation d'indices de diversité spécifique (tels que les indices de Shannon, de Simpson ou le Schao) qui intègrent l'abondance relative des OTUs et estime *l'équitabilité* (*evenness*) des différentes unités taxonomiques dénombrées.
- L'estimation d'indices de diversité phylogénétique (tel que l'indice PD, Faith 1992), qui est une généralisation de la richesse spécifique. En plus de considérer le nombre d'unités taxonomiques différentes dans l'échantillon, ces indices prennent en compte la proximité évolutive des unités considérées. Ces indices peuvent selon leurs propriétés prendre en compte l'abondance relative des unités taxonomiques ou leur entropie (Webb *et al.* 2000).





In part **a** of the figure, the taxonomic thresholds are as follows: genera 94.5% (yellow), families 86.5% (green), orders 82.0% (purple), classes 78.5% (blue) and phyla 75.0% (dark green). For species, thresholds of 98.7% (red), 99.0% (dark grey) and 99.5% (light grey) were used.

Les courbes de raréfaction représentent la richesse spécifique (ici des OTUs) d'un échantillon pour un nombre donné d'observations (ici le nombre de séquences d'ADN produit, également appelé profondeur de séquençage). Lorsque la richesse spécifique d'une communauté est suffisamment échantillonnée, la courbe de raréfaction forme un plateau indiquant qu'un échantillonnage supplémentaire ne permettra pas de capter plus de richesse spécifique.

3.1.2 – Comparaison des approches OTU et ESV pour caractériser la richesse taxonomique

De nombreuses études ont été réalisées pour évaluer les biais découlant de l'utilisation du *metabarcoding* dans la description de la richesse et de la diversité microbienne *via* la génération d'OTU (Huse *et al.* 2007, Quince *et al.* 2009, Kunin et Hugenholtz, 2010). Il a ainsi été observé que les courbes de raréfaction, qui visent à estimer la richesse spécifique du milieu, pouvaient voir leur profil changer en fonction de la profondeur de séquençage considérée (Roesch *et al.* 2007). Sur la base de sous-ensembles de différentes tailles d'un jeu de séquences initial et une clusterisation en liaison complète, ces auteurs ont observé des courbes de raréfaction plus pentues lorsque le nombre de séquences considérées est grand. Cette observation est en contradiction avec le comportement attendu d'une telle courbe et remet en question le principe fondamental selon lequel la diversité d'une communauté entière peut être estimée à partir d'un échantillon séquencé (He *et al.* 2015). Ces résultats peuvent en partie s'expliquer

par le choix de la méthode de clusterisation (*complete-linkage*) qui, bien que garantissant une identité minimale entre toutes les séquences d'un OTU, s'avère être trop stricte dans un contexte de NGS. Ils ont cependant le mérite de souligner l'impact du choix de la méthode dans la définition des OTUs et des propriétés du jeu de données dans la description de la richesse spécifique d'un milieu.

Si la clusterisation en liaison complète ne permet pas une estimation fiable de la richesse d'un environnement dans un contexte NGS, les méthodes alternatives présentent d'autres inconvénients. Le reproche majeur qui est fait aux méthodes gloutonnes ou non seuillées est qu'elles tendent à agréger les amplicons issus de variants de séquences pouvant correspondre à des unités de diversités proches mais distinctes (espèces ou sous-espèces, voire genre), tout en étant peu efficaces dans l'agrégation de séquences divergentes ayant accumulé des erreurs d'amplification et de séquençage. Ceci à pour effet d'augmenter le nombre d'OTUs produits (Kunin *et al.* 2010), tout en masquant la micro-diversité qui pourrait exister.
L'accroissement des valeurs de richesse spécifique estimées dans les études de *metabarcoding* est pour beaucoup associé à une proportion importante d'OTUs représentés par une unique ou un faible nombre de séquences, correspondant pour la plupart à des déviants de séquençage. Afin de pallier ce problème, les OTUs en faible abondance sont en général supprimés du jeu de données, au risque d'éliminer également des OTUs rares. Il existe plusieurs stratégies pour écarter ces OTUs peu abondants, depuis la suppression des singletons, en passant par la suppression des OTUs représentant moins d'une proportion définie du jeu de séquences initial, soit à l'échelle du jeu de séquences global, soit à l'échelle de chaque échantillon lors de l'analyse de plusieurs échantillons (Pedros-Alio 2006, Sogin *et al.* 2006, Campbell *et al.* 2011, Hugoni *et al.* 2013).

Les ESV permettent de réduire la diversité suspecte de l'échantillon de manière plus efficace que les OTUs (Johnson *et al.* 2019) Cependant, des ESV différents peuvent faire référence à une même espèce (ou sous-espèce) lorsque les génomes des organismes considérés possèdent plusieurs loci polymorphes de l'opéron ARNr. Or il n'est pas rare de retrouver des variants de copies de l'ARNr au sein d'un même génome pour une proportion significative de génomes bactériens (Acinas *et al.* 2004, Sun *et al.* 2013, Stoddard *et al.* 2015) et eukaryotes (Zhu *et al.* 2005). Ainsi, les ESV tendent à restituer une diversité génétique plutôt que qu'une diversité spécifique (Brandt *et al*. 2021) et d'un point de vu conceptuel, la désignation d'une unité de base de diversité à travers les ESV, bien que pour des raisons différentes, ne semble pas plus satisfaisante que la définition historique d'un OTU pour caractériser la diversité spécifique d'un écosystème. Dans leur article de 2019 Callahan et collaborateurs affirment : « nous mettons en garde contre la conclusion selon laquelle la quantification des ESVs est préférable aux approches plus traditionnelles basées sur les OTUs. Cette conclusion suppose que les ESVs représentent une unité taxonomique plus significative que les OTUs. Étant donné que la majorité des isolats bactériens que nous avons séquencés contenaient des copies multiples et variantes du gène 16S dans leur génome, cette hypothèse peut ne pas toujours être correcte » (Callahan *et al.* 2019). Aussi, lorsque des variants génomiques de la séquence de l'ARNr 16S au sein d'un génome sont identifiés, ces auteurs proposent de les regrouper au sein d'un "bin" représentant l'espèce, en analysant À travers différents échantillons la distribution (ou la co-occurence) des OTUs rares et abondants partageant une même affiliation taxonomique (tel que proposé par exemple dans Frøslev *et al*. 2017) On ne peut cependant pas

exclure que des ARNr homologues partagent strictement la même séquence entre souches proches, alors que d'autres peuvent avoir accumulé des SNP (Johnson *et al.* 2019).

3.2 – Apport des unités phylogénétiques de diversité

3.2.1 - Calibration *in situ* des unités taxonomiques opérationnelles (OTU)

Afin de proposer une image plus juste de la structure des communautés microbiennes, Sun et collaborateurs préconisent une validation expérimentale de la caractérisation des OTUs par l'ajout de séquences connues dans le jeu de données (Sun *et al.* 2012). Cette approche a été appliquée dans le cadre d'une collaboration avec Jean-François Mangot (doctorant dans l'équipe) concernant la dynamique à court terme des communautés d'eucaryotes unicellulaires du lac Léman (Mangot *et al.* 2013). Dans cette étude, un témoin interne, produit PCR issu du clonage d'une séquence d'ARNr 18S de *Blastocystis hominis* sous-type 4, a été ajouté à hauteur de 1 % de la quantité d'ADN dans chaque échantillon. Au delà de l'utilisation de ces séquences pour calibrer la normalisation quantitative des différents échantillons, celles-ci nous ont permis i) de caractériser les erreurs de séquençage (ici Roche 454) et ii) de déterminer la méthode et un seuil de clusterisation les plus adaptés aux données produites pour générer les OTUs. La meilleure procédure de clusterisation est celle qui regroupe toutes les séquences de *B. hominis* dans un même OTU. Il est apparu que 64 % des séquences de ce témoin interne constituent des variants de la séquence initiale. Parmi eux, 97 % sont des séquences uniques, résultant pour l'essentiel de substitutions.

Méthode	UCLUST			Mothur FN			Mothur AN			Mothur NN		
Similitude	95	96	97	95	96	97	95	96	97	95	96	97
Nbr OTUs	2	4	9	122	154	208	77	104	149	53	66	98
Nbr Singletons	1	1	1	55	73	113	41	58	98	28	37	66
Nbr seq max par OTU	1947	1826	1788	1194	1131	1077	1285	1250	1233	1308	1305	1294

Tableau 1.2 : Nombre et structure des OTUs générés par différentes méthodes de clusterisation et pour différents seuils de similarité à partir des séquences affiliées à *Blastocystis hominis.* UCLUST implémente une heuristique de clusterisation gloutonne. Mothur permet des clusterisations par liaison simple (NN), liaison complète (FN) ou liaison moyenne (AN) à partir d'une matrice de similarité entre séquences (d'après Taib 2013).

La meilleure approximation de la séquence type de *B. Hominis* a été obtenue avec la méthode UCLUST au seuil de 95 % d'identité sans qu'il ne soit possible de générer un unique OTU (Tableau 1.2). Une séquence présentant une divergence extrême (74 % d'identité) forme en effet un singleton. cependant, une phylogénie réalisée sur le groupe des *Stramenopiles* contenant l'ensemble des séquences de *B. hominis* et la séquence de référence du sous-type 4 a permis de constater que les séquences expérimentales forment un groupe monophylétique incluant la séquence du singleton (Figure 1.21). Ces résultats sont en adéquation avec l'idée générale que les OTUs composés de singletons sont des éléments majeurs associés à l'inflation de la richesse spécifique observées dans les écosystèmes depuis l'avènement des techniques de NGS. Ces résultats suggèrent que l'utilisation d'unité phylogénétique de

diversité, c'est-à-dire des unités taxonomiques définies sur la base de groupes monophylétiques, sont plus à même de décrire la diversité dans les écosystèmes (Debroas *et al.* 2017).

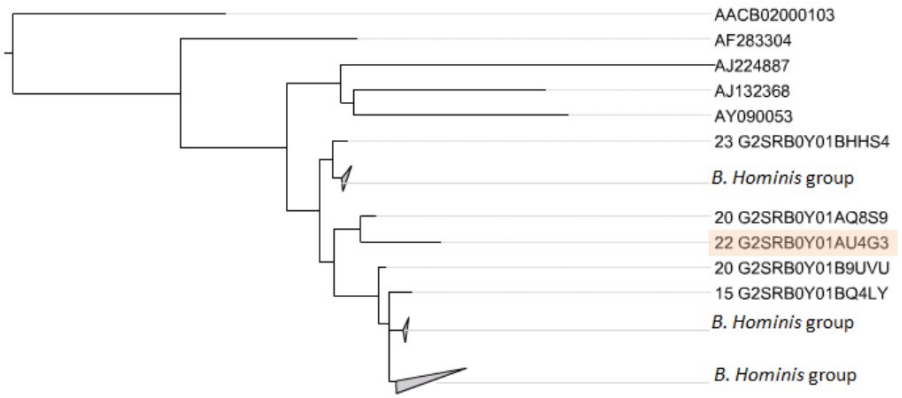


Figure 1.21 : Phylogénie des *Stramenopiles.* 1948 séquences de *Blastocystis hominis* identifiées dans le jeu de données avec les séquences de référence. L'ensemble des séquences de *B hominis* constitue un groupe monophylétiques, incluant la séquence 22_G2SRB0Y01AU4G3 qui forme un singleton au seuil de 95 % d'identité avec l'algorithme UCLUST. Les séquences de référence sont représentées par leur numéro d'accession dans genbank (d'après Taib 2013).

Collaborations
- Jean-François MANGOT - Isabelle DOMAIZON

3.2.2 – Clusterisation emboîtée

Des approches de clusterisations emboîtées ont été proposées pour pallier l'inflation de la richesse observée dans les études de *metabarcoding.* Celles-ci peuvent être basées sur l'application successives d'algorithmes de clusterisation en en OTUs utilisant des seuils décroissants d'identité (Huse *et al.* 2010), la combinaison d'approches de *débruitage* et de clusterisation (Frøslev *et al*. 2017, Brandt *et al.* 2021), ou l'association d'approches de clusterisation ou de *débruitage* avec l'analyse de réseaux de similarité de séquences (Forster *et al*. 2019). Elles permettent des regroupements entre OTUs faiblement abondants ou entre des OTUs rares et des OTUs abondants, sans que cela n'affecte la structure des communautés. L'impact de ces doubles clusterisations semble cependant moins critique que les étapes de pré-traitement des données (nettoyage, suppression des lectures peu abondantes, détection des séquences chimériques – Bonder *et al.* 2012) sur la génération ultérieure des OTUs (Botnen *et al.* 2018). Cette approche peut par contre masquer la micro-diversité dans les écosystèmes. Edgar (Edgar 2016) argumente que les sous-espèces sont relativement rares et que les agréger serait relativement bénin. Cependant, si les variants de séquence reflètent une diversité fonctionnelle ou une spécialisation écologique, l'application de différents seuils pour la délimitation des OTUs peut influencer notre capacité à interpréter l'écologie de ces unités de diversité. En effet, les séquences contenues dans un OTU généré sur la base d'un seuil d'identité ne sont pas toujours monophylétiques , ce qui peut impliquer une hétérogénéité écologique au sein de l'OTU (Koeppel et Wu 2013). Par ailleurs, pour avoir des résultats semblables en termes de richesse ou d'abondance, les seuils de clusterisation à appliquer ne sont pas

les mêmes suivant l'algorithme de clusterisation considéré (Schloss et Handelsman, 2005, Schloss et Westcott, 2011).

3.2.3 – Approche phylogénétique pour l'analyse de la diversité microbienne

S'il semble difficile de détecter toute la diversité présente dans un environnement, sans pour autant produire d'unités de diversité (ESV ou OTU) « suspectes », une solution est d'interpréter cette diversité dans un contexte relatif. Ainsi, une solution alternative pour *étudier la* diversité à partir des amplicons est d'intégrer les OTUs ou ESVs dans des phylogénies afin de caractériser les relations entre ces unités opérationnelles entre elles et avec les espèces connues ou de référence (Taib *et al.* 2013, Yarza *et al.* 2014). Ces approches permettent par ailleurs de prendre en compte la vitesse d'évolution différentielle entre les espèces (Caron *et al.* 2009, Koeppel et Wu 2013, Brown *et al.* 2015). Elles fournissent également une représentation de l'ascendance génétique des unités de diversité comparées et peuvent fournir des indications sur les mécanismes qui façonnent les communautés locales, tels que le filtrage environnemental et la compétition (Webb *et al.* 2002, Cavender-Bares *et al.* 2009). L'intégration d'OTUs au sein de phylogénies permet la détection de clades locaux écologiquement significatifs au sein d'une topologie fiable et la définition d'unités phylogénétiques de diversité (*Operational phylogenetic unit* – OPU ; Rossello-Mora et Lopez-Lopez 2008). Une telle approche a été implémentée dans PANAM (Taib *et al.* 2013), à laquelle a été associé le calcul de divers indices de diversité phylogénétique (PD, NRI, NTI - Faith 1992, Webb *et al.* 2000) et a été recommandée en 2014 pour la définition d'unités taxonomiques candidates (CTU) et engager une réflexion sur la définition des rangs taxonomiques de ces CTU dans un contexte de systématique microbienne (Yarza *et al.* 2014).

4 - Perspectives

Les travaux de *metabarconding* pour la caractérisation des communautés microbiennes de l'environnement ont profité ces quinze dernières années des développements technologiques de séquençage de seconde génération. Celles-ci, de par la profondeur de séquençage qu'elles autorisent, ont permis de façonner l'image d'une diversité jusqu'ici inconnue, au dépend d'une perte de précision du signal permettant la délimitation et l'affiliation des unités de diversité.

Les technologies de séquençage de troisième génération (TGS) permettent de produire des séquences de grande longueur à haut débit, à des profondeurs de séquençage cependant plus modestes. C'est le cas notamment de la technologie PacBio – CSS dont le taux d'erreur est maintenant suffisamment faible pour envisager de l'utiliser dans le contexte d'analyse de diversité (Wagner *et al.* 2016, Heeger *et al.* 2018, Tedersoo *et al.* 2018, Jamy *et al.* 2019, Okasaki *et al.* 2021). Les technologies TGS vont permettre d'obtenir des données de *metabarcoding* avec un signal phylogénétique accru, et une caractérisation plus précise des communautés. Dans ce contexte l'approche implémentée dans PANAM, est tout à fait adaptée et directement utilisable pour l'analyse de séquences complètes de l'ARNr 16S et 18S. Ceci n'empêche cependant pas d'envisager l'amélioration de cet outil, en le rendant notamment plus efficace dans le traitement des séquences de grande longueur. On peut ainsi envisager l'ajout de fonctionnalités (par exemple, améliorer le lien entre les outils de production d'ESV et l'annotation phylogénétique par

PANAM) et le remplacement de certains outils par des outils de nouvelle génération (notamment l'utilisation d'outils HMM optimisés (Zheng *et al.* 2018), ou le remplacement de FastTree par Very Fast-Tree - Piñeiro *et al.* 2020) et l'extension de sa base de référence à une plus large gamme de marqueurs. Ce dernier point mérite cependant une réflexion plus fine, dans la mesure où il s'agit alors d'organiser une information biologique qui nécessite une expertise approfondie. Par ailleurs, alors que l'information biologique qui accompagne les séquences de l'ARNr 16/18S est riche, les données de référence associées aux marqueurs alternatifs (ITS, ARNr 28S) le sont beaucoup moins et sont, dans certains cas, restreintes à certains groupes taxonomiques (Okasaki *et al.* 2021). Enfin les variations de la longueur de la région ITS entre les taxa et la présence d'introns à l'extrémité de l'ARNr 18S chez les eucaryotes risque de complexifier l'exploitation de ces données (Tedersoo et Anslan 2019).
Le développement des technologies TGS offre également l'opportunité d'enrichir et de densifier les bases de données de référence pour ces différents marqueurs. Dès lors, il apparaît primordial de réfléchir à la conception d'un système ou d'une base de référence qui pourra s'enrichir de ces informations. Comme il a été souligné, différents marqueurs (ou différentes régions d'un marqueur) peuvent amener à différentes lectures de la diversité d'un écosystème. Au delà de l'enrichissement de la base de référence de PANAM en séquences pleine longueur, il apparaît important de s'assurer de la cohérence des informations fournies à l'utilisateur. Dans ce contexte, et dans le cadre d'un projet de séquençage TGS (en cours), l'analyse comparative de données issues de séquençage NGS et TGS pour étudier la diversité des eucaryotes unicellulaires lacustres (volet *metabarcoding* du projet MICROSTORE, partie III), nous permettra d'évaluer l'apport relatif de ces différentes technologies en terme de diversité captée (Illumina permet de capter plus de diversité rare), mais également en terme de précision des affiliations produites ou dans la définition et la représentativité des OTUs/ESVs.

Concernant la caractérisation des unités de diversité pertinentes d'un point de vue méthodologique ou écologique, il me semble nécessaire de continuer à travailler sur la définition des unités phylogénétiques de diversité. En effet, bien que permettant de regrouper des OTUs minoritaires avec les OTUs plus abondants dont ils sont issus du fait d'erreurs de séquençage, la définition des limites de ces unités n'est pas totalement résolue. S'il est possible d'utiliser des critères statistiques, comme par exemple le *bootstrap* pour en désigner les limites, rien ne garantit qu'un OTU divergent « réponde de manière satisfaisante » à un tel critère. Par ailleurs, le positionnement des OTUs au sein d'une phylogénie ne résout pas la question de la paraphylie des écotypes constatée par Koeppel et Wu lorsque les séquences sont préalablement clusterisées en OTUs, alors même que ceux-ci s'imposent pour réduire la taille des données à traiter. Une piste pour avancer sur ces questions pourrait consister à travailler les unités de diversité phylogénétique en sélectionnant des groupes monophylétiques larges pour définir des unités de diversité grossières qui seraient ensuite réévaluées individuellement. On pourrait alors considérer l'ensemble des types de séquences incluses dans chaque groupe monophylétique et chercher à produire le meilleur partitionnement de ces séquences. L'exploitation des informations phylogénétiques propres aux séquences types, associées aux informations d'abondance, de co-occurence ou l'exploration des propriétés des graphes produits par l'analyse de ces séquences types pourrait ainsi offrir une alternative « éclairée » aux approches de clusterisation gloutonnes.

PARTIE II

APPROCHE CELLULE UNIQUE

DYNAMIQUE EVOLUTIVE DU PANGENOME A L'ÉCHELLE D'UNE POPULATION NATURELLE

Préambule

Les développements méthodologiques présentés ci-avant et en particulier la définition d'OTUs sur des jeux de données issus de séquençage de deuxième génération m'a conduite, comme d'autres, à m'interroger sur la notion d'unité taxonomique chez les micro-organismes. Dans les faits, la définition d'un OTU basé sur un seuil de similarité entre les séquences du gène de l'ARNr (16S et 18S) ou de l'ITS (pour les *Fungi* ou certains groupes bactériens comme par exemple le genre *Prochlorococcus*), bien que couramment utilisé, peut conduire à la génération d'unités taxonomiques dont les contours peuvent varier en fonction des méthodes de clusterisation utilisées (He *et al.* 2015). De plus, ces marqueurs ont la particularité d'évoluer, soit plus lentement (ARNr), soit plus rapidement (ITS) que le reste du génome (Huse *et al.* 2008, Liu *et al.* 2008). Ceux-ci ne sont donc pas nécessairement représentatifs de la divergence des espèces à l'échelle des génomes entiers (Daubin *et al.* 2002). Le fait de cibler des régions spécifiques de ces marqueurs peut également avoir un impact sur la taxonomie restituée par ces approches (Taib *et al.* 2013, Mysara *et al.* 2017). Les unités taxonomiques obtenues à l'issu d'une analyse de *metabarcoding* ne reflètent par ailleurs pas nécessairement des unités biologiques cohérentes soutenues par des mécanismes de spéciation clairs, en particulier pour représenter des unités de diversité spécifique de niches écologiques particulières. Alors que l'analyse d'amplicons sur la base de modèles d'évolution écotypique produit des groupes monophylétiques correspondant à des écotypes avérés (Koeppel *et al.* 2008), les OTUs construits à partir des mêmes jeux de données regroupent des séquences associées à des écotypes différents (Koeppel et Wu 2013). Ces résultats sont retrouvés pour des seuils de clusterisation allant de 97 à à 99 % d'identité.

Au delà de ces limites, l'éclairage fonctionnel apporté par les approches de *metabarcoding* n'est qu'indirect dans la mesure où l'inférence d'un rôle fonctionnel pour une unité taxonomique donnée est conditionnée par sa proximité phylogénétique avec un groupe taxonomique caractérisé. Cette caractérisation indirecte peut cependant apparaître limitante, en particulier si aucun groupe taxonomique proche des unités taxonomiques étudiées n'a de rôle fonctionnel connu.

Alors que le *metabarcoding* permet une caractérisation taxonomique des communautés par l'analyse d'un gène marqueur universel, la métagénomique permet une caractérisation fonctionnelle des communautés grâce au séquençage de l'ensemble des génomes présents dans un échantillon environnemental (Handelsman *et al.* 1998, Rondon *et al.* 2000, Rodriguez-Valera 2002, Venter *et al.* 2004, Debroas *et al.* , 2009). La métagénomique peut être complétée par le séquençage de génomes issus de cellules uniques (SAG) isolées de l'environnement. Les SAGs permettent alors non seulement de caractériser le potentiel fonctionnel des organismes, mais également de comprendre ce qui se passe à l'échelle de la cellule en ciblant son transcriptome (Stuart et Satija 2019, Liu *et al.* 2019, Kuchina *et al.* 2021). Ainsi, la génomique environnementale permet à travers la métagénomique et l'acquisition de SAGs, d'étudier la diversité taxonomique et fonctionnelle des communautés naturelles, depuis l'échelle individuelle jusqu'à la communauté, en passant par l'étude des populations (Kashtan *et al.* 2014, Rinke *et al.* 2013). Le séquençage de cellules uniques permet par ailleurs d'accéder à la diversité génomique de populations libres de l'environnement et d'apporter un éclairage populationnel et évolutif nouveau, *in situ*, complémentaire aux travaux développés sur les espèces modèles ou sur les souches classiquement étudiées en laboratoire.

Les approches de génomique environnementale ont permis de confirmer que la plupart des espèces bactériennes ne sont pas clonales (Venter *et al.* 2004, Vergin *et al.* 2007, Rosen *et al.* 2015) et se caractérisent par une diversité de contenu en gènes au sein des génomes d'une même « espèce microbienne » (Coleman *et al.* 2006, Bhaya *et al.* 2007, Biller *et al.* 2014), associée à une compartimentation spatiale de ceux-ci. Ces observations ne sont pas sans conséquences sur notre compréhension du fonctionnement et de la dynamique des communautés microbiennes et impliquent de comprendre les processus associés à la mise en place et à l'évolution des gènes qui sous-tendent les fonctions biologiques dans les écosystèmes. Pour cela, il est nécessaire :

- De caractériser les unités de diversité pertinentes dans l'écosystème et de comprendre comment elles se constituent et se maintiennent ;
- D'étudier, à l'échelle de ces unités de diversité, les facteurs évolutifs qui impactent l'organisation, l'expression et la dynamique évolutive des génomes microbiens.

A l'échelle des micro-organismes, les temps de génération sont courts, de sorte que les fréquences alléliques dans les populations peuvent changer suffisamment rapidement pour affecter des interactions écologiques, ces deux processus évoluant sur une échelle de temps similaire (Messer *et al. 2016*, Good *et al.* 2017). Le fait, que chez les micro-organismes, les processus écologiques (changement de l'abondance des individus dans le temps) dans les communautés microbiennes se superposent aux processus évolutifs (changement de la fréquence des gènes dans le temps) (Shapiro 2018) a conduit de nombreux auteurs à argumenter que la génomique des populations microbiennes ne peut être séparée de leur écologie. Les interactions entre les processus écologiques, populationnels et évolutifs dans les populations naturelles restent cependant très mal connues.

Le travail de recherche que je présente dans cette deuxième partie se positionne sur ces questions émergentes en génomique évolutive et environnementale. Il concerne plus précisément l'étude des processus évolutifs qui gouvernent la dynamique des génomes et des unités de diversité microbiennes – vus à l'échelle de différentes populations appartenant à un même genre microbien – dans les communautés aquatiques naturelles. L'approche utilisée est basée sur l'exploitation bio-informatique de génomes issus du séquençage de cellules uniques (SAGs) rendus disponibles à la communauté scientifique. Le modèle d'étude considéré ici est l'écotype HLII du genre *Prochlorococcus*. Dans la suite du manuscrit, je présenterai d'abord le concept de pangénome. Sera ensuite présentée l'étude réalisée par Hélène Gardon (doctorante dont j'ai en charge l'encadrement) sur des populations de l'écotype HLII des BATS (*Bermuda Atlantic Time-series Study*), station de suivi océanographique situé au large des Bermudes dans la mer des Sargasses (http://bats.bios.edu/about/). Enfin, ce chapitre sera clôturé par les perspectives envisagées à partir des résultats obtenus au cours de ce travail.

1 - Le concept de pangénome

C'est au début des années 2000, sur la base de travaux de génomique comparative, que l'idée d'une relation 1:1 entre une espèce bactérienne et son génome a été remise en cause (Perna *et al.* 2001, Welch *et al.* 2002, Tettelin *et al.* 2005).

1.1 - De l'espèce phénotypique au pangénome

Au cours d'une étude pour concevoir un vaccin contre l'infection des nouveaux nés par la bactérie *Streptococcus agalactia* (ou *Streptococcus* du groupe B – GBS), Tettelin et collaborateurs ont constaté que seuls 80 % des gènes identifiés étaient partagés par le génome des huit isolats GBS étudiés. Par ailleurs, une part importante des gènes qui n'étaient pas partagés par tous les génomes se concentrait dans 14 régions génomiques et se caractérisait par une forte densité en gènes de virulence et en éléments génétiques mobiles (Tettelin *et al.* 2005, Medini *et al.* 2020). Après modélisation du nombre total de gènes associés au GBS, ils ont suggéré que la prise en compte de tout nouveau génome conduit à l'identification de nouveaux gènes d'où l'idée que la taille du potentiel génétique de GBS était infinie. Sur la base de ce constat, Tettelin et collaborateurs ont développé le concept de pangénome, défini par trois composantes :

- un génome *core* (ou génome central), correspondant aux gènes retrouvés dans tous les isolats de l'espèce considérée ;
- un génome accessoire (dispensable), qui fait référence aux gènes présents dans plusieurs mais pas tous les isolats ;
- les gènes spécifiques à une souche, échantillonnés dans un seul isolat.

Celles-ci ont par la suite été redéfinies sous les noms de génome *core*, génome *shell* et génome *cloud* (Koonin et Wolf 2008).

Si le concept de pangénome est adapté à la définition du potentiel fonctionnel d'une espèce bactérienne, la structure, la taille et les caractéristiques de celui-ci peuvent être très différentes en fonction des espèces (McInerney *et al.* 2017a; Medini *et al.* 2020). Il est ainsi possible de distinguer les pangénomes dit *fermés* et les pangénomes *ouverts* (Figure 2.1). Alors que les premiers sont reconnaissables à leur génome *core* important associé à un génome flexible réduit et relativement peu variable, le pangénome *ouvert* est caractérisé par sa grande taille et un génome flexible en « expansion », dans la mesure où celui-ci s'accroît avec la prise en compte d'un nombre croissant de génomes (McInerney *et al.* 2017a). La nature (ouverte ou fermée) du pangénome aura un impact sur la structure des génomes des espèces. Ainsi, les espèces avec un pangénome fermé seront caractérisées par des génomes dont la structure est relativement peu diversifiée, alors que les espèces avec un pangénome ouvert afficheront une grande diversité dans la structure et la composition de leur génome.

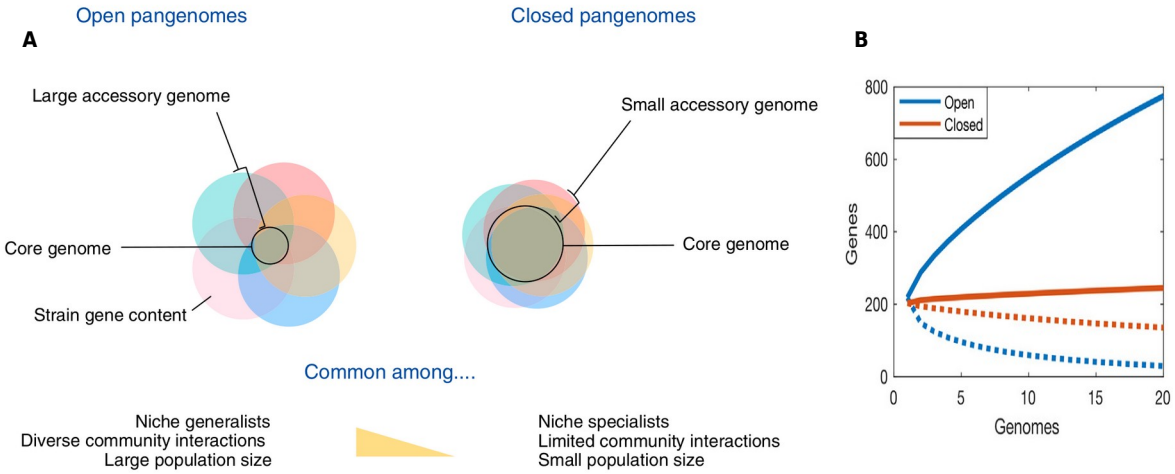


Figure 2.1 : Le concept du Pangénome. (A) Pangénome ouvert *versus* fermé (d'après Brockhurst *et al.* 2019). (B) Courbes d'accumulation de l'ensemble des gènes (lignes pleines) en fonction que le pangénome soit ouvert (bleu) ou fermé (orange). Les lignes en pointillées représentent les courbes d'appauvrissement en gènes *core* en fonction du pangénome. Ainsi pour un pangénome ouvert (bleu), le nombre de gènes *core* diminue rapidement avec l'ajout de génomes, tandis que sa taille augmente linéairement (d'après Domingo-Sananes & McInerney, 2021).

1.2 – Pangénome, transfert de gènes et unité de diversité

Le corollaire à la découverte du pangénome est la démonstration de l'ampleur du phénomène de transferts horizontaux de gènes (HGT) dans le monde microbien, c'est-à-dire la diffusion de matériel génétique entre organismes, y compris entre organismes n'appartenant pas à la même espèce, voire au même règne (Struhl *et al.* 1976, Syvanen 1985, Médigue *et al.* 1991, Doolittle 1999, Ochman *et al.* 2000 – Figure 2.2). L'acquisition ou le remplacement de gènes par HGT et la perte de gènes sont maintenant considérés comme des processus évolutifs-clé chez les procaryotes, plus encore que les mutations ponctuelles (Koonin *et al.* 2021).

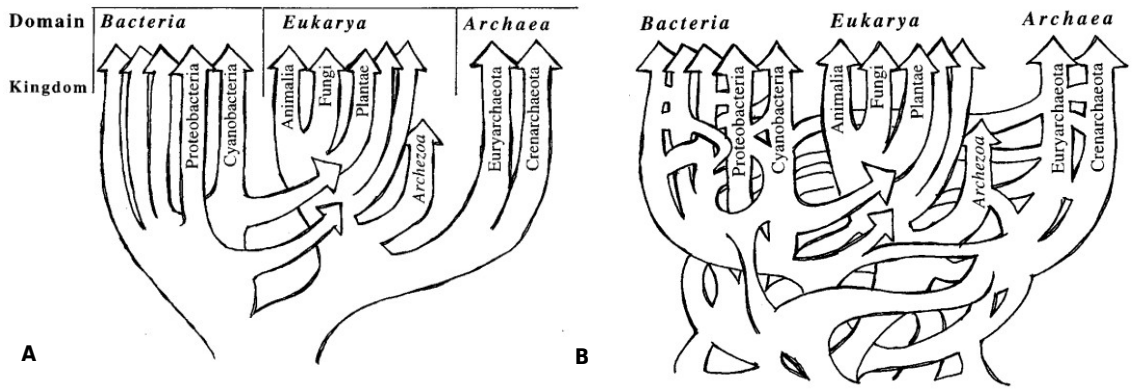


Figure 2.2 : Phylogénies universelles (A) Arbre ramifié illustrant une transmission verticale de l'information génétique (les endosymbioses à l'origine de la mitochondrie et du chloroplaste sont représentées par les flèches traversantes depuis le règne des bactéries vers celui des eucaryotes). (B) arbre réticulé pour la représentation des relations d'évolutions incluant l'idée de transferts horizontaux de gènes (d'après Doolittle 1999).

L'existence de pangénomes et de HGT à une grande échelle de diversité a également un impact sur la définition de la notion d'espèce microbienne. En effet, s'il existe des échanges de matériel génétique au-delà d'une population circonscrite par l'émergence d'une barrière aux flux de gènes, alors la notion d'espèce biologique ne peut s'appliquer aux micro-organismes. La notion d'espèce, vue comme une lignée évolutive résultant d'une transmission verticale du matériel génétique, ne tient pas mieux dans ce contexte (Doolittle 1999). Les phylogénies microbiennes doivent alors être considérées comme « une tendance statistique au sein d'une forêt d'arbres de gènes individuels décrivant des histoires évolutives distinctes » (Koonin *et al.* 2021). Ainsi, la question de la définition d'une unité fondamentale de diversité microbienne ne peut être considérée indépendamment de la question de comprendre les modalités de l'émergence, de la dynamique, de la structuration et du maintien des pangénomes.

1.3 – Dimension écologique du pangénome

Si d'un point de vue mécanique, les échanges de matériel génétique sont possibles (il existe de nombreux mécanismes d'acquisition ou de facilitation du transfert de matériel génétique mais n'ont pas lieu d'être décrits ici - voir Soucy *et al.* 2015 par exemple), la probabilité de ces événements dépend aussi de facteurs populationnels liés à la probabilité de rencontre entre individus donneurs et receveurs. Ces échanges ont en retour un impact sur le rôle fonctionnel des individus porteurs des gènes échangés au sein de l'écosystème. Un tel lien entre diversité génétique et rôle fonctionnel est bien décrit pour les souches de bactéries pathogènes. Par exemple, les variations dans la distribution des régions variables observées chez *Escherichia. coli* font partie d'un schéma plus large de variation phénotypique à l'origine de la nomenclature des différents pathotypes de cette espèce (Dobrindt *et al.* 2004). Toutefois, cette similitude à l'échelle des génomes n'est pas systématiquement corrélée et le lien entre génome et phénotype est certainement plus complexe (Kumar *et al.* 2015, Medini *et al.* 2020).

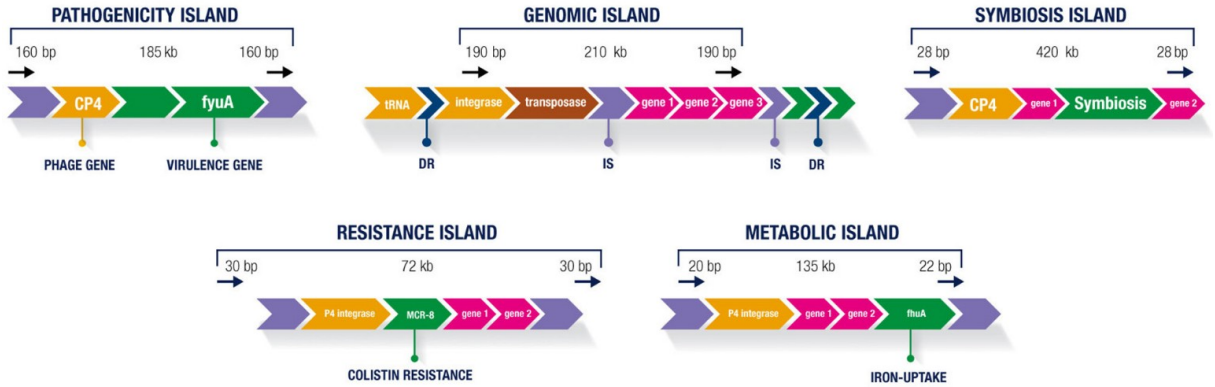
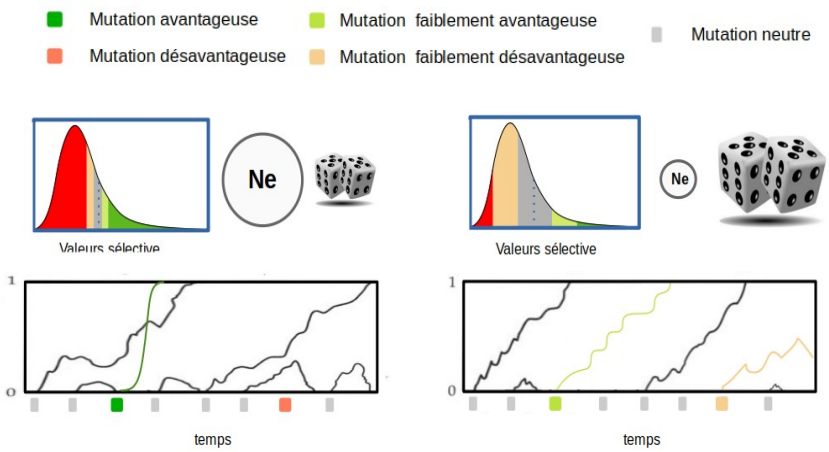


Figure 2.3 : Présentation des types majeurs d'îlots génomiques observés chez les bactéries et leurs principales caractéristiques. En jaune : gènes liés à l'intégration d'ADN ; en mauve : régions cibles d'insertion d'éléments mobiles ; en vert : gènes transférés porteurs d'une « fonction d'intérêt » ; en rose : autres gènes . DR : répétions directes, IS = site d'insertion (d'après da Silva Filho *et al.* 2018).

De nombreux gènes accessoires des pangénomes ont été décrits comme ayant des fonctions écologiques. Ceux-ci sont souvent concentrés dans des régions génomiques variables appelées îlots génomiques (Figure 2.3). C'est le cas des gènes de résistance aux antibiotiques, des gènes de virulence (Ochman *et al.* 1996, Gal-Mor *et al.* 2006), souvent associés à des éléments génétiques mobiles (Partridge *et al.* 2018, Cordero et Polz 2014), mais aussi des gènes adaptatifs d'intérêt écologique. Il a ainsi été montré une grande diversité génétique au sein des groupes bactériens majeurs dans les océans (*Prochlorococcus*, *Pelagibacter, Vibrio*), associée à une distribution écologique différenciée et une organisation de la diversité génétique en îlots, dont certains semblent associés à une différenciation écologique (assimilation du phosphore – Coleman et Chisholm 2010 ; résistance aux infections virales – Avrani *et al.* 2011). C'est par exemple le cas de populations de *Vibrio*, caractérisées par des régions génomiques affichant un polymorphisme nucléotidique (« éco-SNP ») statistiquement liées à la nature du substrat sur lesquelles vivaient ces populations, suggérant que ces régions étaient impliquées dans une adaptation au substrat (Shapiro *et al.* 2012).

La dimension écologique semble être également un déterminant majeur au fait de partager des gènes. Il a ainsi été montré un lien entre la fréquence des HGT et le fait de partager les mêmes caractères écologiques (gradient d'oxygène ou relations symbiotiques - Smillie *et al.* 2011), suggérant que les transferts de gènes sont probablement conditionnés (et possiblement sélectionnés) par le fait de partager les mêmes « préférences » écologiques. Par conséquent, la taille et la diversité du pool génétique, à partir duquel se dessine le génome d'une espèce, devraient dépendre de la diversité de la communauté dans lequel cette espèce évolue (Hooper *et al.* 2009).

BOX 5 – Sélection, mutation, dérive et taille efficace des populations



La mutation est la source de la variation individuelle. Lorsqu'elle n'impacte pas les individus dans la production de leur descendance, elle est qualifiée de neutre, dans le cas contraire, elle est soumise à une processus de sélection. La sélection correspond à un différentiel reproductif entre individus génétiquement différents, elle est exprimée en terme de valeur sélective (ou *fitness*) quantifié à travers un coefficient de sélection. Une sélection positive traduit l'idée d'un différentiel reproductif favorable à l'individu, une sélection négative correspond à un déficit de reproduction d'un individu par rapport au reste de la population. Une mutation avec un coefficient de sélection positif sera qualifiée d'avantageuse. Une mutation avec un coefficient de sélection négatif sera désavantageuse. Dans des populations à grand effectif (grande taille efficace : *Ne* grand – à gauche), la mutation « pourra exprimer pleinement » sa valeur sélective et la distribution de celle-ci est déséquilibrée en faveur des mutations neutres à désavantageuses. Dans les populations de petite taille (*Ne* petit – à droite), la dérive génétique induit une modification aléatoire de la fréquence des mutations d'une génération à l'autre. Elle a également pour effet de « réduire le potentiel » des valeurs sélectives des mutations. Une mutation avantageuse dans une grande population sera moins avantageuse dans une petite population. La théorie neutre de l'évolution prédit que :

- le devenir d'une mutation est gouvernée par son coefficient de sélection lorsque celui-ci surpasse l'intensité de la dérive (s>>1/Ne) ;
- les mutations avantageuses se fixent plus rapidement dans les populations de grande taille alors que les mutations neutres se fixent plus vite dans les populations de petite taille :
- la diversité génétique neutre est proportionnelle à la taille efficace des populations.

2 – Controverse sur l'origine du pangénome

2.1 – Pourquoi un pangénome ?

Les successions ou les modifications du matériel génétique *via* des mutations adaptatives ou des HGT peuvent conférer un avantage adaptatif. Il apparaît cependant que les taux de gain et perte de gènes sont supérieurs aux taux de substitution par site chez de nombreuses espèces bactériennes (Hao et Golding 2006, Marri *et al*. 2006, Touchon *et al*. 2009, Nowell *et al*. 2014). Par ailleurs, ces HGT ont tendance à avoir, en moyenne, un impact plus important en termes de fitness comparativement à une substitution (que ce soit en termes d'augmentation ou de diminution de fitness - Vos et Eyre-Walker 2015). Si un nombre conséquent d'exemples de HGT adaptatifs peut être avancé, il semble cependant que la plupart des changements, en termes de contenu en gènes, ont une durée de vie courte (Hao et Golding 2010, Didelot *et al.* 2012). En effet, le nombre d'événements de HGT détecté ne décroît pas

linéairement avec l'augmentation des temps de divergence des taxa considérés, à l'exception des gènes peu conservés et peu contraints (Bolotin et Hershberg 2016). Ceci suggère que le pool de gènes accessoires n'est pas purement adaptatif, la plupart des HGT pouvant être soit délétères, soit présenter un avantage sélectif transitoire (c'est-à-dire présenter un avantage à un moment, puis le perdre suite à un glissement de conditions environnementales), soit neutres.
La question du caractère adaptatif ou non des gènes constituant le génome accessoire au sein du pangénome (et par extension le caractère adaptatif des HGT) est centrale pour comprendre l'origine et le maintien des pangénomes. En effet, selon que l'on considère les gènes accessoires comme essentiellement adaptatifs ou non, les mécanismes à même d'expliquer le maintien du pangénome sont différents.

Ainsi, McInerney *et al.* (2017a) font l'hypothèse que le pangénome résulte de l'effet combiné de la taille efficace des populations (Ne) et de la migration des individus vers de nouvelles niches. Ces auteurs avancent que dans le pangénome, les séquences doivent être neutres pour être modelées par la dérive et que celles-ci seraient purgées par le biais naturel à la délétion des séquences dans les génomes bactériens (Kuo et Ochman 2009). Les travaux de Bolotin et Hershberg (2016) suggèrent cependant que le glissement des conditions de l'environnement n'est pas nécessaire pour expliquer la majorité des pertes de gènes observées dans les génomes, en particulier pour les gènes peu contraints.
A contrario, Vos et collaborateurs (Vos *et al.* 2015) suggèrent, conformément à la théorie neutre de l'évolution moléculaire (Kimura 1983), que les populations de grande taille doivent porter un nombre plus important d'allèles quasi neutres (BOX 5) et argumentent qu'un modèle neutre d'évolution du pangénome, gouverné essentiellement par du processus de dérive génétique, est plus parcimonieux que le modèle proposé par McInerney. Andreani et collaborateurs n'ont cependant pas trouvé de lien entre la taille des populations microbiennes et la teneur des génomes accessoires en séquences quasi-neutres (Andreani *et al.* 2017).
En analysant l'impact de la relation entre dérive génétique et taille des populations sur le répertoire complet de gènes de différentes espèces bactériennes, Bobay et collaborateurs montrent que la taille du pangénome,plus que celle du génome, est affectée par l'efficacité de la sélection (Bobay et Ochman 2018). Ils avancent l'hypothèse de la barrière à la dérive comme mécanisme de régulation de la dynamique du pangénome. Cette hypothèse, initialement proposée pour expliquer comment l'augmentation de la dérive génétique rend inefficace la sélection pour moduler les taux de mutation (Sung *et al.* 2012), prédit qu'une petite taille efficace de population augmente la perte aléatoire de gènes accessoires et que les espèces à faible taille efficace de population ne conserveraient que les gènes accessoires (les plus) bénéfiques. Les gènes avec une fitness moins évidente pourront, quant à eux, être conservés dans des espèces pour lesquelles la taille efficace des populations est grande.

2.2 – Une organisation génomique en îlots

Alors qu'ils sont caractérisés par une grande diversité compositionnelle, du fait de leur part flexible, à même de refléter une grande plasticité fonctionnelle et structurale, les génomes microbiens affichent des organisations spatiale et fonctionnelle qui ne semblent pas aléatoires. Ainsi, les gènes au sein des génomes bactériens et archéens sont organisés en opérons, c'est-à-dire en unités de transcription

contenant un ensemble de gènes co-régulés appartenant à un même processus métabolique ou de signalisation (Jacob *et al.* 1960). Chez les bactéries, la répartition des gènes le long du génome semble par ailleurs dépendre de leur distance par rapport à l'origine de réplication et du brin sur lequel ils se trouvent (Rocha 2008). Alors que la majorité des gènes de ménage sont dans la partie proximale de l'origine de réplication, les gènes impliqués dans les stress sont localisés de manière prépondérante près du terminus (Krogh *et al.* 2018). De même, chez *E. coli*, le positionnement des gènes codant pour les protéines NAP (qui modulent la structure du nucléoïde) le long du chromosome suggère un motif spatio-temporel de leur expression durant le cycle de croissance bactérienne (Sobetzko *et al.* 2012). Enfin, il existe une compartimentation spatiale du pangénome microbien avec la distinction d'un génome squelette enrichi en gènes *core*, entrecoupé d'îlots génomiques enrichis en gènes flexibles. Alors que les îlots génomiques sont souvent associés à des événements de HGT, il apparaît que moins de 2 % d'entre eux accumulent plus de 50 % des HGT. Ainsi, l'intégration des HGT pourrait dépendre de leur localisation, mais également de leur fonction et de leur niveau d'expression (Oliveira *et al.* 2017).

2.3 – Etudier le pangénome à partir de populations naturelles

Comme nous venons de le voir, les modèles et hypothèses proposés pour expliquer la dynamique du pangénome et les causes sous-jacentes à cette dynamique sont controversés. Ceci est notamment entretenu par des biais dans l'échantillonnage des groupes taxonomiques représentés dans les études de génomique comparative et par le choix des échelles de diversité considérées pour évaluer les différentes hypothèses : « La comparaison d'un petit nombre de génomes bactériens échantillonnés dans de nombreuses niches est susceptible de produire une abondance de gènes accessoires rares, mais ceux-ci pourraient représenter soit des gènes accessoires adaptatifs qui sont localement abondants mais globalement rares, soit des gènes accessoires délétères qui sont à la fois localement et globalement rares » (Brockhurst *et al.* 2019).

Contrairement à l'analyse des pangénomes à partir de la comparaison de génomes issus d'isolats bactériens ou de souches maintenues en laboratoire, la génomique comparative appliquée à des génomes issus d'une même population (différents individus d'une même espèce échantillonnés en même temps au même endroit) doit permettre d'identifier les éléments clés de la dynamique des génomes au sein des populations. En effet, la comparaison de génomes individuels échantillonnés *in situ*, à partir de populations naturelles, peut permettre d'observer « un instantané » des gènes partagés ou non au sein du génome accessoire de la population alors même que ceux-ci ne sont pas totalement purgés par les processus de sélection. Ce type d'analyse en complément de la comparaison de génomes d'isolats, permet d'évaluer à des échelles de temps contrastées les processus populationnels et évolutifs qui influent l'occurrence et le devenir des HGT. Si l'on considère par ailleurs des échantillons environnementaux, plutôt que la comparaison de génomes maintenus en culture, il est alors envisageable d'étudier le lien entre les facteurs gouvernant la dynamique des pangénomes, les caractères écologiques des populations considérées et le métagénome de la communauté dans laquelle ces populations évoluent.

Afin de mieux comprendre les fondements évolutifs de la différenciation des populations bactériennes dans l'environnement en lien avec la dynamique de leur pangénome à l'échelle populationnelle, une caractérisation de la dynamique évolutive d'un pangénome a été réalisée en prenant comme modèle d'étude les sous-populations cooccurrentes de l'écotype HLII de *Prochlorococcus,* isolées de l'océan Atlantique (site BATS) telles qu'elles ont été proposées par Kashtan et collaborateurs (Kashtan *et al.* 2014*)*.

Ce travail a été réalisé dans le cadre de la thèse d'Hélène Gardon. Les objectifs de ces travaux étaient de comparer ces différentes sous-populations au niveau de :

- la structuration de leurs génomes et de l'organisation des génomes *core* et flexible :
- la composition de leur génome flexible, incluant une comparaison en termes d'occurrence (présence / absence), de potentiel fonctionnel et d'origine taxonomique ;
- leurs signatures évolutives en fonction des compartiments génomiques identifiés.

3 - Analyse pangénomique d'une population bactérienne environnementale

Grâce au développement des approches de génomique environnementale, nous disposons maintenant de données génomiques pour des espèces majeures de l'environnement comme, par exemple, *Prochlorococcus marinus*, une des espèces photosynthétiques les plus abondantes de la zone euphotique océanique (c'est-à-dire la zone de pénétration de la lumière), puisqu'elle est responsable de près de 10% de la production primaire marine (Partensky *et al.* 1999, Flombaum *et al.* 2013).

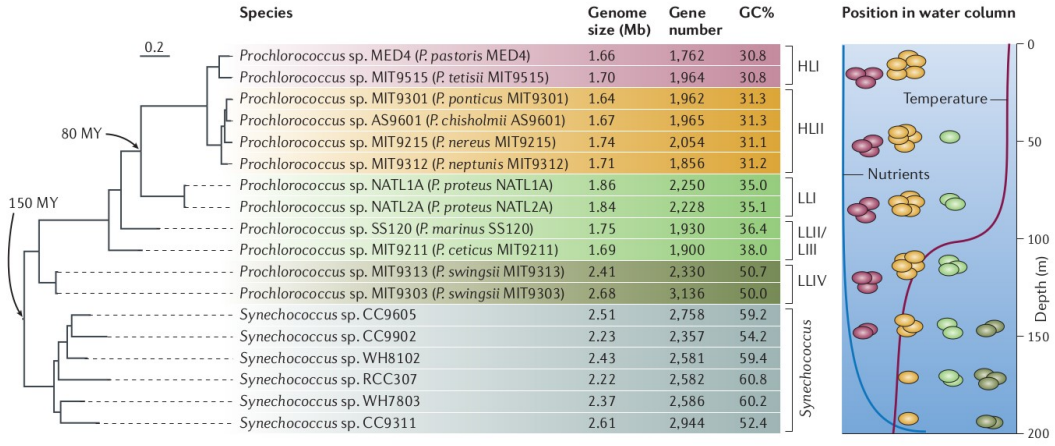


Figure 2.4 : Phylogénie, statistiques sur les génomes et préférences écologiques des écotypes de *Prochlorococcus* (from : Batut *et al.* 2014).

Bien qu'elle soit caractérisé par une identité du gène de l'ARNr 16S supérieure à 96 %, cette espèce présente une diversité génétique importante, avec une identité nucléotidique inférieure à 70 % à l'échelle du génome (Zhaxybaveva *et al.* 2009), associée à un pangénome ouvert (Kettler *et al.* 2007). Cette diversité intra-spécifique est structurée en deux écotypes majeurs associés à une adaptation à la lumière (HL : haute lumière et LL : basse lumière ; Moore *et al.* 1998). Ces écotypes sont eux-mêmes organisés en écotypes plus fins (Figure 2.4), répondant différemment aux conditions environnementales, notamment à la température, la disponibilité en nutriments ou la concentration en fer dans le milieu (Moore *et al.* 1998, Rocap *et al.* 2003, Malmstrom *et al.* 2010, Berude *et al.* 2014). La distribution

géographique de ces écotypes à l'échelle de la planète et le long de la colonne d'eau suggère une répartition stable de ces groupes, écologiquement distincts en niches (Kent *et al.* 2016, Larkin *et al.* 2016).

Sur la base d'une approche de séquençage de cellule unique à grande échelle, il a récemment été proposé que les populations de *Prochlorococcus* sont composées de centaines de sous-populations (Kashtan *et al.* 2014, 2017). Une diversité de séquences à l'échelle plus fine de l'écotype HLII a été mise en évidence pour ces sous-populations coexistantes avec notamment la caractérisation de compartiments génomiques incluant un génome squelette (*backbone*), principalement composé de gènes *core* - également conservés à l'échelle du genre *Prochlorococcus* - et des îlots génomiques (ISLs) principalement composés de gènes flexibles. Cette structuration de la diversité de l'écotype HLII en sous-populations pourrait refléter une spécialisation de niche. Ceci est étayée par i) la prédominance d'allèles au niveau des gènes *core*, fixés au sein des sous-populations mais différents entre sous-populations et ii) d'ensembles distincts de gènes flexibles entre les différentes populations.

3.1 – Quelle représentativité d'un pangénome populationnel ?

Nous avons analysé 87 SAGs de l'écotype HLII au sein du genre *Prochlorococcus*. Ces SAGs sont répartis au sein de trois groupes (cN2, c9301 et cN1), tels que définis par l'analyse phylogénétique du marqueur ITS (Kashtan *et al.* 2014). Sur la base d'une phylogénie génomique, les auteurs ont montré que ces SAGs étaient par ailleurs répartis en sept sous-populations (C1 à C5, C8, C9), également appelées clades dans le présent document. L'ensemble des gènes décrits sur les SAGs ont été organisés en groupes de gènes orthologues (REF) ou COGs. Le pangénome associé aux SAGs analysés est constitué de 7 125 COGs. Au total, 1 410 COGs sont retrouvés dans les 13 génomes de souches cultivées de *Prochlorococcus,* écotype HLII et sont associés au génome *core*, conformément à la classification définie dans Kashtan *et al.* 2014. Quatre-vingt-trois pour cent des COGs restants sont présents dans au moins 10 SAGs, et 70 % dans au moins 50 SAGs. L'accroissement du nombre de COGs, avec le nombre de SAGs considérés est caractéristique d'un pangénome ouvert (Figure 2.5 A), en adéquation avec la nature de celui de *Prochlorococcus,* tel que décrit à des échelles de diversité plus large (Kettler *et al.* 2007).

La complétude des SAGs étudiés varie de 8,6 à 97,4 % pour des longueurs d'assemblage variant de 0,37 à 2,62 Mb. Les segments communs à l'ensemble des populations, chacune représentée par leur SAG le plus complet, totalisent 1,33 Mb (de 1,35 Mb pour C1 à 1,47 Mb pour C3) soit 76 % de la taille du génome de MIT9312 utilisé ici comme génome de référence. Le nombre de COGs associés à chaque SAG dépend de la complétude de ces derniers, ce qui impacte la définition du pangénome à l'échelle des sous-populations. L'absence de relation entre la complétude des SAGs et le nombre de COGs spécifiques à une sous-population suggère par ailleurs une grande variabilité de la part flexible des génomes séquencés. Cette absence de corrélation ne permet pas de quantifier le biais induit par le séquençage partiel de certains SAGs (Figure 2.5 B).

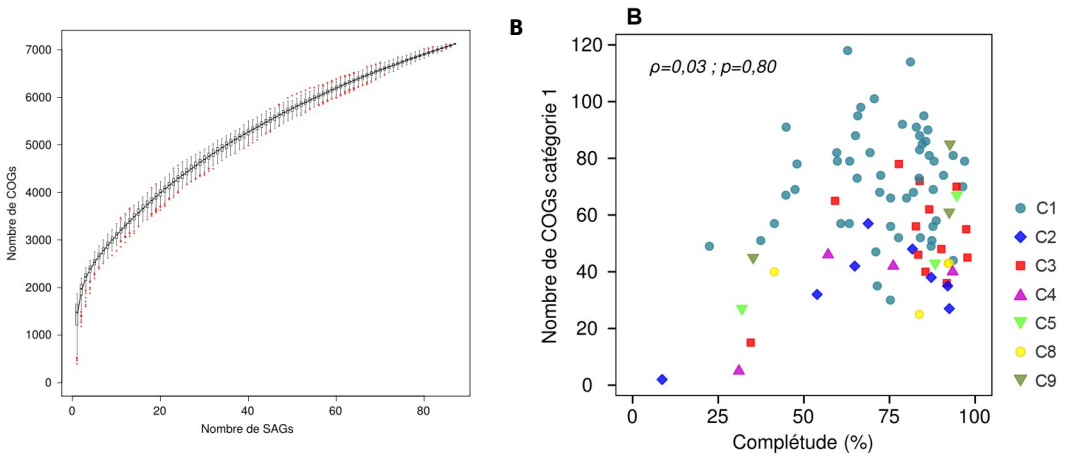


Figure 2.5 : Pangénome et complétude (A) Évolution du nombre de COGs flexibles spécifiques des SAGs en fonction du nombre de SAGs. La distribution du nombre de COGs pour *k* génomes est représentée par un diagramme en boîte et les *outliers* par des points.; (B) : Corrélation entre complétude et COGs flexibles spécifiques d'une sous-population. Chaque point correspond à un SAG. Les SAGs d'une sous-population donnée sont représentés par une couleur et une forme qui leurs sont propres. Les résultats des tests de corrélation de Spearman (*ρ*) et les *p-value* associées sont indiqués.

Une analyse du pangénome des sept sous-populations, restreinte aux seuls SAGs présentant une complétude supérieure à 90 % (soit 18 SAGs), réduit de 38 % le nombre de COGs associés au génome flexible, avec pour conséquence une perte d'information concernant la distribution des COGs flexibles au sein des sous-populations (Figure 2.6), mais également une réduction de 20 % du nombre de COGs associés à une unique sous-population (passant de 66 à 47 % des COGs).

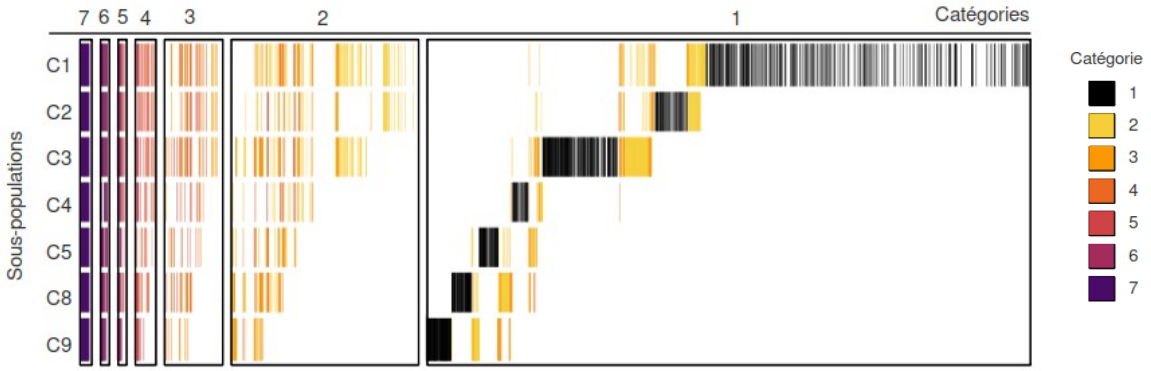


Figure 2.6 : Distribution des COGs SAG-spécifiques pour les 18 SAGs quasi-complets (3 286 au total) en fonction des sous-populations (C1 à C5, C8 et C9) et de leur occurrence dans une à sept populations (catégories 1 à 7, *i.e.,* COGs spécifiques d'une sous-population (1) à droite aux COGs communs à toutes les sous-populations (7) à gauche). Une couleur est par ailleurs attribuée à chaque COG en fonction de son occurrence au sein des sous-populations dans le jeu de données incluant les 87 SAGs (catégories 1 à 7, *i.e.,* COGs spécifiques d'une sous-population en noir à communs à toutes en violet). Chaque barre verticale correspond à un COG. Les catégories, exception faite de la catégorie 7, sont caractérisées par plusieurs couleurs du fait de la réduction du jeu de données et le changement de catégorie de nombreux COGs par rapport au jeu de données complet.

Ainsi, bien que ne permettant pas de fournir une description exhaustive du pangénome des sous-populations étudiées, il est essentiel de conserver l'information portée par l'ensemble des SAGs pour l'étude des pangénomes, notamment pour estimer au mieux la part des COGs partagés entre sous-populations.

3.2 – Organisation du pangénome des sous-populations de *Prochlorococcus marinus* HLII

Sur la base d'une phylogénie génomique, Kashtan et collaborateurs (2014) ont montré que les SAGs étaient répartis en sept sous-populations notées de C1 à C5, C8 et C9. La phylogénie basée sur l'alignement de 1 202 gènes *core*, incluant le génome de la souche cultivée MIT9312 comme groupe externe, permet de retrouver la délimitation des clades / sous-populations, ainsi que le regroupement des clades C8 et C3 (Figure 2.7). Ces résultats suggèrent que le groupe cN2 tel que défini à partir des ITS ne serait pas monophylétique. La même démarcation des sous-populations est également retrouvée à travers l'analyse de l'identité nucléotidique moyenne (ANI) à l'échelle du génome (ANI inter-clade : 97 % en moyenne sauf entre C1 et C2 ; ANI intra-clade > 98 % sauf pour C8 : 97 % et C9 : 96%).

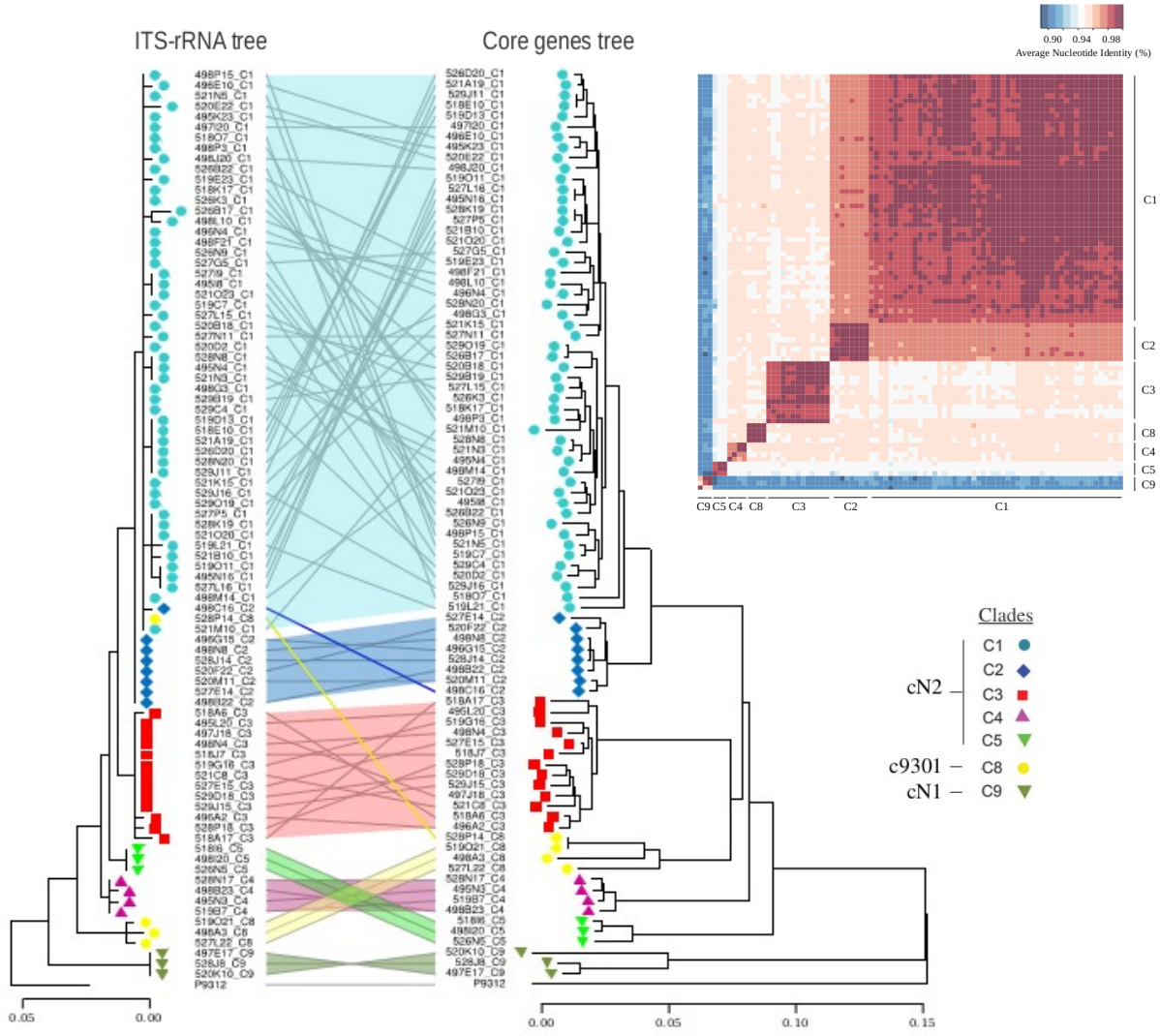


Figure 2.7 : Arbres phylogénétiques au maximum de vraisemblance basé sur l'ITS (à gauche) et la concaténation des alignements des gènes *core* présents en copie unique (à droite). Les gènes *core* (1 202 au total) sont partagés par 13 génomes de souches cultivées de l'écotype HLII de *Prochlorococcus*. Le génome de référence MIT9312 a été utilisé pour enraciner l'arbre. Les différents clades sont représentés par des formes et des couleurs qui leur sont propres et qui sont positionnés sur chaque feuille de l'arbre. Le modèle d'évolution utilisé est le GTR.

La caractérisation de la synthénie entre les génomes représentatifs des sept sous-populations étudiées montre une organisation génomique semblable à celle observée pour les souches composant l'écotype HLII de *Prochlorococcus* (Yan *et al.* 2018, Avrani *et al.* 2011). On retrouve ainsi six îlots génomiques caractérisés par une similarité réduite entre les SAGs, dispersés au sein du génome squelette (ou *backbone*). L'ensemble des COGs spécifiques des SAGs étudiés et sans contrepartie dans le génome de référence MIT9312 ont été assignés à un compartiment génomique (*backbone* ou îlot) sur la base d'une analyse de leur voisinage. Ainsi, au sein d'un COG, chaque gène est affecté à un compartiment génomique sur la base du compartiment associé aux gènes *core* qui l'entourent. Chaque COG est ensuite affecté au compartiment majoritairement restitué par les gènes qui le composent. Dans le cas où aucune information claire ne se dégage, les COGs sont attribués à un compartiment virtuel noté *ambigu*. Près de 63 % des COGs spécifiques des sous-populations étudiées ont ainsi pu être assignés au *backbone*, 8 % et 15 % ont été attribués respectivement aux îlots ISL3 et ISL4, et 9 % ont été répartis sur les autres îlots. Cinq pour cent des COGs spécifiques des clades ont été assignés au compartiment ambigu.

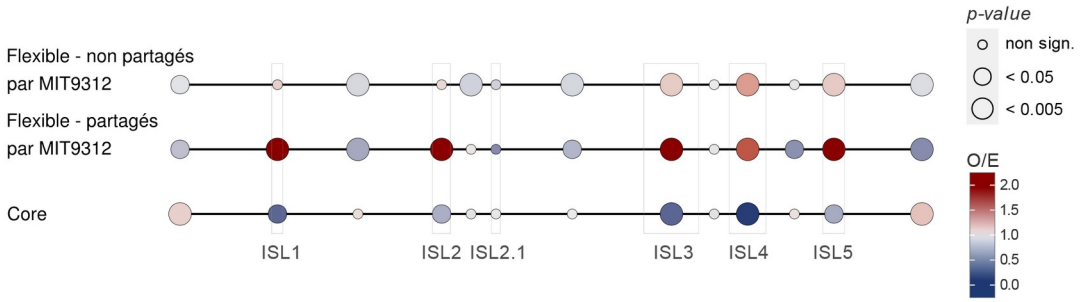


Figure 2.8 : Densité de distribution des COGs le long du génome exprimée en fonction des différences entre effectifs observés et théoriques (nombre de COGs [*core*, flexibles partagés ou non par MIT9312] attendus pour un compartiment donné en fonction du nombre total de COGs pour ce même compartiment ; O/E) et la différence de densité de distribution sur l'ensemble des compartiments génomiques (*core* – flexibles ; *backbone* – îlots) a été évaluée par un test du χ^2 (p<0,001). Des tests du χ^2 ont également été effectués pour chacun des groupes de COGs (*core*, flexibles partagés ou non par MIT9312) afin de tester la significativité du ratio O/E pour un compartiment donné (*backbone* – îlots). La taille des cercles est fonction de la *p*-value du test du χ^2 associée.

Le génome flexible spécifique des SAGs étudiés et localisé au niveau des îlots est majoritairement constitué de COGs spécifiques à une unique population. Tous les îlots ne partagent cependant pas les mêmes caractéristiques (Figure 2.8). Si les îlots sont en moyenne 2,5 fois plus dense en COGs flexibles que le *backbone*, les îlots ISL2 et ISL4 sont respectivement appauvris et enrichis en COGs flexibles relativement à la densité moyenne des îlots. Par ailleurs, les COGs spécifiques d'une unique sous-population sont particulièrement enrichis dans les îlots ISL3 et ISL4, alors que ceux communs à au moins cinq populations sont appauvris dans l'îlot ILS4 (Figure 2.9).

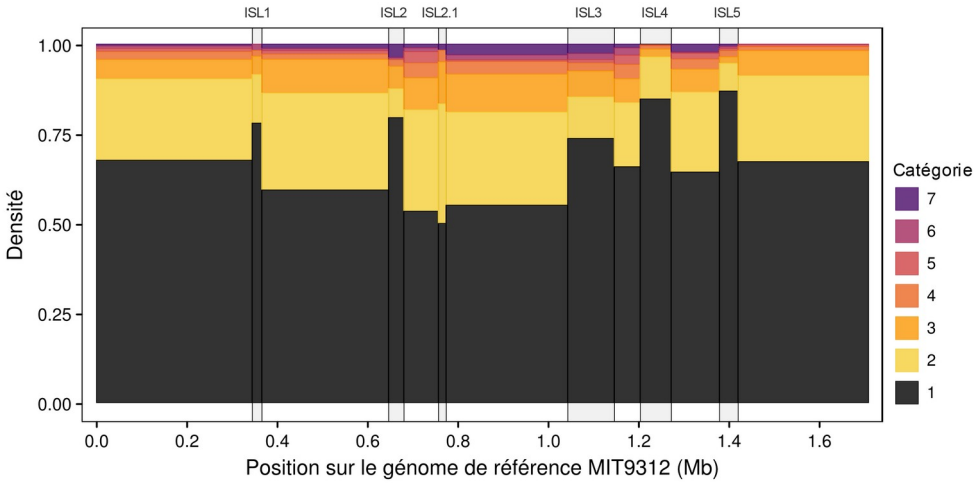
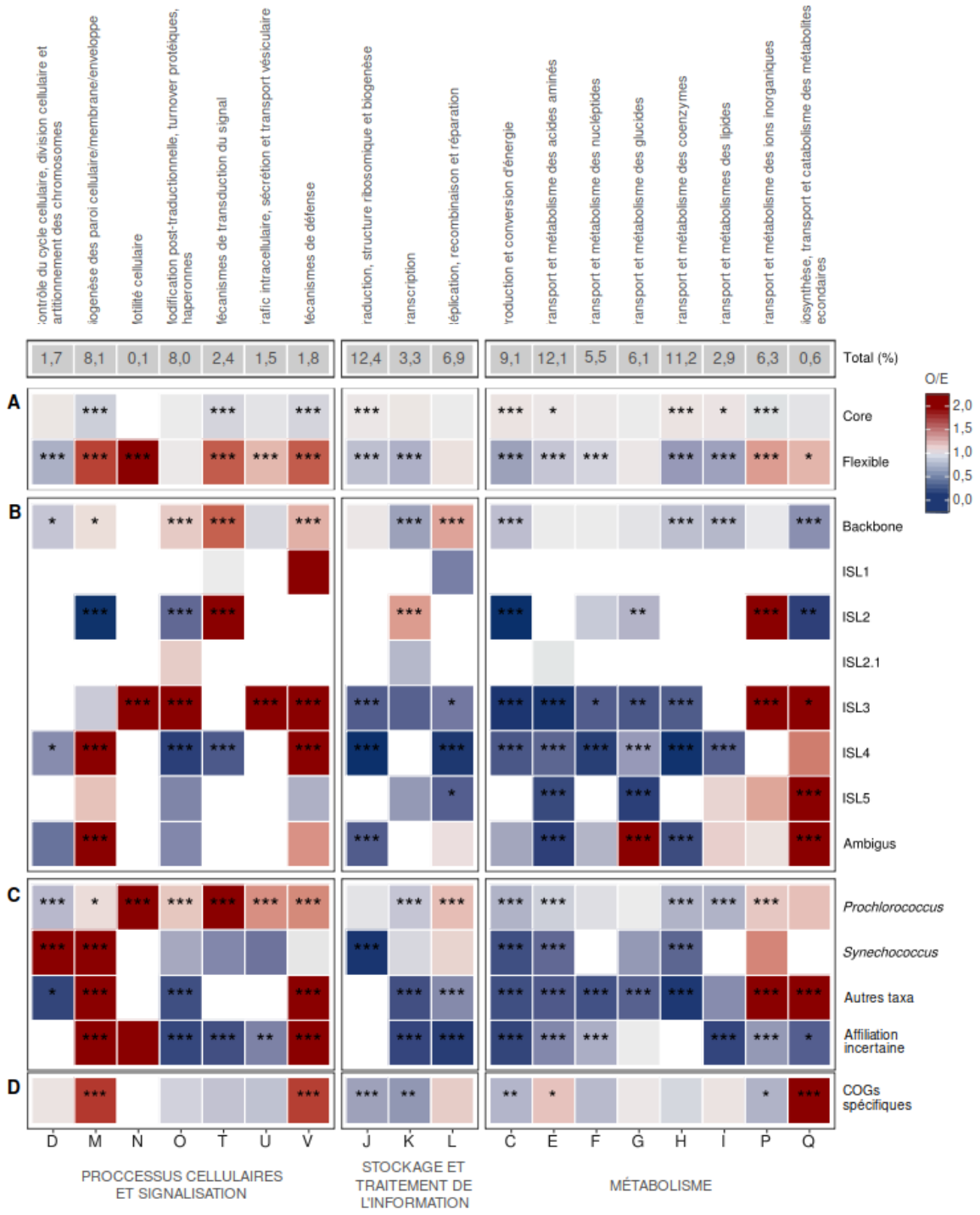


Figure 2.9 : Densité de distribution des COGs flexibles SAG-spécifiques le long du génome en fonction de leur représentativité au sein des sous-populations (catégories 1 à 7, *i.e.,* COGs spécifiques d'une sous-population en noir à communs à toutes en violet).

3.3 – Caractérisation des COGs flexibles spécifiques des sous-populations

Le potentiel fonctionnel des COGs flexibles a été étudié à l'échelle des sous-populations et comparé entre les compartiments reflétant l'organisation génomique des SAGs. Sur la base d'une analyse comparative des gènes contre la base de données EggNOG (Huerta-Cepas *et al.* 2016), 47 % des COGs (*core* et flexibles) ont été annotés fonctionnellement. Aucune différence dans la distribution des catégories fonctionnelle de ces COGs n'est observée entre les populations, hormis une surreprésentation de la catégorie fonctionnelle « Motilité cellulaire » dans les clades C1 et C9. La différentiation des sous-populations ne semble donc pas associée à une spécialisation fonctionnelle à cette échelle de description.

La distribution des catégories fonctionnelles associées aux COGs flexibles par rapport à celle des SAGs dans leur ensemble, est caractérisée par une sous-représentation des hiérarchies « Stockage et traitement de l'information » et « Métabolisme » (Figure 2.10). Cet appauvrissement touche principalement les catégories fonctionnelles associées aux mécanismes de transcription et de traduction, à la production d'énergie et au transport et métabolisme des nucléotides, acides aminés, coenzymes et lipides. Les processus cellulaires et de signalisation sont globalement surreprésentés au sein des COGs flexibles, particulièrement dans l'îlot ISL3 et le *backbone*. Les fonctions liées à la biogenèse de la paroi cellulaire et aux mécanismes de défense sont surreprésentées dans tous les clades et sont associées à des COGs situés principalement dans les îlots ISL4, à ISL3. Ces COGss sont le plus souvent spécifiques à une sous-population (Figure 2.10 D).



gure 2.10 : Distribution des enrichissements fonctionnels des COGs flexibles comparativement aux COGs *core* (A), et selon (B) leur position sur le génome, (C) leur affiliation, et (D) leur appartenance à une unique sous-population. Les enrichissements sont illustrés par les rapports observés/théoriques (O/E) pour chaque catégorie fonctionnelle. Les valeurs observées correspondent au nombre de gènes attribués à chaque catégorie dans chaque ensemble de données. Les valeurs théoriques (E) ont été obtenues en multipliant le nombre de gènes (*core* ou flexibles en fonction de leur localisation génomique, de leur affiliation taxonomique ou de leur appartenance à une unique sous-population) par le pourcentage de gènes totaux de chaque catégorie fonctionnelle. Les différences de distribution pour chaque type de comparaison ont été testées avec un χ^2 ($p < 0,005$ pour les groupements A, B et C). Des tests du χ^2 ont également été effectués pour chaque ligne du graphique (chaque catégorie contre toutes les autres) afin de tester la significativité de l'enrichissement pour un compartiment donné, une taxonomie donnée, ou la spécificité des sous-populations. Test du χ^2 : *, *p-value* < 0,05 ; **, *p-value* < 0,01 ; *, *p-value* < 0,001 (d'après Gardon *et al.* 2020).**

La comparaison des séquences des gènes contre la base de données EggNOG a permis d'exploiter l'information taxonomique restituée pour 48,0 % des COGs flexibles. Alors qu'une large majorité d'entre eux sont affiliés au genres *Prochlorococcus* (94,4 % de HLII, 3,3 % de HLI et 2,3 % de LL) et *Synechococcus* (6,0 %), près de 4,5 % des COGs flexibles spécifiques des clades ont une affiliation incertaine (c'est-à-dire que tous les gènes du COG ne sont pas affiliés aux genres *Prochlorococcus* ou *Synechococcus*), et 13 % présentent des similitudes avec des gènes affiliés à des taxa autres, majoritairement des protéobactéries et des cyanobactéries (Figure 2.11).

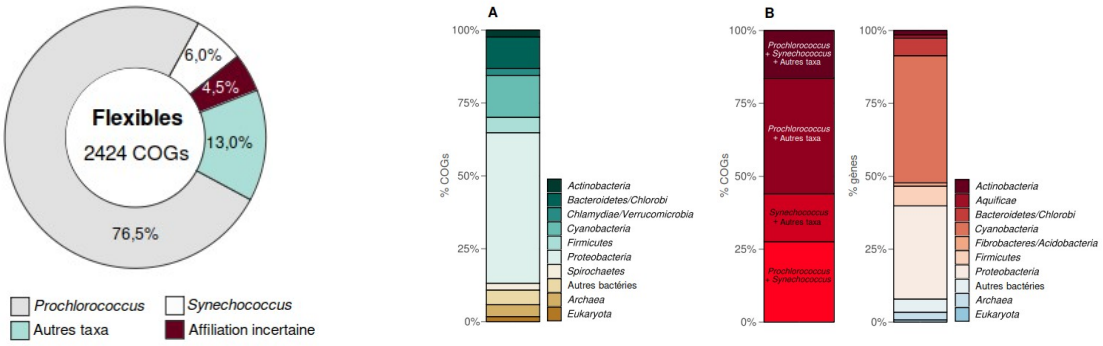
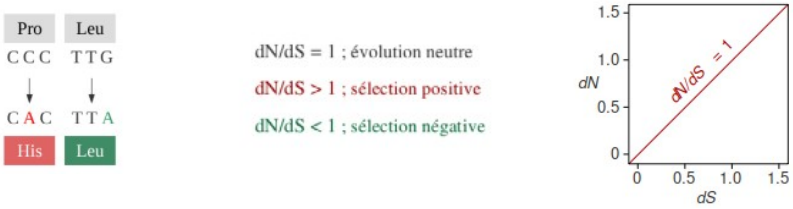


Figure 2.11 : Affiliations taxonomiques des COGs flexibles. Les COGs sont affiliés à *Prochlorococcus* (gris), *Synechococcus* (blanc), divers groupes taxonomiques incluant *Prochlorococcus* ou *Synechococcus* (affiliation incertaine – rouge pourpre) ou à des taxa autres que *Prochlorococcus* et *Synechococcus* (vert). (A) Distributions taxonomiques observées pour les COGs flexibles affiliés au taxon « autre ». (B) Distributions taxonomiques des COGs dont l'affiliation est incertaine. La catégorie « Autres bactéries » regroupe les taxa dont l'abondance est inférieure à 1 %.

Lorsqu'ils sont partagés par un petit nombre de sous-populations, les COGs flexibles spécifiques des clades ont plutôt une affiliation incertaine, tandis que les COGs retrouvés dans au moins six populations sont essentiellement affiliés au genre *Prochlorococcus*. Le *backbone* est également enrichi en COGs affiliés au genre *Prochlorococcus*, alors que cette affiliation est sous-représentée dans les îlots ISL3 à ISL5. L'îlot ISL4 et le compartiment *ambigu* sont particulièrement enrichis en COGs dont l'affiliation est incertaine ou autre que les genres *Prochlorococcus* et *Synechococcus* (Gardon *et al*. 2020).

L'hétérogénéité des affiliations taxonomiques retrouvées suggère l'existence de flux de gènes d'origines différentes, y compris entre les sous-populations. Cela semble être particulièrement le cas pour les COGs impliqués dans des fonctions liées à la biogenèse de la paroi cellulaire, aux mécanismes de défense et à la biosynthèse, le transport et le catabolisme de métabolites secondaires. Au contraire, la localisation génomique et l'affiliation taxonomique des COGs associés aux mécanismes de transduction du signal suggèrent l'existence de voies de régulation partagées entre les sous-populations.

Box 6 – Substitution synonymes, non synonymes et pression de sélection



La structure du code génétique assurant la représentation des acides aminés au niveau de l'ADN par un triplet nucléotidique (codon) permet de distinguer des sites synonymes (pour lesquels une mutation nucléotidique n'induit pas de modification de l'acide aminé codé – en vert sur la figure) et des sites non synonymes pour lesquels les mutations à l'échelle nucléotidique ont pour conséquence de modifier l'acide aminé codé (en rouge). La sélection qui opère sur les protéines pour maintenir / garantir leur fonction, s'applique également sur les sites non synonymes, contrairement aux sites synonymes qui évoluent selon un processus neutre. Le taux relatif de substitutions sur les sites synonymes (dS) par rapport aux sites non synonymes (dN) permet d'évaluer l'effet relatif de la sélection sur la protéine étudiée, par rapport à une évolution neutre. Lorsque les valeurs du rapport dN/dS est autour de 1, la protéine évolue selon un processus neutre. Lorsque le rapport dN/dS < 1, les substitutions non synonymes sont moins fréquentes que les substitutions neutres et la protéine est soumise à une pression de sélection négative d'autant plus forte que la valeur du rapport dN/dS est faible. Lorsque le rapport dN/dS > 1, la protéine est soumise à une pression de sélection positive, elle tend à accumuler plus de mutation que le nombre de mutations attendues dans le cas d'un processus neutre.

4- Dynamique évolutive des compartiments génomiques

Afin d'évaluer les pressions de sélection qui s'appliquent au sein du pangénome, les taux de substitutions synonymes (*dS*) et non synonymes (*dN*) moyens ont été estimés pour les COGs retrouvés dans au moins deux sous-populations (BOX 6).

Les rapports de *dN/dS* estimés sur les COGs *core* et flexibles sont globalement inférieurs à 1, indiquant qu'ils sont essentiellement soumis à une pression de sélection négative. Celle-ci est moins forte sur les COGs flexibles, et est inégale le long du *backbone* et entre îlots (*p<0,001,* test Kruskal-Wallis, (Figure 2.12). Dans le *backbone*, la région située entre les îlots ISL2 et ISL2.1 présente ainsi le rapport de *dN/dS* moyen le plus faible (0,29) tandis que celle située entre les îlots ISL4 et ISL5 est la plus élevée (0,73) .

Pour ce qui concerne les îlots, ISL1, ISL2 et ISL2.1 affichent des pressions de sélection négatives similaires à celles observées pour les gènes c*ore* situés dans le *backbone* (et même supérieures pour l'îlot ISL2.1). Par contre, on observe un relâchement des contraintes sélectives dans les îlots ISL3, ISL4 et ISL5 pour les gènes *core* et flexibles (allant de 0,36 pour ISL3 à 0,52 pour ISL5). Les gènes flexibles propres à chaque sous-population, localisés dans l'îlot ISL4, ou assignés au compartiment *ambigu*, présentent les pressions sélectives les plus faibles (avec un rapport *dN/dS* moyen de 0,44 et 0,52 respectivement). Ces fluctuations des valeurs du rapport de *dN/dS*, peuvent être la conséquence d'une variation du taux de substitutions non-synonymes, synonymes ou les deux.

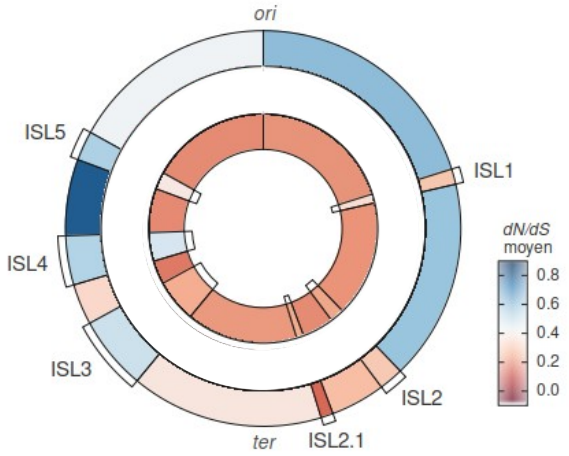


Figure 2.12 : Représentation des rapports de *dN/dS* moyens le long du chromosome (le génome de la souche MIT9312 tient lieu de référence). Le cercle interne représente les rapports de dN/dS estimés pour les COGs *core* et le cercle externe, les COGs flexibles spécifiques des sous-populations étudiées. L'origine (*ori*) et le terminus (*ter*) de réplication sont indiqués. ISL : îlot génomique (Gardon *et al.* 2020).

4.1 - Taux de substitution et signatures des compartiments génomiques

Les valeurs de *dN* et *dS* estimées varient considérablement le long du *backbone*. Bien que la valeur moyenne de *dS* soit faible (0,100 ± 0,091), celles-ci peuvent individuellement dépasser 1,5. Par comparaison, les valeurs de *dN* affichaient des variations moindres (0,026 ± 0,003). Les relations entre les rapports de *dN/dS* et *dN* ou *dS* sont semblables pour les COGs *core* et flexibles et permettent de distinguer trois groupes de gènes. Le premier (regroupant les points jaunes, oranges et rouges sur la Figure 2.13) est caractérisé par des valeurs de *dS* comprises entre 0 et ~ 0,27, des valeurs *dN* faibles (<0,34) et des rapports dN/dS variant de 0 à plus de 1,5. Les valeurs élevées des rapports de *dN/dS* sont pour l'essentiel associées à des *dS* faibles. Le deuxième groupe (points verts, Figure 2.13) inclut des COGs avec des valeurs de *dN* faibles (< 0,326) et des valeurs de *dS* intermédiaires (comprises entre 0,173 et 1,076), traduisant la présence dans ces COGs de séquences plus divergentes que dans le premier groupe. Le troisième groupe (incluant les points bleu foncé, Figure 2.13) est caractérisé par des valeurs de *dS* supérieures à 0,25, des valeurs de *dN* comprises entre 0,17 et 1,08 et des rapports de dN/dS inférieurs à 1.

Les variations de *dN* et *de dS* sont également contrastées entre îlots. Ces valeurs sont faibles et homogènes sur les îlots ISL1, ISL2 et ISL2.1 (*dN* <0,25 et dS <*0*,5) avec des rapports de *dN/dS* globalement faibles, suggérant l'existence d'une pression de sélection négative forte. À l'inverse, les COGs situés dans l'îlot ISL4 ou affectés au compartiment *ambigu* sont caractérisés par des valeurs *dN* et *dS* plus élevées, des rapports de *dN/dS* liés linéairement au *dN* et des valeurs de *dS* pour l'essentiel supérieures à 1. Les COGs situés dans les îlots ISL3 et ISL5, présentent un profil mixte.

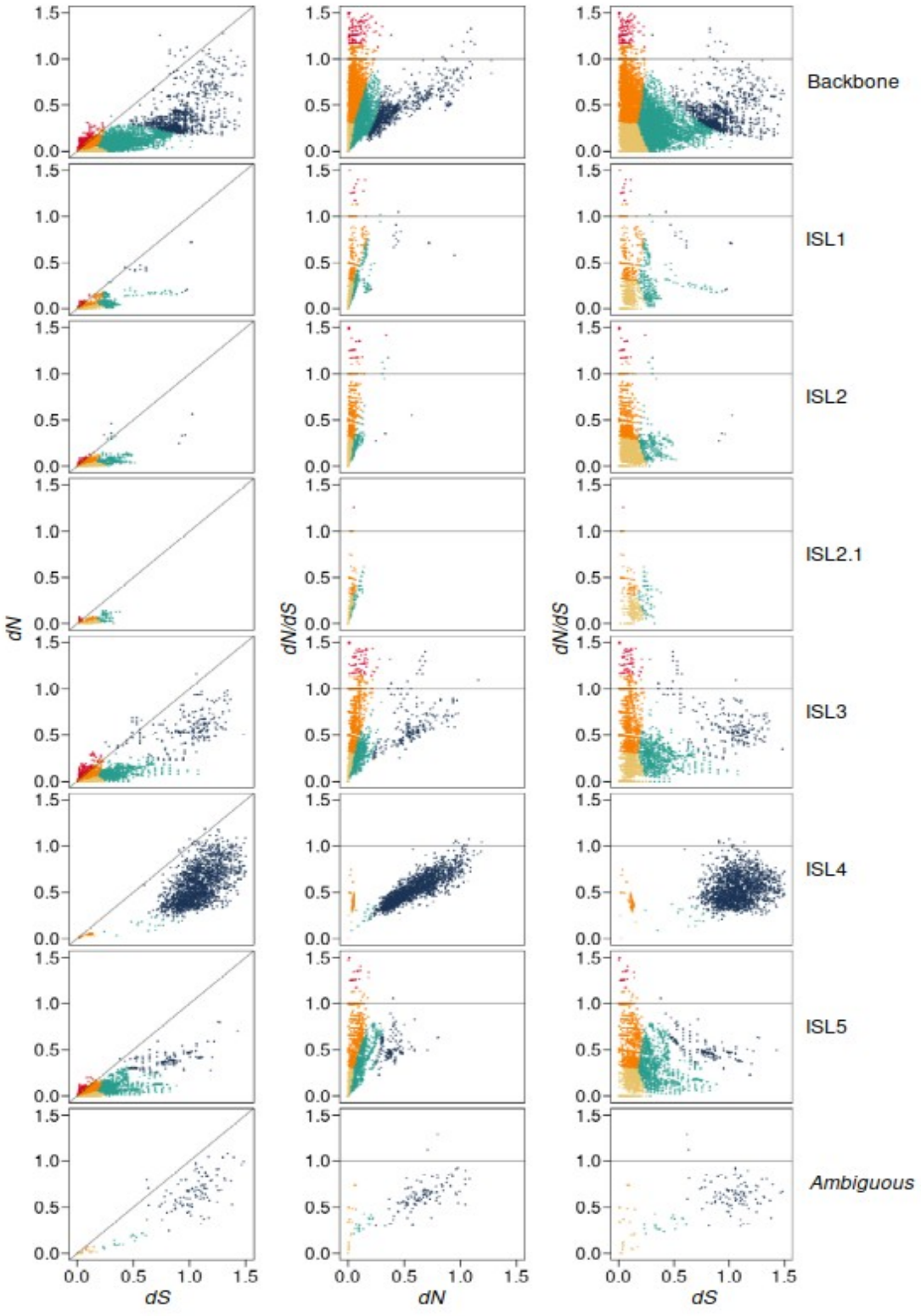


Figure 2.13 : Relations entre les valeurs de *dN*, *dS* et *dN/dS* des COGs flexibles pour les différents compartiments génomiques (*backbone*, ISLs et *ambigu*). Chaque point représente une comparaison deux à deux de gènes appartenant à un même COG et issus de deux sous-populations distinctes. Les groupes de gènes (jaune, orange, rouge, vert et bleu foncé) ont été définis par le partitionnement des données en *k*-moyennes (*k-means*). Le rapport de *dN/dS* égal à 1 est représenté par une ligne continue. Seules les valeurs de *dN*, *dS* et *dN/dS* inférieures à 1,5 sont représentées. Des profils similaires ont été obtenus pour les COGs *core*. ISL : îlot génomique (d'après Gardon *et al.* 2020).

4.2 - Signature évolutive et distribution des COGs dans les sous-populations

Dans la mesure où les COGs flexibles sont retrouvés dans un nombre variable de sous-populations, les fluctuations des valeurs de *dN*, *dS* et du rapport de *dN/dS* ont été analysées au regard de leur distribution au sein de ces dernières. Les COGs flexibles (exception faite de ceux communs à tous les clades pour lesquels peu de COGs affichent des valeurs de *dS* saturés) sont caractérisés par une grande variabilité des valeurs de *dS* (Figure 2.14). Par ailleurs, les valeurs des rapports de *dN/dS* diffèrent significativement en fonction du nombre de clades dans lesquels les COGs ont été trouvés (*p<0,001*, test de Kruskal-Wallis) ; les COGs partagés par un nombre important de clades tendant à avoir de faibles valeurs du rapport de *dN/dS*, suggérant des contraintes sélectives négatives plus fortes pour les COGs présents dans un plus grand nombre de sous-populations.

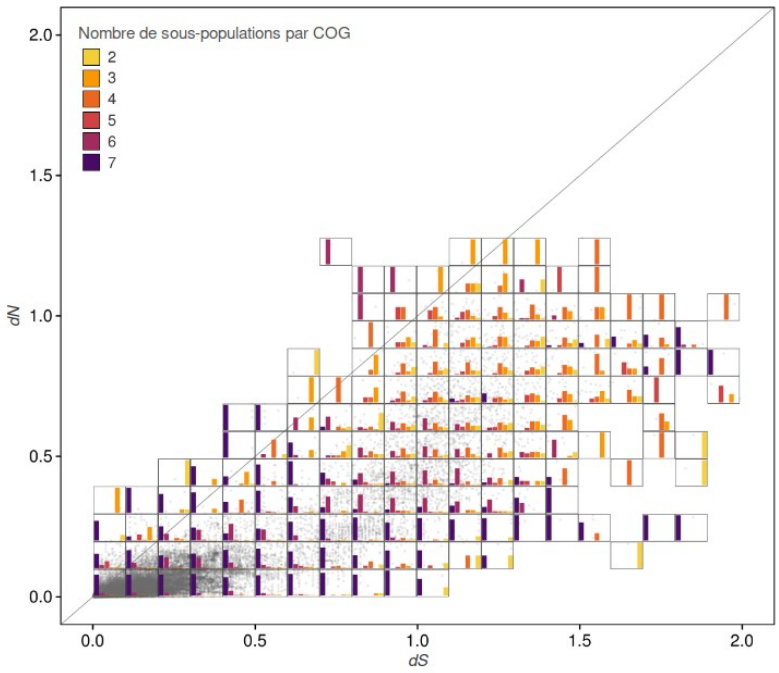


Figure 2.14 : Relations entre les taux de substitution synonymes et non synonymes estimés pour les COGs flexibles spécifiques des clades et leur distribution dans les différentes sous-populations. Le nombre de sous-populations par COG a été reporté sur les estimations des valeurs de *dN* et de *dS* (points gris). Le nombre de sous-populations varie de 2 à 7. Il est égal à 7 lorsqu'un COG est présent chez toutes les sous-populations. Chaque histogramme décrit la distribution du nombre de sous-populations par COG dans une région 2D du graphique correspondant à des valeurs de *dN* et *dS* dans un intervalle de 0,1. La ligne diagonale représente un ratio *dN/dS* égal à 1 (d'après Gardon *et al.* 2020).

Les affiliations taxonomiques des COGs flexibles analysés pour le rapport de *dN/dS* (soit 65,5 % des COGs analysés) ont révélé que ceux ayant de faibles valeurs de *dS* sont, pour l'essentiel, affiliés à *Prochlorococcus* et majoritairement retrouvés dans les îlots ISL3 et ISL5 (Figure 2.15 A). Les COGs avec un *dS* saturé (pour l'essentiel situés dans l'îlot ISL4 et le compartiment *ambigu)* présentent des affiliations multiples (catégorie incertaine), incluant entre autres les genres *Prochlorococcus* ou *Synechococcus* (Figure 2.15 B). Les COGs qui ne sont affiliés, ni à *Prochlorococcus,* ni à *Synechococcus* (autres taxa) sont enrichis dans les îlots ISL3 et ISL4 (*p<0,005,* test chi carré) et sont caractérisés par de faibles valeurs de *dN* et de *dS*.

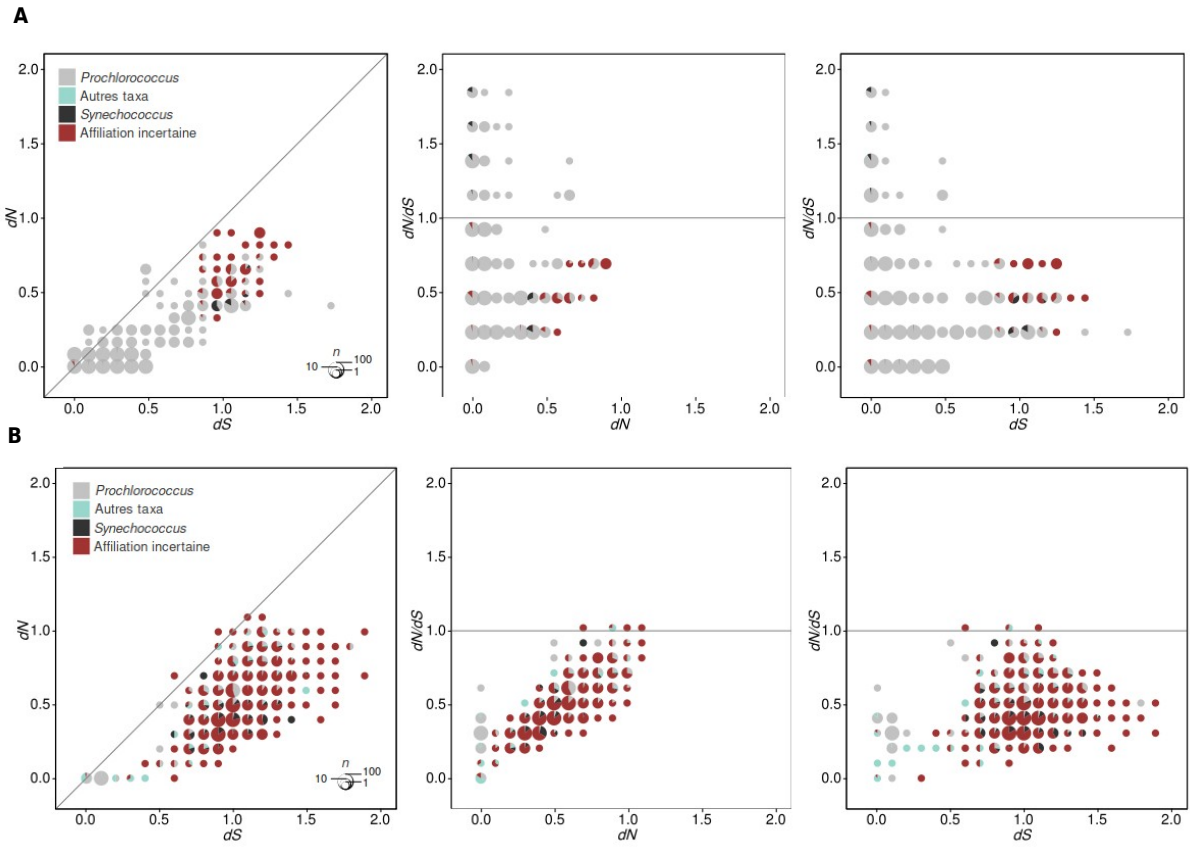
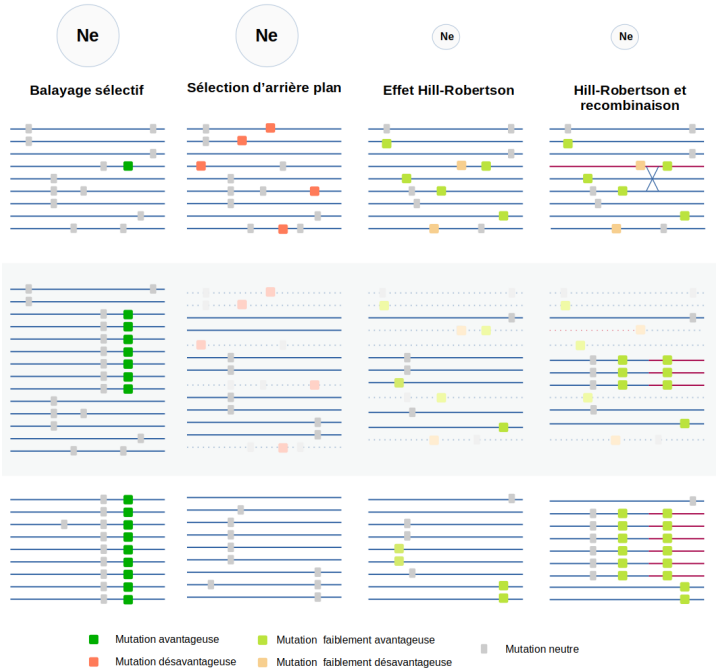


Figure 2.15 : Relations entre taux de substitution estimés pour les COGs flexibles spécifiques des sous-populations et leurs affiliations taxonomiques. (A) COGs localisés dans les îlots ISL3 et ISL5. (B) COGs localisés dans ISL4 ou attribués au compartiment *ambigu*. Les affiliations taxonomiques ont été reportées sur les estimations des *dN, dS* et des rapports de *dN/dS*. Chaque diagramme circulaire illustre la distribution des affiliations dans une région du graphique correspondant à des valeurs de *dN*, *dS* et *dN/dS* dans un intervalle de 0,1. La taille des diagrammes circulaires est proportionnelle au nombre d'observations *n* pour la région 2D considérée (*n* : au moins une observation ; de 10 à 100 observations ; plus de 100 observations). Les lignes horizontales et diagonales représentent un rapport *dN/dS* égal à 1 (d'après Gardon *et al.* 2020).

5 – Bilan et perspectives

Alors que les populations de *Prochlorococcus* décrites dans Kashtan et collaborateurs sont caractérisées par des abondances différentes en fonction des saisons et une différentiation allélique à l'échelle des gènes *core* (Kashtan *et al.* 2014), l'analyse de la distribution des catégories fonctionnelles attribuées aux COGs flexibles, comparativement au potentiel fonctionnel des SAGs ne permet pas de mettre en évidence de différences entre les sous-populations de l'écotype HLII étudiées (Gardon *et al.* 2020). Ceci est en accord avec les travaux de Larkin et collaborateurs qui, à l'exception de la température, ne trouvent pas de lien entre la répartition de cet écotype à l'échelle du globe et différents paramètres biogéochimiques (notamment pH, phosphate, ammonium - Larkin *et al.* 2016). Ainsi, bien qu'étant retrouvé de manière consistante à travers différentes analyses, aucun élément ne permet actuellement d'expliquer la structuration de la population associée à l'écotype HLII de *Prochlorococcus* isolée dans les BATS.

BOX 7 – Balayage sélectif, sélection d'arrière plan et effet Hill-Robertson



L'action conjointe de la sélection et de la dérive génétique à pour effet de réduire la diversité génétique au sein des populations. Cette réduction répond à différentes dénominations en fonction des facteurs qui sont en jeu.

(A) Balayage sélectif : Une mutation avantageuse, sélectionnée positivement, « envahi » la population du fait d'un différentiel reproductif favorable aux porteurs de cette mutation ; la diversité génétique de la population est éliminée à l'exception des mutations génétiquement liées à cette mutation avantageuse.

(B) Sélection d'arrière plan : les individus porteurs de mutations désavantageuses ne participent pas à la génération suivante ; les mutations, neutres pour l'essentiel, liées aux mutations désavantageuses sont « purgées » avec celles-ci, du fait d'une sélection négative, réduisant de fait la diversité génétique de la population.

(C) Effet Hill-Robertson : dans une population de petite taille, la dérive génétique produit un échantillonnage aléatoire des mutations qui sont maintenues à la génération suivante, induisant une perte de diversité génétique des mutations, qu'elles soient neutres, ou soumises à sélection. L'efficacité de la sélection est globalement réduite.

(D) Impact de la recombinaison : dans une population de petite taille soumise à l'effet Hill-Robertson, la recombinaison génère des combinaisons de mutations nouvelles et augmente la diversité génétique de la population. Elle rend possible l'association de mutations avantageuses acquises indépendamment par plusieurs individus et peut permettre d'augmenter la valeur sélective globale de cette combinaison, qui peut potentiellement présenter un avantage sélectif significatif (s>>1/Ne).

Il existe une compartimentation fonctionnelle des COGs flexibles le long du génome des sous-populations étudiées avec une sur-représentation des fonctions associées aux catégories « Processus cellulaires et de signalisation », « Biogenèse de la paroi cellulaire » et « Mécanismes de défense » dans les îlots ISL3 et ISL4, ainsi que dans le compartiment *ambigu*. Les COGs associés aux deux dernières catégories sont par ailleurs affiliés à une grande diversité de taxa et sont, pour l'essentiel, spécifiques à une sous-population. Ces résultats suggèrent qu'ils ont été acquis par transfert horizontal et sont congruents avec les résultats décrits à une échelle de diversité plus large (Coleman *et al.* . 2006). Ceci pourrait signifier que les différences de distribution et d'abondance observées au niveau des écotypes pourraient également s'appliquer à l'échelle des sous-populations.

D'un point de vue évolutif, les résultats obtenus montrent que le pangénome de la population de l'écotype HLII échantillonné dans les BATS est globalement soumis à des contraintes sélectives négatives inégales le long du génome. Les COGs situés dans les îlots ISL1, ISL2, et ISL2.1 sont composés de gènes

affiliés au genre *Prochlorococcus*, communs à tous les clades, et soumis à une sélection négative forte. Ils pourraient être en cours de fixation dans ces sous-populations. A l'inverse, les COGs situés dans les îlots ISL3 à ISL5 affichent une pression de sélection négative relâchée. Les COGs situés en particulier dans les îlots ISL3 et ISL4 sont plutôt affiliés à des taxa autres que les cyanobactéries et tendent à avoir des taux de substitutions synonymes élevés. L'ensemble de ces observations laisse à penser que ces COGs sont issus de transferts horizontaux. Les deux groupes d'îlots identifiés ici s'apparente à la distinction proposée par Lopez-Perez et collaborateurs entre îlots génomiques additifs et îlots de remplacement (Lopez-Perez et al. 2014, Lopez-Perez et al. 2016). Les îlots de remplacement sont caractérisés par la présence de gènes différents mais codant pour des fonctions semblables, tandis que les îlots additifs affichent une grande variabilité de contenu en gènes issus de HGT. Alors que les îlots ISL3 et ISL4 semblent présenter les caractéristiques des îlots génomiques additifs, il n'est pas possible dans le cadre de cette étude d'affirmer que les îlots ISL1, ISL2 et ISL2.1 correspondent à des îlots de remplacement. La caractérisation des COGs permet toutefois de souligner une répartition non aléatoire des fonctions portées par les gènes flexibles le long de ces génomes, en lien avec des contraintes fonctionnelles et évolutives différentes. Schmutzer et Barraclough (2019) ont suggéré qu'en présence de flux de gènes entre des populations divergentes, la concentration de gènes localement adaptatifs dans un nombre réduit de loci pourrait être favorisée, car cela permettrait i) de réduire l'impact négatif des insertions le long du génome de gènes transférés horizontalement et ii) d'augmenter l'efficacité relative de la sélection sur quelques « mega loci » par rapport à une situation avec de nombreux loci dispersés à effet réduit. Dans ce cadre, les îlots ISL1, ISL2 et ISL2.1 pourraient tenir lieu de « mega loci » et concentrer des gènes flexibles essentiels pour l'ensemble des clades constituant la population, à l'image de ce qui a été proposé de manière plus générale pour l'îlot ISL2.1 par rapport à l'écotype HLII (Avrani *et al.* 2011).

L'analyse des signatures sélectives du génome flexible a révélé deux grands ensembles de COGs. Le premier est présent dans tous les compartiments génomiques et regroupe des COGs avec de faibles valeurs de *dN* et des valeurs de *dS* très inférieures aux valeurs moyennes de *dS* estimées sur les COGs *core*, pouvant conduire à des valeurs de rapport de dN/dS supérieures à 1 (clusteur rouge). Ces résultats suggèrent soit une forte sélection d'arrière plan (BOX 7), soit une sélection négative opérant sur les sites synonymes (Parmley et Hurst 2007), soit une homogénéisation de l'information de séquence entre les sous-populations du fait d'événements de recombinaison homologue (Hanage 2016). D'une manière générale, la recombinaison homologue tend à augmenter l'efficacité de la sélection en réduisant l'effet Hill-Robertson (Hill et Robertson 1966). Dans un contexte de populations structurées, la recombinaison entre sous-populations permettrait d'augmenter la taille efficace des populations dans lesquelles circulent les gènes recombinants, tout en contraignant la différenciation des clades au sein d'un mode de divergence guidé par les gènes non recombinants (Marttinen *et al.* 2015).
Le second ensemble (cluster bleu foncé), retrouvé dans le *backbone*, les îlots ISL3, ISL4 et le compartiment ambigu est caractérisé par des COGs présentant un relâchement des contraintes sélectives, un taux de substitutions synonymes (*dS*) élevé et des affiliations à des taxa autres que les cyanobactéries ce qui pourrait laisser supposer des transferts horizontaux au sein de ce pool de gènes (Castillo-Ramírez *et al.* 2011). Bien que les gènes ayant à la fois une affiliation taxonomique et fonctionnelle ne soient pas assez nombreux pour permettre de tester cette hypothèse, la

surreprésentation des COGs impliqués dans les mécanismes de défense, ou la biogenèse de la paroi cellulaire en particulier dans les îlots ISL3 et ISL4, pourrait refléter l'acquisition de gènes transitoirement adaptatifs (pendant les périodes d'infection par les phages) (Avrani *et al.* 2011 ; Coleman *et al.* 2006 ; Kettler *et al.* 2007). Il n'est cependant pas certain que tous les gènes transférés présentent un tel avantage. Dans le cas de populations structurées, comme observé ici, l'évolution d'un gène quasi neutre acquis à partir d'une lignée distante pourrait être limitée au clade dans lequel il a été introduit. Ainsi, le HGT distant pourrait évoluer dans un contexte de taille efficace faible. Ce schéma pourrait expliquer les caractéristiques de l'îlot ISL4. D'un autre côté, la taille efficace associée aux gènes transférés pourrait également être élargie par le biais, par exemple, d'événements de recombinaison homologue locaux, favorisant l'empreinte de la sélection et la persistance de gènes acquis avec un effet marginal. L'apparition à la fois de HGT et de gènes sélectionnés dans les îlots ISL3 et ISL5 pourrait être le résultat de tels processus. Dans l'ensemble, ces résultats sont en accord avec le modèle de la barrière de dérive dans le cadre d'une population structurée (Gardon *et al*. 2020). En outre, la structuration de l'information génétique le long du génome pourrait dépendre de la dynamique des flux de gènes entre clades au sein d'une population structurée, en particulier pour les gènes flexibles.

Si l'on peut faire l'hypothèse que différents mécanismes (recombinaison, sélection…) participent au modelage du génome de manière à limiter les perturbations de son organisation et de son fonctionnement en concentrant les flux de gènes dans un nombre limité d'endroits permissifs (Oliveira *et al.* 2017), les questions i) de la part relative de ces différents processus dans la structuration des génomes microbiens, ii) le caractère aléatoire ou non de l'acquisition des gènes transférés par rapport à leur emplacement génomique et iii) la probabilité de rétention différentielle des gènes transférés comme une conséquence de la fluctuation de la taille efficace le long du génome, ne sont pas résolues et constituent les perspectives envisagées sur cette thématique.

Structure des génomes microbiens et potentiel fonctionnel : quel est le modèle évolutif du pangénome ?

La répartition non aléatoire des gènes flexibles au sein des génomes avait jusqu'ici été rapporté sur la base de comparaisons de génomes distants, ne prenant en compte que les événements de recombinaison ou de transfert effectivement maintenus dans les génomes après un temps d'évolution (Vos et Eyre-Walker 2017). Ici, les comparaisons sont réalisées entre génomes d'individus cooccurrents, impliquant une unité de lieu et de temps, et pour lesquelles on peut espérer observer le résultat de flux de gènes récents, incluant des événements qui n'ont pas en*core* été « purgés ». Bien que l'échelle d'organisation génomique analysée ici ne soit pas suffisante pour distinguer les flux de gènes récents et anciens, ce modèle d'étude semble prometteur pour tenter de caractériser plus finement les mécanismes d'acquisition de gènes et de leur maintien dans les populations. Une des perspectives de ce travail est, bien entendu, d'explorer à cette échelle fine la distribution des gènes recombinés ou transférés entre populations, voire depuis des groupes taxonomiques plus distants en fonction de leur localisation sur le génome. Ainsi, en premier lieu le caractère aléatoire de l'insertion des gènes transférés entre les sous-populations doit être évalué, ainsi que l'impact relatif de la sélection et de la

recombinaison sur leur trajectoire évolutive. On pourra en particulier chercher àe répondre aux questions suivantes :

- Est ce que les transferts horizontaux de gènes se produisent dans des points chauds au sein des génomes (plus fréquents dans les îlots par exemple) ?
- Y a t il une relation entre la répartition des gènes accessoires et l'âge des événements de transferts horizontaux inférés ?
- En quoi la répartition des événements de transfert dans les compartiments influence-t-elle le devenir évolutif et le taux d'évolution des gènes transférés ?
- Y a t il un lien entre les taux de gain / perte de gènes, le compartiment génomique et les pressions de sélections appliquées sur les gènes transférés ?
- Quel est le coût associé à ce génome accessoire et qu'est ce qui justifie son maintien ?

La mise à disposition de SAGs à l'échelle de populations (voire de communautés), en complément de données de métagénomique plus classiques permet d'évaluer à des échelles de temps contrastées les processus populationnels et évolutifs susceptibles d'avoir une influence sur l'occurrence et le devenir des transferts horizontaux de gènes. Si on considère le métagénome comme un pool infini de gènes disponibles au transfert, on peut élargir notre questionnement aux points suivants :

- Est-ce que le transfert de gènes est un processus neutre, c'est à dire que n'importe quel ADN du métagénome a les mêmes chances d'être recruté dans le cadre d'un transfert de gènes ou non. Si c'est le cas on ne devrait pas trouver de différences de répartition des groupes de gènes orthologues ou de catégories fonctionnelles dans le pool de gènes transférés par rapport au potentiel fonctionnel du métagénome.
- Quel est le coût du génome accessoire et plus largement l'intérêt ou le coût évolutif de tels flux de gènes et sur quelle unité biologique est-il pertinent de calculer ce coût ? Cette question peut être abordée en analysant le profil de gain / perte de gènes le long de la phylogénie en intégrant des niveaux de diversité extrêmement fins (populations et sous-populations) en relation avec les métagénomes et les conditions environnementales.

PARTIE III

« LA MATIERE NOIRE EUCARYOTE »

ET AUTRES TRAVAUX

1 – Le projet MICROSTORE

Ces dernières années, les approches de métagénomique et/ou de *metabarcoding* ont permis de faire émerger de nouvelles connaissances sur la diversité des microorganismes lacustres et ont notamment confirmé le gouffre qui sépare le nombre d'espèces eucaryotes unicellulaires cultivées de la diversité moléculaire retrouvée dans les écosystèmes (Debroas *et al.* 2017). Elles ont également révélé une biosphère d'eucaryotes unicellulaires rare et active, aussi bien dans la colonne d'eau des lacs (Debroas *et al.* 2015, Lepère *et al.* 2016, Li *et al.* 2017), que dans les sédiments (Capo *et al.* 2016). Celle-ci peut représenter jusqu'à 77% de la diversité phylogénétique présente. Il a également été mis en évidence une diversité importante d'eucaryotes unicellulaires actifs dans la zone anoxique des lacs, correspondant à des OTUs dont les séquences sont distantes de celles retrouvées dans les banques de données, soulevant ainsi des questions sur la nature de ces OTUs et sur leur métabolismes (Lepère *et al.* 2016). Au sein des groupes taxonomiques restitués par ces études, un grand nombre inclut des espèces potentiellement parasites, dont les cycles de développement et les hôtes sont encore largement méconnus. On trouve notamment de nombreux OTUs affiliés à la DMF (*Dark Matter Fungi*) (Grossart *et al.* 2016), en référence à différents groupes comme par exemple les *Rozellomycota* et les *Chytridiomycota*, qui émergent à la racine des *Fungi.* Celle-ci est omniprésente dans les environnements *lacustres*, mais ne possède à l'heure actuelle, que très peu de représentants cultivés. Ces champignons sont généralement considérés comme saprophytes ou parasites. Cependant, leurs rôles écologiques restent très peu connus. On retrouve également le phylum des *Perkinsozoa,* appartenant au clade des Alvéolés (*Alveolata*), qui inclut la plus grande diversité de parasites eucaryotes connus (Cavalier-Smith 1993, Mangot *et al.* 2011). Détectés pour la première fois dans des lacs en 2005 (Lefranc *et al.* 2005), des études plus récentes ont montré leur présence récurrente et l'importance quantitative de leurs

zoospores dans les lacs (Lepere *et al.* 2008, Brate *et al.* 2010, Lepère *et al.* 2010, Mangot *et al.* 2009, Mangot *et al.* 2013), ainsi que leur activité (Debroas *et al.* 2015). Alors que les espèces de *Perkinsozoa* marine sont connues pour infecter des bivalves (Mackin *et al.* 1950), ou impacter la prolifération de dinoflagellés toxiques (Lepelletier *et al.* 2014), l'écologie des *Perkinsozoa* lacustres est encore peu caractérisée. Les traits écologiques de leur espèces sœurs marines, associés à des observations en microscopie, suggèrent cependant que ces espèces pourrait également avoir un mode de vie parasitaire (Magot *et al.* 2013, Jobard *et al.* 2020 – Figure 3.1).

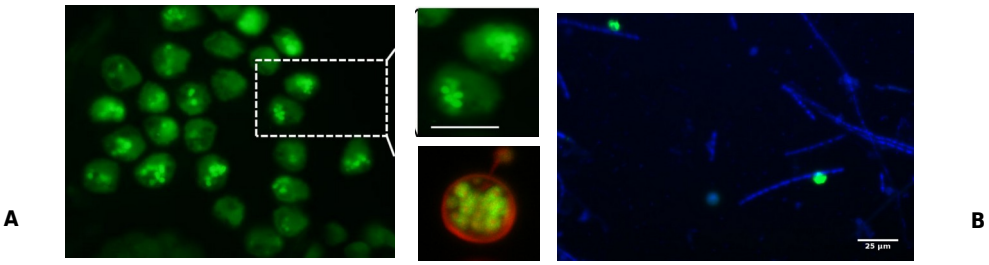


Figure 3.1 : Observation de cellules potentiellement infectées par des *Perkinsozoa* lacustres. (A) Observations cellules de *Perkinsozoa* ciblées par TSA-FISH à l'intérieur de chlorophycées coloniales lacustres (Mangot *et al.* 2013, Jobard *et al.* 2020) ; *Perkinsozoa* attachés à des cyanobactéries en filament (Jobard *et al.* 2020).

Alors qu'il a été mis en évidence une diversité d'eucaryotes microbiens aquatiques au moins égale à la diversité observée pour les bactéries et les archées (Lopez-Garcia et al.2001; Not et al.2009), les études la reliant aux aspects fonctionnels des communautés sont rares (Del Campo *et al.* 2014)

Dans ce contexte, le projet MICROSTORE ("Omic"approaches to decipher MICrobial eukaRyOteS funcTiOn in fReshwater lake Ecosystems), initié en 2018 en réponse à l'appel d'offre France Génomique « Grand projet de génomique » a pour objectif de caractériser la diversité fonctionnelle des eucaryotes microbiens dans les lacs par des approches *omiques (metabarcoding*, metagénomique, metatranscriptomique et génomique sur cellule unique (SAG)). Ce projet, porté par l'équipe, comporte deux aspects :

- Aspect écologique : l'objectif est ici de générer un catalogue de référence de gènes et de génomes d'eucaryotes microbiens lacustres, pour un large éventail de groupes taxonomiques, de définir les fonctions des gènes exprimés par cette communauté en fonction des conditions environnementales, de caractériser leurs potentiels métaboliques et de détecter leurs associations potentielles (symbiotiques ou parasitaires) ;
- Aspect évolutif : l'objectif est de proposer une résolution phylogénomique des clades eucaryotes unicellulaires lacustres ainsi que l'étude des facteurs qui gouvernent l'évolution de groupes d'eucaryotes unicellulaires particuliers, ou l'émergence de fonctions particulières en lien avec les caractéristiques biologiques et écologiques des groupes considérés.

Pour caractériser le potentiel fonctionnel des communautés microbiennes eucaryotes, il est plus intéressant de cibler son métatranscriptome (c'est-à-dire l'étude de l'ensemble des transcriptomes d'un milieu donné) que le métagénome. En effet, le séquençage des ARN messagers permet une caractérisation des gènes exprimés, en évitant le séquençage de grandes régions intergéniques, d'ADN

intronique ou d'ADN répété, fréquent dans les génomes eucaryotes. Cette approche facilite par ailleurs les étapes ultérieures d'assemblage et de prédiction des gènes (Worden et Allen, 2010). À partir des séquences codantes prédites après l'assemblage du métatranscriptome, il est possible d'identifier les voies métaboliques actives dans les communautés microbiennes. En mettant ces voies en perspectives des paramètres environnementaux qui caractérisent les différents échantillons, il est par exemple possible d'associer leur activation à des conditions environnementales particulières. Une telle approche a été mise en place dans le cadre du programme TARA OCEAN (Karsenti *et al.* 2011) et a permis de produire des catalogues de gènes eucaryotes exprimés en milieu marin à l'échelle du globe (Carradec *et al.* 2018).

Alors que les métatranscriptomes constituent le matériel le plus pertinent pour évaluer le potentiel fonctionnel des communautés naturelles, ceux-ci seront difficiles à analyser dans la mesure où l'annotation de ces transcrits passe par la comparaison des séquences environnementales aux bases de données publiques. Or, celles-ci contiennent particulièrement peu de séquences de microorganismes au sein des eucaryotes, l'effort de séquençage portant pour l'essentiel sur les eucaryotes pluricellulaires et leurs parasites (Del Campo *et al.* 2014). Pour contourner le manque de séquences de références et de représentativité de la diversité taxonomique des communautés microbiennes, le séquençage de l'ARNm environnemental peut être associé à celui de transcriptomes issus d'organismes maintenus en culture, afin de construire une base de données de référence pour des espèces microbiennes plus proches. Cette approche a permis de réduire le biais de représentativité des bases de données publiques dans le cadre de l'analyse des métatranscriptomes du programme TARA océans. Ceux-ci ont en effet profité des informations produites à partir du séquençage de 650 transcriptome d'eucaryotes unicellulaires issus de cultures marines (Marine Microbial Eukaryotes Transcriptome Sequencing Project - MMETSP) (Keeling *et al.* 2014).

Le projet MICROSTORE est donc basé sur le séquençage massif issu de lignées cultivées et environnementales échantillonnées dans le lac Pavin. Concernant le volet *culture*, le transcriptome de quatre-vingt-treize souches d'eucaryotes unicellulaires (LMGE, Thonon Culture Collection – TCC et Culture Collection of Algea et Protozoa – CCAP) a été réalisé. Concernant le volet *in situ*, des échantillons d'eau du lac Pavin ont été prélevés à quatre saisons et deux profondeurs (5 m et 80 m) de jour et de nuit. Après filtration, deux fractions de taille ont été retenues (<10 µm et 10-50 µm). Pour les trente deux conditions considérées, l'ADN et l'ARN a été extrait, divisé en deux duplicats (lorsque la quantité de matériel le permettait) et séquencé. Ceci a permis de générer cinquante-sept jeux de *metabarcoding,* trente six métagénomes et autant de métatranscriptomes, soit 35 Tb de données.

Dans le cadre de ce projet je suis plus particulièrement impliquée dans la mise en place des ressources bioinformatiques permettant l'exploitation des données produites et l'analyse la résolution phylogénomique de la diversité des eucaryotes microbiens lacustres.

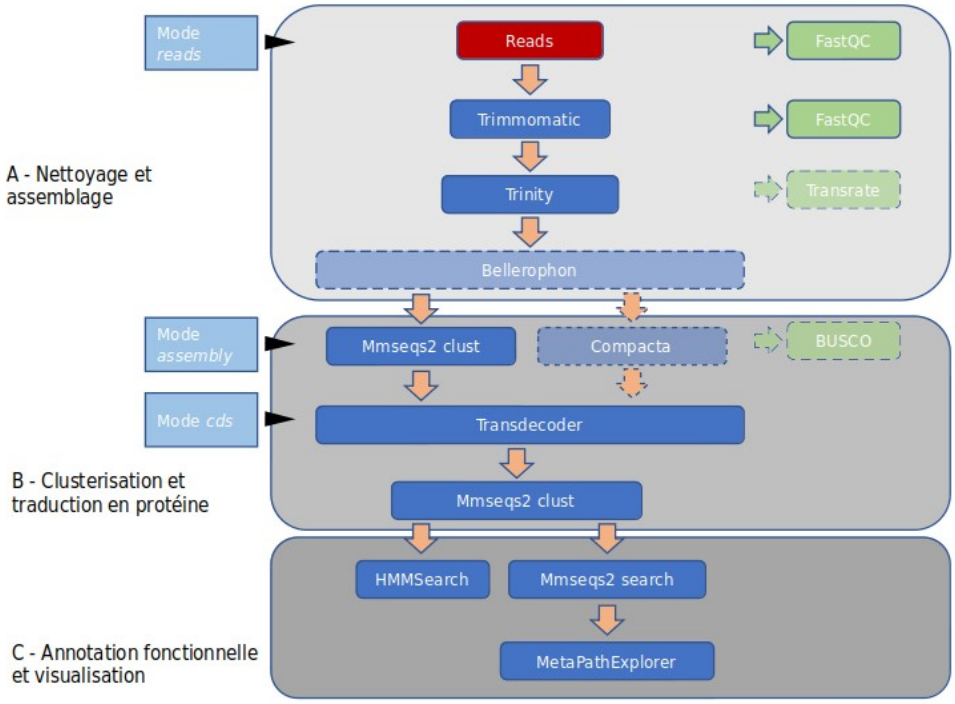


Figure 3.2 : Chaîne de traitement du pipeline KRYPTON. (A) Après nettoyage avec FastQC et Trimmomatic, les lectures sont assemblées à l'aide de Trimmomatic. (B) Les transcripts prédits sont clusterisés avec Mmeseq2 clust, puis les CDS sont prédits et traduits en protéines avec l'outil Transdecoder. (C) Les protéines prédites sont comparées d'une part à Pfam-A pour rechercher des domaines fonctionnels connus et d'autre part comparé à la base de données Uniref90. Les protéines identifiées sont cartographiées sur les cartes métaboliques de la base de données KEGG à 'aide de l'outil MetaPathExplorer. Sont représentés en transparence les outils qu'il est envisagé d'inclure dans cette chaîne de traitement (Transrate, BUSCO, Bellerophon, Compacta).

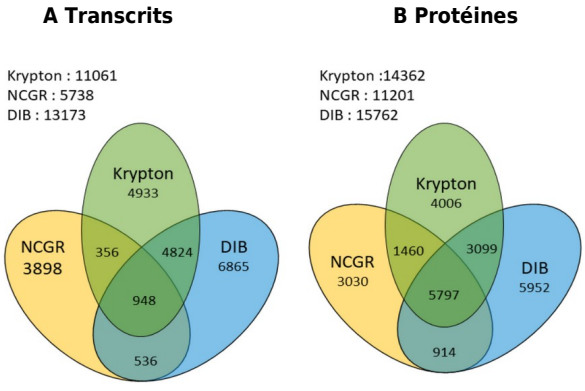


Figure 3.3 : Comparaison des assemblage NCGR, KRYPTON et DIB (NKD) sur le transcriptome MMETSP1161. (A) Comparaison du nombre de transcrits entre les trois assemblages (NKD). Seul 948 transcrits des transcrits sont communs aux trois pipelines, soit respectivement 16,5 %, 8,5 % et 7,2 % des assemblages NCGR, KRY et DIB. 1 304 transcrits sont communs à KRYPTON et NCGR et 5 772 sont communs à DIB et KRYPTON. (B) Comparaison du nombre de protéines prédites entre NKD. 5 797 protéines sont aux trois assemblages NKD , soit 51,7%, 40,3 % et 36,8 % des assemblages NCGR, KRY et DIB. 7 252 protéines sont communes à NCGR et KRYPTON. 8 896 protéines sont communes à KRYPTON et DIB (d'après Milisavljevic 2020).

1.1 – Mise en place d'une chaîne de traitement pour l'annotation des transcriptomes de cultures

L'annotation des transcriptomes issus de cultures est un travail préalable à l'étude des données *méta-omiques* acquises dans le cadre de MICROSTORE dans la mesure où elles permettent de fournir un catalogue de gènes d'espèces d'eucaryotes unicellulaires lacustres qui permettra d'enrichir les bases de références publiques utilisées pour l'annotation et l'analyse des données issues du lac Pavin.

La plupart des travaux publiés concernant l'analyse de transcriptomes sur des organismes non modèles sont réalisés sur la base de chaînes de traitement *ad hoc* (par exemple Strassert *et al.* 2019). C'est par exemple le cas pour les données MMETSP (Keeling *et al.* 2014). Quelques chaînes de traitement pour l'assemblage *de novo* et l'annotation de transciptomes ont été publiées, cependant, celle-ci sont plutôt dédiés à l'annotation des transcriptomes procaryotes (Mgnify – Mitchell *et al.* 2020) ou eucaryotes pluricellulaires (par exemple DRAP - Cabau *et al.* 2017, Seq2Fun - Liu *et al.* 2021, Banerjee *et al.* 2021) et profitent de l'existence de génomes de référence, qui, bien que pouvant appartenir à des espèces distantes, demeurent phylogénétiquement proche (animaux, plantes, champignons). L'une des difficultés que nous rencontrons dans le cadre du projet MICROSTORE, est que nous sommes amenés à caractériser des transcriptomes pouvant appartenir à des groupes taxonomiques sous-représentés dans les bases de données publiques, possiblement rares et représentatifs d'une diversité qui émergent à proximité de la racine de l'arbre des eucaryotes (Logares *et al.* 2021). Les approches basées sur la comparaison avec des références proches ne sont donc pas envisageables. Une chaîne de traitements pour l'assemblage et l'annotation de ces transcriptomes est donc en cours de développement (notamment à travers deux stages de M2 que j'encadre). Une première version de cette chaîne de traitements a été implémentée par Baptiste Milisavljevic (M2 2019-2020). Elle s'articule autour de 3 axes principaux, le contrôle qualité des lectures, ainsi que l'assemblage du transcriptome, la clusterisation des transcrits et leur traduction en protéines et l'annotation fonctionnelle des protéines prédites (Figure 3.2).

Après un contrôle qualité (FastQC), les lectures sont nettoyées avec Trimmomtic (Bolger *et al.* 2014). L'assemblage du transcriptome est réalisé avec Trinity v2.9.1 (Grabherr et.al 2011) à partir des lectures nettoyées. Une première étape de clusterisation avec MMseqs2 (Steinegger et Söding 2017) permet de réduire cette redondance liée au polymorphisme allélique, aux variants d'épissage ou à de possibles duplications (Vijay *et al.* 2013, Davidson et Oshlack 2014, Chang *et al.* 2015, Razo-Mendvil *et al.* 2020). Transdecoder (Haas *et al.* 2014) permet de prédire les structures codantes (CDS) sur les transcrits assemblés, qui sont ensuite traduits en protéines. Une seconde clusterisation est appliquée sur les protéines prédites.
L'annotation fonctionnelle des protéines prédite est réalisée d'une part à travers la recherche de domaines fonctionnels contre la base de données Pfam-A et d'autre par la recherche de séquences homologues dans la base de données UniRef90 via l'outil MMseqs2. L'association entre les accessions de Uniref90 et KEGG, calculée en amont permet, d'identifier les groupes d'orthologues de la base de données KEGG associés des protéines prédites affiliées à Uniref90 et de leur associer une voie métabolique avec l'outil MetaPathExplorer (Hochart et Debroas 2017). Cet outil fourni également des

images de cartes métaboliques dans lesquelles les groupes d'orthologue de KEGG correspondant à des protéines retrouvées dans le transcriptome sont explicitement représentés. L'implantation de KRYPTON sur le cluster de calcul du mésocentre de l'UCA est en cour de réalisation (travail réalisé par Anthony Auclair, M2 2020-2021).

KRYPTON a été testé sur trois transcriptomes de *Cholorophytes* extraits du projet MMEPTSP (Keeling *et al.* 2014). Ils ont été analysés à partir des lectures brutes (fichiers fastq) et de leurs assemblages, tel que produits dans leur publication de Keeling et collaborateurs (NCGR), ou réassemblés (DIB) via l'assembleur Trinity (Johnson *et al.* 2019).

La comparaison de la couverture relative des trois assemblages à l'échelle des transcrits est faible. Par exemple, pour la souche MMETSP1161 (Figure 3.3), les transcrits produits à l'issue de KRYPTON couvrent 23 % des CDS de l'assemblage NCGR et 44 % de l'assemblage DIB, alors même que ce dernier a été réalisé avec Trinity (v2.2.0). A l'échelle des protéines, les résultats sont plus encourageants, puisque les protéines retenues à l'issue du traitement des données par KRYPTON couvrent le deux tiers des protéines prédites à partir de l'assemblage NCGR et 56 % des protéines prédites à partir de l'assemblage DIB. Ces résultats sont en accord avec ceux de Jonhson et collaborateurs, qui soulignent que si l'assemblage DIB couvrent environ 71 % des transcrits NCGR, ces derniers ne couvrent pas plus de 34 % des transcrits de DIB (Johnson *et al.* 2019b). L'obtention de résultats divergents lors de l'assemblage de transcriptomes n'est pas rare dans la littérature (Voshall et Moriyana 2018, Salzberg 2019, Hölzer et Marz 2019, Banerjee *et al.* 2021), ces approches étant particulièrement sensibles à la nature de l'assembleur et à des paramètres utilisés. Les transcrits restitués sont par ailleurs très similaires entre eux, ce qui rend difficile la distinction des assemblages corrects parmi les artefacts (Banerjee *et al.* 2021). Parmi les améliorations envisagées sur la chaîne de traitement KRYPTON, on peut noter l'ajout des outils Bellerophon (Kerkvliet *et al.* 2019) et Conpacta (Razo-Mendivil *et al.* 2020). Le premier outil est spécifiquement dédié à la recherche et la suppression des chimères dans les assemblages *de novo* de transcriptomes. Le second permet de réduire la redondance des assemblages par une approche basée sur l'estimation de la part de lectures partagées par différents contigs et pourrait constituer une alternative à la clusterisation basée sur la similitude entre les séquences.

La caractérisation d'un transcriptome s'avère être une tache complexe pour ses aspects techniques, auquel il faut ajouter la dimension totalement nouvelle et exploratoire des données produites. En effet, des travaux antérieurs portant sur l'analyse de métatranscriptomes marins ont révélé que près de la moitié des transcrits produits n'avaient pas de correspondance dans les bases de données publiques, ce qui suggère qu'ils sont trop divergents pour permettre l'identification de leurs homologues potentiels ou qu'ils pourraient correspondre à de nouveaux gènes. Ces gènes pourraient conférer des fonctions qui différencient les lignées eucaryotes entre elles et avec les procaryotes (Van Etten et Bhattacharya 2020). Ainsi, au-delà de l'analyse des transcrits via KRYPTON et leur comparaison avec des bases de référence, il sera nécessaire d'envisager des traitements alternatifs pour explorer la part des données pour lesquelles il n'existe à l'heure actuelle aucune contrepartie décrite dans la littérature, telle que ceux proposés par Vanni et collaborateurs par exemple (Vanni *et al.* 2021).

1.2 – Perspectives : Phylogénomique des eucaryotes unicellulaires lacustres

Le domaine des eucaryotes englobe une très grande diversité, (estimée autour de 76 millions d'espèces (de Vargas *et al.* 2015)) portée pour l'essentiel par des organismes unicellulaires, dont la classification demeure un problème complexe. En effet, une grande partie de cette diversité est cryptique, au sens où des cellules morphologiquement indiscernables peuvent être séparées par une grande divergence moléculaire ou fonctionnelle (Keeling 2019). Ainsi, c'est à travers le développement de la phylogénomique et à l'étude de groupes de protistes clés pour l'évolution que la phylogénie des eucaryotes s'est affinée et enrichie ces 15 dernières années (Burki 2020). La phylogénomique est basée sur la concaténation de gènes cores partagés par l'ensemble des eucaryotes, pour tenter de dégager un signal commun à même de refléter cette diversité phylogénétique. Cependant, l'exploration de l'influence des gènes qui ne seraient pas partagés par l'ensemble des eucaryotes en complément du signal restitué par les gènes cores eucaryotes peut aider à déterminer plus précisément où et comment le signal phylogénétique est distribué (Brown *et al.* 2013, Shen *et al.* 2017). De manière plus fine encore, l'exploitation phylogénomique du pangénome dans certains groupes taxonomiques peut permettre d'apporter un éclairage fonctionnel et évolutif sur leur fonctionnement ou leur rôle potentiel dans l'environnement. Une telle analyse du pangénome des *Mamiellales* (12 500 gènes - Worden *et al.* 2009), a ainsi permis de définir un core génome (d'un peu plus de 7000 gènes) dans lequel il a été retrouvé des gènes impliqués dans les processus de la méiose. Ceci était inattendu puisque les *Mamiellales* sont considérés comme asexués, ce qui suggère que ces gènes étaient des vestiges de leur ancêtre ou qu'ils possèdent un type de sexualité qui n'a pas encore été décrit (Qui Richard cite-t-il dans pagenome euk 2020).

Dans ce contexte, les données de transcriptomes constituent un apport intéressant dans la caractérisation de la diversité phylogénétique des eucaryotes. En effet, les transcriptomes, sont dominés par des gènes de ménage fortement exprimés dans les cellules (Lahr *et al.* 2019, Lax *et al.* 2018), qui constituent une part importante des gènes utilisés dans les approches de phylogénomique (Burki 2020). Ainsi, on peut penser qu'à l'échelle d'un écosystème, les (méta)transcriptomes issus d'échantillons naturels devraient restituer une bonne couverture des données utilisées en phylogénomique et permettre une résolution fine de la diversité des eucaryotes propres à cet écosystème. Les travaux de l'équipe sur la caractérisation de clades eucaryotes lacustres à partir de données de *metabarcoding* (ARNr 18S) ont permis de définir des clades spécifiques des milieux lacustres au sein de différents taxa eucaryotes. Les données de transcriptomique et de métatranscriptomique produitent dans le cadre du projet MICROSTORE, devraient nous permettre d'accéder à une information phylogénétique plus fine pour caractériser ces clades. Ainsi, une de mes perspectives de recheche est de développer une approche phylogénomique sur les données de (meta)trancriptomique issues de MICROSTORE afin de proposer une résolution plus fine de ces clades. Ce travail nécessitera l'analyse préalable du potentiel génique des différentes clades, incluant la caractérisation de leur pangénome lorsque cela sera possible et l'étude de l'évolution des génomes propres à certains groupes.

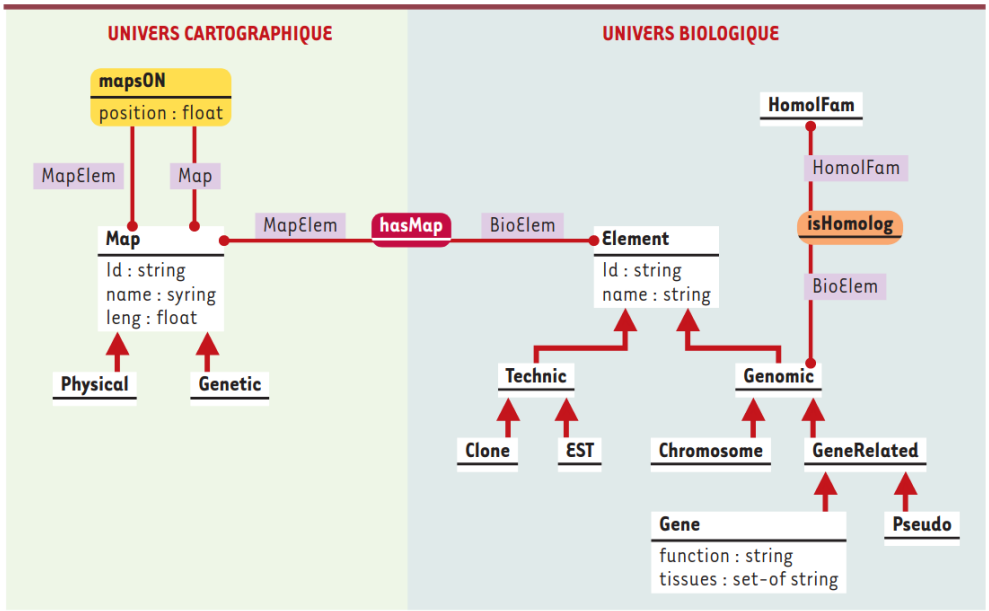


Figure 3.4 : Modélisation des connaissances sur la cartographie comparée : modèle conceptuel. Les entités manipulées en cartographie comparée sont de deux ordres, d'une part les cartes et marqueurs cartographiques et d'autre part les entités biologiques. Ces dernières peuvent être considérées selon leurs propriétés biologiques ou cartographiques lorsqu'elles sont manipulées en tant que marqueurs cartographiques (d'après Bronner *et al.* 2002b).

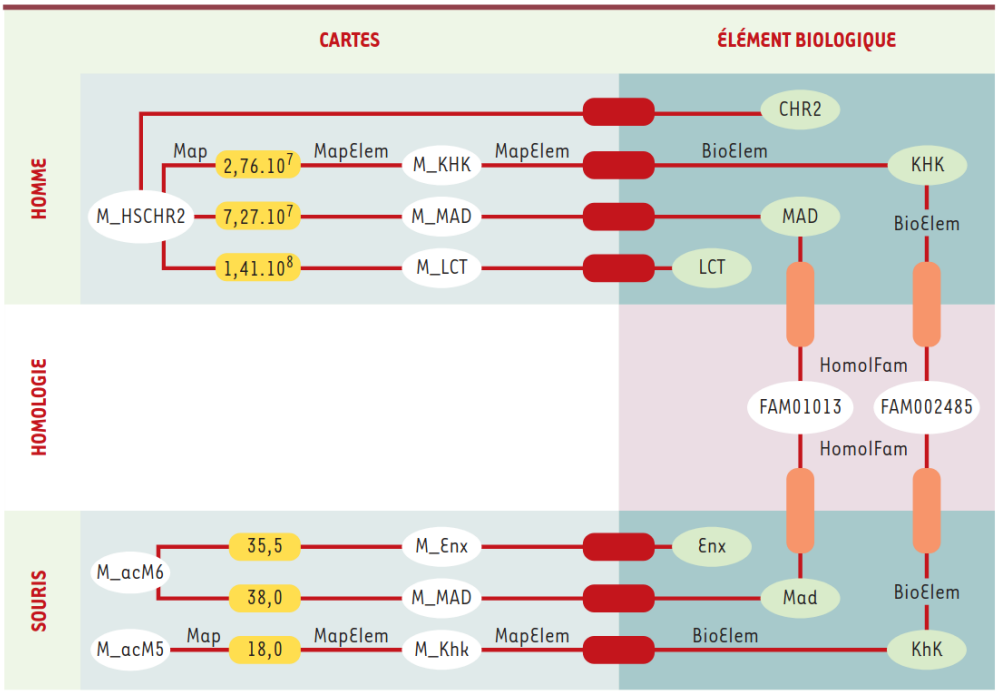


Figure 3.5 : Modélisation des connaissances sur la cartographie comparée : diagramme d'instances. Ce diagramme d'instance illustre l'organisation au sein du système GeM (genomic mapping) des informations relatives à la description de la carte comparée du chromosome 2 de l'homme avec la souris (d'après Bronner *et al.* 2002b).

2 – Représentation des connaissances en cartographie comparée des génomes de mammifères

La caractérisation des génomes de vertébrés a permis de constater la superposition de structures spatiales hétérogènes de différentes natures (fonctionnelles, compositionnelles, évolutives) et à différentes échelles (pour revue, voir Gautier 2000, Hurst et Eyre-Walker 2000) souvent conservées entre espèces. L'analyse conjointe de l'organisation spatiale des génomes entre espèces par des approches de cartographie comparée permet d'étudier les facteurs à l'origine de ces structures et du fonctionnement des génomes. Au delà de ces questions fondamentales, la cartographie comparée est un outil qui rend possible, sur la base de leur localisation génomique, le transfert d'informations entre espèces. Dans ce contexte, mon travail de thèse a consisté en la conception d'un système à base de connaissances destiné à la représentation de données plurispécifiques hétérogènes pour la cartographie comparée des génomes.

Une difficulté de la représentation des concepts de la biologie en termes informatiques réside dans leur grand nombre de facettes. En effet, ceux-ci peuvent souvent être considérés selon différents points de vue, un même terme pouvant recouvrir différents aspects, parfois non chevauchants, du concept. L'explicitation des objets biologiques représentés en termes informatiques dans le cadre d'une base de connaissances implique alors la définition d'une ontologie[1]. Celle qui a été développée ici se décline à travers trois hiérarchies de classes et six associations majeures (Figure 3.4). Les concepts de la biologie et de la génétique ont été représentés indépendamment de leur structure moléculaire. Ainsi, un « fragment génomique », entité « fondamentale » de la cartographie génomique, a été décrite à travers trois hiérarchies dont les racines sont *Element*, *Sequence* et *Map*. Cette modélisation présente l'avantage de distinguer les trois univers intégrés par la cartographie comparée et permet une modélisation spécifique de chacun d'eux. Les associations portent sur la description des relations d'homologie, d'appartenance d'un objet cartographique à une carte. Les relations spatiales entre les éléments d'une carte sont décrites à travers une adaptation de l'algèbre de Allen pour les cartes génomiques (Allen 1983, Schmeltzer 1995). Deux relations entre les classes *Element* et *Sequence* d'une part et *Element* et *Map* d'autre part garantissent l'intégrité du « fragment génomique » (Bronner *et al.* 2002a).

Une base de connaissances GeMCore reposant sur l'implémentation de ce modèle a également été réalisée au titre de preuve de concept (Figure 3.5). Elle inclue une procédure de mise à jour des données et d'une API (*Application Programing Interface*) java qui permet l'interrogation du système par d'autres programmes (Bronner *et al.* 2001). Ce travail à été poursuivi dans le cadre d'une thèse et à donné lieu à la réalisation d'une base de connaissances opérationnelle support à publications dans le domaine médical (Aouacheria *et al.* 2005, Aouacheria *et al.* 2006).

[1] En informatique, une ontologie est un modèle de travail décrivant les entités et les interactions d'un domaine (Stevens *et al.* 2000). Celle-ci doit être comprise par des individus et être interprétée de façon non ambigüe par des logiciels, c'est-à-dire qu'elle doit à la fois être décrite à travers le langage naturel et des formalismes de représentation adaptés à l'utilisateur et être exprimée à travers un modèle de données « interprétables » par un système informatique.

3 - Biodiversité et génomique végétale

J'ai intégré au 1er septembre 2003 l'unité Amélioration et Santé des Plantes (UMR 1095 UBP-INRA) au titre de Maître de conférence, dans l'équipe Résistance du tournesol aux pathogènes. Mes premiers travaux ont porté notamment sur l'annotation d'un fragment génomique de tournesol d'environ 112 kb lié à la résistance de tournesol au mildiou (*Plasmopara halstedii*) un pathogène eucaryote unicellulaire du clade des oomycète (Franchel *et al*. 2011).

Au sein de l'équipe, ou à travers des collaborations, j'ai également participé à des études associant l'organisation et la diversité de l'information génétique (les familles multigéniques en particulier) au sein des génomes à leur évolution (spécialisation) et leur expression, en relation avec les contraintes de l'environnement. Ainsi, dans le cadre d'une collaboration avec l'équipe MECA du laboratoire de Physique et Physiologie Intégrative de l'Arbre Fruitier et Forestier (PIAF ; UMR 547 UBP-INRA), nous avons caractérisé les signatures génomique et évolutive des facteurs de transcription à doigt de zinc C2H2 de type Q chez le peuplier et leur expression en réponse à divers stress mécaniques (blessure, stress salin, stress osmotique, gel). En effet ces facteurs de transcription jouent un rôle clé dans l'acclimatation des plantes tant à des stress biotiques qu'abiotiques (Ciftci-Yilmaz et Mittler 2008) bien que le rôle et les éléments régulant l'expression de ces différentes isoformes reste peu caractérisés. Il a pu être établi que les doigt de zinc C2H2 de type Q chez les plantes se divisaient en deux groupes monophylétiques dont les séquences se caractérisent par des signatures protéiques spécifiques. Cette séparation est antérieure à la divergence entre les monocotylédones et les dicotylédones, qui eux mêmes possèdent des signatures particulières au sein de chaque groupe. Il n'a pu être montré une quelconque spécificité d'expression entre les sous type de facteurs de transcription chez le peuplier et la nature du stress appliqué, ou l'organe ciblé (Gourcilleau *et al*. 2011).

Dans le cadre d'une collaboration avec Jean-Stéphane Venisse (UMR PIAF) nous nous sommes intéressés à une catégorie de protéines intrinsèques majeures non caractérisées fonctionnellement (protéines intrinsèques X – XIP). Celles-ci ont pu être classées phylogénétiquement en cinq groupes dont quatre correspondent à la divergence de taxons d'angiospermes. Les schémas complets d'expression des gènes de XIP chez le peuplier ont montré que seules deux isoformes étaient transcrites dans des tissus végétatifs avec cependant des schémas contrastés (Lopez *et al*. 2012).

Collaborations

- Catherine LENNE – Nathalie LEBLANC
- Jean-Stéphane VENISSE

4 - Big data, grille et complexité biologique

Les infrastructures de calcul distribué permettent de mobiliser les ressources de plusieurs ordinateurs au sein d'un réseau sur un problème nécessitant un grand nombre de cycles de calcul, ou l'accès à de grandes quantités de données. Les grilles de calculs notamment donnent accès à des ressources distantes de calcul et/ou de stockage via des protocoles et interfaces standardisés au travers d'une couche logicielle (intergiciel ou middleware) permettant un accès transparent pour les utilisateurs.

Le déploiement d'un traitement sur une grille passe par la parallélisation des processus lorsque le traitement des données implique de nombreux calculs indépendants, ou par la parallélisation des données lorsque le processus est divisé en plusieurs sous-tâches qui peuvent être traitées indépendamment. Alors que la parallélisation d'un grand nombre de calculs élémentaires dans le premier cas, ou le paquetage de données pour leur traitement parallèle (comme on le retrouve communément en sciences physiques) dans le second cas, sont relativement bien maîtrisées pour ce qui concerne l'adaptation des traitements aux grilles de calculs et au « big data », la gestion de chaînes de traitements complexes impliquant la mobilisation et l'interrogation de gros volumes de données constituent encore des points limitants dans le traitement des données complexes sur des infrastructures partagées (*big complexity*)

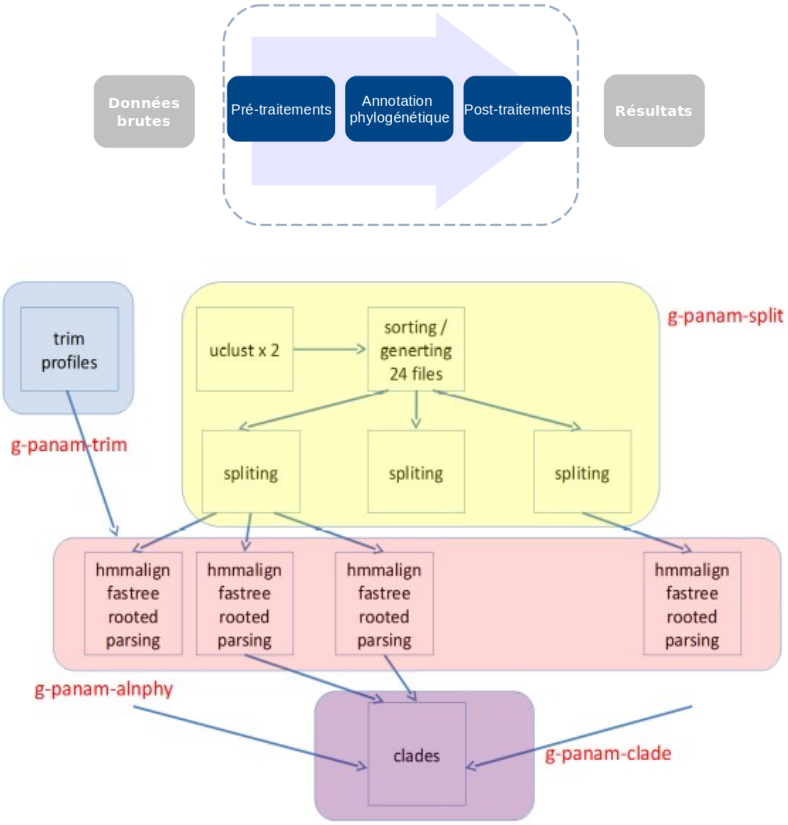


Figure 3.6 : Représentation graphique du découpage des tâches de PANAM en 4 services WPE (d'après Doan *et al.* 2013).

La chaîne de traitement développée dans PANAM s'est avérée tout à fait adaptée à l'investigation de cette notion de « big complexity » dans le contexte des grilles de calculs et a donné lieu à une collaboration avec l'équipe Plateforme de Calcul pour les Sciences du Vivant (PCSV) du Laboratoire de Physique Corpusculaire (LPC UMR 6533) ceci afin :

- d'identifier les limites des technologies dans le contexte d'une chaîne de traitement complexes
- d'évaluer les performances de PANAM sur la grille de calcul

Le portage de PANAM sur la grille de calcul BioMed a été réalisé en faisant le choix d'une parallélisation par les données, son déploiement sur la grille a été basé sur le logiciel # Wisdom Production Environment # (WPE, Breton *et al*., 2009).

Ces travaux ont permis de montrer l'intérêt des structures de calcul distribué pour de grands jeux de données, confirmant l'efficacité de la grille pour le traitement de données issues des NGS (Pour un jeu de 1 million d'OTU, le temps de traitement total nécessaire à PANAM après déploiement sur la grille est 12h 43min, contre 16 jours pour sa version non parallélisée).

Workflow	Total	g-panam-split	g-panam-alnphy	g-panam-clade
Test107s250k	5h32	3h15	0h20	2h17
Test88s500k	6h22	2h33	0h22	3h48
T53s750k	9h23	8h6	0h23	0h57
T81s1M	12h43	7h18	0h27	0h25

Tableau 3.1 : Temps moyens d'exécution des différents services de PANAM sur la grille de calcul Biomed pour différents jeux de séquences (d'après Doan 2012)

Ils ont également permis de pointer un certain nombre de points limitants, notamment à cause de la dernière étape de PANAM qui implique la fusion de l'ensemble des résultats précédemment obtenus. Ainsi, le traitement d'un jeu de données par PANAM dans l'environnement WPE sera conditionné par la durée de la tâche la plus longue. Si pour une raison ou une autre, l'une des tâches ne se termine pas, le traitement du jeu de données peut en théorie durer un temps infini. Ces observations ont permis de proposer des modifications du WPE qui en ont considérablement amélioré les performances (Doan *et al.* 2013).

Collaborations

- Vincent BRETON (LifeGrid 2012)

Références Bibliographiques

Achtman M, Wagner M. (2009) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol. Jun;6(6):431-40.

Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol. May;186(9):2629-35.

Albanese D, Donati C (2017). Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat Commun. 2017 Dec 22;8(1):2260.

Alberti A, Poulain J, Engelen S, Labadie K, Romac S, Ferrera I, Albini G, Aury JM, Belser *et al.* (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci Data. 2017 Aug 1;4:170093.

Allen, J. F. (1983): Maintaining Knowledge about Temporal Intervals. Comm of the ACM 26(11): 832-43.

Alneberg J, Karlsson CMG, Divne AM, Bergin C, Homa F, Lindh MV, Hugerth LW, Ettema TJG, Bertilsson S, Andersson AF, Pinhassi J (2018). Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. Microbiome. 2018 Sep 28;6(1):173.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. Sep 1;25(17):3389-402.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS One. Jul 27;4(7):e6372. doi: 10.1371/journal.pone.0006372. Erratum in: PLoS One. (12).

Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. (2017) Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems. Mar 7;2(2):e00191-16.

Andreani NA, Hesse E, Vos M. (2017) Prokaryote genome fluidity is dependent on effective population size. ISME J. Jul;11(7):1719-1721.

Aouacheria A, Navratil V, Barthelaix A, Mouchiroud D, Gautier C (2006). Bioinformatic screening of human ESTs for differentially expressed genes in normal and tumor tissues. BMC Genomics. 2006 Apr 26;7:94.

Aouacheria A, Navratil V, Wen W, Jiang M, Mouchiroud D, Gautier C, Gouy M, Zhang M (2005). In silico whole-genome scanning of cancer-associated nonsynonymous SNPs and molecular characterization of a dynein light chain tumour variant. Oncogene. 2005 Sep 8;24(40):6133-42.

Aparicio, S., *et al.* (2002) : Whole-Genome shotgun assembly and analysis of the genome of Fugu rubripes. Science, 297: 1301-1310.

Auguet JC, Triado-Margarit X, Nomokonova N, Camarero L,Casamayor EO (2012) Vertical segregation and phylogenetic characterization of ammonia-oxidizing Archaea in a deep oligotrophiclake. ISME J 6:1786–1797

Auguet JC, Casamayor EO (2013) Partitioning of Thaumarchaeota populations along environmental gradients in high mountain lakes.FEMS Microbiol Ecol 84:154–164

Avrani S, Wurtzel O, Sharon I, Sorek R, Lindell D. (2011) Genomic island variability facilitates Prochlorococcus-virus coexistence. Nature. Jun 29;474(7353):604-8.

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE (2004). Hotspots of mammalian chromosomal evolution. Genome Biol. 2004;5(4):R23. Epub 2004 Mar 8.

Banerjee S, Bhandary P, Woodhouse M, Sen TZ, Wise RP, Andorf CM. (2021) FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. BMC Bioinformatics. 2021 Apr 20;22(1):205.

Batut B, Knibbe C, Marais G, Daubin V. (2014) Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol. Dec;12(12):841-50.

Behnke A, Engel M, Christen R, Nebel M, Klein RR, Stoeck T. (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. Environ Microbiol. Feb;13(2):340-9.

Beman JM, Popp BN, Francis CA (2008) Molecular and biogeochemical evidence for ammonia oxidation by marine Crenarchaeota in the Gulf of California. ISME J 2:429–441

Bernhard AE, Tucker J, Giblin AE, Stahl DA (2007) Functionally distinct communities of ammonia-oxidizing bacteria along an estuarine salinity gradient. Environ Microbiol 9:1439–1447

Bernhard AE, Bollmann A (2010) Estuarine nitrifiers: new players,patterns and processes. Estuar Coast Shelf Sci 88:1–11

Berube PM, Biller SJ, Kent AG, Berta-Thompson JW, Roggensack SE, Roache-Johnson KH, Ackerman M, Moore LR, Meisel JD, Sher D, Thompson LR, Campbell L, Martiny AC, Chisholm SW. (2015) Physiology and evolution of nitrate acquisition in Prochlorococcus. ISME J. May;9(5):1195-207.

Berube PM, Biller SJ, Hackl T, Hogle SL, Satinsky BM, Becker JW, Braakman R, Collins SB, Kelly L, Berta-Thompson J, Coe A, Bergauer K, Bouman HA, Browning TJ, De Corte D, Hassler C, Hulata Y, Jacquot JE, Maas EW, Reinthaler T, Sintes E, Yokokawa T, Lindell D, Stepanauskas R, Chisholm SW. (2018) Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. Sci Data. Sep 4;5:180154.

Bhaya D, Grossman AR, Steunou AS, Khuri N, Cohan FM, Hamamura N, Melendrez MC, Bateson MM, Ward DM, Heidelberg JF. (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J. Dec;1(8):703-13.

Biller SJ, Berube PM, Berta-Thompson JW, Kelly L, Roggensack SE, Awad L, Roache-Johnson KH, Ding H, Giovannoni SJ, Rocap G, Moore LR, Chisholm SW. (2014) Genomes of diverse isolates of the marine cyanobacterium Prochlorococcus. Sci Data. Sep 30;1:140034.

Bik HM, Sung W, De Ley P, Baldwin JG, Sharma J, Rocha-Olivares A, Thomas WK. (2012) Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. Mol Ecol. Mar;21(5):1048-59.

Bobay LM, Ochman H (2017). The Evolution of Bacterial Genome Architecture. Front Genet. 2017 May 30;8:72.

Bobay LM, Ochman H. (2018) Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol. Oct 12;18(1):153.

Bolger AM, Lohse M, Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. Aug 1;30(15):2114-20.

Bolotin E, Hershberg R. (2016) Bacterial intra-species gene loss occurs in a largely clocklike manner mostly within a pool of less conserved and constrained genes. Sci Rep. Oct 13;6:35168.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, et. Al (2019). Reproducible, interactive, scalable and extensiblemicrobiome data science using QIIME 2. Nat Biotechnol. Aug;37(8):852-857.. Erratum in: Nat Biotechnol. 2019 Sep;37(9):1091

Bonder MJ, Abeln S, Zaura E, Brandt BW. (2012) Comparing clustering and pre-processing in taxonomy analysis. Bioinformatics. Nov 15;28(22):2891-7.

Botnen SS, Davey ML, Halvorsen R, Kauserud H. (2018) Sequence clustering threshold has little effect on the recovery of microbial community structure. Mol Ecol Resour. Apr 19.

Boussarie G, Bakker J, Wangensteen OS, Mariani S, Bonnin L, Juhel JB, Kiszka JJ, Kulbicki M, Manel S, Robbins WD, Vigliola L, Mouillot D. (2018) Environmental DNA illuminates the dark diversity of sharks. Sci Adv. 2;4(5):eaap9661.

Brandt MI, Trouche B, Quintric L, Günther B, Wincker P, Poulain J, Arnaud-Haond S. (2021) Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and

eukaryotic metabarcoding. Mol Ecol Resour. 2021 Apr 9. doi: 10.1111/1755-0998.13398. Epub ahead of print.

Bråte J, Logares R, Berney C, Ree DK, Klaveness D, Jakobsen KS, Shalchian-Tabrizi K. (2010) Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA. ISME J. Sep;4(9):1144-53.

Breton V, da Costa AL, de Vlieger P, Kim YM, Maigne L, Reuillon R, Sarramia D, Truong NH, Nguyen HQ, Kim D, et al (2009). Innovative in silico approaches to address avian u using grid technology. Infectious Disorders-Drug Targets, 9(3) :358365, 2009.

Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. (2019) The Ecology and Evolution of Pangenomes. Curr Biol. Oct 21;29(20):R1094-R1103.

Bronner G, Spataro B, Page M, Gautier C, Rechenmann F. (2002a) Modeling comparative mapping using objects and associations. Comput Chem. Jul;26(5):413-20.

Bronner G, Spataro B, Gautier C (2002b) Comparative genomic mapping in mammals. MS - Medecine Science. 18: 767-774

Brown MW, Sharpe SC, Silberman JD, Heiss AA, Lang BF, Simpson AG, Roger AJ. (2013) Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. Proc Biol Sci. Aug 28;280(1769):20131755.

Brown EA, Chain FJ, Crease TJ, MacIsaac HJ, Cristescu ME. (2015) Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? Ecol Evol. Jun;5(11):2234-51.

Burki F, Roger AJ, Brown MW, Simpson AGB. (2020) The New Tree of Eukaryotes. Trends Ecol Evol. Jan;35(1):43-55.

Cabau C, Escudié F, Djari A, Guiguen Y, Bobe J, Klopp C. (2017) Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. PeerJ. Feb 16;5:e2988.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. Jul;13(7):581-3.

Callahan BJ, McMurdie PJ, Holmes SP. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 11(12):2639-2643.

Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK. (2019) High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Res. 10;47(18):e103.

Campbell BJ, Yu L, Heidelberg JF, Kirchman DL. (2011) Activity of abundant and rare bacteria in a coastal ocean. Proc Natl Acad Sci U S A. Aug 2;108(31):12776-81.

Campbell BJ, Kirchman DL. (2013) Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. ISME J. Jan;7(1):210-20.

Capo E, Debroas D, Arnaud F, Guillemot T, Bichet V, Millet L, Gauthier E, Massa C, Develle AL, Pignol C, Lejzerowicz F, Domaizon I. (2016) Long-term dynamics in microbial eukaryotes communities: a palaeolimnological view based on sedimentary DNA. Mol Ecol. Dec;25(23):5925-5943.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Gonzalez Pena A *et al.* (2010) Qiime allows analysis of high-throughput community sequencing data. Nature methods, 7(5) :335336, 2010.

Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, Lima-Mendez G, Rocha F *et al.* (2018). A global ocean atlas of eukaryotic genes. Nat Commun. 2018 Jan 25;9(1):373.

Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, Moorthi SD, Dennett MR, Moran DM, Jones AC. (2009) Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. Appl Environ Microbiol. Sep;75(18):5797-808.

Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, Frischkorn KR, Tringe SG, Singh A, Markillie LM, Taylor RC, Williams KH, Banfield JF. (2015) Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr Biol. Mar 16;25(6):690-701.

Castelle CJ, Banfield JF. (2018) Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell. Mar 8;172(6):1181-1197.

Castillo-Ramírez S, Harris SR, Holden MT, He M, Parkhill J, Bentley SD, Feil EJ. (2011) The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog. Jul;7(7):e1002129.

Cavalier-Smith T. (1993) Kingdom protozoa and its 18 phyla. Microbiol Rev. Dec;57(4):953-94.

Cavender-Bares J, Kozak KH, Fine PV, Kembel SW. (200ç) The merging of community ecology and phylogenetic biology. Ecol Lett. Jul;12(7):693-715.

Chaffron S, Rehrauer H, Pernthaler J, von Mering C. (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res. Jul;20(7):947-59.

Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer CL, Huang X. (2015) Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. Feb 11;16(1):30.

Ciftci-Yilmaz S, Mittler R (2008). The zinc finger network of plants. Cell Mol Life Sci. 2008 Apr;65(7-8):1150-60.

Cohan FM (2001). Bacterial species and speciation. Syst Biol. 2001 Aug;50(4):513-24.

Cohan FM (2002). Sexual isolation and speciation in bacteria. In Genetics of Mate Choice : From Sexual Selection to Sexual Isolation, pages 359370. Springer, 2002.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, Chisholm SW. (2006) Genomic islands and the ecology and evolution of Prochlorococcus. Science. Mar 24;311(5768):1768-70.

Coleman ML, Chisholm SW. (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A. Oct 26;107(43):18634-9.

Cordero OX, Polz MF. (2014) Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol. Apr;12(4):263-73.

Costantini M, Musto H (2017). The Isochores as a Fundamental Level of Genome Structure and Organization: A General Overview. J Mol Evol. 2017 Mar;84(2-3):93-103

da Silva Filho AC, Raittz RT, Guizelini D, De Pierri CR, Augusto DW, Dos Santos-Weiss ICR, Marchaukoski JN. (2018) Comparative Analysis of Genomic Island Prediction Tools. Front Genet. Dec 12;9:619.

Daubin V, Gouy M, Perrière G. (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. Jul;12(7):1080-90.

Davidson NM, Oshlack A. (2014) Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol. Jul 26;15(7):410.

De Queiroz K. (2007) Species concepts and species delimitation. Syst Biol. Dec;56(6):879-86.

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury JM, Bittner L, Chaffron S, Dunthorn M, Engelen S, Flegontova O, Guidi L, Horák A, Jaillon O, Lima-Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S; Tara Oceans Coordinators, Acinas SG, Bork P, Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E. (2015) Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science. May 22;348(6237):1261605.

Debroas D, Humbert JF, Enault F, Bronner G, Faubladier M, Cornillot E. (2009) Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget--France). Environ Microbiol. Sep;11(9):2412-24.

Debroas D, Hugoni M, Domaizon I. (2015) Evidence for an active rare biosphere within freshwater protists community. Mol Ecol. Mar;24(6):1236-47.

Debroas D, Domaizon I, Humbert JF, Jardillier L, Lepère C, Oudart A, Taïb N (2017). Overview of freshwater microbial eukaryotes diversity: a first analysis of publicly available metabarcoding data. FEMS Microbiol Ecol. 2017 Apr 1;93(4).

del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I. (2014) The others: our biased perspective of eukaryotic genomes. Trends Ecol Evol. May;29(5):252-9.

Didelot X, Méric G, Falush D, Darling AE. (2012) Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC Genomics. 2012 Jun 19;13:256.

Doan TT Breton V, Bronner G (2013). Experience of porting a metagenomics pipeline on the grid using DIRAC framework. EGI Technical Forum 2013. 16-20 Septembre 2013, Madrid

Dobrindt U, Hochhut B, Hentschel U, Hacker J. (2004) Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol. May;2(5):414-24.

Domingo-Sananes MR, McInerney JO. (2021) Mechanisms That Shape Microbial Pangenomes. Trends Microbiol. Jan 7:S0966-842X(20)30321-8.

Doolittle WF. (1999) Phylogenetic classification and the universal tree. Science. Jun 25;284(5423):2124-9.

Doolittle WF. (2012) Population genomics: how bacterial species form and why they don't exist. Curr Biol. Jun 5;22(11):R451-3.

Ducklow, H. (2008). Microbial services: Challenges for microbial ecologists in a changing world. Aquatic Microbial Ecology. 53, 13–19.

Dunthorn M, Klier J, Bunge J, Stoeck T. (2012) Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. J Eukaryot Microbiol. Mar-Apr;59(2):185-7.

Eddy SR. (1998) Profile hidden Markov models. Bioinformatics.(9):755-63.

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct 1;26(19):2460-1.

Edgar RC (2016) UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. October bioRxiv /doi.org/10.1101/081257

Elbrecht V, Vamos EE, Steinke D, Leese F. (2018) Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ. Apr 9;6:e4644.

Ellegren H, Galtier N. (2016) Determinants of genetic diversity. Nat Rev Genet. Jul;17(7):422-33.

Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG (2017). Archaea and the origin of eukaryotes. Nat Rev Microbiol. 2017 Nov 10;15(12):711-723.

Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. (2015) Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J. Mar 17;9(4):968-79.

Escalas A, Hale L, Voordeckers JW, Yang Y, Firestone MK, Alvarez-Cohen L, Zhou J. (2019) Microbial functional diversity: From concepts to applications. Ecol Evol. Oct 2;9(20):12000-12016.

Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, Maman S, Hernandez-Raquet G, Combes S, Pascal G. (2018) FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics. Apr 15;34(8):1287-1294.

Faith, DP. (1992) Conservation evaluation and phylogenetic diversity. Biol. Conserv.61, 1–10.

Falkowski PG, Fenchel T, Delong EF. (2008) The microbial engines that drive Earth's biogeochemical cycles. Science. May 23;320(5879):1034-9.

Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jiao N, Karl DM, Li WK, Lomas MW, Veneziano D, Vera CS, Vrugt JA, Martiny AC. (2013) Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. Proc Natl Acad Sci U S A. Jun 11;110(24):9824-9.

Forster D, Lentendu G, Filker S, Dubois E, Wilding TA, Stoeck T. (2019) Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. Environ Microbiol. Nov;21(11):4109-4124.

Franchel J, Bouzidi MF, Bronner G, Vear F, Nicolas P, Mouzeyar S. (2013) Positional cloning of a candidate gene for resistance to the sunflower downy mildew, Plasmopara halstedii race 300. Theor Appl Genet. Feb;126(2):359-67.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. (2009) The bacterial species challenge: making sense of genetic and ecological diversity. Science. Feb 6;323(5915):741-6.

Freudenstein JV, Broe MB, Folk RA, Sinn BT. (2017) Biodiversity and the Species Concept-Lineages are not Enough. Syst Biol. Jul 1;66(4):644-656.

Frøslev TG, Kjøller R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, Hansen AJ. (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nat Commun. Oct 30;8(1):1188.

Gaidos E, Rusch A, Ilardo M. (2011) Ribosomal tag pyrosequencing of DNA and RNA from benthic coral reef microbiota: community spatial structure, rare members and nitrogen-cycling guilds. Environ Microbiol. May;13(5):1138-52.

Galand PE, Gutierrez-Provecho C, Massana R, Gasol J,Casamayor EO (2010) Inter-annual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea. Limnol Oceanogr 55:2117–2125

Gal-Mor O, Finlay BB. (2006) Pathogenicity islands: a molecular toolbox for bacterial virulence. Cell Microbiol. Nov;8(11):1707-19.

Gardon H, Biderre-Petit C, Jouan-Dufournel I, Bronner G. (2020) A drift-barrier model drives the genomic landscape of a structured bacterial population. Mol Ecol. Nov;29(21):4143-4156.

Gautier, C. (2000): Compositional bias in DNA. Curr Opin Genet Dev 10(6): 656-61.

Giller PS, Hillebrand H, Berninger UG, Gessner MO, Hawkins S, Inchausti P, Inglis C, Leslie H, Malmqvist B, Monaghan MT, Morin PJ, , O'Mullan G (2004). Biodiversity effects on ecosystem functioning: Emerging issues and their experimental test in aquatic environments. Oikos, 104(3), 423–436.

Good, B.H., McDonald, M.J., Barrick, J.E., Lenski, R.E., and Desai, M.M. (2017) The dynamics of molecular evolution over 60,000 generations. *Nature* 551: 45–50.

Goodwin S, McPherson JD, McCombie WR. (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 17;17(6):333-51.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. Jan;57(Pt 1):81-91.

Gourcilleau D, Lenne C, Armenise C, Moulia B, Julien JL, Bronner G, Leblanc-Fournier N. (2011) Phylogenetic study of plant Q-type C2H2 zinc finger proteins and expression analysis of poplar genes in response to osmotic, cold and mechanical stresses. DNA Res. Apr;18(2):77-92.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. May 15;29(7):644-52.

Grossart HP, Wurzbacher C, James TY, Kagami M (2016) Discovery of dark matter fungi in aquatic ecosystems demands a reappraisal of the phylogeny and ecology of zoosporic fungi. Fungal Ecology 19:28-38.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. Aug;8(8):1494-512.

Hacker J, Carniel E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep. May;2(5):376-81.

Hamady M and Knight R (2009). Microbial community proling for human microbiome projects : Tools, techniques, and challenges. Genome research, 19(7) :11411152, 2009.

Hanage WP. (2016) Not So Simple After All: Bacteria, Their Population Genetics, and Recombination. Cold Spring Harb Perspect Biol. Jul 1;8(7):a018069.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. Oct;5(10):R245-9.

Hao W, Golding GB. (2006) The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res. May;16(5):636-43.

Hao W, Golding GB. (2010) Inferring bacterial genome flux while considering truncated genes. Genetics. Sep;186(1):411-26.

He Y, Caporaso JG, Jiang XT, Sheng HF, Huse SM, Rideout JR, Edgar RC, Kopylova E, Walters WA, Knight R, Zhou HW. (2015) Stability of operational taxonomic units: a important but neglected property for analyzing microbial diversity. Microbiome. 20;3:20. . Erratum in: Microbiome. 2015;3:34.

Heeger F, Bourne EC, Baschien C, Yurkov A, Bunk B, Spröer C, Overmann J, Mazzoni CJ, Monaghan MT. (2018) Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. Mol Ecol Resour. Nov;18(6):1500-1514.

Hill WG, Robertson A. (1966) The effect of linkage on limits to artificial selection. Genet Res. Dec;8(3):269-94.

Hochart, C and Debroas, D (2017) MetaPath Explorer: predicting and visualizing metabolic functions from high-throughput sequencing data. Available online at:https://github.com/meb-team/MetaPathExplorer

Hölzer M, Marz M. (2019) De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience. May 1;8(5):giz039.

Hooper SD, Mavromatis K, Kyrpides NC. (2009) Microbial co-habitation and lateral gene transfer: what transposases can tell us. Genome Biol. ;10(4):R45.

Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, Andersson AF (2015). Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol. 2015 Dec 14;16:279.

Hugoni M (2013) Structure et activité des Archaea planctoniques dans les écosystèmes aquatiques. Thèse de doctorat : Ecologie Microbienne : Clermont-Ferrand 2

Hugoni M, Etien S, Bourges A, Lepere C, Domaizon I et al(2013 a) Dynamics of ammonia-oxidizing Archaea andBacteria in contrasted freshwater ecosystems. Res Microbiol164:360–370

Hugoni M, Taib N, Debroas D, Domaizon I, Jouan Dufournel I, Bronner G, Salter I, Agogué H, Mary I, Galand PE. (2013 b) Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. Proc Natl Acad Sci U S A. Apr 9;110(15):6004-9.

Hugoni M, Agogué H, Taib N, Domaizon I, Moné A, Galand PE, Bronner G, Debroas D, Mary I. (2015) Temporal Dynamics of Active Prokaryotic Nitrifiers and Archaeal Communities from River to Sea. Microb Ecol. Aug;70(2):473-83.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. May 23;320(5879):1081-5.

Hurst, L. D. and A. Eyre-walker (2000): Evolutionary genomics: reading the bands. Bioessays 22(2): 105-107.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8(7):R143.

Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Relman DA, Sogin ML. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. 2008 Nov;4(11):e1000255.

Huse SM, Welch DM, Morrison HG, Sogin ML. (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol. 2010 Jul;12(7):1889-98

Jaob F, Perrin D, Sanchez C, Monod J. (1960) L'Opéron: groupe de gènes à expression coordonnées par un opérateur C R Hebd Seances Acad Sci. Feb 29;250:1727-9.

Jamy M, Foster R, Barbera P, Czech L, Kozlov A, Stamatakis A, Bending G, Hilton S, Bass D, Burki F. (2020) Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. Mol Ecol Resour. Mar;20(2):429-443.

Jobard M, Wawrzyniak I, Bronner G, Marie D, Vellet A, Sime-Ngando T, Debroas D, Lepère C (2020) Freshwater Perkinsea: diversity, ecology and genomic information, Journal of Plankton Research,. 42(1): 3–17

Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, Leopold SR, Hanson BM, Agresta HO, Gerstein M, Sodergren E, Weinstock GM. (2019a) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019 6;10(1):5029.

Johnson LK, Alexander H, Brown CT. (2019b) Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. Gigascience. Apr 1;8(4):giy158.

Jones SE, Lennon JT. (2010) Dormancy contributes to the maintenance of microbial diversity. Proc Natl Acad Sci U S A. Mar 30;107(13):5881-6.

Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, Sullivan M, Arendt D, Benzoni F, Claverie JM, Follows M, Gorsky G, Hingamp P, Iudicone D, Jaillon O, Kandels-Lewis S, Krzic U, Not F, Ogata H, Pesant S, Reynaud EG, Sardet C, Sieracki ME, Speich S, Velayoudon D, Weissenbach J, Wincker P; Tara Oceans Consortium. (2011) A holistic approach to marine eco-systems biology. PLoS Biol. Oct;9(10):e1001177

Kashtan N, Roggensack SE, Berta-Thompson JW, Grinberg M, Stepanauskas R, Chisholm SW (2017). Fundamental differences in diversity and genomic population structure between Atlantic and Pacific Prochlorococcus. ISME J. 2017 Sep;11(9):1997-2011.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 2014 Apr 25;344(6182):416-20.

Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, Sullivan M, Arendt D, Benzoni F, Claverie Jm *et al.* (2011) Tara Oceans Consortium. A holistic approach to marine eco-systems biology. PLoS Biol. 2011 Oct;9(10):e1001177.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM *et al.* (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. Jun 24;12(6):e1001889.

Keeling PJ, Burki F. (2019) Progress towards the Tree of Eukaryotes. Curr Biol. Aug 19;29(16):R808-R817.

Keeling PJ. (2019) Combining morphology, behaviour and genomics to understand the evolution and ecology of microbial eukaryotes. Philos Trans R Soc Lond B Biol Sci. Nov 25;374(1786):20190085.

Kent AG, Dupont CL, Yooseph S, Martiny AC. (2016) Global biogeography of Prochlorococcus genome diversity in the surface ocean. ISME J. Aug;10(8):1856-65.

Kerkvliet J, de Fouchier A, van Wijk M, Groot AT. (2019) The Bellerophon pipeline, improving de novo transcriptomes and removing chimeras. Ecol Evol. Aug 17;9(18):10513-10521.

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW. (2007) Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. PLoS Genet. Dec;3(12):e231.

Kimura, M. (1983) The Neutral Theory of Molecular Evolution. Cambridge Univ. Press, pp 367

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, Nevo E, Cohan FM. (2008) Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci U S A. Feb 19;105(7):2504-9.

Koeppel AF and Wu M (2013). Surprisingly extensive mixed phylogenetic and ecological signals among bacterial operational taxonomic units. Nucleic acids research, 2013.

Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA.,(2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature. Sep,22;437(7058):543-6.

Konstantinidis KT, Tiedje JM. (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. Feb 15;102(7):2567-72.

Koskella B, Hall LJ, Metcalf CJE. (2017) The microbiome beyond the horizon of ecological and evolutionary theory. Nat Ecol Evol. Nov;1(11):1606-1615.

Koonin EV, Wolf YI. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. Dec;36(21):6688-719.

Koonin, E.V., and Wolf, Y.I. (2009) Is evolution Darwinian or/and Lamarckian? *Biol Direct* 4: 42.

Koonin EV, Makarova KS, Wolf YI. (2021) Evolution of Microbial Genomics: Conceptual Shifts over a Quarter Century. Trends Microbiol. Feb 1:S0966-842X(21)00007-X.

Krogh TJ, Møller-Jensen J, Kaleta C. (2018) Impact of Chromosomal Architecture on the Function and Evolution of Bacterial Genomes. Front Microbiol. Aug 27;9:2019.

Kuchina A, Brettner LM, Paleologu L, Roco CM, Rosenberg AB, Carignano A, Kibler R, Hirano M, DePaolo RW, Seelig G. (2021) Microbial single-cell RNA sequencing by split-pool barcoding. Science. Feb 19;371(6531):eaba5257.

Kumar N, Lad G, Giuntini E, Kaye ME, Udomwong P, Shamsani NJ, Young JP, Bailly X. (2015) Bacterial genospecies that are not ecologically coherent: population genomics of Rhizobium leguminosarum. Open Biol. Jan;5(1):140133.

Kunin V et Hugenholtz P (2010). Pyrotagger : A fast, accurate pipeline for analysis of rrna amplicon pyrosequence data. The Open Journal, 1(1), 2010.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ Microbiol. 2010 Jan;12(1):118-23.

Kuo CH, Ochman H. D(2009) Deletional bias across the three domains of life. Genome Biol Evol. Jun 27;1:145-52.

Lahr DJG, Kosakyan A, Lara E, Mitchell EAD, Morais L, Porfirio-Sousa AL, Ribeiro GM, Tice AK, Pánek T, Kang S, Brown MW. (2019) Phylogenomics and Morphological Reconstruction of Arcellinida Testate Amoebae Highlight Diversity of Microbial Eukaryotes in the Neoproterozoic. Curr Biol. Mar 18;29(6):991-1001.e3.

Lapierre P, Gogarten JP. (2009) Estimating the size of the bacterial pan-genome. Trends Genet. Mar;25(3):107-10.

Larkin AA, Blinebry SK, Howes C, Lin Y, Loftus SE, Schmaus CA, Zinser ER, Johnson ZI. (2016) Niche partitioning and biogeography of high light adapted Prochlorococcus across taxonomic ranks in the North Pacific. ISME J. Jul;10(7):1555-67.

Lax G, Eglit Y, Eme L, Bertrand EM, Roger AJ, Simpson AGB. (2018) Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. Nature. Dec;564(7736):410-414.

Lecointre et Le Guyader, (2017) Classification phylogénétique du Vivant 4e édition, éditions Belin 2017 pp 584)

Lefranc M, Thénot A, Lepère C, Debroas D. (2005) Genetic diversity of small eukaryotes in lakes differing by their trophic status. Appl Environ Microbiol. Oct;71(10):5935-42.

Lehours AC, Bardot C, Thenot A, Debroas D, Fonty G (2005). Anaerobic microbial communities in Lake Pavin, a unique meromictic lake in France. Appl Environ Microbiol. 2005 Nov;71(11):7389-400.

Lennon JT, Jones SE. (2011) Microbial seed banks: the ecological and evolutionary implications of dormancy. Nat Rev Microbiol. Feb;9(2):119-30.

Lepère C, Boucher D, Jardillier L, Domaizon I, Debroas D. (2006) Succession and regulation factors of small eukaryote community composition in a lacustrine ecosystem (Lake Pavin). Appl Environ Microbiol. Apr;72(4):2971-81.

Lepère C, Domaizon I, Debroas D. (2008) Unexpected importance of potential parasites in the composition of the freshwater small-eukaryote community. Appl Environ Microbiol. May;74(10):2940-9.

Lepère C, Masquelier S, Mangot JF, Debroas D, Domaizon I. (2010) Vertical structure of small eukaryotes in three lakes that differ by their trophic status: a quantitative approach. ISME J. Dec;4(12):1509-19.

Lepère C, Domaizon I, Hugoni M, Vellet A, Debroas D. (2016) Diversity and Dynamics of Active Small Microbial Eukaryotes in the Anoxic Zone of a Freshwater Meromictic Lake (Pavin, France). Front Microbiol. Feb 10;7:130.

Lepelletier F, Karpov SA, Le Panse S, Bigeard E, Skovgaard A, Jeanthon C, Guillou L. (2014) Parvilucifera rostrata sp. nov. (Perkinsozoa), a novel parasitoid that infects planktonic dinoflagellates. Protist. Jan;165(1):31-49.

Li W, Godzik A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. Jul 1;22(13):1658-9.

Li S, Bronner G, Lepère C, Kong F, Shi X. (2017) Temporal and spatial variations in the composition of freshwater photosynthetic picoeukaryotes revealed by MiSeq sequencing from flow cytometry sorted samples. Environ Microbiol. Jun;19(6):2286-2300.

Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. 35(18):e120.

Liu Z, DeSantis TZ, Andersen GL, Knight R. (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. 36(18):e120.

Liu Y, Jeraldo P, Jang JS, Eckloff B, Jen J, Walther-Antonio M. (2019) Bacterial Single Cell Whole Transcriptome Amplification in Microfluidic Platform Shows Putative Gene Expression Heterogeneity. Anal Chem. Jul 2;91(13):8036-8044.

Liu P, Ewald J, Galvez JH, Head J, Crump D, Bourque G, Basu N, Xia J.(2021) Ultrafast functional profiling of RNA-seq data for nonmodel organisms. Genome Res. Apr;31(4):713-720.

Lopez D, Bronner G, Brunel N, Auguin D, Bourgerie S, Brignolas F, Carpin S, Tournaire-Roux C, Maurel C, Fumanal B, Martin F, Sakr S, Label P, Julien JL, Gousset-Dupont A, Venisse JS. (2012) Insights into Populus XIP aquaporins: evolutionary expansion, protein functionality, and environmental regulation. J Exp Bot. Mar;63(5):2217-30.

López-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D. (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature. Feb 1;409(6820):603-7.

López-Pérez M, Martin-Cuadrado AB, Rodriguez-Valera F. (2014) Homologous recombination is involved in the diversity of replacement flexible genomic islands in aquatic prokaryotes. Front Genet. May 22;5:147.

Rodriguez-Valera F, Martin-Cuadrado AB, López-Pérez M. (2016) Flexible genomic islands as drivers of genome evolution. Curr Opin Microbiol. Jun;31:154-160.

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D (2015). ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol. 2015 Oct;33(10):1045-52.

Mackin JG, Owen HM, Collier A. (1950) Preliminary Note on the Occurrence of a New Protistan Parasite, Dermocystidium marinum n. sp. in Crassostrea virginica (Gmelin). Science. Mar 31;111(2883):328-9.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2014) Swarm: robust and fast clustering method for amplicon-based studies. PeerJ. 25;2:e593.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2015) Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ. 10;3:e1420.

Malmstrom RR, Coe A, Kettler GC, Martiny AC, Frias-Lopez J, Zinser ER, Chisholm SW. (2010) Temporal dynamics of Prochlorococcus ecotypes in the Atlantic and Pacific oceans. ISME J. Oct;4(10):1252-64.

Mangot JF, Lepère C, Bouvier C, Debroas D, Domaizon I. (2009) Community structure and dynamics of small eukaryotes targeted by new oligonucleotide probes: new insight into the lacustrine microbial food web. Appl Environ Microbiol. Oct;75(19):6373-81.

Mangot JF, Debroas D, Domaizon I (2011) Perkinsozoa, a well-known marine protozoan flagellate parasite group, newly identified in lacustrine systems: a review. Hydrobiologia 659(1):37-48

Mangot JF, Domaizon I, Taib N, Marouni N, Duffaud E, Bronner G, Debroas D. (2013) Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. Environ Microbiol. Jun;15(6):1745-58.

Mangot JF, Logares R, Sánchez P, Latorre F, Seeleuthner Y, Mondy S, Sieracki ME, Jaillon O, Wincker P, Vargas C, Massana R (2017). Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. Sci Rep. 2017 Jan 27;7:41498.

Mani K, Taib N, Hugoni M, Bronner G, Bragança JM, Debroas D. (2020) Transient Dynamics of Archaea and Bacteria in Sediments and Brine Across a Salinity Gradient in a Solar Saltern of Goa, India. Front Microbiol. Aug 13;11:1891.

Marin B, Melkonian M. (2010) Molecular phylogeny and classification of the Mamiellophyceae class. nov. (Chlorophyta) based on sequence comparisons of the nuclear- and plastid-encoded rRNA operons. Protist. Apr;161(2):304-36.

Marri PR, Hao W, Golding GB. (2006) Gene gain and gene loss in streptococcus: is it driven by habitat? Mol Biol Evol. Dec;23(12):2379-91.

Marttinen P, Croucher NJ, Gutmann MU, Corander J, Hanage WP. (2015) Recombination produces coherent bacterial species clusters in both core and accessory genomes. Microb Genom. Nov 5;1(5):e000038.

Mayr E. (1942) Systematics and the origin of species, from the viewpoint of a zoologist . Number 13. Harvard University Press.

Mayr, E. (1942). Systematics and the origin of species. Annals of the Entomological Society of America, 36(1), 138–139.

McInerney J, Pisani D, O'Connell MJ (2015). The ring of life hypothesis for eukaryote origins is supported by multiple kinds of data. Philos Trans R Soc Lond B Biol Sci. 2015 Sep 26;370(1678):20140323.

McInerney JO, McNally A, O'Connell MJ (2017a). Why prokaryotes have pangenomes. Nat Microbiol. 2017 Mar 28;2:17040.

McInerney JO, McNally A, O'Connell MJ (2017b). Reply to 'The population genetics of pangenomes'. Nat Microbiol. 2017 Dec;2(12):1575.

Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A. (1991) Evidence for horizontal gene transfer in Escherichia coli speciation. J Mol Biol. Dec 20;222(4):851-6.

Medini D, Donati C, Rappuoli R, Tettelin H. (2020) The Pangenome: A Data-Driven Discovery in Biology. 2020 May 1. In: Tettelin H, Medini D, editors. The Pangenome: Diversity, Dynamics and Evolution of Genomes [Internet]. Cham (CH): Springer; 2020.

Messer, P.W., Ellner, S.P., and Hairston, N.G. (2016) Can population genetics adapt to rapid evolution? *Trends Genet* 32L: 408–418.

Milisavljevic B (2020) Assemblage et annotation fonctionnelle de transcriptomes d'eucaryotes microbiens des écosystèmes lacustres. Rapport de stage de M2 Bioinformatique – Université Clermont Auvergne.

Mincer TJ, Church MJ, Taylor LT, Preston C, Karl DM, DeLong EF. (2007) Quantitative distribution of presumptive archaeal and bacterial nitrifiers in Monterey Bay and the North Pacific Subtropical Gyre. Environ Microbiol. May;9(5):1162-75.

Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR, Kale V, Potter SC, Richardson LJ, Sakharova E, Scheremetjew M, Korobeynikov A, Shlemov A, Kunyavskaya O, Lapidus A, Finn RD. (2020) MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res. Jan 8;48(D1):D570-D578.

Mitri S (2019) : The Evolutionary Ecology of Microbes. Encyclopedia of Microbiology (Fourth Edition), Editor(s): Thomas M. Schmidt, Academic Press, Pages 416-422, ISBN 9780128117378,

Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature. Dec 19-26;420(6917):806-10.

Moore LR, Rocap G, Chisholm SW. (1998) Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. Nature. Jun 4;393(6684):464-7.

Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F *et al.* (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science. Jul 22;309(5734):613-7.

Mysara M, Vandamme P, Props R, Kerckhof FM, Leys N, Boon N, Raes J, Monsieurs P. (2017) Reconciliation between operational taxonomic units and species boundaries. FEMS Microbiol Ecol. Apr 1;93(4):fix029.

Nadeau, J. H. and Sankoff, D. (1998): Counting on comparative maps. Trends in Genetics 14 (12): 495-501.

National Research Council (US) Committee on Noneconomic and Economic Value of Biodiversity. (1999) Perspectives on Biodiversity: Valuing Its Role in an Everchanging World. Washington (DC): National Academies Press (US). 168 p.

Nakahara S, Zacca T, Huertas B, Neild AFE, Hall JPW, Lamas G, Holian, LA, Espeland M, Willmott, KR. (2018) Remarkable sexual dimorphism, rarity and cryptic species: a revision of the '*aegrota* species group' of the Neotropical butterfly genus *Caeruleuptychia* with the description of three new species (Lepidoptera, Nymphalidae, Satyrinae). Insect Systematics & Evolution. Apr 49(2):130-182.

Nearing JT, Douglas GM, Comeau AM, Langille MGI. (2018) Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ. Aug 8;6:e5364.

Not F, del Campo J, Balagué V, de Vargas C, Massana R. (2009) New insights into the diversity of marine picoeukaryotes. PLoS One. Sep 29;4(9):e7143.

Nowell RW, Green S, Laue BE, Sharp PM. (2014) The extent of genome flux and its role in the differentiation of bacterial lineages. Genome Biol Evol. Jun 12;6(6):1514-29.

O'Brien, S. J., M. Menotti-Raymond, W. J. Murphy, W. G. Nash, J. Wienberg, R. Stanyon, N. G. Copeland, N. A. Jenkins, J. E. Womack and J. A. Marshall Graves (1999): The promise of comparative genomics in mammals. Science 286(5439): 458-62, 479-81.

Ochman H, Soncini FC, Solomon F, Groisman EA. (1996) Identification of a pathogenicity island required for Salmonella survival in host cells. Proc Natl Acad Sci U S A. Jul 23;93(15):7800-4.

Ochman H, Lawrence JG, Groisman EA. (2000) Lateral gene transfer and the nature of bacterial innovation. Nature. May 18;405(6784):299-304.

Okazaki Y, Fujinaga S, Salcher MM, Callieri C, Tanaka A, Kohzu A, Oyagi H, Tamaki H, Nakano SI. (2021) Microdiversity and phylogeographic diversification of bacterioplankton in pelagic freshwater systems revealed through long-read amplicon sequencing. Microbiome. Jan 22;9(1):24.

Oliveira PH, Touchon M, Cury J, Rocha EPC. (2017) The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun. Oct 10;8(1):841.

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. 2017 Nov;2(11):1533-1542.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P (2017). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018 Nov;36(10):996-1004.

Parmley JL, Hurst LD. (2007) How common are intragene windows with KA > KS owing to purifying selection on synonymous mutations? J Mol Evol. Jun;64(6):646-55.

Partensky F, Hess WR, Vaulot D. (1999) Prochlorococcus, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev. Mar;63(1):106-27.

Partridge SR, Kwong SM, Firth N, Jensen SO. (2018) Mobile Genetic Elements Associated with Antimicrobial Resistance. Clin Microbiol Rev. Aug 1;31(4):e00088-17.

Pedrós-Alió C. (2006) Marine microbial diversity: can it be determined? Trends Microbiol. Jun;14(6):257-63.

Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR. (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature. Jan 25;409(6819):529-33.

Piñeiro C, Abuín JM, Pichel JC. (2020) Very Fast Tree: speeding up the estimation of phylogenies for large alignments through parallelization and vectorization strategies. Bioinformatics. Nov 1;36(17):4658-4659.

Price MN, Dehal PS, Arkin AP. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. Mar 10;5(3):e9490.

Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. (2020) Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS One. Jan 16;15(1):e0227434.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35(21):7188-96.

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods. Sep;6(9):639-41.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics. Jan 28;12:38.

Razo-Mendivil FG, Martínez O, Hayano-Kanashiro C. (2020) Compacta: a fast contig clustering tool for de novo assembled transcriptomes. BMC Genomics. Feb 11;21(1):148.

Reeder J, Knight R. (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. Nat Methods. Sep;7(9):668-9.

Restrepo-Ortiz CX, Auguet JC, Casamayor EO (2013) Targeting spatio temporal dynamics of planktonic SAGMGC-1 and segregation of ammonia-oxidizing thaumarchaeota ecotypes by newly designed primers and quantitative polymerase chain reaction. EnvironMicrobiol 16:689–700

Richard GF. (2020) Eukaryotic Pangenomes. May 1. In: Tettelin H, Medini D, editors. The Pangenome: Diversity, Dynamics and Evolution of Genomes - Springer.

Richter M, Rosselló-Móra R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. Nov 10;106(45):19126-31.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013 Jul 25;499(7459):431-7.

Rocap G, Distel DL, Waterbury JB, Chisholm SW. (2002) Resolution of Prochlorococcus and Synechococcus ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol. Mar;68(3):1180-91.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW. (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature. Aug 28;424(6952):1042-7.

Rocha EP. (2008) The organization of the bacterial genome. Annu Rev Genet. 42:211-33

Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. Aug;1(4):283-90.

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J, Goodman RM. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol. Jun;66(6):2541-7.

Rodríguez-Valera F. (2002) Approaches to prokaryotic biodiversity: a population genetics perspective. Environ Microbiol. Nov;4(11):628-33.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016 8;4:e2584.

Rosen MJ, Callahan BJ, Fisher DS, Holmes SP. (2012) Denoising PCR-amplified metagenome data. BMC Bioinformatics. 31;13:283.

Rosen MJ, Davison M, Bhaya D, Fisher DS. (2015) Microbial diversity. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science. May 29;348(6238):1019-23.

Rosselló-Mora R, López-López A. 2008. The Least Common Denominator: Species or Operational Taxonomic Units?, In Zengler K (ed), Accessing Uncultivated Microorganisms. ASM Press, Washington, DC.p 117-130.

Salzberg SL. (2019) Next-generation genome annotation: we still struggle to get it right. Genome Biol. May 16;20(1):92.

Santoferrara L, Burki F, Filker S, Logares R, Dunthorn M, McManus GB. (2020) Perspectives from Ten Years of Protist Studies by High-Throughput Metabarcoding. J Eukaryot Microbiol. Sep;67(5):612-622.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. Mar 31;43(6):e37.

Schloss PD, Handelsman J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol. Mar;71(3):1501-6.

Schloss PD, Westcott LW, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, et al (2009). Introducing mothur : open-source, platform-independent, communitysupported software for describing and comparing microbial communities. Applied and environmental microbiology, 75(23) :75377541, 2009.

Schloss PD, Westcott SL. (2011) Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol. May;77(10):3219-26.

Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. (2016) Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ. Mar 28;4:e1869.

Schloss PD. Reintroducing mothur: 10 Years Later. Appl Environ Microbiol. 2020 Jan 7;86(2):e02343-19.

Schmalenberger A, Schwieger F, Tebbe CC. (2001) Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. Appl Environ Microbiol. Aug;67(8):3557-63.

Schmeltzer, O. (1995): Modélisation des cartes génomiques, une formalisation et un algorithme de construction fondé sur le raisonnement temporel. thèse de 3e cycle, Joseph Fourier: 202.

Schmutzer M, Barraclough TG. (2019) The role of recombination, niche-specific gene pools and flexible genomes in the ecological speciation of bacteria. Ecol Evol. Apr 4;9(8):4544-4556.

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ (2012). Population genomics of early events in the ecological differentiation of bacteria. Science. 2012 Apr 6;336(6077):48-51.

Shapiro BJ, Polz MF. (2014) Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. May;22(5):235-47.

Shapiro B.J. (2018) What Microbial Population Genomics Has Taught Us About Speciation. In: Polz M., Rajora O. (eds) Population Genomics: Microorganisms. Population Genomics. Springer, Cham

Shen XX, Hittinger CT, Rokas A. (2017) Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol. Apr 10;1(5):126.

Sintes E, Bergauer K, De Corte D, Yokokawa T, Herndl GJ (2013) Archaeal amoA gene diversity points to distinct biogeography of ammonia-oxidizing Crenarchaeota in the ocean. Environ Microbiol15:1647–1658

Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature. Oct 30;480(7376):241-4.

Sobetzko P, Travers A, Muskhelishvili G. (2012) Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. Proc Natl Acad Sci U S A. Jan 10;109(2):E42-50.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A. 8;103(32):12115-20.

Soucy SM, Huang J, Gogarten JP. (2015) Horizontal gene transfer: building the web of life. Nat Rev Genet. Aug;16(8):472-82.

Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG. (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. May 14;521(7551):173-179.

Stackebrandt E et Goebel BM (1994). Taxonomic note : a place for dna-dna reassociation and 16s rrna sequence analysis in the present species de#nition in bacteriology. International Journal of Systematic Bacteriology , 44(4) :846-849, 1994.

Steinegger M, Söding J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. Nov;35(11):1026-1028.

Stevens, R., C. A. Goble and S. Bechhofer (2000): Ontology-based knowledge representation for bioinformatics. Brief. Bioinform. 1(4): 398-414.

Stoddard SF, Smith BJ, Hein R, Roller BR, Schmidt TM. (2015) rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res. Jan;43(Database issue):D593-8.

Strassert JFH, Jamy M, Mylnikov AP, Tikhonenkov DV, Burki F. (2019) New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life. Mol Biol Evol. Apr 1;36(4):757-765.

Struhl K, Cameron JR, Davis RW. (1976) Functional genetic expression of eukaryotic DNA in Escherichia coli. Proc Natl Acad Sci U S A. May;73(5):1471-5.

Stuart T, Satija R. (2019) Integrative single-cell analysis. Nat Rev Genet. 2019 May;20(5):257-272.

Sun Y, Cai Y, Huse SM, Knight R, Farmerie WG, Wang X, and Mai F (2012). A large-scale benchmark study of existing algorithms for taxonomy independent microbial community analysis. Briengs in bioinformatics, 13(1) :107121, 2012.

Sun DL, Jiang X, Wu QL, Zhou NY. (2013) Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. Appl Environ Microbiol. Oct;79(19):5962-9.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. (2012) Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A. Nov 6;109(45):18488-92.

Syvanen M. (1985) Cross-species gene transfer; implications for a new theory of evolution. J Theor Biol. Jan 21;112(2):333-43.

Taib N (2013) Analyse de la diversité microbienne par séquençage massif : Méthodes et Applications. Thèse de doctorat : Ecologie Microbienne : Clermont-Ferrand 2

Taib N, Mangot JF, Domaizon I, Bronner G, Debroas D. (2013) Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: new insights into the freshwater protist diversity. PLoS One. ;8(3):e58950.

Tarbe A, Stenuite S, Balagu V, Sinyinza D, Descy J, et al. (2011) Molecular characterisation of the small-eukaryote community in a tropical Great Lake (Lake Tanganyika, East Africa). Aquat Microb Ecol 62: 177–190.

Tedersoo L, Tooming-Klunderud A, Anslan S. (2018) PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytol. Feb;217(3):1370-1385.

Tedersoo L, Anslan S. (2019) Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences. Environ Microbiol Rep. Oct;11(5):659-668.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 2005 Sep 27;102(39):13950-5.

Tikhonov M, Leach RW, Wingreen NS. (2015) Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. ISME J. Jan;9(1):68-80.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EP, Denamur E. (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. Jan;5(1):e1000344.

Tragin M, Vaulot D. (2019) Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. Sci Rep. 2019 26;9(1):5190.

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N (2017). Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 2017 Apr;27(4):626-638.

Van Etten J, Bhattacharya D. (2020) Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? Trends Genet. Dec;36(12):915-925.

Vanni C, Schechter M, Acinas S, Barberán A, Buttigieg PL, Casamayor EO,Delmont TO, Duarte CM, Eren AM, Finn R, Mitchell A, Sanchez P, Siren K, Steinegger M, Glöckner FO, Fernandez-Guerra A (2021) Light into the darkness: Unifying the known and unknown coding sequence space in microbiome analyses. Feb 18 bioRxiv /doi.org/10.1101/2020.06.30.180448

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004 Apr 2;304(5667):66-74.

Vergin KL, Tripp HJ, Wilhelm LJ, Denver DR, Rappé MS, Giovannoni SJ. (2007) High intraspecific recombination rate in a native population of Candidatus pelagibacter ubique (SAR11). Environ Microbiol. Oct;9(10):2430-40.

Vijay N, Poelstra JW, Künstner A, Wolf JB. (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Mol Ecol. Feb;22(3):620-34.

Violle C, Navas ML, Vile D, Kazakou E, Fortunel C, Hummel I, Garnier E (2007) Let the concept of trait be functional! Oikos 116: 882-892, 2007

Vissers EW, Anselmetti FS, Bodelier PL, Muyzer G, Schleper C, Tourna M, Laanbroek HJ. (2013) Temporal and spatial coexistence of archaeal and bacterial amoA genes and gene transcripts in Lake Lucerne. Archaea. 2013:289478.

Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 2009 Feb;3(2):199-208.

Vos M, Hesselman MC, Te Beek TA, van Passel MWJ, Eyre-Walker A. (2015) Rates of Lateral Gene Transfer in Prokaryotes: High but Why? Trends Microbiol. Oct;23(10):598-605.

Vos M, Eyre-Walker A (2017). Are pangenomes adaptive or not? Nat Microbiol. 2017 Dec;2(12):1576.

Voshall A and Moriyama EN (2018). Next-Generation Transcriptome Assembly: Strategies and Performance Analysis, Bioinformatics in the Era of Post Genomics and Big Data, Ibrokhim Y. Abdurakhmonov, IntechOpen, ; Available from: https://www.intechopen.com/books/bioinformatics-in-the-era-of-post-genomics-and-big-data/next-generation-transcriptome-assembly-strategies-and-performance-analysis

Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. (2016) Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. BMC Microbiol. Nov 14;16(1):274.

Webb CO. (2000) Exploring the Phylogenetic Structure of Ecological Communities: An Example for Rain Forest Trees. Am Nat. Aug;156(2):145-155.

Webb CO, Ackerly DD, McPeek MA, Donoghue MJ (2002) Phylogenies and community ecology. Ann. Rev. Ecol. Syst. 33 : 475-505.

Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A. Dec 24;99(26):17020-4.

Worden AZ, Allen AE. (2010) The voyage of the microbial eukaryote. Curr Opin Microbiol. Oct;13(5):652-60.

Worden AZ, Lee JH, Mock T, Rouzé P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, Foulon E, Grimwood J, Gundlach H, Henrissat B, Napoli C, McDonald SM, Parker MS, Rombauts S, Salamov A, Von Dassow P, Badger JH, Coutinho PM, Demir E, Dubchak I, Gentemann C, Eikrem W, Gready JE, John U, Lanier W, Lindquist EA, Lucas S, Mayer KF, Moreau H, Not F, Otillar R, Panaud O, Pangilinan J, Paulsen I, Piegu B, Poliakov A, Robbens S, Schmutz J, Toulza E, Wyss T, Zelensky A, Zhou K, Armbrust EV, Bhattacharya D, Goodenough UW, Van de Peer Y, Grigoriev IV. (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science. Apr 10;324(5924):268-72.

Yan W, Wei S, Wang Q, Xiao X, Zeng Q, Jiao N, Zhang R. (2018) Genome Rearrangement Shapes Prochlorococcus Ecological Adaptation. Appl Environ Microbiol. Aug 17;84(17):e01178-18.

Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. (2014)Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014 Sep;12(9):635-45.

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. Jan;42(Database issue):D643-8.

Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP. (2009) Intertwined evolutionary histories of marine Synechococcus and Prochlorococcus marinus. Genome Biol Evol. Sep 2;1:325-39.

Zheng Q, Bartow-McKenney C, Meisel JS, Grice EA. (2018) HmmUFOtu: An HMM and phylogenetic placement based ultra-fast taxonomic assignment and OTU picking tool for microbiome amplicon sequencing studies. Genome Biol. Jun 27;19(1):82.

Zhu F, Massana R, Not F, Marie D, Vaulot D. (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. FEMS Microbiol Ecol. Mar 1;52(1):79-92.