Étude du pangénome d'une population bactérienne structurée : vers une nouvelle compréhension de l'origine des variations intra-génomiques

Hélène Gardon

Université Clermont Auvergne

Laboratoire Microorganismes : Génome et Environnement, UMR CNRS 6023

Thèse de Doctorat

Spécialité Bioinformatique

Présentée par

Hélène Gardon

Pour obtenir le grade de

DOCTEUR de l'Université Clermont Auvergne

Étude du pangénome d'une population bactérienne structurée : vers une nouvelle compréhension de l'origine des variations intra-génomiques

---

Soutenue publiquement le 03 décembre 2021

Membres du Jury

Gwenaël PIGANEAU *DR CNRS, Sorbonne Université*	Rapportrice
Eduardo ROCHA *DR CNRS, Institut Pasteur*	Rapporteur
Eric BAPTESTE *DR CNRS, Sorbonne Université*	Examinateur
Didier DEBROAS *Professeur, Université Clermont Auvergne*	Examinateur
François ENAULT *MCF, Université Clermont Auvergne*	Invité
Gisèle BRONNER *MCF, Université Clermont Auvergne*	Co-Directrice de thèse
Corinne PETIT *CR CNRS, Université Clermont Auvergne*	Co-Directrice de thèse

À Corinne et Gisèle

Sa voix, je l'entends qui me dit : « Debout, viens-t-en
Vite, accours jusqu'à moi
La triste saison est finie, c'est le printemps
Vite, rejoins-moi »

Sa voix, je l'entends qui me dit : « Debout, viens-t-en
Vite jusqu'à moi
La saison des pluies est finie, c'est le printemps
Vite, rejoins-moi »

Il est pentu, le long chemin qui mène
À ta vertu, à l'eau de ta fontaine
Mais je le ferai jusqu'au bout
Pour arriver à toi
Jusqu'à toi

Cantique
Feu! Chatterton

Remerciements

Il est des gens qu'on ne rencontre qu'une fois, et dont on ne peux plus se passer.
Frédéric II de Prusse ; Lettre à Voltaire, 16 novembre 1737

À toutes ces personnes qui m'ont encadrées, soutenues et encouragées au cours de ces années de thèse[1], à qui je témoigne une profonde gratitude, je tiens à vous dire sincèrement merci.

Vous êtes bien trop nombreuses et nombreux, et pour être sûre de n'oublier personne, je ne citerai aucun nom. Exception faite de Gisèle et Corinne, mes deux directrices de thèse, sans qui celle-ci n'aurait pu voir le jour et, sous bien des aspects, n'aurait certainement pas abouti.
Merci à vous deux pour cette grande générosité professionnelle et personnelle dont vous avez fait preuve, qui m'a portée toutes ces années. Merci pour votre confiance, et pour toutes ces fois où vous aurez essayé de me convaincre que je pouvais croire, ne serait-ce qu'un peu, en moi. Et pour toutes les difficultés et complexités que j'ai apportées à votre quotidien, je suis désolée…

Un immense merci à l'ensemble de l'équipe, ou devrais-je dire la famille, MEB. Les moments partagés, votre bienveillance à toutes et tous, votre convivialité, et parfois même votre morosité…, m'ont apporté beaucoup de bien, de douceur et m'ont permis de toujours aller de l'avant.

Aux stagiaires, doctorant.e.s et contractuel.le.s, et toutes celles et ceux[2] que j'ai croisé au détour d'une boisson (souvent bien fraîche !), passés ou présents, je vous remercie pour cette vie sociale somme toute très satisfaisante dont j'ai bénéficié grâce à vous :)

Et pour finir, je pense surtout à celles et ceux qui me sont tellement cher.e.s. Pas toujours comprise, à l'image de cette question souvent posée, « pourquoi ne travailles-tu pas sur le covid ?! », mais encouragée, je remercie ma famille, et notamment mes neveu et nièce, cette nouvelle génération, déjà passionnée de sciences, avec qui je l'espère je pourrai très prochainement discuter *évolution* !

1. NB. – Isa, Gigi & Co, ma plus grande satisfaction au cours de ces années de thèse aura été de publier un papier 100 % féminin……
2. À mes deux *co-buralistes* des tout premiers débuts : « Quel beau métier, professeur ! »

Liste des figures

Liste des tableaux

Liste des abréviations

2D	2 dimensions
ACP	Analyse en Composantes Principales
ADN	Acide DéoxyriboNucléique
AIC	Akaike Information Criterion
ANI	Average Nucleotide Identity
ARNr	Acide RiboNucléique ribosomal
ARNt	Acide RiboNucléique de transfert
ARNtm	Acide RiboNucléique de transfert-messager
BAPS	Bayesian Analysis of Population Structure
BATS	Bermuda-Atlantic Times-series Study
BF	Bayes Factor
BLAST(P, N)	Basic Local Alignment Search Tool (Proteic, Nucleic)
BSC	Biological Species Concept
CDS	Coding Sequences
cf.	se référer à
COG	Cluster of Orthologous Genes
DDH	DNA–DNA hybridization
dN	taux de substitutions non synonymes
dS	taux de substitutions synonymes
e.g.	par exemple
EggNOG	Evolutionary genealogy of genes: Non-supervised Orthologous Groups
EMBOSS	European Molecular Biology Open Software Suite
ESC	Ecological Species Concept
fGI	îlots génomiques flexibles
F_{ST}	indice de fixation
Gt	Gigatonne
GTR	General Time Reversible
H	entropie de Shannon
HGT	Horizontal Gene Transfer
HL	high-light
HMM	Hidden Markov Model

i.e.	à savoir
IMG	Infinitely Many Genes
ISL	genomic island
ITS	Internal Transcribed Spacer
Kpb	Kilo paires de bases
LCB	Locally Collinear Blocks
LL	low-light
LMA	Local Multiple Alignments
MAFFT	Multiple Alignment using Fast Fourier Transform
Mpb	Méga paires de bases
MDA	Multiple Displacement Amplification
MEPS	Minimally Efficient Processing Segments
NCBI	National Center for Biotechnology Information
NGS	New Generation Sequencing
NSS	Neighbor Similarity Score
nt	nucléotide
O/E	observé sur théorique
ori	origine de réplication
OTU	Operational Taxonomic Unit
PAML	Phylogenetic Analysis by Maximum Likelihood
pb	paires de bases
PHI	Pairwise Homoplasy Index
RH	Recombinaison Homologue
SAG	Single-Amplified Genome
SCG	Single-Cell Genomics
SGT	Stationary Genome on Tree
SNP	Single-Nucleotide Polymorphism
ter	terminus de réplication
χ^2	chi2

Table des matières

1. Introduction

1.1. Concepts d'espèce bactérienne et spéciation

Selon la définition générale de l'espèce proposée par Mayr (Mayr, 1942), les populations résultent de discontinuités de flux génétiques au sein d'une espèce, conduisant à des unités génétiquement cohésives qui peuvent être distinguées selon les caractéristiques de leur génome. Alors que chez les animaux et les plantes, cette cohésion génétique peut être définie par leur capacité à produire une descendance viable et fertile, ceci est difficilement applicable chez les procaryotes pour lesquels des flux de gènes se produisent non seulement entre parents proches mais aussi entre parents éloignés. Dans la pratique, des unités de diversité microbienne ont pu être définies sur la base de traits écologiques et génétiques, donnant lieu à la proposition de plusieurs concepts pour expliquer la notion d'espèces chez les bactéries.

1.1.1. Définition opérationnelle de l'espèce bactérienne

La notion d'espèce est controversée chez les bactéries (Krause and Whitaker, 2015). D'abord définie phénotypiquement (notamment à partir de caractérisations métaboliques), l'intégration d'une dimension moléculaire dans cette définition a permis de proposer une démarcation empirique et normalisée, afin de présenter une taxonomie microbienne cohérente et opérationnelle (Cohan and Perry, 2007). Ainsi, en l'absence d'un cadre théorique relatif à un modèle évolutif, des méthodes, basées sur des similarités de séquences génétiques ou génomiques, ont permis de regrouper les micro-organismes en unités taxonomiques opérationnelles (*Operational Taxonomic Units* ou OTUs) plutôt qu'en espèces *stricto sensu*. Les similarités entre les micro-organismes ont d'abord été évaluées sur la base du pourcentage d'hybridation moléculaire ADN-ADN (*DNA-DNA hybridization* ou DDH) de leurs génomes, permettant ainsi d'assigner deux individus à une même espèce dès lors que la DDH était supérieure à 70 % (Wayne *et al.*, 1987). Par la suite, l'émergence et l'essor des techniques de biologie moléculaire et de séquençage ont rendu possible le développement d'approches telles que celles basées sur l'analyse des identités nucléotidiques entre les gènes de l'ARN ribosomal (ARNr) 16S, marqueur géné-

tique universel bactérien (Woese and Fox, 1977). Il a longtemps été admis que le seuil de 70 % par DDH correspondait approximativement à une identité de 97 % entre ces gènes (Stackebrandt and Goebel, 1994). Cependant, les génomes microbiens peuvent porter plusieurs copies du gène de l'ARNr 16S dont les séquences sont plus ou moins divergentes. Bien que leur dissimilarité soit rarement supérieure à 1 %, elle peut atteindre 6 % chez certaines espèces et donc compliquer la définition des OTUs (Acinas *et al.*, 2004; Větrovský and Baldrian, 2013). Par ailleurs, ces gènes ont la particularité d'être très conservés et d'évoluer plus lentement que le reste du génome. Ces propriétés font qu'ils sont pertinents pour tracer les relations anciennes entre groupes taxonomiques distants, mais génèrent un signal limité pour caractériser des relations récentes entre organismes. La quantité de sites informatifs partagés entre ces gènes peut alors être insuffisante pour permettre la démarcation des OTUs (Kettler *et al.*, 2007).

Pour pallier à la définition des OTUs sur la base d'un marqueur unique, non représentatif de la divergence à l'échelle des génomes, une approche prenant en compte une combinaison de gènes universels (gènes de ménage présents en copie unique) a été proposée (Santos and Ochman, 2004; Roux *et al.*, 2011). Plus récemment encore, il a été proposé de définir un OTU à partir de l'estimation des identités nucléotidiques moyennes (*Average Nucleotide Identity* ou ANI) entre génomes complets. Dans ce cas, deux OTUs sont assignés à une même espèce lorsque leur identité nucléotidique moyenne à l'échelle du génome est supérieure à 95 % (Konstantinidis and Tiedje, 2005; Goris *et al.*, 2007; Richter and Rosselló-Móra, 2009). Ces différentes approches restent cependant guidées par une dimension « pratique » de la définition de clusters assimilés à des espèces microbiennes, basée sur la considération de seuils de similitude pour la délimitation des groupes, dont il est sous-entendu qu'ils répondent aux mêmes contraintes de différenciation. Or, les forces qui gouvernent la différenciation des espèces microbiennes peuvent être différentes en fonction des systèmes biologiques considérés. Par exemple, la dynamique évolutive d'un micro-organisme parasite sera contrainte par le nombre d'hôtes infectés et l'espace cellulaire disponible pour la croissance, ce qui n'est pas le cas pour les micro-organismes libres de l'environnement (Toft and Andersson, 2010). Ainsi, qu'importe la méthode utilisée et le seuil de similarité choisi, il est toujours possible que la définition de certains OTUs ne correspondent pas à de véritables unités de diversité au sein des systèmes biologiques étudiés (Figure 1.1).

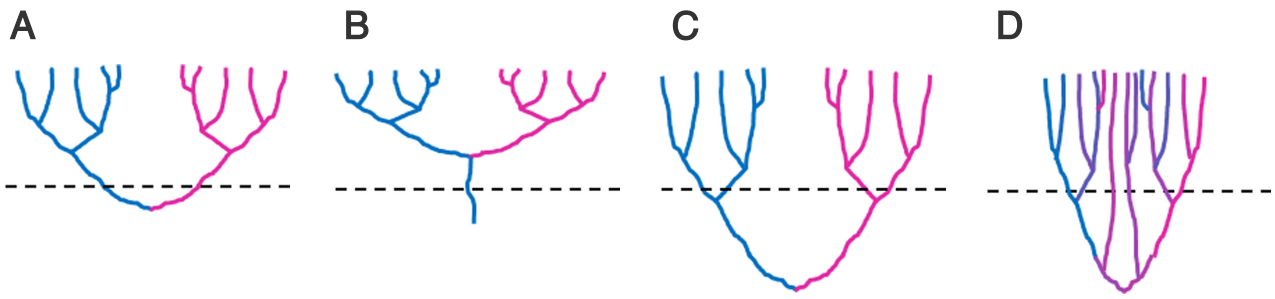


Figure 1.1. Seuils d'identité et spéciation (Whitman, 2015). Soit une méthode permettant la définition d'unités taxonomiques opérationnelles (OTUs) basée sur un seuil de similarité moléculaire défini (lignes pointillées) et son application pour quatre scénarios hypothétiques de diversification des espèces microbiennes. Les lignées présentant des propriétés biologiques différentes sont représentées par des couleurs distinctes. **(A)** Le seuil établi permet de distinguer deux lignées ayant des propriétés différentes. **(B)** Le seuil ne permet pas de distinguer les deux lignées, qui sont alors regroupées au sein d'une unique espèce. **(C)** Le seuil définit quatre lignées distinctes, séparant ainsi les deux groupes ayant des propriétés biologiques similaires. **(D)** L'histoire évolutive de certaines lignées est tellement complexe (espèces hybrides, transferts de gènes) que l'utilisation de seuils de similarité est peu utile.

Dès lors, il devient nécessaire d'établir un cadre théorique intégrant des dimensions écologiques et biologiques associées aux souches analysées.

1.1.2. Concept écologique de l'espèce

Le concept écologique de l'espèce – *Ecological Species Concept* ou ESC – est un concept théorique soutenant que, dans le cas d'un système de reproduction essentiellement clonal (*i.e.*, en l'absence d'échanges de matériel génétique entre individus), la spéciation est ponctuée par des événements périodiques de sélection positive en faveur du génotype qui se reproduit « le plus efficacement » dans la population (Cohan, 2001). Chaque espèce est alors adaptée à une niche écologique spécifique (Schluter, 2009) et s'apparente à un écotype. Des écotypes divergents peuvent être identifiés, sans même connaître les dimensions écologiques de leur divergence, dès lors qu'ils sont supportés par un modèle d'évolution sur lequel se fonde le concept d'espèce (Cohan and Perry, 2007; Wiedenbeck and Cohan, 2011). Parmi les différents modèles existants, celui de l'écotype stable avance l'idée que la compétition pour une même ressource va conduire à la fixation d'un génotype donné, c'est-à-dire que la combinaison allélique de l'ensemble

des gènes qui constituent le génotype sélectionné remplacera *in fine* la diversité génotypique de la population. Ce processus est connu en génétique des populations sous la dénomination de balayage sélectif. Du fait de la liaison génétique de l'ensemble des gènes portés sur un génome bactérien et de l'absence d'échanges de matériel génétique entre individus, ce balayage sélectif porte ici sur l'ensemble du génome (Figure 1.2A). La diversité au sein de l'écotype, au cours de son histoire évolutive, est alors éliminée par un processus de sélection périodique, à la faveur de l'émergence d'une combinaison allélique plus efficiente. Ceci explique la cohésion génétique qui est associée à ce modèle d'écotype stable. L'émergence d'un nouvel écotype est possible lorsqu'une lignée est en mesure de coloniser une nouvelle niche, utilisant des ressources différentes de l'écotype initial, et qui ne sera pas purgé par sélection.

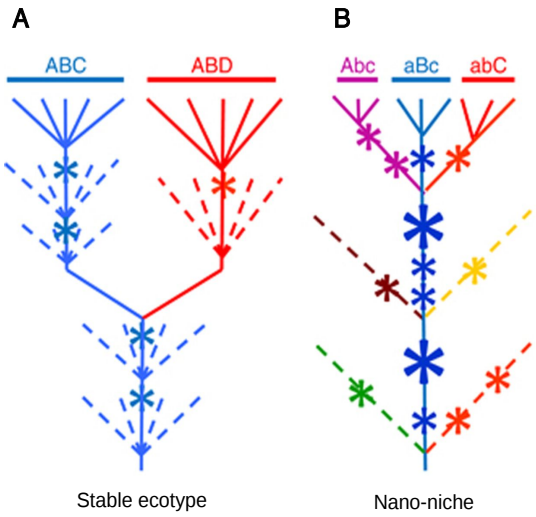


Figure 1.2. Présentation de deux modèles de spéciation bactérienne (Cohan and Perry, 2007; Wiedenbeck and Cohan, 2011). Les écotypes sont représentés par des couleurs différentes ; les événements de sélection périodique sont indiqués par des astérisques et les lignées éteintes sont représentées par des lignes pointillées. Les lettres représentent les ressources que chaque groupe d'organismes peut utiliser. Dans le cas où les écotypes utilisent des ressources équivalentes, mais en différentes proportions, la ressource majoritaire est représentée par une lettre capitale. **(A)** Modèle de l'écotype stable caractérisé par un taux bien plus important d'événements de sélection périodique que la formation d'écotypes en elle-même, de sorte que chaque écotype subit de nombreux événements de sélection périodique au cours de son existence. Les écotypes peuvent coexister indéfiniment car chacun d'entre eux possède une ressource non partagée avec l'autre. **(B)** Modèle de la nano-niche. Dans cet exemple, trois écotypes (Abc, aBc et abC) utilisent les mêmes ressources, mais en différentes proportions. Chaque écotype peut ainsi coexister avec les deux autres dans la mesure où leurs ressources sont partitionnées quantitativement. Cependant, puisque les écotypes partagent tous leurs ressources, chacun est plus susceptible de subir les conséquences d'une éventuelle

mutation se produisant dans les autres écotypes. Il pourrait s'agir de mutations qui accroîtraient l'efficacité de l'utilisation des ressources. Ces mutations, représentées par un grand astérisque, entraînent l'extinction des autres écotypes.

Des variations à ce modèle ont été proposées (Cohan and Perry, 2007). C'est notamment le cas du modèle de la nano-niche (Figure 1.2B) pour lequel de nombreuses niches écologiques sont présentes et conditionnent l'existence de sous-groupes au sein d'un écotype. L'adaptation de chaque sous-groupe à son micro-habitat est nuancée et rendue possible grâce à l'acquisition de gènes accessoires spécialisés dans l'utilisation des ressources. Chaque sous-groupe peut également subir ses propres événements de sélection périodique.

1.1.3. Concept biologique de l'espèce

Bien que les modèles associés à l'ESC expliquent théoriquement l'émergence et le maintien d'une cohésion génétique au sein d'un écotype, des approches de génomique comparative ont montré que le balayage sélectif concerne souvent une partie des gènes et non les génomes dans leur totalité. Dans ce cas, seuls les gènes (ou allèles) qui confèrent un avantage sélectif au sein de l'écosystème seront fixés dans la population (Cadillo-Quiroz *et al.*, 2012; Shapiro *et al.*, 2012; Bendall *et al.*, 2016). Les allèles peuvent par ailleurs être échangés entre les individus au sein de la population. Ces observations sont en contradiction avec l'hypothèse de l'ESC selon laquelle les échanges de matériel génétique entre individus sont trop peu fréquents pour limiter la divergence induite par l'adaptation à une niche écologique.

Ainsi, le fondement même du concept biologique de l'espèce – *Biological Species Concept* ou BSC – repose sur l'existence de flux de matériel génétique entre populations, à l'instar de ce qui se passe au cours de la recombinaison méiotique lors de la reproduction sexuée chez les eucaryotes. Bien que les bactéries ne réalisent pas strictement de reproduction sexuée, il est admis qu'une grande majorité d'entre elles échange du matériel génétique par le biais de mécanismes de recombinaisons homologues ou hétérologues (Smith *et al.*, 1993; Vos and Didelot, 2009; Bobay and Ochman, 2017). Dans le contexte du BSC, c'est donc l'interruption des flux de gènes au sein des populations qui initie le processus de spéciation.

1.1.4. Barrières à la recombinaison *versus* sélection

Les concepts d'espèces présentés ci-dessus pour caractériser les unités de diversité microbienne mettent en avant des mécanismes de spéciation différents, à savoir la sélection *via* le processus de balayage sélectif à l'échelle du génome (ESC) et l'interruption des flux de gènes (BSC). Si l'ESC est en accord avec l'émergence de clusters écologiques ou écotypes tels que ceux décrits dans la littérature (Hunt *et al.*, 2008; Koeppel *et al.*, 2008; Preheim *et al.*, 2011), il est en contradiction avec l'observation de l'envahissement de certaines populations par des gènes spécifiques *via* des processus de balayage sélectif à l'échelle des gènes (Guttman and Dykhuizen, 1994; Cohan and Perry, 2007; Papke *et al.*, 2007; Shapiro *et al.*, 2009; Coleman and Chisholm, 2010; Denef *et al.*, 2010). Ce modèle n'est donc pas totalement satisfaisant. D'un autre côté, alors que le processus de balayage sélectif à l'échelle des gènes fait appel à des mécanismes autorisant les flux de gènes entre populations (comme la recombinaison), la détermination de l'origine des barrières aux flux de gènes (*i.e.,* les causes de l'interruption de flux de gènes) à même d'initier un processus de spéciation dans ce contexte reste une question ouverte.

Les données expérimentales concernant les flux de gènes, notamment par le biais de la recombinaison homologue, suggèrent l'existence de similarités entre les séquences d'ADN transférées du génome donneur vers le génome receveur (Figure 1.3) (Zawadzki *et al.*, 1995; Vulić *et al.*, 1997; Majewski and Cohan, 1999; Majewski *et al.*, 2000). Ainsi, l'accumulation de mutations au cours du temps et l'augmentation progressive de la divergence entre les séquences auraient en théorie pour conséquence de créer une barrière à la recombinaison homologue.

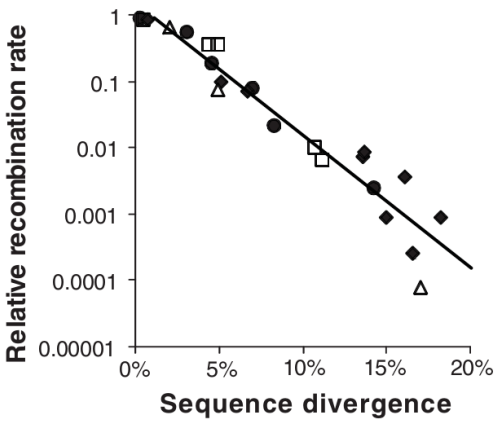


Figure 1.3. Taux de recombinaison entre génomes donneurs et receveurs lors de la recombinaison homologue médiée par RecA en fonction de la divergence des MEPS (*Minimally Efficient Processing Segments*) (Fraser *et al.*, 2007). Les données ont été obtenues expérimentalement pour les espèces bactériennes *Bacillus subtilis* (cercle), *B. mojavensis* (carré), *Streptococcus pneumoniae* (losange) et *Escherichia coli* (triangle) (Zawadzki *et al.*, 1995; Vulić *et al.*, 1997; Majewski *et al.*, 2000). La courbe de régression log-linéaire est montrée pour l'ensemble des observations.

Ceci pourrait alors conduire à la formation d'unités génétiques distinctes – telles que décrites par le concept BSC – en l'absence d'un processus de sélection. Cependant, d'après une étude théorique, ce processus neutre de spéciation ne serait envisageable que si les taux de recombinaison entre ces unités diminuent rapidement et considérablement (Fraser *et al.*, 2007). Or, la vitesse d'accumulation des mutations fait qu'il est peu probable que celles-ci représentent une barrière efficace aux flux de gènes. En outre, une corrélation positive entre flux de gènes et similarité de séquences n'est pas systématique comme le montre, par exemple, l'événement de fusion décrit entre les espèces *Campylobacter coli* et *C. jejuni* (Sheppard *et al.*, 2008), illustrant ainsi la possibilité d'un échange de matériel génétique entre espèces « distantes ». Des corrélations négatives entre identité nucléotidique à l'échelle du génome et taux de recombinaison ont par ailleurs été observées pour différentes espèces bactériennes, *e.g.*, *Pseudomonas putida, C. coli, B. subtilis* et *Vibrio cholerae* (Figure 1.4) (Bobay and Ochman, 2017). Ainsi, la divergence entre les séquences nucléotidiques ne suffirait pas à expliquer à elle seule l'interruption de flux de gènes à même d'initier un processus de spéciation.

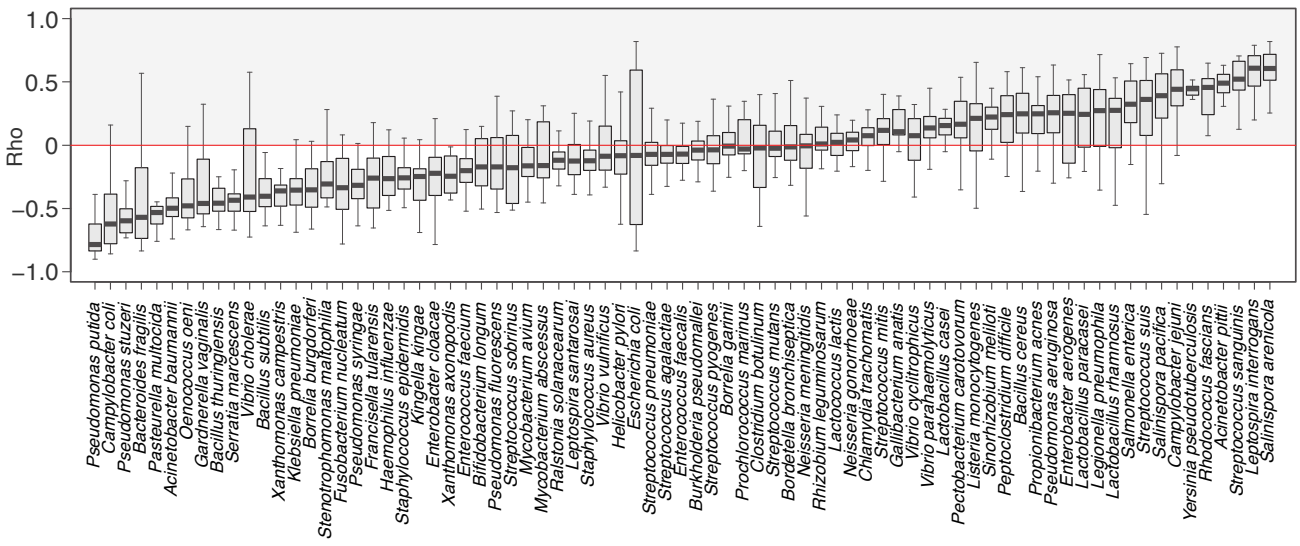


Figure 1.4. Effet de la divergence des séquences sur les flux de gènes chez différentes espèces bactériennes (Bobay and Ochman, 2017). Le rapport *r/m* (effet relatif de la recombinaison par rapport aux mutations) pour chacune des combinaisons de souches sous-échantillonnées a été comparé à l'identité moyenne des séquences à l'échelle des génomes *core*. La distribution des coefficients de corrélation de Spearman – rho, calculés entre ces deux métriques – est représentée sous forme de diagramme en boîte à moustache pour chaque es-

pèce. Des valeurs positives de rho indiquent une corrélation positive entre l'identité des séquences nucléotidiques et les flux de gènes (*i.e.*, les échanges de matériel génétique se produisent préférentiellement entre des souches plus semblables), tandis que des valeurs négatives indiquent que les flux de gènes se produisent entre des souches divergentes.

En ce qui concerne la part de la sélection dans les processus de spéciation, une étude portant sur les données génomiques de deux populations marines proches appartenant à l'espèce *Vibrio cyclitrophicus* (gènes de l'ARNr 16S identiques, plus de 99 % d'identité protéique), mais caractérisées par des niches écologiques distinctes (*i.e.,* une associée à des fractions particulaires types phyto/zooplancton et l'autre à des particules organiques en suspension) suggère que leur divergence résulte d'un processus sélectif caractérisé par une signature restreinte à quelques régions génomiques couvrant ~1 % du génome *core* (Shapiro *et al.*, 2012). D'après les auteurs, ces régions génomiques proviendraient d'événements de recombinaison (probablement à partir d'une population éloignée). Bien que des événements de recombinaison anciens entre les deux populations aient été détectés, les auteurs ont constaté une plus grande proportion d'événements de recombinaison récents au sein des populations qu'entre elles, suggérant que les flux de gènes surviennent préférentiellement au sein des populations. Ainsi, l'acquisition de gènes en lien avec une niche pourrait initier une spécialisation pour celle-ci (hôte ou habitat), entraînant une diminution des flux de gènes entre l'ensemble des populations et l'émergence de barrières à la recombinaison (Figure 1.5) (Shapiro *et al.*, 2012). La question des mécanismes à même de générer une barrière à la recombinaison entre populations sympatriques (en dehors de l'idée de réduction des probabilités de rencontre entre individus spécialisés / spatialisés sur des habitats différents) n'est cependant pas résolue à l'échelle de cette étude.

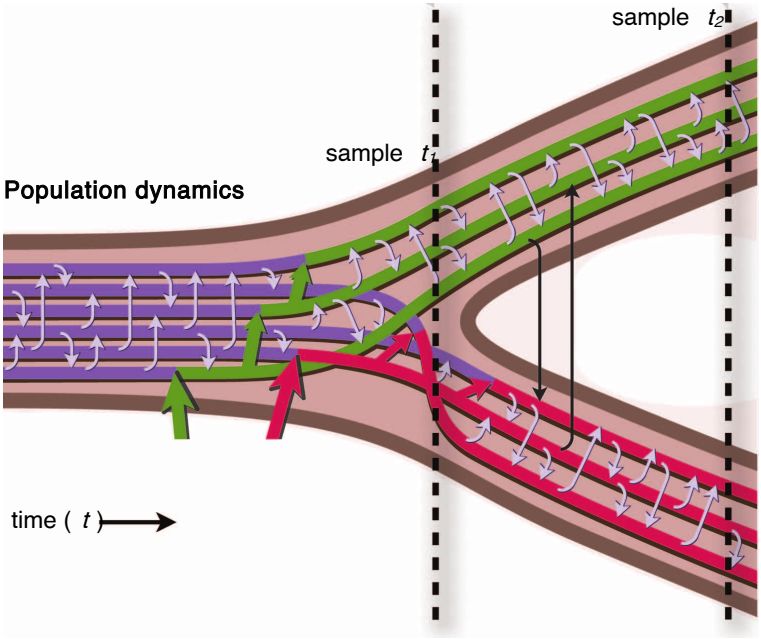


Figure 1.5. Modèle de différenciation écologique de populations microbiennes recombinantes (Shapiro *et al.*, 2012). La population de couleur violette est une population ancestrale, uniforme d'un point de vue écologique, et recombinante. L'acquisition de gènes accessoires (flèches rouges et vertes), spécifiques d'une niche, initie une spécialisation pour une niche donnée et entraîne une diminution des flux de gènes entre les populations. Les flèches grises et noires représentent les recombinaisons au sein ou entre les populations. Les flèches épaisses et colorées représentent l'acquisition d'allèles adaptatifs pour les niches représentées en rouge et vert.

Ainsi, les modèles d'évolution associés aux concepts BSC et ESC ne sont pas fondamentalement opposés. Les populations bactériennes clonales, pour lesquelles les taux de recombinaison sont très faibles, voire inexistants, verront leur cohésion génétique maintenue par des processus de sélection, telle que décrite par l'ESC. À l'inverse, les études montrant l'importance des balayages sélectifs de gènes uniques sont relativement nombreuses (Croucher *et al.*, 2011; Cadillo-Quiroz *et al.*, 2012; Shapiro *et al.*, 2012; Bao *et al.*, 2016; Bendall *et al.*, 2016), impliquant des flux de matériel génétique au sein des populations, ce qui entre dans le cadre du BSC.

1.2. Mécanismes évolutifs et dynamique des génomes

La diversification des populations bactériennes libres de l'environnement, et *in fine* la spéciation, résulteraient d'une perte de la cohésion génétique entre les populations. Cependant, la compréhension complète des forces à l'origine de cette différenciation repose également sur l'évaluation de la dynamique de leur génome, en considérant les mutations et la recombinaison à la lumière des processus populationnels tels que la dérive génétique et la sélection.

1.2.1. Mutation, sélection, dérive génétique et dynamique des populations

Les mutations sont sources de variations individuelles. Survenues dans un gène et transmises à la descendance, elles coexistent dans la population avec le génotype d'origine du gène ciblé, entraînant ainsi des polymorphismes. Les différents allèles de ce gène circulent donc simultanément au sein de la population. La dynamique et l'évolution de cette dernière sont dès lors le résultat d'un changement des fréquences alléliques, certains allèles étant éliminés ou fixés au cours du temps. La fixation des mutations sur le long terme va dépendre de la taille efficace des populations *Ne* – qui se réfère à la taille d'une population théorique pour laquelle la fluctuation du polymorphisme serait équivalente à celle effectivement observée au sein de la population – et de l'avantage adaptatif apporté par les allèles circulants.

La variation des fréquences alléliques, dans le cas où *Ne* est faible, est principalement conditionnée par la dérive génétique. Celle-ci fait référence au processus d'échantillonnage des allèles lors de leur transmission à la génération suivante et conduit à une fluctuation aléatoire des fréquences alléliques d'autant plus grande que les populations sont de petite taille. La dérive génétique est particulièrement présente chez les bactéries intracellulaires qui montrent une dynamique de populations bien différente de celle des bactéries libres (Moran, 1996; Andersson and Kurland, 1998). En effet, la taille efficace des populations intracellulaires est contrainte par le nombre de cellules et l'espace dispo-

nible pour leur croissance. Ceci peut créer un goulot d'étranglement et réduire considérablement la diversité génétique au sein des générations suivantes (Didelot *et al.*, 2016) (Figure 1.6). Les bactéries libres de l'environnement présentent généralement des *Ne* beaucoup plus grandes (Kuo *et al.*, 2009; Batut *et al.*, 2014).

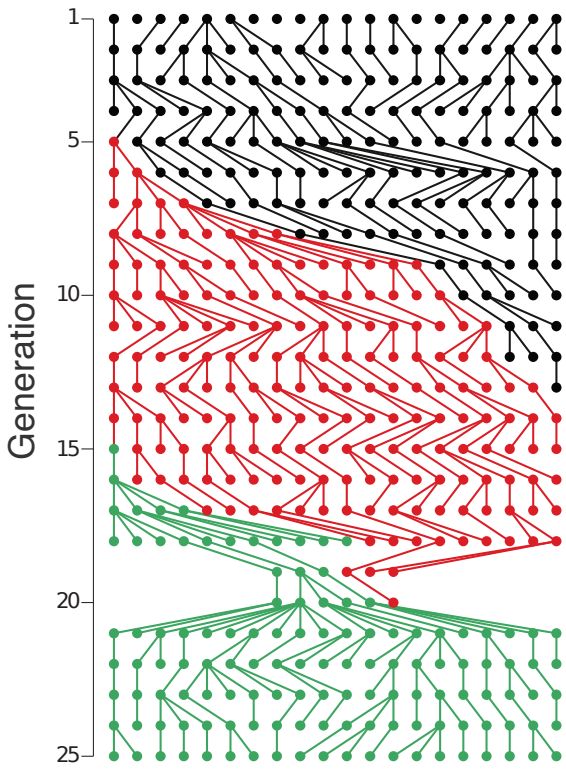


Figure 1.6. Processus de dérive génétique (Didelot *et al.*, 2016). La dérive génétique est le processus selon lequel les fréquences alléliques varient aléatoirement au cours du temps au sein d'une population. La population initiale a une taille effective de 20 bactéries, et chaque génération (lignes horizontales) est formée par la sélection aléatoire des parents de la génération précédente (les relations de parenté sont représentées par des liens entre deux générations). Aux générations 5 et 15, deux individus subissent un événement de mutation (en rouge et vert respectivement). Ces nouvelles mutations, qui augmentent sensiblement la fitness des individus, seront transmises aux descendants, et leur fréquence sera de plus en plus importante dans la population. Ceci entraîne dans un premier temps l'extinction de l'allèle sauvage (noir) à la 14ème génération. Au cours des générations 19 et 20 se produit un goulot d'étranglement qui augmente l'effet de la dérive génétique. Ceci augmente la vitesse de fixation de l'allèle muté vert et l'extinction de l'allèle muté rouge à la 21ème génération.

Les mutations à l'origine de l'émergence d'un nouvel allèle dans une population peuvent être caractérisées par leur effet sur la valeur sélective (ou *fitness*) des individus qui les portent. Lorsqu'une mutation n'induit pas d'effet sur la capacité d'un individu à produire une descendance, celle-ci est qualifiée de neutre et sa valeur sélective est nulle. Dans le cas contraire, la mutation peut être qualifiée d'avantageuse (ou adaptative) ou de désavantageuse, selon qu'elle induit un différentiel reproductif favorable ou défavorable aux individus porteurs de la mutation par rapport à l'allèle originalement présent dans la population. Il est à noter que cette valeur sélective n'est ni absolue, ni figée. Elle est conditionnée par le contexte environnemental, génétique et populationnel dans lequel la

mutation se produit. La taille efficace de la population en particulier conditionne « l'impact potentiel » des valeurs sélectives des mutations. Celui-ci sera d'autant plus grand que *Ne* est importante (Figure 1.7).

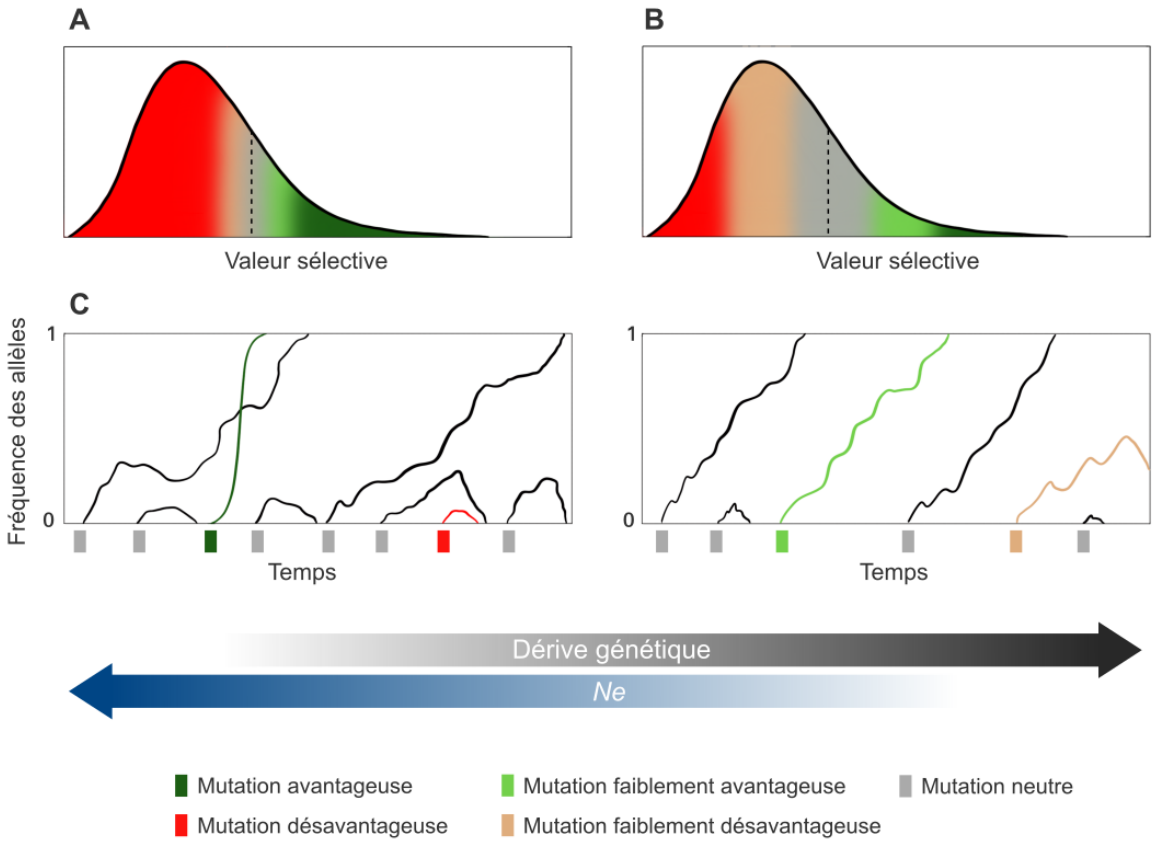


Figure 1.7. Conséquence de la taille efficace des populations (*Ne*) sur le différentiel reproductif entre individus au sein d'une population. La sélection peut être assimilée à un différentiel reproductif entre individus génétiquement différents au sein d'une population, et elle est exprimée en termes de valeur sélective (ou *fitness*). Une sélection positive traduit l'idée d'un différentiel reproductif favorable à la descendance de l'individu, une sélection négative correspond à un déficit de reproduction d'un individu par rapport au reste de la population. Dans les populations de grande taille, la distribution des valeurs sélectives est déséquilibrée en faveur des mutations neutres à désavantageuses **(A)**. Dans les populations de petite taille, la dérive génétique a également pour effet de « réduire le potentiel » des valeurs sélectives des mutations **(B)**. Une mutation avantageuse dans une grande population le sera moins dans une petite population. La théorie de l'évolution prédit également que :

- le devenir d'une mutation est gouverné 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 rapidement dans celles de petite taille **(C)** ;
- la diversité génétique neutre est proportionnelle à *Ne*.

1.2.2. Balayage sélectif, sélection d'arrière-plan et effet Hill-Robertson

Alors que les mutations sont source de diversité génétique au sein des populations, l'action conjointe de la sélection et de la dérive génétique a pour effet de réduire cette diversité (Figure 1.8). Cette réduction répond à différentes dénominations selon l'importance relative des facteurs en jeu.

Ainsi, on appelle balayage sélectif (Smith and Haigh, 1974) une situation dans laquelle une mutation sous sélection positive « envahit » la population, avec pour conséquence une élimination de la diversité au sein de la population, à l'exception des mutations génétiquement liées à cette mutation avantageuse. La sélection d'arrière-plan (Charlesworth *et al.*, 1993) est liée aux mutations sous sélection négative. Les individus porteurs de mutations désavantageuses ne participent pas à la génération suivante et le polymorphisme, neutre pour l'essentiel, lié à ces mutations désavantageuses, est purgé avec celles-ci, réduisant de fait la diversité génétique de la population. L'effet Hill-Robertson (Hill and Robertson, 1966) fait référence à un processus plus complexe observé dans des populations de petite taille, pour lesquelles la dérive génétique induit une perte de diversité génétique puisque les mutations, qu'elles soient neutres ou soumises à sélection, ne sont pas retenues dans la génération suivante. L'efficacité de la sélection est alors globalement réduite. Ce phénomène peut cependant être limité par la recombinaison génétique entre individus qui rend possible l'association de mutations avantageuses acquises indépendamment par plusieurs individus. En associant des mutations faiblement avantageuses, la recombinaison peut permettre d'augmenter la valeur sélective globale d'un génotype dont l'avantage sélectif peut potentiellement atteindre un niveau significatif au regard de la taille de la population considérée ($s >> 1/Ne$).

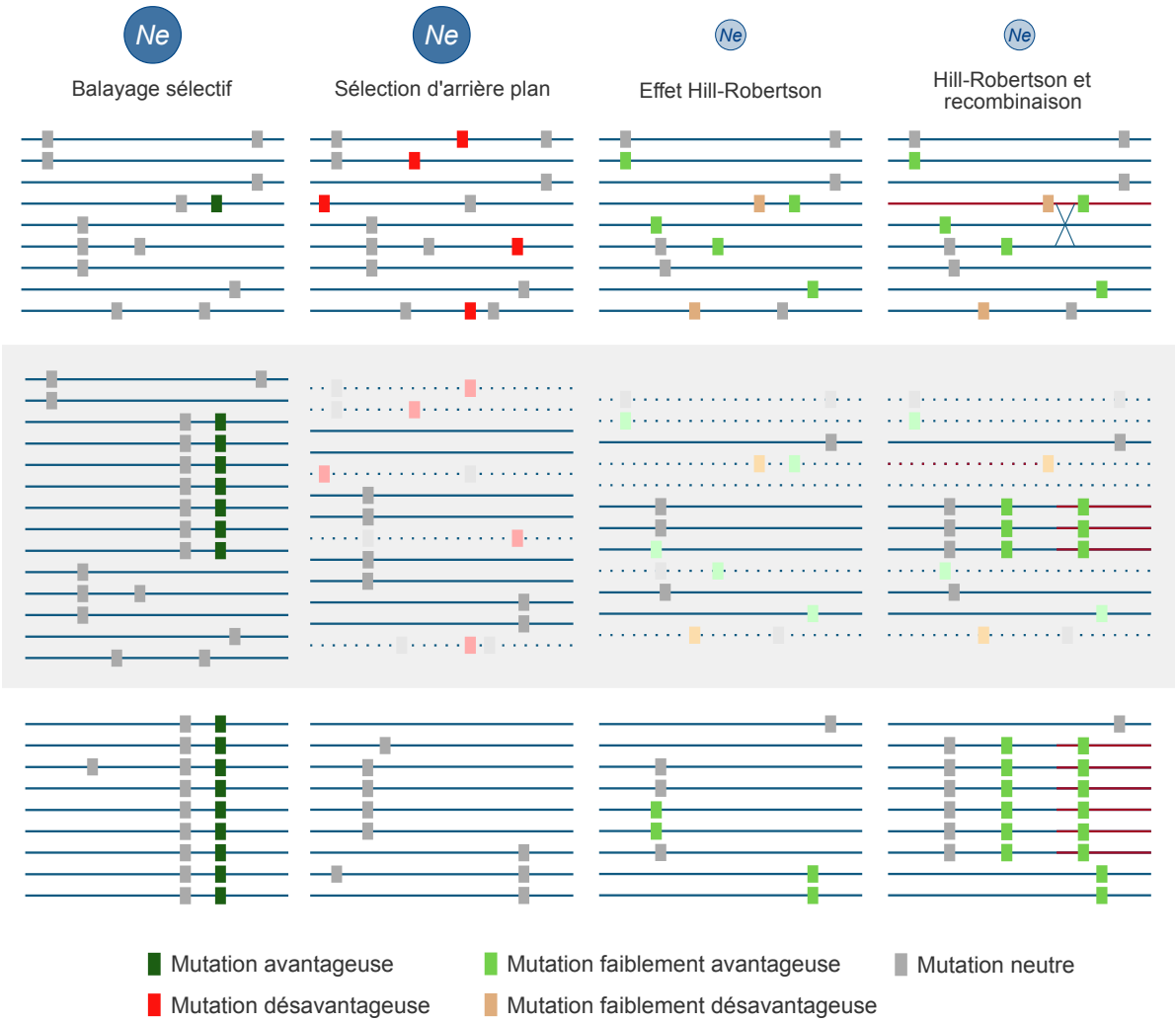


Figure 1.8. Action conjointe de la sélection et de la dérive génétique sur la diversité. La réduction de la diversité génétique au sein des populations est le résultat d'un équilibre entre sélection et dérive génétique, fonction de la taille efficace des populations *Ne*. Différents mécanismes évolutifs entrent ainsi en jeu, tels que, de gauche à droite, (i) le balayage sélectif (sélection des mutations fortement avantageuses conduisant à la perte du polymorphisme génétiquement lié à celles-ci), (ii) la sélection d'arrière-plan (sélection négative opérée sur les mutations désavantageuses conduisant à la perte du polymorphisme génétiquement lié à celles-ci), et (iii) l'effet Hill-Robertson (perte de diversité par dérive génétique).

1.2.3. Mécanismes de transferts horizontaux de gènes (HGTs) et conséquences évolutives

Il est communément admis que les populations bactériennes se développent par reproduction clonale (ou asexuée). Cependant, il a été montré que les bactéries étaient capables d'acquérir du matériel génétique de sources externes par différents mécanismes tels que la conjugaison (Sullivan *et al.*, 2002), la transduction (Zinder and Lederberg, 1952) ou encore la transformation (Griffith, 1928; Avery *et al.*, 1944) (Figure 1.9).

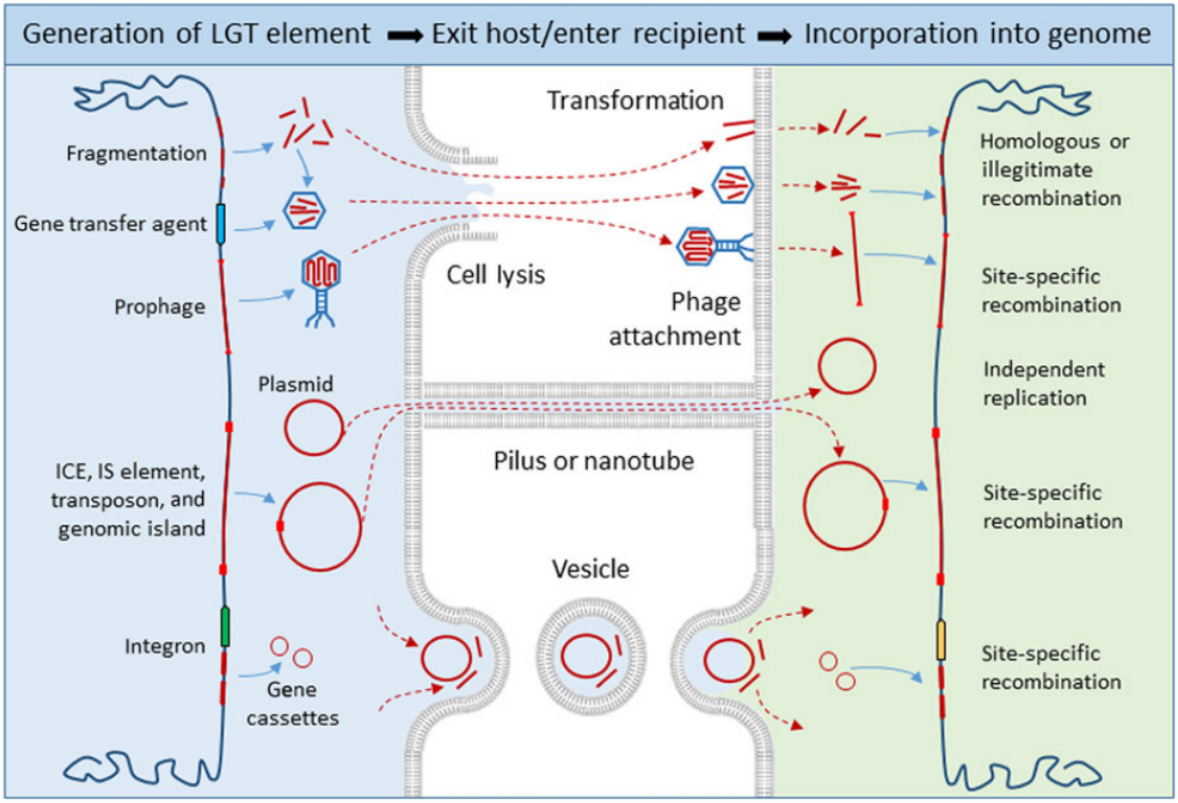


Figure 1.9. Représentation des différents processus à l'origine de transferts horizontaux de gènes (Gillings, 2016). Chacune des séries d'événements physiques qui aboutissent au transfert réussi d'ADN entre les cellules est schématisée. À gauche : production d'unités transférables (flèches bleues). Cela peut impliquer l'excision d'ADN du chromosome, la fragmentation de l'ADN ou l'expression de gènes résultant de l'assemblage de phages ou d'agents de transfert de gènes. Au centre **:** processus impliqués dans le transfert physique (lignes pointillées rouges) comprenant le transport *via* (i) un pilus ou un nanotube (mécanisme de transfert alors appelé conjugaison), (ii) une vésicule (vésiduction ; Soler and Forterre, 2020), (iii) la libération dans l'environnement *via* une lyse cellulaire suivie d'une transformation ou (iv) la fixation d'un phage (transduction). À droite **:** après son entrée dans

la cellule réceptrice, l'ADN transféré horizontalement doit être intégré (flèches bleues) par recombinaison, qui peut être homologue ou hétérologue, ou impliquer des sites de reconnaissance spécifiques.

Ces transferts horizontaux de gènes (HGTs) permettent un réarrangement de l'information génétique entre deux individus, et peuvent même impliquer des individus d'espèces différentes. Le transfert horizontal de gènes homologues entre bactéries apparentées se produit principalement par conjugaison ou transformation (Smith, 1991; Lorenz and Wackernagel, 1994). Cela peut introduire une variation génétique par recombinaison allélique, créant ainsi de nouvelles combinaisons d'allèles (Feil, 2004). Ce processus de recombinaison homologue permet, entres autres, d'éliminer de la population des mutations désavantageuses (Treangen *et al.*, 2008). *A contrario*, le transfert de gènes non-homologues par transduction, conjugaison ou transformation induit l'apparition de caractères nouveaux et de fonctions nouvelles. Ceci permet notamment une adaptation rapide des bactéries à une nouvelle niche écologique lors de changements ou stress environnementaux. Un exemple largement documenté au cours des dernières décennies est le transfert de gènes de résistance aux antibiotiques (Huddleston, 2014; von Wintersdorff *et al.*, 2016; Vrancianu *et al.*, 2020), accentué en présence de pressions sélectives exercées par les antibiotiques (Pelgrift and Friedman, 2013).

Les HGTs ont un impact considérable sur les communautés bactériennes en éliminant les allèles délétères et en permettant la dissémination rapide de fonctions. Ils impactent également la structure et l'organisation des génomes, puisque de nombreux gènes transférés sont retenus, entraînant ainsi une expansion des familles de gènes (Treangen and Rocha, 2011). La taille du génome bactérien dépend alors de l'équilibre entre la délétion de gènes pour lesquels le coefficient de sélection est faible (Mira *et al.*, 2001) et les processus d'acquisition de gènes, notamment *via* les HGTs (Touchon and Rocha, 2016). Bien que la taille des génomes au sein d'une même espèce bactérienne soit homogène, ces flux de gènes ont pour conséquence d'alimenter le répertoire de gènes d'une population bactérienne, d'une espèce, voire d'un genre et participent au remodelage de leur pangénome.

1.3. Le pangénome microbien

La démocratisation des approches par séquençage au début des années 2000 et le développement de la génomique comparative, en plus de mettre en évidence des variations nucléotidiques entre les séquences génomiques de différents isolats d'une même espèce bactérienne, ont permis de constater un grande hétérogénéité de leur contenu en gènes (Perna *et al.*, 2001; Welch *et al.*, 2002; Tettelin *et al.*, 2005; Kettler *et al.*, 2007; Bakker *et al.*, 2010). L'avènement des technologies de séquençage à haut débit (NGS) a ensuite confirmé et permis de mieux évaluer l'ampleur des variations en termes de contenu en gènes entre les individus d'une même espèce (Vernikos *et al.*, 2015; McInerney *et al.*, 2017; Brockhurst *et al.*, 2019). De ce constat est née la notion de pangénome (Medini *et al.*, 2005; Tettelin *et al.*, 2005). Celui-ci est défini comme regroupant l'ensemble des gènes présents chez une espèce donnée. Il est communément admis que sa taille dépend essentiellement de processus d'acquisitions et de pertes de gènes, l'acquisition étant souvent attribuée à des mécanismes de HGTs (Treangen and Rocha, 2011; Puigbò *et al.*, 2014; Vos *et al.*, 2015), tandis que des processus de sélection ou la dérive génétique conditionnent par la suite la fixation, ou *a contrario* la perte, des gènes acquis. Par conséquent, les hypothèses évolutives décrivant l'évolution des pangénomes sont multiples et dépendent de la part relative de ces processus.

1.3.1. Définitions et propriétés

Un pangénome est la collection des familles de gènes pour une espèce donnée (Figure 1.10) (Tettelin *et al.*, 2005). Cette collection est classiquement divisée en deux catégories de gènes. Les gènes *core* sont retrouvés chez tous les représentants de l'espèce étudiée ; ils constituent le génome *core*. Les gènes flexibles, également nommés gènes accessoires ou *dispensables*, ont une distribution restreinte à seulement une partie des individus de l'espèce, voire pour certains à un unique individu ; ils constituent le génome flexible (Figure 1.10) (Welch *et al.*, 2002; Tettelin *et al.*, 2005).

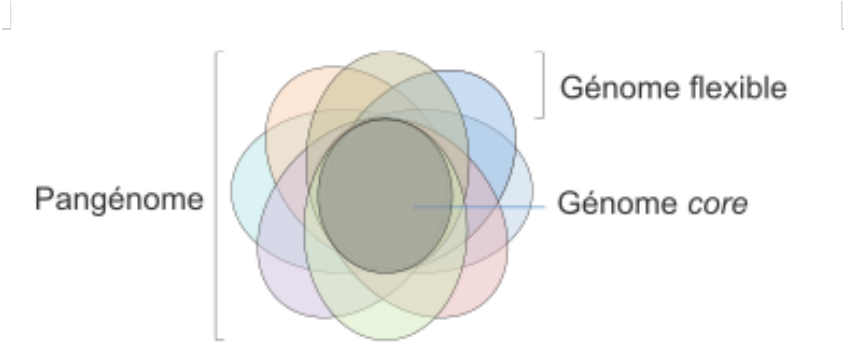


Figure 1.10. Représentation de la structure d'un pangénome sous forme d'un diagramme de Venn (McInerney *et al.*, 2017). Les gènes communs à l'ensemble des individus d'une espèce constituent le génome *core* alors que ceux retrouvés uniquement chez une partie des individus ou spécifiques d'un individu constituent le génome flexible.

Si la notion de pangénome a d'abord été appliquée à la caractérisation du *pool* génétique associé à une espèce, cette notion peut cependant se décliner à différents niveaux taxonomiques. En se plaçant aux niveaux taxonomiques plus élevés, on peut observer que la répartition des gènes au sein des différents génomes prend une forme en U caractéristique (Figure 1.11) (Makarova *et al.*, 2007; Koonin and Wolf, 2008; Lapierre and Gogarten, 2009). Cette distribution peut être approximée par trois fonctions exponentielles parmi lesquelles, une représente les gènes du génome *core* [forte *commonalité* selon la définition de Sela *et al.* (2021)], alors que les autres distinguent deux groupes de gènes au sein du génome flexible :

- un groupe de gènes de proportion deux fois supérieure à celle des gènes *core* et présents que chez certains génomes. Ces gènes, considérés comme retenus sur de longues périodes au sein des génomes, constituent le génome *shell ;*
- un groupe de gènes constituant une proportion importante du pangénome mais qui sont très faiblement partagés entre les génomes, voire spécifiques d'un unique génome. Ces gènes, considérés comme soumis à un *turnover* important et à une élimination rapide, constituent le génome *cloud*.

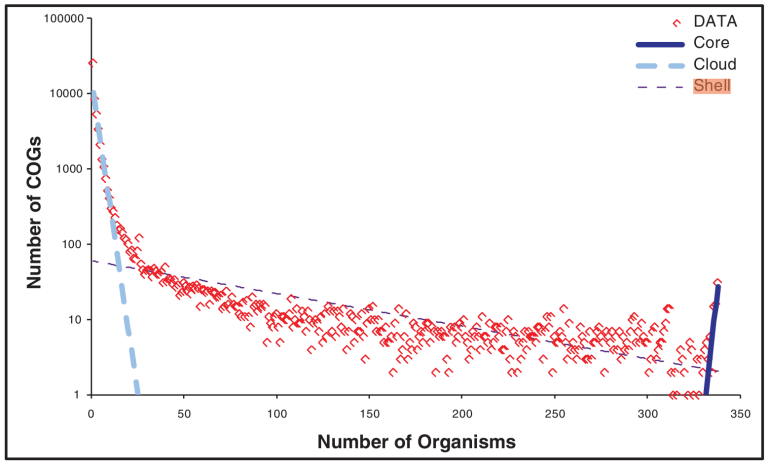


Figure 1.11. Représentation du nombre de clusters de gènes orthologues (COGs) en fonction du nombre d'organismes considérés (Koonin, 2011). Si les génomes sont assimilés à un « univers » ou espace génomique (Koonin and Wolf, 2008), les COGs montrent une distribution unique au travers de ceux-ci qui peut être approximée avec trois fonctions exponentielles qui partitionnent les gènes en trois catégories distinctes, *i.e.*, les gènes *core*, *shell* et *cloud*. Dans ce schéma : le nombre maximal de procaryotes considérés est de 338 provenant de la collection EggNOG.

La structure, la taille et les caractéristiques du pangénome peuvent varier d'une espèce à l'autre (McInerney *et al.*, 2017; Medini *et al.*, 2005) et être appréhendées à travers l'analyse d'un nombre croissant de génomes pour une espèce donnée. En effet, s'il est possible de définir la part du génome *core* et du génome flexible à partir de la comparaison de deux génomes *a minima*, la part relative de ces deux compartiments varie à mesure que le nombre de génomes considérés pour l'espèce croît. L'acquisition indépendante de gènes par HGTs dans les génomes de différents isolats a en effet pour conséquence directe d'augmenter le nombre de gènes flexibles, mais aussi de diminuer le nombre de gènes *core* car certains, initialement considérés comme *core,* ne sont alors plus partagés par tous les génomes. Actuellement, deux profils types de pangénomes sont distingués :

- les pangénomes dits fermés, caractéristiques des espèces pour lesquelles la variabilité du contenu en gènes est relativement faible – *i.e.*, avec un génome *core* important et un génome flexible de petite taille (McInerney *et al.*, 2017) (Figure 1.12A). Le nombre de nouveaux gènes flexibles identifiés par l'ajout de génomes analysés est ainsi très faible alors que les gènes partagés par un maximum d'individus sont largement majoritaires (Figure 1.12B).

C'est ce qui est observé par exemple pour l'espèce pathogène intracellulaire, *Chlamydia trachomatis*, dont la taille du génome *core* constitue 84 % du pangénome ;

- les pangénomes qualifiés d'ouverts, caractérisés par un génome flexible montrant une « expansion » avec l'ajout de génomes. Ceci est observé, par exemple, pour l'espèce *E. coli* pour laquelle l'analyse de 2 000 génomes a révélé que son pangénome se composait d'approximativement 3 200 gènes *core* (communs à 95 % des génomes) et de 90 000 familles de gènes différentes (Land *et al.*, 2015), d'où une importante variabilité du contenu en gènes entre les individus.

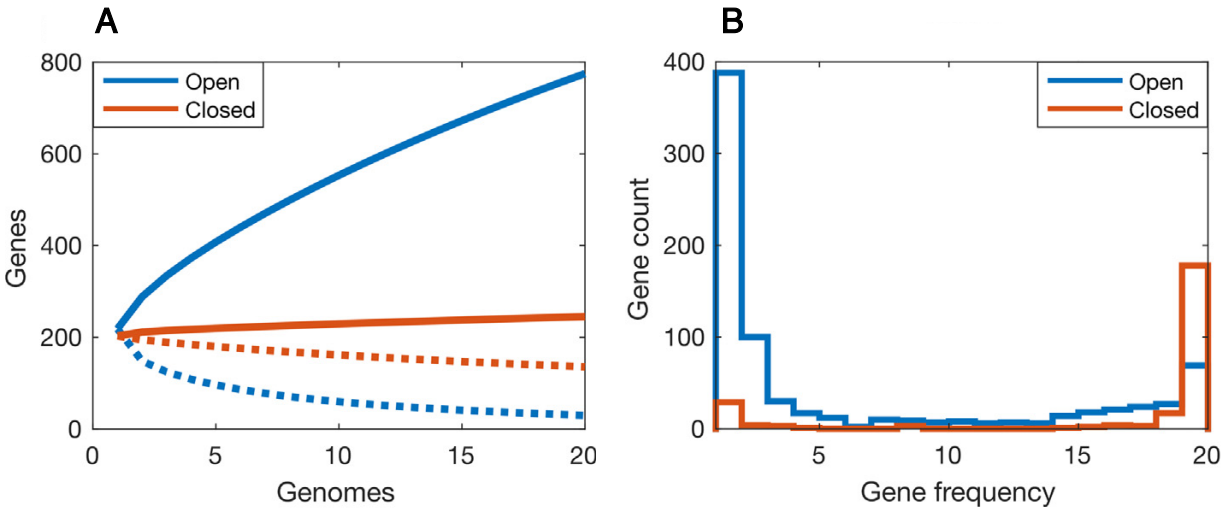


Figure 1.12. Pangénomes ouverts *versus* fermés (Domingo-Sananes and McInerney, 2021). (A) Courbes d'accumulation de l'ensemble des gènes (lignes pleines) en fonction d'un pangénome ouvert (bleu) ou fermé (orange). Les lignes pointillées représentent les courbes d'appauvrissement en gènes *core* en fonction 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. L'inverse est observé pour un pangénome fermé (orange). **(B)** Distribution de la fréquence des gènes selon que le pangénome est ouvert (bleu) ou fermé (orange).

Ainsi, ces dernières décennies, les observations portant sur l'analyse des pangénomes de différents groupes microbiens ont fait émerger certaines questions concernant :

- la nature finie ou infinie du pangénome associée à l'évolution de sa taille à mesure que de nouveaux génomes sont séquencés ;

- la définition des propriétés liées à l'échantillonnage des représentants de l'espèce (en termes de nombre de représentants mais également par rapport à la nature de la diversité considérée – comparaison d'isolats *versus* comparaison d'individus d'une même population) afin d'évaluer la taille du pangénome. Brockhurst *et al*. (2019) avancent que la comparaison d'un petit nombre de génomes bactériens échantillonnés dans de nombreuses niches est susceptible de produire une abondance de gènes accessoires rares. Cependant, ceux-ci pourraient représenter des gènes accessoires adaptatifs localement abondants mais globalement rares, ou des gènes délétères à la fois localement et globalement rares ;
- l'évaluation et la validation des différentes théories avancées pour expliquer l'existence des pangénomes chez les procaryotes mais aussi les mécanismes évolutifs impliqués dans leur émergence et leur maintien.

Des développements théoriques ont été réalisés afin de comprendre la dynamique évolutive du pangénome, les modalités de la répartition des gènes au sein des génomes *core* et flexible et ce qui distingue les groupes de gènes *shell* et *cloud*.

1.3.2. Modélisation de la dynamique du pangénome

Modéliser la dynamique d'un pangénome implique de décrire les facteurs qui gouvernent les génomes *core* et flexible. Ces deux composantes ne sont toutefois pas forcément soumises aux mêmes processus et sont susceptibles de ne pas répondre avec la même dynamique. Parmi les modèles développés, les modèles d'évolution neutre de pangénome sont basés sur les hypothèses de gains et pertes de gènes aléatoires (modèle *birth and death*). Ils constituent des modèles « nuls » d'évolution au sens où ils proposent la description d'un *pattern* en l'absence de mécanismes tels que la sélection ou la recombinaison. Ils constituent un point de comparaison par rapport à des modèles moins parcimonieux (impliquant notamment la sélection). Tous ces modèles sont ensuite mis en perspective d'analyses de génomique comparative.

Les premiers modèles décrivant un processus d'évolution neutre du pangénome ont été proposés par Baumdicker *et al.* (2012), d'une part, et Lobkovsky *et al.* (2013),

d'autre part. Dans ces modèles, le gain de gènes peut résulter d'un HGT depuis « l'extérieur » ou d'une mutation sur un gène existant conduisant à un nouveau gène (Baumdicker *et al.*, 2012). Pour ce qui concerne la perte de gènes, celle-ci peut par exemple représenter une perte de fonction sur un gène suite à une mutation, conduisant à sa disparition dans la descendance de l'individu qui la porte. Le modèle de Lobkovsky *et al.* (2013) est basé sur l'idée que les génomes conservent une taille constante à l'échelle de l'arbre phylogénétique (modèle S*tationary Genome on Tree* ou SGT). Dans ce cas, la perte de gènes est décrite par un processus de remplacement de manière à maintenir constante la taille du génome. L'évolution du pangénome est décrite à travers la définition de taux de renouvellement des gènes. L'évaluation de l'adéquation de ce modèle avec la répartition des gènes au sein des génomes pour différents groupes bactériens indique que le pangénome dans son ensemble n'est pas soumis à une stricte évolution neutre. Par ailleurs, cette analyse souligne que les répartitions des gènes observées sont mieux décrites par un modèle avec des catégories de taux de renouvellement des gènes, qui représentent finalement le génome *core* et le génome flexible. Le modèle de Baumdicker *et al.* (2012), appelé *Infinitely Many Genes* (ou IMG), repose pour sa part sur la généalogie de génomes individuels appartenant à une population. Il modélise la dérive génétique ainsi que les gains et pertes de gènes flexibles appartenant à un réservoir de gènes potentiellement infini. En ce point, le modèle IMG s'apparente au modèle des sites infinis bien connu en génétique des populations (Kimura, 1969). Du fait de la nature même du réservoir de gènes, le modèle IMG est basé sur l'hypothèse selon laquelle chaque gène ne peut être acquis qu'une seule fois. Il s'apparente ainsi à un modèle neutre d'évolution du pangénome. Ce modèle offre un cadre théorique permettant de prédire, à partir de données observées, la taille moyenne du génome d'un individu au sein d'une espèce, la taille du pangénome de cette espèce, ou la fréquence des gènes au sein du génome flexible, sous l'hypothèse d'un processus neutre d'évolution. Il a été testé pour les genres bactériens *Prochlorococcus* et *Synechococcus*. La distribution observée de la fréquence des gènes flexibles au sein de 11 génomes pour chacun de ces deux genres a été comparée à la distribution attendue de la fréquence de ces mêmes gènes sous l'hypothèse du modèle IMG (test du χ^2). Il est apparu que celle pour *Prochlorococcus* n'était pas significativement différente de la distribution sous le modèle IMG et que l'hypothèse nulle d'une évolution neutre du génome flexible était acceptée pour ce genre. A l'inverse, elle est rejetée pour

Synechococcus, suggérant une déviation au modèle nul et une évolution non neutre pour ce dernier.

Le modèle IMG sous-entend l'existence de deux classes de gènes (1D+E) représentant respectivement le génome *core* (E, gènes essentiels) et le génome flexible (D, gènes accessoires – *dispensables*). Une évolution de ce modèle par Collins and Higgs (2012) avec la considération de deux classes de gènes accessoires caractérisées par des taux de gains et pertes de gènes différents (2D+E) permet cependant un meilleur ajustement aux données observées, tant pour l'estimation de la taille du pangénome que l'adéquation à la distribution en U de la fréquence des gènes (Figure 1.13) (Collins and Higgs, 2012). Ce modèle s'apparente à la distinction des génomes *core*, *shell* et *cloud* (Koonin and Wolf, 2008).

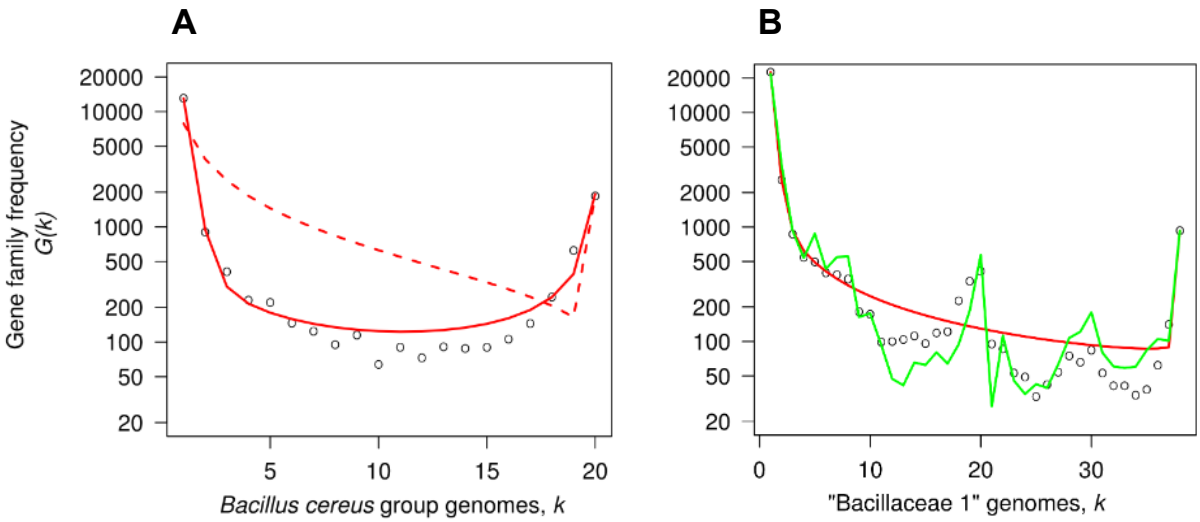


Figure 1.13. Spectre de fréquence des familles de gènes (Collins and Higgs, 2012). **(A)** Spectre calculé pour 20 génomes séquencés de *Bacillus cereus* (cercles). Le spectre de fréquence a été calculé avec les paramètres optimisés à partir de l'ajustement du modèle 1D+E (ligne pointillée) ou 2D+E (ligne continue) en fonction de la fréquence des familles de gènes *G(k)* en utilisant un arbre coalescent. **(B)** Spectre calculé pour 38 génomes séquencés de *Bacillaceae* (cercles). Le spectre de fréquence a été calculé avec les paramètres du modèle 2D+E sur la base d'un arbre coalescent (ligne rouge) et d'un arbre phylogénétique inféré (ligne verte). D : gènes accessoires – *dispensables* ; E : gènes essentiels.

Ces modélisations ont apporté des éléments de réponses quant aux hypothèses évolutives qui sous-tendent la dynamique des pangénomes. Bien qu'il ait été proposé que les gains et pertes de gènes, et par extension l'évolution du génome flexible, soient neutres (Baumdicker *et al.*, 2012), il semblerait que l'on puisse inclure des forces sélec-

tives dans la description de la dynamique évolutive du pangénome (Collins and Higgs, 2012; Lobkovsky *et al.*, 2013). Celles-ci impliquent que les gènes *core*, et les gènes flexibles ayant une fréquence élevée, seraient en moyenne avantageux et maintenus par sélection, alors que les gènes rares seraient plus susceptibles d'être neutres.

1.3.3. Caractère adaptatif du pangénome ?

Outre les modélisations du pangénome intégrant un cadre théorique neutre, des hypothèses alternatives considèrent que l'évolution des pangénomes est soumise à des processus de sélection. Le débat « neutraliste *versus* sélectionniste » est alimenté par le fait que certains auteurs considèrent que les HGTs sont principalement neutres ou délétères (Gogarten and Townsend, 2005; Vos *et al.*, 2015), alors que d'autres les considèrent comme étant adaptatifs (Sela *et al.*, 2016; McInerney *et al.*, 2017).

Même si de nombreux HGTs adaptatifs ont été relatés, il semble cependant que la plupart des changements induits par ceux-ci ont une durée de vie courte en termes de contenu en gènes (Hao and Golding, 2010; Didelot *et al.*, 2012). En effet, le nombre d'événements de HGTs détectés ne décroît pas linéairement avec l'augmentation des temps de divergence des taxa, à l'exception des gènes peu conservés et peu contraints (Bolotin and Hershberg, 2016). Ceci suggère que le *pool* de gènes accessoires n'est pas purement adaptatif, la plupart des HGTs pouvant (i) être délétères, (ii) 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), ou (iii) être neutres.

Par ailleurs il a été démontré que la fluidité des génomes – définie comme le rapport entre les familles de gènes uniques et la somme des familles de gènes communes à une paire de génomes pour un groupe bactérien donné (Kislyuk *et al.*, 2011) – qui reflète finalement celle du pangénome, est positivement corrélée à la diversité génétique des gènes *core* (Figure 1.14) (Andreani *et al.*, 2017). Dans la mesure où les espèces qui possèdent une plus grande diversité nucléotidique ont une taille efficace des populations *Ne* plus importante (*cf.* partie 1.2), on peut faire l'hypothèse que la variation du contenu en gènes, et donc la diversité du pangénome, est impactée par celle-ci.

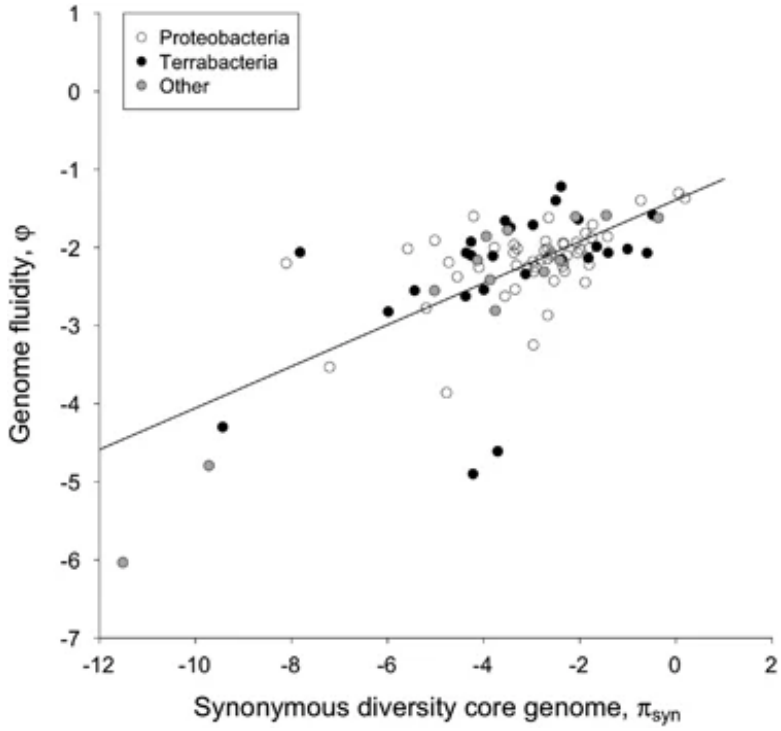


Figure 1.14. Représentation de la fluidité des génomes φ en fonction de la diversité nucléotidique synonyme des gènes *core* (Andreani *et al.*, 2017). La fluidité génomique a été estimée pour 90 espèces procaryotiques telle que décrite dans Kislyuk *et al.* (2011) et est représentée en fonction de la diversité génétique des gènes *core*. Les échelles des axes x et y sont en logarithme népérien. Points blancs : protéobactéries ; points noirs : terrabactéries (actinobactéries, firmicutes et cyanobactéries), points gris : autres taxa.

L'efficacité de la sélection est également affectée par *Ne* (*cf.* partie 1.2). Dans le cas où l'évolution du pangénome serait gouvernée par la sélection, *Ne* pourrait conditionner le nombre de gènes flexibles qui serait effectivement conservé par celle-ci. C'est l'idée principale amenée par l'hypothèse évolutive de barrière à la dérive (Bobay and Ochman, 2018), telle que proposée pour l'évolution des taux de mutations (Sung *et al.*, 2012). Basée sur la théorie quasi-neutraliste de l'évolution, celle-ci prédit que dans le cas d'une population dont la taille efficace est petite, la perte de gènes, bien que légèrement avantageux, serait induite par la dérive génétique (Figure 1.15). Dans ce cas, seuls les gènes accessoires (les plus) bénéfiques seraient conservés, alors que les populations ayant une *Ne* importante pourraient voir maintenus dans leurs génomes des gènes dont la valeur sélective est moins prononcée et donc présenter un nombre de gènes flexibles relativement important. Ceci est soutenu par l'existence d'une relation positive entre la taille du pangénome et *Ne* (Bobay and Ochman, 2018).

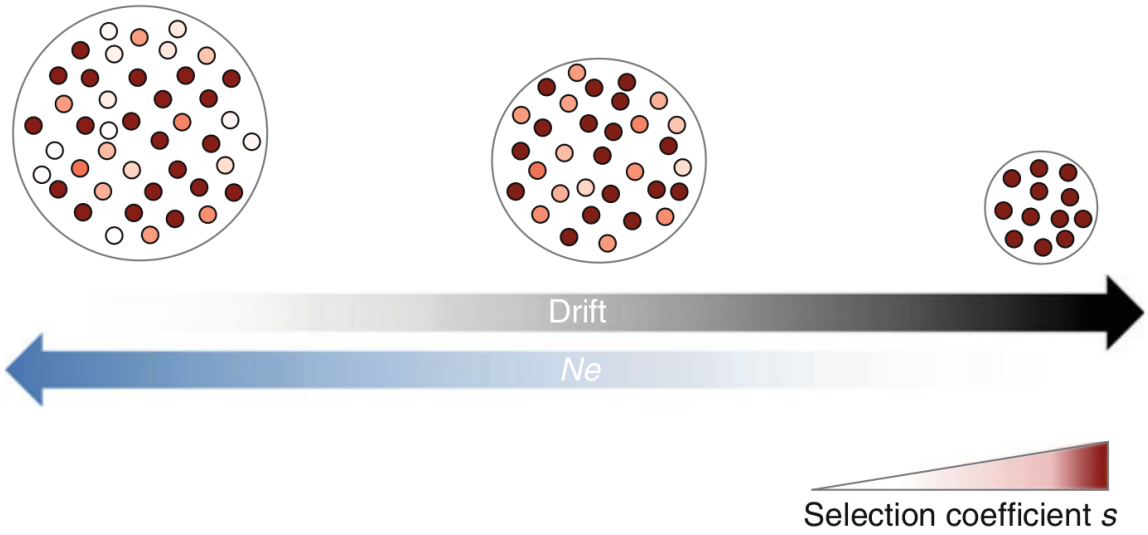


Figure 1.15. Modèle de barrière à la dérive de l'évolution du pangénome (Bobay, 2020). Sous le modèle de barrière à la dérive, la taille des pangénomes est une fonction de la taille efficace des populations *Ne*. Les espèces pour lesquelles *Ne* est importante sont moins sujettes à la dérive et peuvent ainsi conserver des gènes dont l'avantage adaptatif est faible (gauche). Lorsque *Ne* diminue, les gènes de faible valeur adaptative seront perçus comme neutres et perdus par dérive (centre). En cas de forte dérive, comme attendu lorsque les espèces ont une très petite *Ne*, seuls les gènes les plus avantageux seront conservés par sélection, ce qui se traduira par des pangénomes de petite taille composés principalement de gènes *core* (droite). Chaque grand cercle représente un pangénome et reflète sa taille, au sein duquel les petits cercles représentent des gènes individuels. Le gradient de couleur reflète le coefficient de sélection des gènes.

1.4. *Prochlorococcus* – un taxon modèle

Avec une population globale estimée à 10^{27} cellules environ, les cyanobactéries marines du genre *Prochlorococcus* forment un groupe de micro-organismes photosynthétiques parmi les plus abondants de la zone euphotique océanique (Figure 1.16) (Partensky *et al.*, 1999; Flombaum *et al.*, 2013). Ainsi, ce genre joue un rôle majeur dans le cycle du carbone puisqu'il fixe au moins quatre Gt de CO_2 atmosphérique chaque année et est responsable de près de 10 % de la productivité primaire des océans (Flombaum *et al.*, 2013). Par ailleurs, *Prochlorococcus* est présent dans la quasi-totalité des océans, de l'équateur aux pôles, de la surface jusqu'à presque 200 m de profondeur, reflétant ainsi une très forte capacité d'adaptation à des niches écologiques contrastées.

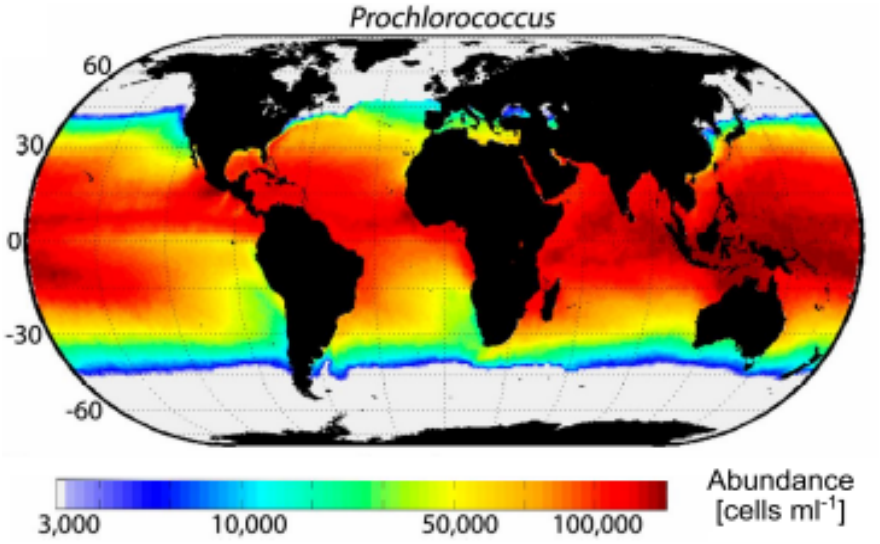


Figure 1.16. Répartition mondiale de l'abondance de la bactérie photosynthétique *Prochlorococcus marinus* (Flombaum *et al.*, 2013). L'abondance est calculée en cellules/ml.

La diversité génétique de *Prochlorococcus,* structurée en clades caractérisés par des ensembles de gènes fonctionnels en lien avec des contextes environnementaux particuliers, suggère un cloisonnement stable de groupes écologiquement distincts, également appelés écotypes (Kent *et al.*, 2016 ; Larkin *et al.*, 2016). L'étude de souches cultivées de ce genre et appartenant à différents écotypes a par ailleurs mis en évidence des génomes

de taille inférieure à 1,7 Mpb pour certaines d'entre-elles (Rocap *et al.*, 2003), faisant ainsi partie des plus petits génomes de bactéries libres de l'environnement. De part ses caractéristiques écologiques, physiologiques et évolutives particulières, *Prochlorococcus* a donc suscité un vif intérêt parmi la communauté scientifique et de nombreux travaux ont été développés dans le but d'élucider les mécanismes à l'origine de son succès écologique. Il a notamment été montré que, bien que les représentants de ce genre présentaient une identité nucléotidique de leur gène de l'ARNr 16S supérieure à 96 %, ces souches avaient une divergence nucléotidique importante à l'échelle de leur génome, certaines d'entre elles partageant moins de 70 % d'identité (Zhaxybayeva *et al.*, 2009). Par conséquent, les questions liées à la définition de l'espèce s'appliquent également au genre *Prochlorococcus* ainsi que les conséquences directes de cette définition sur l'étude et les limites du pangénome.

1.4.1. Diversité écologique

Les premières études de diversité pour le genre *Prochlorococcus,* basées sur l'analyse des gènes de l'ARN 16S, ont révélé l'existence de deux groupes phylogénétiques adaptés à des conditions d'intensités lumineuses distinctes (Moore *et al.*, 1998), avec un écotype présent dans les zones de fortes intensités (écotype HL – *high-light*) et un autre, dans celles de faibles intensités (écotype LL – *low-light*). La lumière est donc l'un des paramètres forts à l'origine de la diversification de *Prochlorococcus,* ce qui est cohérent avec son développement à différentes profondeurs de la zone euphotique océanique. L'analyse de ces écotypes a permis de mettre en évidence, à une échelle plus fine, 12 clades environnementaux (Figure 1.17) qui se distinguent sur plusieurs critères, comme leurs caractéristiques photo-physiologiques (données de laboratoire) (Moore *et al.*, 1998), la séquence de leurs espaceurs internes transcrits (*Internal Transcribed Spacer* ou ITS) (Rocap *et al.*, 2003), leurs informations génomiques (Biller *et al.*, 2014a; Kashtan *et al.*, 2014; Larkin *et al.*, 2016) et leurs répartitions le long de la colonne d'eau (Malmstrom *et al.*, 2010). Par ailleurs, d'un point de vue phylogénétique, les clades appartenant à l'écotype HL (HLI à HLVI) forment un groupe monophylétique alors que ceux appartenant à l'écotype LL (LLI à LLVII) sont paraphylétiques (Figure 1.17) (Biller *et al.*, 2014b).

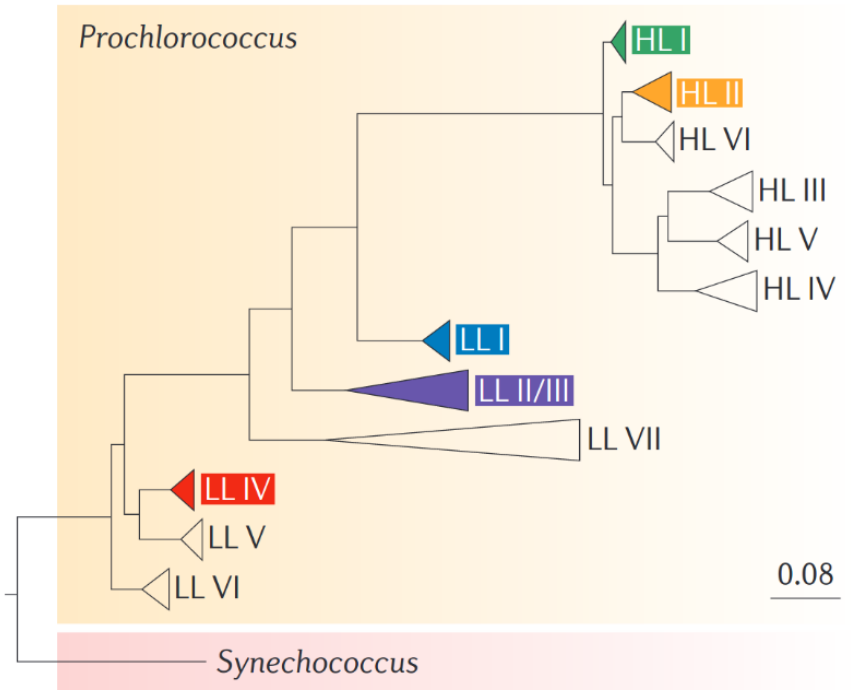


Figure 1.17. Distribution phylogénétique des écotypes de *Prochlorococcus high-light* (HL) et *low-light* (LL) (Biller *et al.*, 2014b). La phylogénie a été construite à partir de l'analyse des séquences des espaceurs internes transcrits (*Internal Transcribed Spacer* ou ITS) de l'opéron ribosomique. Seuls cinq des 12 clades environnementaux caractérisés possèdent un représentant cultivé en laboratoire (en couleur sur la phylogénie ; HLI, HLII, LLI, LLII/III et LLIV). Le genre *Synechococcus* est utilisé comme groupe externe.

Au sein de l'écotype HL, plusieurs paramètres peuvent en partie expliquer la diversification des clades. Ainsi, la température joue un rôle clé dans la distinction des clades HLI et HLII (Figure 1.18), le premier étant adapté à de plus faibles températures (Johnson *et al.*, 2006) alors que la concentration en fer soutient la distinction des clades HLIII et HLIV, leur abondance étant plus importante dans les environnements présentant une plus faible concentration (Rusch *et al.*, 2010; West *et al.*, 2011; Huang *et al.*, 2012). Pour les clades LL, moins de données sont disponibles concernant leur répartition. Cependant, il a été montré que ces clades sont retrouvés à des profondeurs plus importantes (Figure 1.18), où la concentration en nutriments est plus élevée. Ainsi, le clade LLI se localise préférentiellement près de la nutricline (zone moyenne de la zone euphotique, localisée sous les eaux de surface dominées par les clades HL), tandis que le clade LLIV est

plus présent dans la partie inférieure de la zone euphotique (Ahlgren *et al.*, 2006; Zinser *et al.*, 2007).

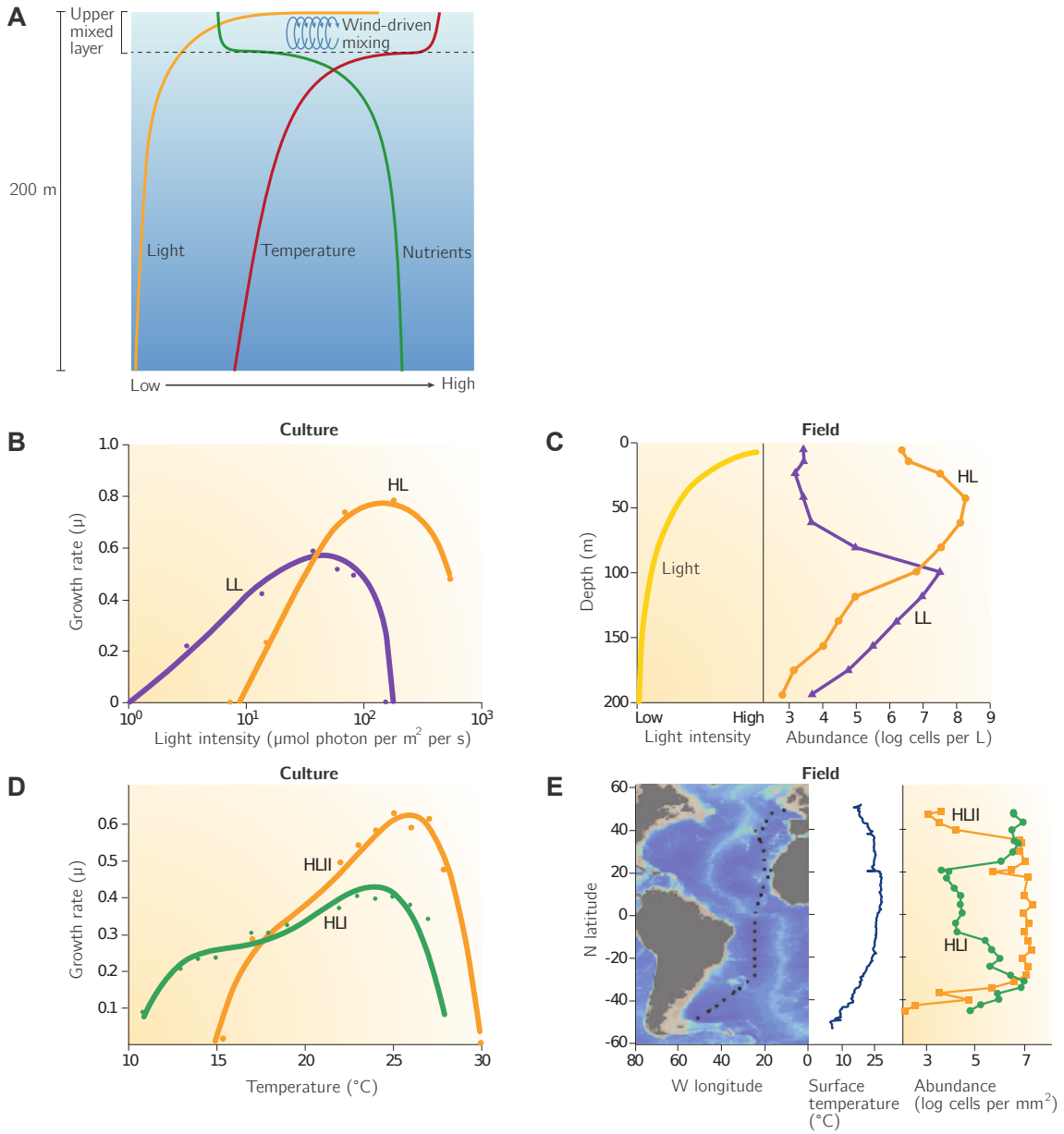


Figure 1.18. Répartition géographique des écotypes de *Prochlorococcus* (Biller *et al.*, 2014b). (A) *Prochlorococcus* est distribué sur l'ensemble de la zone euphotique (entre 0 et 200 m de profondeur), zone caractérisée par des gradients de lumières, de températures et de nutriments. La colonne d'eau se divise en une couche supérieure « mixte » (la turbulence due au vent et à la chaleur homogénéise la distribution des nutriments et des cellules) et des eaux profondes stratifiées, où des gradients de nutriments se forment en raison de l'activité biogéochimique. Les niveaux de lumières diminuent de façon exponentielle ; les gradients de températures et de nutriments sont largement similaires au sein de la couche mixte, tandis que la

température diminue et les concentrations en nutriments augmentent avec la profondeur. **(B-E)** Relations entre les optimums de croissance des souches cultivées en laboratoire et leur abondance dans l'environnement. Ces relations ont été plus clairement démontrées par le partitionnement longitudinal et en profondeur des écotypes en fonction de la lumière **(B et C)** ou de la température **(D et E)**.

1.4.2. Diversité génomique et pangénome

Dès 2003, le séquençage des génomes de souches de différents clades de *Prochlorococcus* a permis de constater des variabilités au niveau de leur taille, de leur architecture et de leur contenu en gènes (Dufresne *et al.*, 2003; Rocap *et al.*, 2003). Ces premiers travaux ont révélé des tailles de génome comprises entre environ 1,7 et 2,7 Mpb (Figure 1.19). De manière générale, les génomes du genre *Prochlorococcus* sont plus petits que ceux de la plupart des autres cyanobactéries. Ceci s'explique par une perte massive de gènes, peu après la divergence des genres *Prochlorococcus* et *Synechococcus* (Sun and Blanchard, 2014). La radiation des différentes lignées de *Prochlorococcus* et leur spécialisation en fonction de niches écologiques sont postérieures à cette réduction de taille. Les génomes des écotypes HL sont généralement plus petits (compris entre 1,66 et 1,71 Mpb), avec un contenu en dinucléotides GC de l'ordre de 30 %, tandis que ceux des écotypes LL ont des tailles plus variables (comprises entre 1,69 et 2,68 Mpb), avec un contenu en GC allant jusqu'à 50 % (Figure 1.19). Pour l'écotype HL, les génomes sont qualifiés de *streamlined* (simplifiés ou rationalisés) du fait de leur petitesse et représentent le premier exemple documenté d'un tel phénomène de réduction pour des organismes libres de l'environnement (Rocap *et al.*, 2003; Dufresne *et al.*, 2005; Giovannoni *et al.*, 2014).

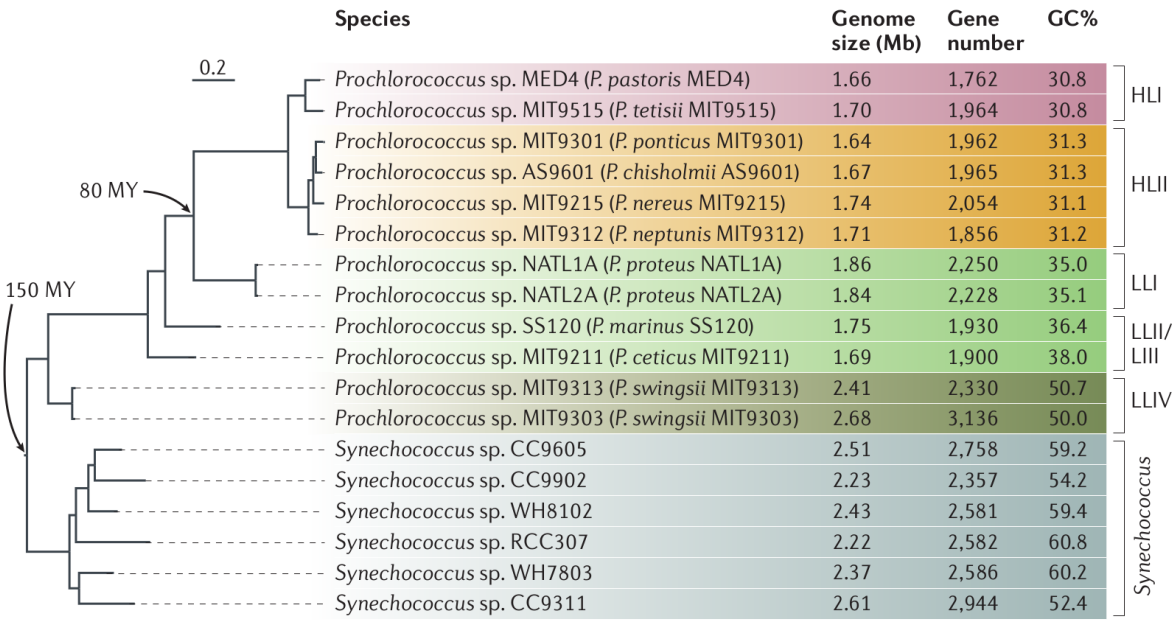


Figure 1.19. Arbre phylogénétique et données génomiques pour différents écotypes du genre *Prochlorococcus* et pour le genre *Synechococcus* (Batut *et al.*, 2014). L'arbre phylogénétique a été reconstruit par la méthode du maximum de vraisemblance à partir d'un ensemble de gènes orthologues présents en copie unique. Les temps de divergence sont issus de l'étude de Dufresne *et al.* (2005). Les données concernant la taille des génomes, le nombre de gènes et le contenu en GC sont issus de la base de données NCBI. Les noms entre parenthèses sont les noms d'espèces telles qu'indiquées par Thompson *et al.* (2013).

La variation de la taille des génomes de *Prochlorococcus* s'accompagne d'une variation de leur contenu en gènes. Ainsi, aux 1 200 gènes, environ, partagés par l'ensemble des souches et qui constituent le génome *core*, s'ajoute un nombre variable de gènes flexibles, propres aux individus, et qui constituent le génome flexible. Ceux-ci sont au nombre de 1 000 à 1 500 au sein des génomes des souches cultivées et définissent des catalogues de gènes et de fonctions variables entre clades. Les gènes flexibles sont répartis sur l'ensemble du génome mais tendent cependant à être plus concentrés dans des régions hypervariables particulières appelées îlots génomiques (ISLs) (Figure 1.20) (Kettler *et al.*, 2007; Dufresne *et al.*, 2008). Ces ISLs sont caractérisés par une densité en gènes flexibles bien supérieure à celle du reste du génome et par un taux en nucléotides GC faible. Ces îlots sont en général bornés par des gènes des ARNt, souvent en association avec des éléments génétiques mobiles qui, dans le cas de *Prochlorococcus*, s'apparentent à des structures cargo appelées *tycheposons* (Hackl *et al.*, 2020). Les gènes flexibles présents dans ces ISLs interviennent notamment dans l'adaptation aux niches environnementales et

dans la défense contre les phages (Coleman *et al.*, 2006; Avrani *et al.*, 2011; Delmont and Eren, 2018). Ainsi, ils contribueraient au succès écologique de chaque lignée au sein d'un environnement spécifique (Coleman *et al.*, 2006; Kettler *et al.*, 2007).

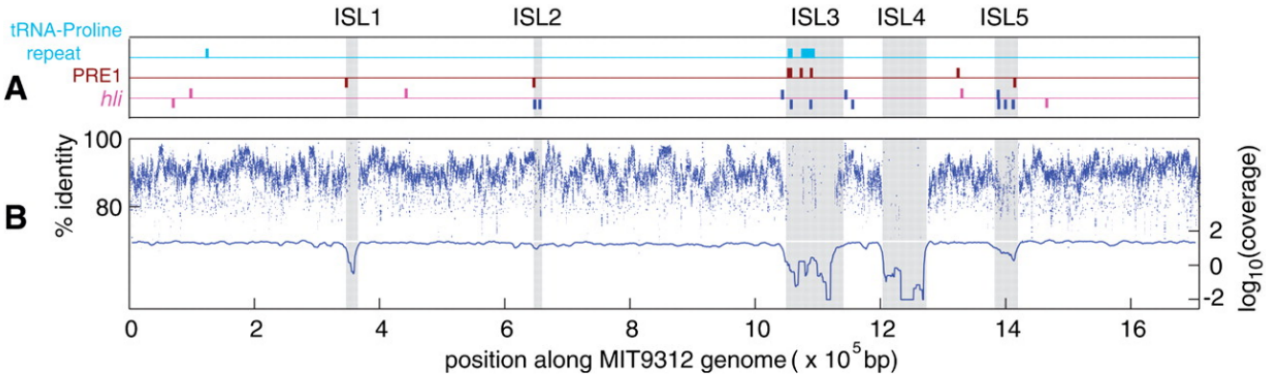


Figure 1.20. Caractéristiques des îlots génomiques (ISLs) présents dans le génome de la souche MIT9312 de *Prochlorococcus* et comparaison avec ceux présents dans le génome de souches environnementales (Coleman *et al.*, 2006). Les ISLs (ISL1 à ISL5) ont été identifiés par analyse de la synténie entre les génomes de souches affiliées à deux clades de l'écotype HL, *i.e.*, MIT9312 (HLII) et MED4 (HLI) **(A)** Localisation d'éléments répétés (ARNt et PRE1) et des gènes *hli (high-light inducible genes)* dans le génome de MIT9312, représentés au-dessus et en-dessous de la ligne verticale, en fonction des brins sens et antisens. **(B)** Les séquences de souches environnementales provenant d'un métagénome de la mer des Sargasses (Venter *et al.*, 2004) ont été alignées contre le génome de MIT9312. Les pourcentages d'identité ainsi que la couverture le long du génome de MIT9312 sont indiqués.

Alors que les souches de *Prochlorococcus* ont des génomes de petite taille, le pangénome qui en découle est relativement important. Ceci est supporté par le fait que l'annotation de chaque nouveau génome donne lieu à la description de 150 nouveaux gènes, en moyenne, encore jamais observés au sein de ce genre (Kettler *et al.*, 2007). L'absence de saturation sur les courbes d'accumulation indique que ce pangénome est ouvert (Figure 1.21).

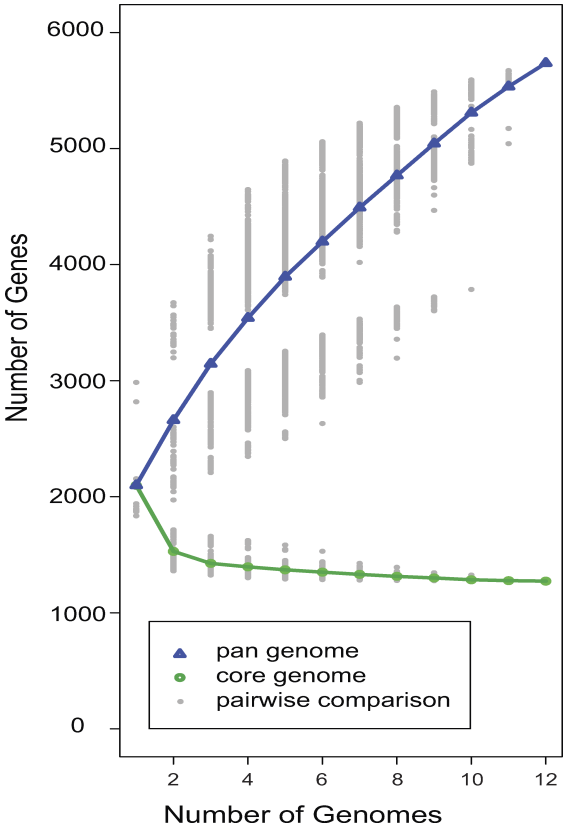


Figure 1.21. Représentation de la taille du pangénome et de celle du génome *core* chez *Prochlorococcus* en fonction du nombre de génomes analysés (Kettler *et al.*, 2007). La taille du pangénome et celle du génome *core* dépendent du nombre de génomes utilisés dans l'analyse. Si k génomes ont été sélectionnés sur un total de 12, il existe $12!/(k!(12\text{-}k)!)$ combinaisons possibles de comparaison de génomes pour les calculs de taille du pangénome et du génome *core*. Chaque sélection possible est représentée par un point gris, et les lignes de couleur représentent la moyenne.

Plus récemment, l'analyse des 41 génomes issus de souches cultivées a permis d'estimer la taille du pangénome de *Prochlorococcus* à plus de 80 000 gènes (Biller *et al.*, 2014). Cette analyse suggère que seule une infime partie des gènes du pangénome de *Prochlorococcus* a été identifiée. Ainsi, l'hypothèse de l'existence d'un *pool* de gènes bien plus important dans l'environnement, et par extension la présence de lignées potentiellement inconnues, est avancée. Au cours des dernières années, le développement de la métagénomique (Venter *et al.*, 2004; Yooseph *et al.*, 2007; Biller *et al.*, 2014b) ainsi que celui de la génomique sur cellule-unique (*Single-Cell Genomics* ou SCG) (Kashtan *et al.*, 2014, 2017; Berube *et al.*, 2018; Pachiadaki *et al.*, 2019) ont amélioré l'échantillonnage de la diversité génétique dans l'environnement, en lien avec des variables écologiques ou géographiques. Ainsi, par le biais de la métagénomique, une corrélation a été montrée entre la diversité phylogénétique du génome *core* des populations de *Prochlorococcus* présentes dans les différents océans et la distribution du contenu en gènes flexibles dans le génome de ces mêmes populations. Ces résultats suggèrent une structuration phylogénétique du contenu en gènes des populations naturelles, l'existence de facteurs environnementaux

qui agissent de manière similaire sur les génomes *core* et flexible ou la prévalence de HGTs entre lignées proches supérieure à ce qui était envisagé jusqu'ici (Kent *et al.*, 2016).

1.4.3. Fine échelle de diversité

Dans le cadre d'une étude réalisée sur le long terme (différentes saisons) dans l'océan Atlantique (site *Bermuda Atlantic Time-series Study* ou BATS) par une approche SCG, il a récemment été proposé que le genre *Prochlorococcus* se compose de centaines de sous-populations coexistantes résultant probablement d'un ancien cloisonnement de niche (Kashtan *et al.*, 2014). Celles-ci sont inférées sur la base des fluctuations saisonnières de leur abondance relative et la fixation d'allèles différents de leurs gènes *core* qui pourraient conférer une stabilité à ce « collectif » (Figure 1.22). Une étude fine de la diversité génétique des sous-populations de l'écotype HLII a par ailleurs révélé un schéma de diversité génétique caractérisée d'une part par un « squelette génomique » (ou *backbone*) commun aux sous-populations, composé essentiellement de gènes *core* avec des allèles distincts entre sous-populations et d'autre part, de plusieurs ISLs composés majoritairement de gènes flexibles, mais aussi d'allèles spécifiques au niveau des gènes *core* (Biller *et al.*, 2014b; Kashtan *et al.*, 2014). Une structuration des génomes amplifiés sur cellules-uniques (*Single Amplified Genomes* ou SAGs) en sous-populations a par ailleurs été retrouvée par une approche sans *a priori* basée sur la distribution de la longueur des segments génomiques strictement identiques entre couples de génomes, ceux-ci étant vus comme un *proxy* de l'estimation des flux de gènes récents entre génomes (Arevalo *et al.*, 2019). Ainsi, les sous-populations seraient également caractérisées par des flux de gènes qui se produiraient préférentiellement en leur sein, la réduction des flux de gènes entre sous-populations suggérant un processus de différentiation (spéciation) dont il reste à préciser les contours. La répartition des différentes sous-populations dans l'espace et le temps serait ainsi conditionnée par les combinaisons des gènes *core* et flexibles propres à chaque niche environnementale (Kashtan *et al.*, 2014) (Figure 1.22B).

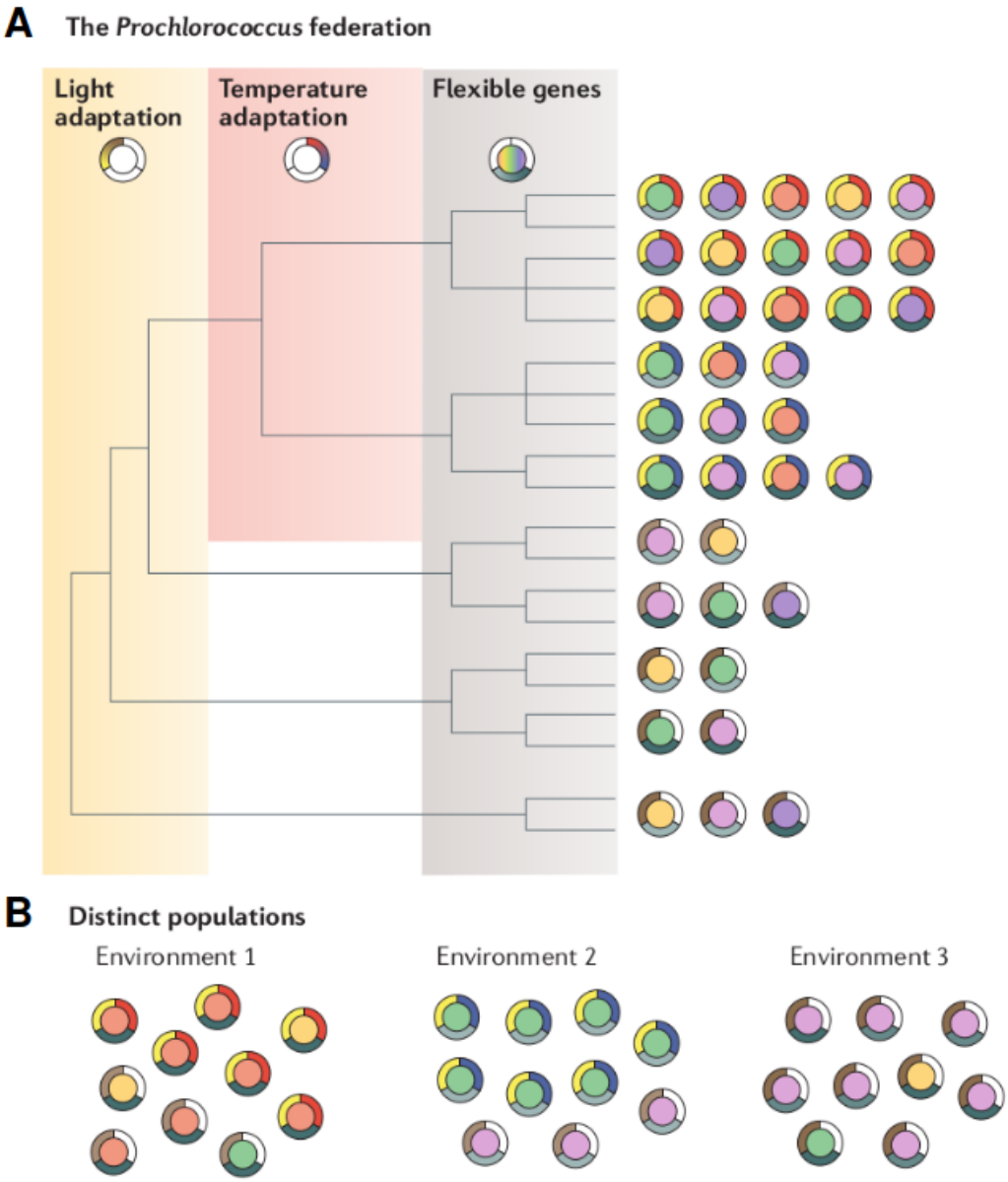


Figure 1.22. *Prochlorococcus* en tant que « fédération » (Biller *et al.*, 2014). (A) *Prochlorococcus* peut être considéré comme une « fédération » de cellules coexistantes. Chaque groupe de cellules représente différents arrangements combinatoires de gènes nécessaires à l'adaptation à des niches écologiques distinctes. Chaque cercle représente une bactérie individuelle ou une lignée clonale. L'anneau coloré extérieur représente le *backbone* (ou squelette génomique, contenant à la fois des gènes *core* et flexibles), alors que le centre représente un ensemble unique de gènes flexibles. Le *backbone* est constitué d'allèles qui déterminent l'adaptation à des caractéristiques basales, liées à une diversification associée à une adaptation ancienne (*e.g.*, lumière ou encore température optimale de croissance) ainsi que d'un sous-ensemble de gènes flexibles contribuant à l'adaptation à des micro-niches. La composition du *backbone* reflète généralement la phylogénie des génomes complets. Cependant, la composition en gènes flexibles peut varier considérablement en fonction de l'environnement

local. **(B)** La diversité au sein de la « fédération » contribue à la stabilité et à la résilience des populations de *Prochlorococcus*, à l'échelle globale, en fournissant un vaste ensemble de caractéristiques diverses qui peuvent être sélectionnées par différentes conditions environnementales.

1.5. Objectifs de thèse

Sur la base des connaissances actuelles abordées dans l'introduction en lien avec les concepts biologique et écologique de l'espèce chez les procaryotes, l'objectif général de mes travaux de recherche visait à reconsidérer les processus à l'origine de la différenciation des populations bactériennes de l'environnement à la lumière des mécanismes évolutifs populationnels et de la dynamique de leur génome. A l'échelle des génomes, la différenciation des populations bactériennes peut être appréhendée en termes de divergence nucléotidique et de dynamique du pangénome. Ainsi la question qui anime plus particulièrement ma thèse est la suivante : quelles sont les forces évolutives à l'origine du pangénome ? Pour répondre à cet objectif, les sous-populations cooccurrentes appartenant à l'écotype HLII de *Prochlorococcus* sont un modèle adapté permettant l'étude des signatures évolutives caractéristiques de la différenciation des sous-populations bactériennes libres de l'environnement. Cet objectif général se décline en plusieurs points, dont certains restent encore à approfondir.

Le premier point repose sur une ré-évaluation du contenu génétique, du potentiel fonctionnel et de l'unité taxonomique du pangénome à l'échelle populationnelle. L'objectif est ici de mieux comprendre les composantes du pangénome qui caractérisent ces différentes sous-populations. Ainsi, pouvons-nous identifier des catalogues de gènes différents en fonction des sous-populations considérées et cela apporte-t-il des réponses quant à leur différenciation ?

Ceci nous amène alors au deuxième point qui porte sur l'évaluation de l'empreinte des mécanismes évolutifs impactant les gènes en fonction de leur caractère essentiel (gènes *core*) ou facultatif (gènes flexibles). Alors que le pangénome est par ailleurs caractérisé par un paysage génomique incluant des régions conservées et variables, la question se pose de savoir si ces empreintes suivent un schéma similaire. Ainsi, existe-t-il une dynamique évolutive différentielle le long du génome ?

Le dernier point s'appuie sur l'observation d'une dynamique évolutive contrastée et une répartition non-aléatoire le long des génomes des fonctions portées par les gènes flexibles. Ceci pourrait traduire l'existence d'événements de recombinaison homologue ou de HGTs et ouvre des questions à différents niveaux concernant ces flux de gènes. Ainsi, peut-on caractériser des échanges de matériel génétique entre différentes sous-populations, voire à

des niveaux de diversité plus larges ? Par ailleurs, existe-t-il des compartiments préférentiels pour l'intégration de ce matériel génétique ?

2. Approches méthodologiques

2.1. Jeux de données et traitements initiaux des séquences

2.1.1. Génomes de *P. marinus* de l'écotype HLII utilisés dans cette étude

2.1.1.1. Génomes issus de séquençage « cellule-unique »

Au total, 87 génomes de la cyanobactérie marine *Prochlorococcus* issus de séquençage « cellule-unique » (SAGs) et appartenant à l'écotype HLII (Tableau 2.1) ont été examinés.

Tableau 2.1. Présentation des caractéristiques des génomes de la cyanobactérie marine *Prochlorococcus* issus de séquençage « cellule-unique » (SAGs) utilisés dans cette étude et appartenant à l'écotype HLII (Kashtan *et al.*, 2014).

Numéro d'accession NCBI	Nom MDA	Sous-population	Complétude (%)	Contamination (%)	Taille d'assemblage (Mb)	GC (%)	nombre de contigs	Longueur moyenne des contigs (kb)	Nombre de CDS
GCF_000634835.1	495I8	C1	22,42	0,36	0,37	32,41	231	1,59	396
GCF_000634575.1	521N3	C1	37,41	1,49	0,62	31,55	442	1,4	638
GCF_000635655.1	498M14	C1	41,38	1,72	0,94	30,91	163	5,79	1 102
GCF_000635875.1	521N5	C1	44,75	1,07	0,83	31,38	425	1,94	939
GCF_000635855.1	521M10	C1	44,83	0	1,14	31,03	261	4,36	1 305
GCF_000635335.1	527L16	C1	47,41	0	0,97	31,09	183	5,31	1 145
GCF_000634315.1	498J20	C1	47,99	1,22	0,86	31,31	351	2,44	948
GCF_000634635.1	526B22	C1	59,59	0,91	0,93	31,62	498	1,87	990
GCF_000634595.1	521O23	C1	59,78	1,46	0,89	31,71	346	2,58	998
GCF_000634495.1	520D2	C1	60,78	0,68	1,01	31,42	290	3,5	1 135
GCF_000634855.1	495N4	C1	62,77	1,27	1,19	31,36	519	2,29	1 344
GCF_000635455.1	529J16	C1	63,27	1,63	0,95	31,76	384	2,47	1 027
GCF_000635635.1	498G3	C1	63,38	0,69	1,05	31,12	325	3,23	1 203
GCF_000634895.1	496E10	C1	65,07	0,33	1,09	31,34	411	2,66	1 231
GCF_000635795.1	520E22	C1	65,53	0,14	1,13	31,19	338	3,35	1 277
GCF_000633975.1	526B17	C1	65,72	0,77	1,1	31,34	378	2,92	1 239
GCF_000633995.1	526K3	C1	66,60	1,18	1,13	31,24	523	2,16	1 278
GCF_000635295.1	527G5	C1	69,31	0,41	1,17	31,05	471	2,48	1 314
GCF_000634695.1	527L15	C1	70,55	1,59	1,19	31,53	299	3,96	1 278
GCF_000634555.2	521A19	C1	70,99	1,68	1,16	31,07	331	3,49	1 345
GCF_000634935.1	497I20	C1	71,42	0,66	1,02	31,56	136	7,45	1 156
GCF_000634015.1	526N9	C1	72,13	0,54	1,27	31,17	173	7,3	1 487
GCF_000634755.1	528N20	C1	72,42	1,18	1,14	31,36	225	5,06	1 311
GCF_000635315.1	527I9	C1	75,25	0,27	1,14	31,59	159	7,17	1 300
GCF_000635175.1	519E23	C1	75,32	0,14	1,21	31,53	409	2,97	1 366
GCF_000635775.1	519O11	C1	75,33	1,04	1,28	31,17	238	5,37	1 425
GCF_000635275.1	526D20	C1	77,63	0,54	1,31	31,36	288	4,56	1 501
GCF_000635035.1	498L10	C1	78,76	0,82	1,34	31,07	381	3,53	1 565
GCF_000635695.1	498P3	C1	79,98	1,49	1,43	31,13	258	5,52	1 654
GCF_000634775.1	528N8	C1	81,11	0,63	1,37	31,25	332	4,13	1 572
GCF_000635075.1	518E10	C1	81,97	0,14	1,32	31,49	381	3,47	1 423

GCF_000635135.1	518O7	C1	82,84	0,14		1,46	31,19	231	6,33	1 703
GCF_000635115.1	518K17	C1	83,61	1,04		1,4	31,30	335	4,18	1 577
GCF_000635615.1	498F21	C1	83,79	1,31		1,39	31,14	420	3,32	1 550
GCF_000634475.1	520B18	C1	83,79	1,45		1,43	31,31	474	3,01	1 600
GCF_000635435.1	529J11	C1	84,01	1,04		1,33	31,50	205	6,49	1 514
GCF_000634055.1	527N11	C1	84,56	1,12		1,47	31,30	321	4,58	1 703
GCF_000635735.1	519D13	C1	85,01	1,36		1,4	31,35	426	3,3	1 579
GCF_000634235.1	495N16	C1	85,60	0,86		1,43	31,26	264	5,41	1 646
GCF_000635195.1	519L21	C1	86,19	1,36		1,43	31,19	309	4,61	1 614
GCF_000634715.1	528K19	C1	86,59	1,09		1,44	31,25	337	4,28	1 687
GCF_000635475.1	529O19	C1	87,14	0,54		1,3	31,67	149	8,7	1 502
GCF_000635155.1	519C7	C1	87,36	1,04		1,36	31,40	317	4,29	1 522
GCF_000635255.1	521O20	C1	87,68	0,68		1,51	31,15	139	10,83	1 794
GCF_000635375.1	527P5	C1	87,76	1,00		1,46	31,28	324	4,5	1 702
GCF_000635415.1	529B19	C1	88,00	1,18		1,48	31,46	300	4,91	1 678
GCF_000635235.1	521K15	C1	88,63	0,54		1,48	31,26	329	4,49	1 697
GCF_000635555.1	496N4	C1	90,76	0,95	†	1,45	31,44	377	3,84	1 670
GCF_000635835.1	521B10	C1	93,57	1,09	†	1,55	31,17	218	7,1	1 811
GCF_000635675.1	498P15	C1	93,61	0,82	†	1,46	31,45	143	10,14	1 698
GCF_000634155.1	529C4	C1	96,38	0,27	†	1,6	31,26	187	8,54	1 899
GCF_000635495.1	495K23	C1	96,92	0,27	† ‡	1,62	31,34	124	12,96	1 899
GCF_000634535.1	520M11	C2	8,62	0		0,4	31,30	81	4,88	474
GCF_000634355.1	498N8	C2	53,86	0,69		0,9	31,14	445	2,03	1 024
GCF_000634075.1	528J14	C2	64,86	0,95		1,18	31,06	326	3,61	1 298
GCF_000634035.1	527E14	C2	68,76	0,28		1,1	31,30	271	4,07	1 219
GCF_000634915.1	496G15	C2	81,70	2,17		1,37	31,18	296	4,64	1 535
GCF_000635815.1	520F22	C2	87,18	0,41		1,35	31,46	245	5,5	1 505
GCF_000634975.1	498B22	C2	91,98	0,63	†	1,47	31,20	187	7,85	1 702
GCF_000634995.1	498C16	C2	92,45	0,23	† ‡	1,47	31,19	281	5,23	1 746
GCF_000634675.1	527E15	C3	34,48	0		0,57	31,46	84	6,81	663
GCF_000635215.1	521C8	C3	59,11	2,26		0,96	31,29	439	2,19	1 051
GCF_000635755.1	519G16	C3	77,72	1,02		1,31	31,26	357	3,65	1 520
GCF_000634275.1	497J18	C3	82,70	1,14		1,4	31,32	266	5,25	1 611
GCF_000635055.1	518A6	C3	83,38	1,36		1,37	31,40	278	4,92	1 559
GCF_000635095.1	518J7	C3	83,80	1,31		1,38	31,50	426	3,23	1 537
GCF_000634335.1	498N4	C3	85,51	0,54		1,41	31,34	126	11,12	1 781
GCF_000635395.1	528P18	C3	86,60	0,56		1,46	31,26	295	4,94	1 681
GCF_000634175.1	529D18	C3	90,13	0,68	†	1,51	31,41	170	8,83	1 763
GCF_000634255.1	496A2	C3	91,71	2,17	†	1,43	31,44	242	5,92	1 646
GCF_000634795.1	529J15	C3	94,57	0,27	†	1,6	31,23	180	8,84	1 849
GCF_000634375.1	518A17	C3	97,39	0,77	† ‡	1,62	31,30	136	11,86	1 909
GCF_000634215.1	495L20	C3	97,74	0,54	†	1,56	31,34	143	10,89	1 804
GCF_000634435.1	519B7	C4	31,03	0		0,69	31,48	97	7,13	769
GCF_000635515.1	495N3	C4	57,07	1,27		0,9	31,30	368	2,44	967
GCF_000635595.1	498B23	C4	76,09	1,09		1,23	31,13	314	3,92	1 398
GCF_000634735.1	528N17	C4	93,43	0,41	† ‡	1,53	31,28	144	10,55	1 767
GCF_000634655.1	526N5	C5	31,90	0		0,48	31,09	187	2,6	524
GCF_000635715.1	518I6	C5	88,29	0,28		1,36	31,47	308	4,42	1 534
GCF_000635015.1	498I20	C5	94,63	0	† ‡	1,58	31,25	167	9,46	1 892
GCF_000634135.1	528P14	C8	41,38	0		1,1	31,08	281	3,91	1 301
GCF_000634455.1	519O21	C8	83,70	0,54		1,26	31,58	205	6,13	1 419
GCF_000634295.1	498A3	C8	92,07	0,59	†	1,47	31,33	255	5,75	1 715
GCF_000635355.1	527L22	C8	92,44	0,63	† ‡	1,51	31,28	206	7,33	1 696
GCF_000635575.1	497E17	C9	35,19	0,82		0,55	31,84	154	3,56	589
GCF_000634515.1	520K10	C9	92,44	1,22	†	1,45	31,49	237	6,1	1 630
GCF_000634095.1	528J8	C9	92,62	1,39	† ‡	1,57	31,27	166	9,44	1 821

MDA : Multiple Displacement Amplification
CDS : Séquences codantes (coding sequences)

Complétude et Contamination estimées avec CheckM (Parks *et al.*, 2015)
† SAGs considérés comme quasi-complets (complétude ≥ 90% et contamination ≤ 5%)
‡ SAGs représentatifs de chaque clade utilisés pour l'alignement multiple des génomes complets, obtenu *via* l'utilisation de progressiveMauve (Darling *et al.*, 2010)

Ces SAGs correspondent à un sous-ensemble d'un jeu de données initialement composé de 96 SAGs d'individus collectés sur le site BATS, au cours de trois campagnes d'échan-

tillonnage entre novembre 2008 et avril 2009 (Kashtan *et al.*, 2014). En effet, seuls les SAGs appartenant à des sous-populations (également nommées clades) constituées *a minima* de trois individus ont été sélectionnés afin d'étudier la dynamique évolutive d'une population bactérienne environnementale. Par ailleurs, a également été exclu un SAG du clade C1 (*i.e.,* 518D8), jugé contaminé puisque contenant une proportion importante de séquences affiliées aux protéobactéries. Les sous-populations utilisées dans cette étude sont délimitées phylogénétiquement et réparties dans trois grands groupes définis à 98 % d'identité des séquences de leurs ITS (Kashtan *et al.*, 2014), *i.e.*, C1 à C5 dans le groupe cN2, C8 dans le groupe c9301 et C9 dans le groupe cN1 (Figure 2.1).

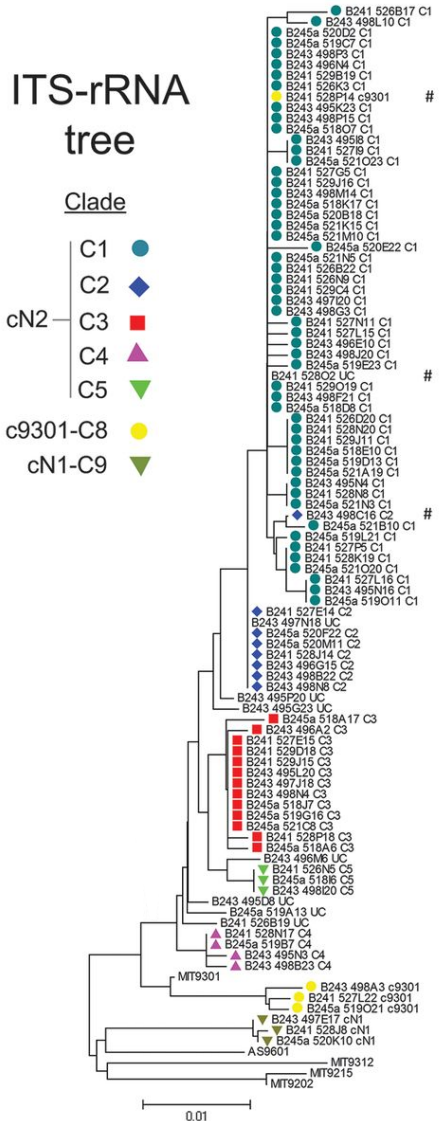


Figure 2.1. Phylogénie des séquences internes transcrites (ITS) du gène codant l'ARNr 16S inférée par la méthode du *Neighbor-Joining* (Kashtan *et al.*, 2014). L'arbre phylogénétique a été reconstruit à partir de l'alignement multiple des séquences des ITS de 96 SAGs (90 cN2, trois cN1 et trois c9301-*ribotypes*) ainsi que cinq séquences génomiques de souches cultivées de l'écotype HLII (MIT9301 ; AS9601 ; MIT9312 ; MIT9215 ; MIT9202). Une forme et une couleur, propres à chaque sous-population (C1 à C5, C8 et C9), sont attribuées à chaque feuille de l'arbre. La divergence a été estimée par la distance *p* (fonction du nombre de substitutions observées au niveau des sites homologues). Les unités de distance correspondent donc à des substitutions par site nucléotidique.

Les séquences génomiques des SAGs ont été téléchargées depuis le *National Center for Biotechnology Information* (NCBI) (numéros d'accession des *BioProject* : PRJNA239833, PRJNA239872 et PRJNA239873). Leur taille d'assemblage est comprise entre 0,37 et 1,62 Mpb, avec un pourcentage moyen en dinucléotides GC de l'ordre de 31 %. Leur complétude et leur contamination ont été estimées *via* l'utilisation de l'outil CheckM (Parks *et al.*, 2015) (Figure 2.2).

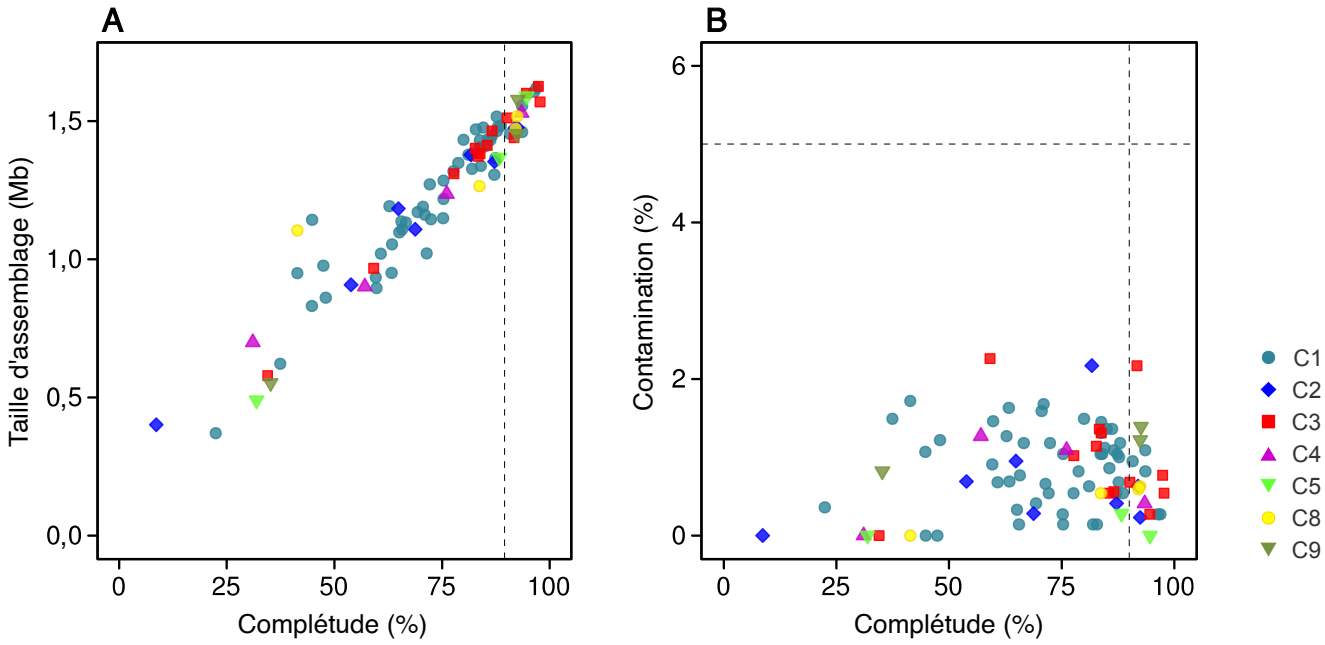


Figure 2.2. Estimation de la complétude et de la contamination pour les 87 SAGs de *Prochlorococcus* étudiés. (A) Corrélation entre la taille d'assemblage (Mpb) et la complétude (%) pour chacun des SAGs. La ligne pointillée correspond à une complétude égale à 90 %. **(B)** Corrélation entre les estimations de complétude (%) et de contamination (%). Les lignes pointillées définissent les critères requis pour qu'un SAG soit considéré comme quasi-complet (complétude ≥ 90% et contamination ≤ 5%; Parks *et al.*, 2015). Ceci est vrai pour 18 SAGs (496N4 ; 521B10 ; 498P15 ; 529C4 ; 495K23 ; 498B22 ; 498C16 ; 529D18 ; 496A2 ; 529J15 ; 518A17 ; 495L20 ; 528N17 ; 498I20 ; 498A3 ; 527L22 ; 520K10 ; 528J8) répartis dans les sept sous-populations. Une forme et une couleur sont attribuées à chaque sous-population (C1 à C5, C8 et C9).

Ces estimations sont calculées sur la base de la présence / absence d'un ensemble de gènes marqueurs spécifiques d'une lignée, théoriquement présents en copie unique. Tous les SAGs, quelle que soit leur complétude, ont une contamination inférieure à 2,3 % (Figure 2.2B). Les valeurs de complétude se situent entre 8,6 % et 97,4 %, ce qui a permis de définir quatre catégories de SAGs (Parks *et al.*, 2015). Ceux considérés comme :

- partiels ont une complétude inférieure à 50 %. Cette catégorie englobe près de 14 % des SAGs et montre une sur-représentation de ceux affiliés au clade C1 (7 sur les 12 SAGs de cette gamme de complétude) ;
- substantiels ont des complétudes comprises entre 50 à 70 % et incluent 45 % des SAGs ;
- modérés ont des complétudes comprises entre 70 et 90 %. Ils représentent 19 % des SAGs ;
- quasi-complets ont une complétude supérieure à 90 %. Ils sont au nombre de 18 (21 % des SAGs) et sont équitablement distribués dans tous les clades, y compris les clades faiblement représentés en termes de nombre de SAGs.

2.1.1.2. Choix d'un génome de référence

En raison de la forte synténie entre les génomes des souches cultivées de *Prochlorococcus* de l'écotype HLII (Yan *et al.*, 2018), et pour limiter la complexité de l'information traitée ici, un unique génome de référence a été choisi pour toutes les analyses menées par la suite, *i.e.*, la souche MIT9312 (Figure 2.3).

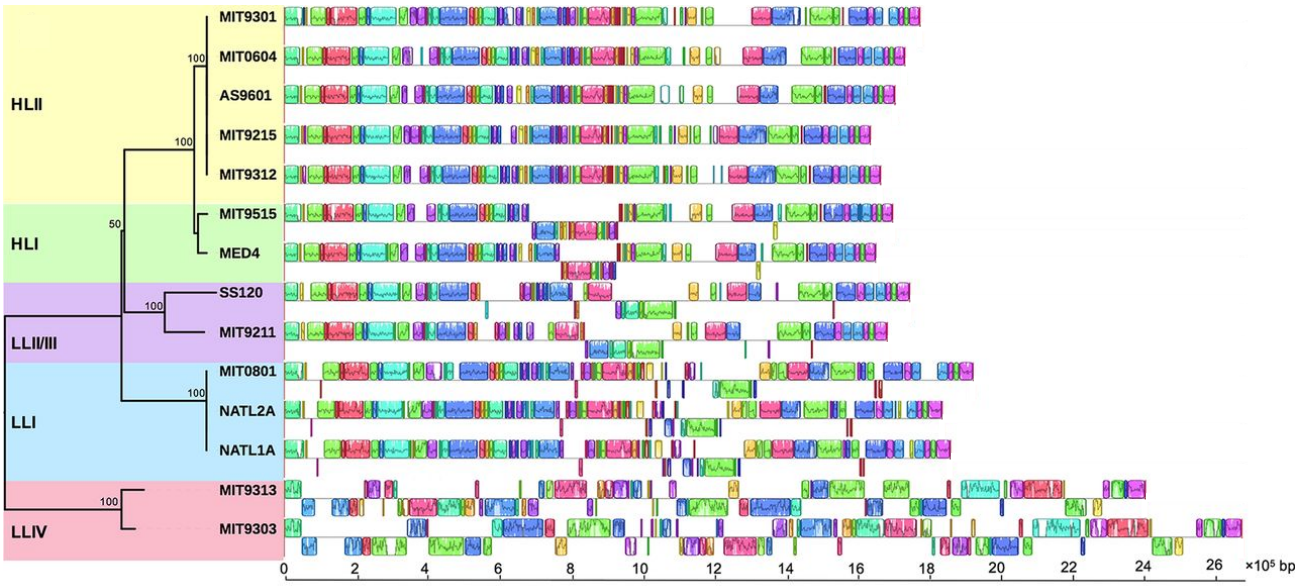


Figure 2.3. Comparaison de 14 génomes complets de *Prochlorococcus* appartenant aux écotypes *high-light* (HL) et *low-light* (LL) (Yan *et al.*, 2018). Les séquences génomiques ont été comparées à l'aide de Mauve (Darling *et al.*, 2004). La figure montre les réarrangements des blocs localement colinéaires (*locally collinear blocks* ou LCBs) entre les génomes, identifiés par différentes couleurs. Les LCBs sous les lignes horizontales référençant chaque génome ont une orientation inverse par rapport au génome MIT9301 situés en haut de la figure. Les régions dépourvues de LCBs (régions blanches) sont spécifiques du génome considéré. L'arbre phylogénétique a été reconstruit sur la base des permutations de 69 LBCs par la méthode du maximum de vraisemblance. Les valeurs de *bootstrap* (estimées par 1 000 ré-échantillonnages) apparaissent au niveau des nœuds.

Les principaux critères justifiant le choix du génome de référence reposent sur le fait :

- que ce génome ne partage aucun ancêtre commun direct avec l'un des clades étudiés. Ceci a conduit à exclure les souches MIT9301 et AS9601, partageant un ancêtre commun direct avec les clades C8 et C9 (Figure 2.1), ainsi que toutes les souches proches de ces deux souches (Figure 2.4) (Kent *et al.*, 2019) ;
- qu'il partage une relation étroite avec l'ensemble des clades, sans que celle-ci ne soit trop étroite tout de même. Ainsi les trois souches les plus distantes, *i.e.*, MIT9107, MIT9123 et MIT9116 (Kent *et al.*, 2019), ont été écartées (Figure 2.4) ;
- finalement, parmi les trois génomes restants (*i.e.*, MIT9302, MIT9311 et MIT9312), la souche *P. marinus* str. MIT9312 (numéro d'accession

ABB49062.1) a été choisie car elle représente historiquement l'écotype eMIT9312 / HLII (Biller *et al.*, 2014b).

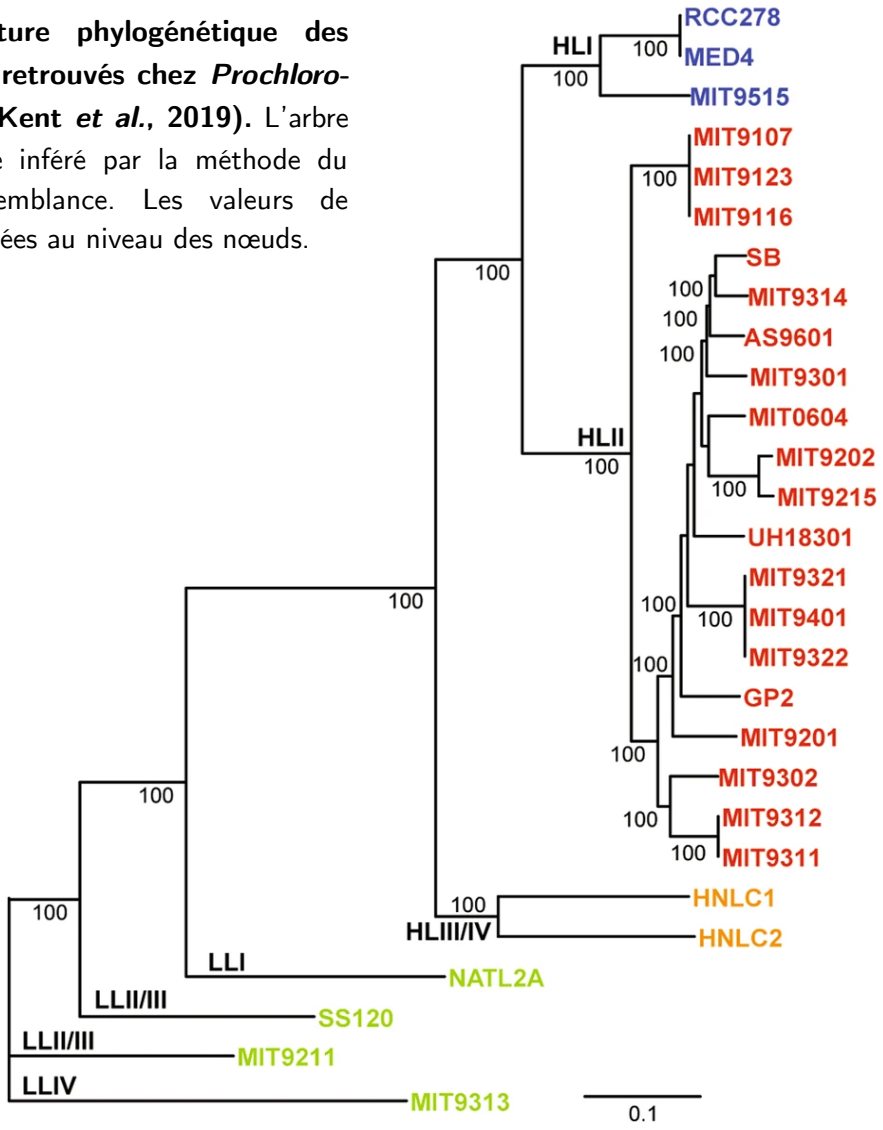


Figure 2.4. Structure phylogénétique des écotypes HL et LL retrouvés chez *Prochlorococcus* (adapté de Kent *et al.*, 2019). L'arbre phylogénétique a été inféré par la méthode du maximum de vraisemblance. Les valeurs de *bootstrap* sont indiquées au niveau des nœuds.

Le génome de MIT9312, de taille 1,71 Mpb et de pourcentage de GC moyen 31 %, contient 1 962 séquences codantes (*coding sequences* ou CDS) et six ISLs dispersés le long de son *backbone* (Avrani *et al.*, 2011) (Tableau 2.2).

Tableau 2.2. Caractéristiques des six îlots génomiques (ISLs) dispersés le long du *backbone* chez MIT9312 (numéro d'accession : ABB49062.1).

	Positions nucléotidiques		Longueur (Kb)	Nombre de CDS
	Début	Fin		
ISL1	344 270	365 710	21,44	23
ISL2	646 201	680 105	33,91	61
ISL2.1	756 519	773 586	17,07	25
ISL3	1 042 883	1 146 036	103,15	132
ISL4	1 203 215	1 271 749	68,54	60
ISL5	1 378 720	1 419 857	41,14	77

CDS : séquences codantes (*coding sequences*)
ISL : îlot génomique (*genomic island)*

2.1.2. Clusters de gènes orthologues

2.1.2.1. Histoires d'homologie

Il est possible, à l'échelle des gènes, d'évaluer les processus évolutifs lignée-spécifiques qui peuvent conduire *in fine* à la différenciation de populations bactériennes. Il est alors nécessaire de caractériser des gènes d'intérêt partageant une origine évolutive commune entre individus appartenant à différents taxa, depuis l'échelle inter-spécifique jusqu'à l'échelle intra-populationnelle. Ces gènes, dit homologues (Fitch, 1970), sont issus d'événements évolutifs distincts dont nous pouvons reconstruire l'histoire évolutive qui repose sur différents scénarios (Snel *et al.*, 2000; Kunin and Ouzounis, 2003; Mirkin *et al.*, 2003) :

- la spéciation. Les gènes homologues ayant divergé depuis un unique gène ancestral, appartenant au dernier ancêtre commun aux deux espèces comparées, sont dits orthologues (Figure 2.5) ;
- la duplication de gènes. Cet événement conduit, au sein d'un même génome, à l'émergence de gènes paralogues (Figure 2.5) ;

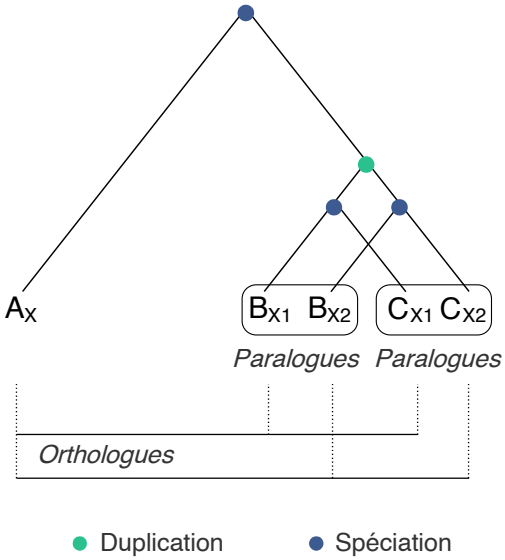


Figure 2.5. Relations d'homologie, notions d'orthologie et de paralogie. Les gènes retrouvés chez les espèces A, B et C sont dits homologues lorsqu'ils partagent un ancêtre commun, et donc une origine évolutive commune. Les gènes A_X, B_{X1} et C_{X1}, ainsi que A_X, B_{X2} et C_{X2} sont orthologues puisque issus d'événements de spéciation. Les gènes B_{X1} et B_{X2} sont paralogues, de même que les gènes C_{X1} et C_{X2}, puisque issus d'un événement de duplication.

- la fusion et la fission de gènes. Dans ce cas, un gène codant une protéine constituée de plusieurs domaines sera orthologue à d'autres gènes indépendants codant individuellement les domaines protéiques respectifs ;
- les HGTs. Le déplacement d'un gène depuis un taxon donné vers un individu d'une lignée distante, entraîne la formation de gènes dits xénologues (Figure 2.6) ;

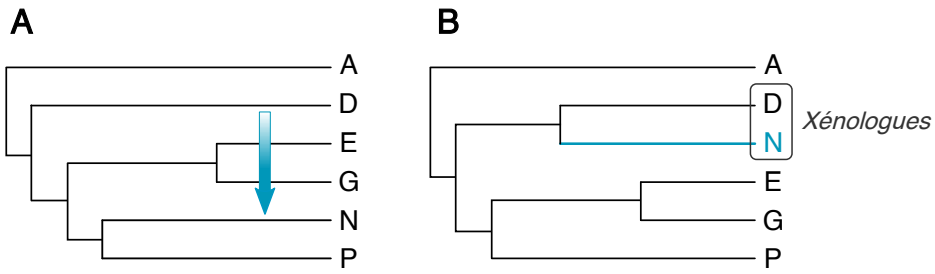


Figure 2.6. Impact d'un transfert horizontal de gènes (HGT) sur les relations d'homologie. L'histoire évolutive des espèces A, D, E, G, N et P est représentée en **(A).** L'hypothétique transfert d'un gène depuis l'espèce D vers l'espèce N conduit à une altération des relations d'homologie pour le gène en question **(B)**. Les gènes incriminés chez les espèces D et N sont dits xénologues.

- la perte de gènes. Cet événement, initié par l'apparition d'un codon stop (substitution) ou d'un décalage du cadre de lecture (insertion / délétion), conduit à la perte progressive du-dit gène et affecte les relations d'homologie entre espèces.

2.1.2.2. Clusters de gènes orthologues *core* et flexibles

Au cours de ces travaux de thèse, 7 125 clusters de gènes orthologues (COGs) ont été analysés. Ceux-ci ont été définis lors d'une étude précédente (Kashtan *et al.*, 2014) au cours de laquelle les auteurs ont déduit les relations d'homologie deux à deux par la méthode décrite par Kelly *et al.* (2012). Brièvement, ces relations ont, dans un premier temps, été définies entre les séquences protéiques déduites de chaque gène sur la base des meilleurs résultats réciproques d'un BLASTP (*e-value* ≤ 1e-5 ; identité de séquences > 35% ; longueur de l'alignement > 75% de la longueur de la protéine la plus courte des deux comparées). Les gènes ont ensuite été regroupés en COGs de manière transitive ; *i.e.*, si les gènes A et B sont homologues et les gènes B et C sont homologues, alors les gènes A, B et C sont regroupés au sein d'un même COG. Dans un deuxième temps, des profils de Markov cachés (*Hidden Markov Model* ou HMM en anglais) (Eddy, 2009) ont été construits pour chaque COG précédemment défini afin d'intégrer les gènes dont les relations d'homologie lointaine n'avaient pu être caractérisées par l'approche BLAST. Cette approche a permis de décrire des relations d'orthologie, mais également de paralogie, puisque pour un génome donné, certains gènes sont présents en plusieurs copies et appartiennent au même cluster. Mais par abus de langage, le terme de cluster de gènes orthologues sera conservé dans la suite de cette thèse.

Par définition, le génome *core* est le *pool* de COGs communs à l'ensemble des génomes pour un groupe taxonomique donné (Medini *et al.*, 2005; Tettelin *et al.*, 2005). Dans le cadre de l'étude de Kashtan *et al.* (2014), ces COGs ont été définis tels que communs aux 13 génomes de souches cultivées de *Prochlorococcus* décrites au moment de l'étude et appartenant à l'écotype HLII (*i.e.*, MIT9311, MIT9314, MIT9401, MIT9301, MIT9312, MIT9107, MIT9201, MIT9321, MIT9202, MIT9215, SB, GP2 et AS9601) (Figure 2.7). Dans la mesure où les SAGs sont incomplets, ils ne peuvent être intégrés *stricto sensu* dans la définition du génome *core.* Mais il est possible de détecter, dans les SAGs, des gènes homologues à ces gènes *core*. Sur les 7 125 COGs, 1 410 ont été identi-

fiés comme appartenant au génome *core* (Figure 2.8), parmi lesquels 1 397 sont présents en copie unique.

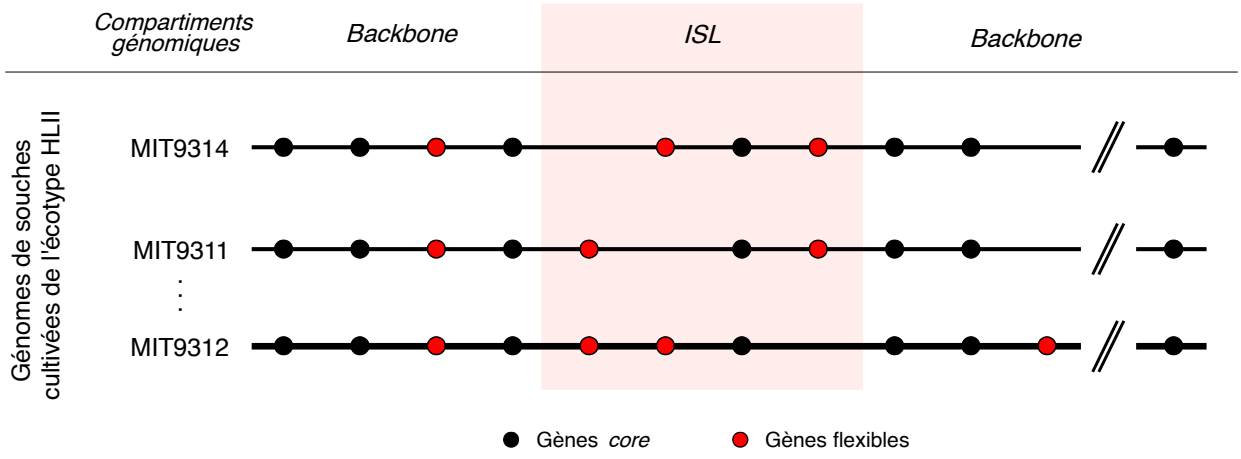


Figure 2.7. Schéma illustrant l'organisation chromosomique entre des génomes de souches cultivées de *Prochlorococcus* appartenant à l'écotype HLII. Les gènes partagés par l'ensemble des souches cultivées (cercles noirs) définissent les COGs appartenant au génome *core*. Les gènes spécifiques d'une ou plusieurs souches correspondent aux COGs flexibles (cercles rouges). ISL ; îlot génomique

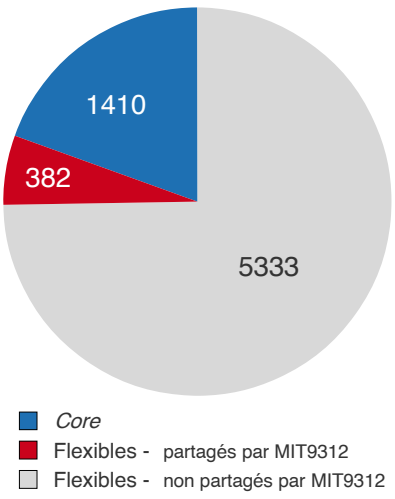


Figure 2.8. Nombre de COGs en fonction de leur classification en COGs *core* (1 410 COGs), COGs flexibles partagés (382 COGs) et non partagés (5 333 COGs) par le génome de référence MIT9312.

A contrario, les COGs flexibles (Figure 2.7) correspondent à ceux partagés par certains génomes de souches cultivées et/ou SAGs, ou spécifiques d'un unique SAG. Les gènes spécifiques des génomes des souches cultivées, *i.e.*, qui n'ont été retrouvés sur aucun SAG, n'ont pas été intégrés à l'étude. Ainsi, 5 715 COGs issus de SAGs sont considérés comme flexibles. Il a été choisi de distinguer les COGs flexibles communs à MIT9312 (382 COGs, dont 372 en copie unique) de ceux qui sont SAG-spécifiques (5 333 COGs, dont 5 290 en copie unique) (Figure 2.8). Ceci permet de réaliser une analyse comparative au regard du génome de référence MIT9312 tout en garantissant l'homogénéité de la comparaison par rapport à tous les SAGs, quel que soit leur clade d'origine. Il est à noter que parmi les COGs flexibles qui ne sont pas partagés par MIT9312, 678 (dont 653 en copie unique) présentent des homologies avec des gènes d'un ou plusieurs génomes issus des 12 autres souches cultivées ayant servi à définir le génome *core*. Plus de 50 % de ces COGs sont partagés avec un seul génome de souche cultivée (Figure 2.9). Compte tenu de leur faible nombre (comparativement au nombre de COGs SAG-spécifiques), de la proximité relative de certaines souches cultivées par rapport aux clades étudiés et pour limiter la complexité de l'analyse qui découlerait de la multiplication des génomes de référence (*cf.* partie 2.1.1.2), ces COGs ont été inclus dans la catégorie des COGs flexibles non partagés par MIT9312, sans distinction supplémentaire.

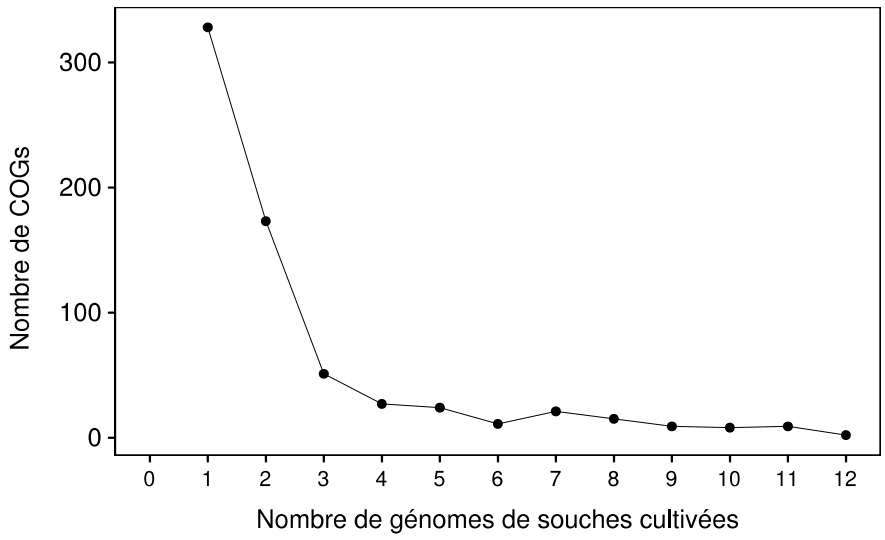


Figure 2.9. Distribution des 678 COGs flexibles absents de MIT9312 en fonction de leur présence dans les autres génomes de souches cultivées de *Prochlorococcus* de l'écotype HLII. Certains COGs flexibles, absents du génome de référence MIT9312, peuvent

néanmoins être présents *a minima* dans l'un des 12 autres génomes de souches cultivées (*i.e.,* MIT9311, MIT9314, MIT9401, MIT9301, MIT9107, MIT9201, MIT9321, MIT9202, MIT9215, SB, GP2 et AS9601).

2.1.2.3. Localisation génomique des COGs, un second niveau d'intégration

Chaque COG est susceptible de présenter sa propre histoire évolutive, notamment en fonction du compartiment génomique auquel il est rattaché (*core versus* flexible ; *backbone versus* ISL). Ainsi, pour les COGs flexibles détectés *a minima* dans un SAG mais absents du génome de référence MIT9312, il devenait nécessaire de les affecter à un compartiment (*backbone* ou ISL). À cet égard, le contexte génomique des gènes flexibles de chaque COG a été étudié, par comparaison à la localisation chromosomique des gènes de MIT9312 qui les encadrent (Figure 2.10). L'assignation des COGs non partagés par MIT9312 a été déterminée sur la base d'intervalles génomiques définis par :

- deux gènes contigus présents chez MIT9312 et appartenant à un même compartiment génomique. Le ou les gènes flexibles SAG-spécifiques intégrés dans cet intervalle appartiennent à ce même compartiment ;
- deux gènes contigus présents chez MIT9312 mais appartenant à des compartiments distincts. Il ne peut y avoir d'attribution à un compartiment, la position de ce gène est alors considérée comme ambiguë ;
- un gène présent chez MIT9312 et à l'extrémité d'un *contig*. La position du gène compris dans cet intervalle est également considérée comme ambiguë.

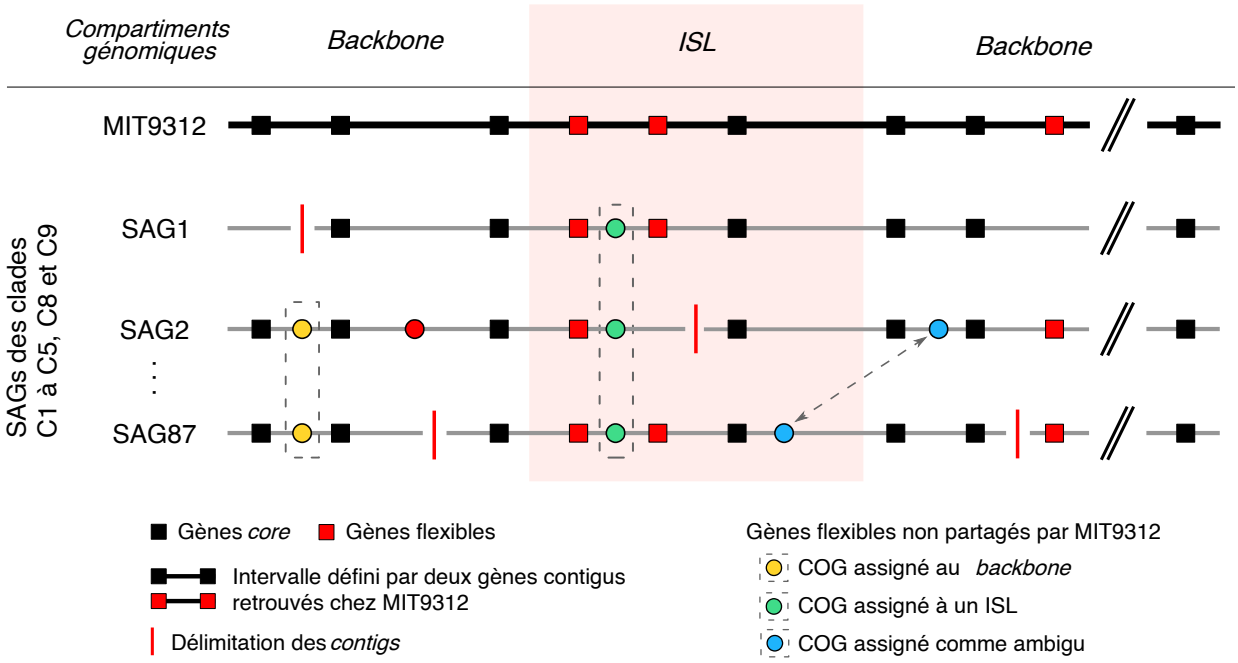


Figure 2.10. Représentation de la méthode d'assignation à un compartiment génomique des COGs flexibles non partagés par le génome de référence MIT9312. Le génome de MIT9312 a été utilisé comme référence afin d'attribuer un compartiment génomique (*i.e.*, *backbone* ou ISL) à chaque COG flexible présent dans au moins un SAG mais absent de MIT9312. Cette compartimentation est définie selon les intervalles déterminés par deux gènes contigus appartenant à MIT9312. Cercles jaunes : gènes spécifiques des SAGs assignés au *backbone* dans la mesure où les gènes contigus les plus proches, trouvés chez MIT9312, appartiennent au *backbone*. Cercles verts : pour le SAG2, le gène est situé à l'extrémité du *contig* et ne peut donc être attribué à un compartiment (ambigu). La définition d'un compartiment est cependant possible à l'échelle du COG puisque la localisation des gènes orthologues est possible dans les deux autres SAGs (SAG1 et SAG87). Ainsi, le COG est assigné à un ISL. Cercles bleus : le gène du SAG2 est assigné au *backbone*, tandis qu'il est considéré comme ambigu pour le SAG87 puisque les deux gènes contigus les plus proches sont attribués à différents compartiments (*backbone* et ISL). L'assignation d'un compartiment à l'échelle du COG ne peut être réalisée (COG dont la localisation est ambiguë). ISL : îlot génomique.

L'assignation à un compartiment génomique a ensuite été réalisée à l'échelle du COG. Pour cela, l'entropie de Shannon (Shannon, 1948), qui se rapporte à la quantité d'information délivrée par un système d'information, a été utilisée. Les COGs sont alors considérés comme des systèmes constitués de gènes qui peuvent être attribués à différents compartiments. Il est ainsi envisageable d'apprécier la quantité d'information relative à un COG *via* l'estimation de l'entropie de Shannon comme suit : $H(X) = -\sum_{i=1}^{n} P_i \log_2 P_i$; où n correspond au nombre de compartiments génomiques (*i.e.,* ISL1, ISL2, ISL2.1, ISL3, ISL4, ISL5, *backbone,* ambigu) et P_i est la proportion de gènes provenant du compartiment génomique i au sein du COG considéré. Ceci permet d'évaluer :

- la variabilité des attributions des compartiments génomiques à l'échelle des gènes pour un COG donné. La valeur d'entropie varie entre 0 – tous les gènes sont attribués au même compartiment génomique – et $\log_2(n)$ – entropie maximale pour n compartiments en proportions égales (Figures 2.11A et 2.11C) ;
- la précision de l'assignation à un compartiment à l'échelle d'un COG par rapport aux gènes qui le composent. L'entropie est plus faible lorsqu'un compartiment est plus fortement représenté au sein d'un COG, et a tendance à augmenter lorsque la proportion des différents compartiments s'équilibre (Figure 2.11B et 2.11C).

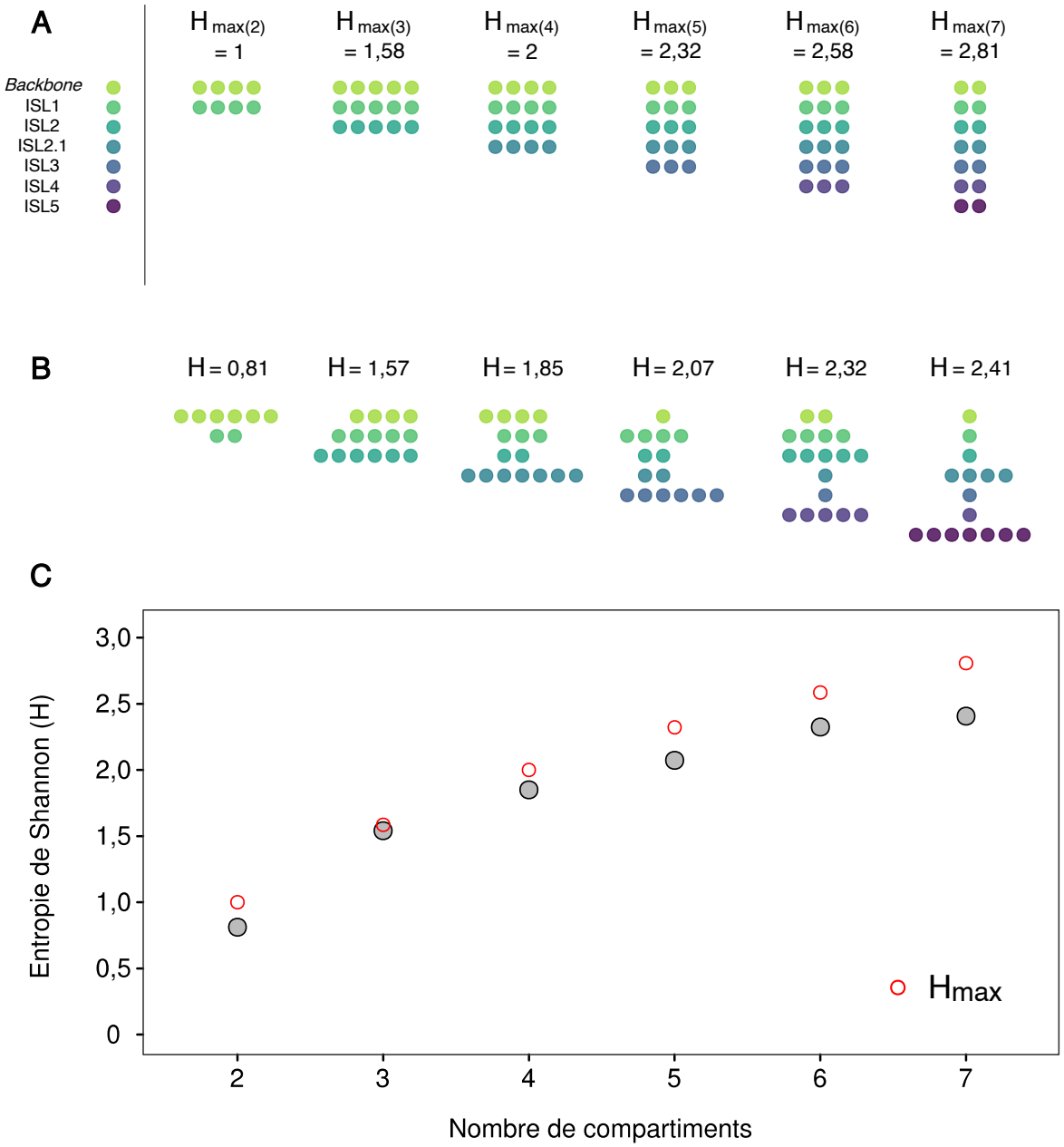


Figure 2.11. Entropie de Shannon comme outil d'assignation de compartiments génomiques. L'entropie de Shannon, notée H, a été utilisée dans le but d'analyser, au sein de chaque COG flexible non partagé par MIT9312, les proportions des gènes assignés aux différents compartiments génomiques. L'entropie est dépendante du nombre de compartiments présents pour un COG donné, mais également de leur proportion respective. La formule de l'entropie de Shannon est la suivante : $H(X) = -\sum_{i=1}^{n} P_i \log_2 P_i$; où n correspond au nombre de compartiments génomiques (*i.e.,* ISL1, ISL2, ISL2.1, ISL3, ISL4, ISL5, *backbone,* ambigu) et P_i est la proportion de gènes provenant du compartiment génomique i au sein du COG considéré. **(A)** Lorsque les compartiments sont représentés dans les mêmes proportions au sein d'un COG, l'entropie prend sa valeur maximale (H_{max}), bornée par $\log_2(n)$. À titre

d'exemple, un COG constitué de 15 gènes dont l'assignation est équilibrée entre trois compartiments distincts (*backbone*, ISL1 et ISL2) aura une entropie égale à 1,58. **(B)** Lorsque les proportions entre les différents compartiments diffèrent, la valeur d'entropie H diminue par rapport au H_{max}. **(C)** Pour chaque exemple donné en **(A)** et en **(B)**, les valeurs de H et H_{max} sont représentées en fonction du nombre de compartiments.

2.2. Analyse pangénomique

L'un des objectifs de cette thèse repose sur l'évaluation des relations existant entre le pangénome et l'évolution d'une population bactérienne environnementale. Ceci implique l'estimation de la taille du pangénome, à une échelle de diversité intra-populationnelle, qui diffère de celle communément utilisée – *i.e.*, à l'échelle de l'espèce.

2.2.1. Phylogénie basée sur le génome *core*

Une approche phylogénomique, basée sur l'information contenue dans le génome *core*, a été adoptée pour l'analyse phylogénétique des différentes sous-populations. L'approche utilisée ici est celle du « super-alignement », largement utilisée aujourd'hui et produisant des phylogénies de haute résolution (Dutilh *et al.*, 2007). Pour chaque COG *core* présent en copie unique et *a minima* dans un SAG de chaque sous-population, les alignements des séquences protéiques ont été réalisés avec MAFFT v7.271 (Katoh and Standley, 2013) (alignement local effectué *via* l'option linsi), sur lesquels les séquences nucléotidiques ont été apposées (tranalign, EMBOSS v6.6.0.0) (Rice *et al.*, 2000). Les brèches (ou *gaps* en anglais) ont été traitées avec Gblocks (Castresana, 2000), améliorant ainsi le signal phylogénétique (Talavera and Castresana, 2007). Seules les positions pour lesquelles plus de 50 % des séquences présentaient un *gap* ont effectivement été traitées comme tel et n'ont pas été conservées dans l'alignement final. Les alignements nettoyés ont ensuite été concaténés. Les séquences pour lesquelles aucun orthologue n'a été identifié dans un SAG donné sont considérées comme des *gaps*.

Le choix du modèle de substitutions des séquences nucléotidiques a été fait sur la base d'un test statistique probabiliste qui est le critère d'information d'Akaiké ($AIC = 2k - 2L$; où k est le nombre de paramètres et L est la log-vraisemblance du modèle). Le ou les modèles ayant des valeurs minimales de l'AIC sont considérés comme les plus appropriés aux données analysées. Différents modèles, et différents paramètres associés, ont ainsi été testés avec l'outil jModelTest v2.1.10 (Darriba *et al.*, 2012). Le modèle d'évolution minimisant l'AIC apparaît être le *General Time Reversible* (GTR) (Tavaré, 1986). L'arbre

phylogénétique au maximum de vraisemblance a été inféré avec PhyML v3.0 (Guindon *et al.*, 2010), avec une analyse de la robustesse basée sur 100 *bootstraps*.

2.2.2. Comparaisons des sous-populations à l'échelle génomique

L'ANI permet de délimiter des OTUs, non plus sur la base de l'analyse de similarité des séquences des gènes codant les ARNr, mais à l'échelle des séquences génomiques dans leur ensemble (Varghese *et al.*, 2015).

Les ANI ont été calculées simultanément entre paires de SAGs appartenant à un même clade et à des clades distincts. La méthode ANIb, telle que décrite dans l'étude de Richter and Rosseló-Móra (2009), implémentée dans le module python pyani (Pritchard *et al.*, 2016), a été utilisée. Les séquences génomiques sont, dans un premier temps, artificiellement morcelées en fragments nucléotidiques d'une longueur de 1 020 nt (Goris *et al.*, 2007). La longueur de ces fragments correspond approximativement à celle obtenue suite à la fragmentation de l'ADN génomique au cours d'expérience d'hybridation moléculaire de l'ADN (DDH). Dans un deuxième temps, les valeurs d'identités nucléotidiques sont calculées avec l'outil BLASTN+ (Altschul *et al.*, 1990; Camacho *et al.*, 2009). Au cours d'une expérience de DDH, les valeurs d'identités réciproques entre deux génomes ne sont pas symétriques (Johnson and Whitman, 2007; Tindall *et al.*, 2010). Puisque le concept de l'ANI résulte de la DDH, pour une paire de SAGs donnée, les identités sont estimées réciproquement entre les fragments de l'un des SAGs – dit séquence requête – et la séquence génomique complète de l'autre SAG – dit séquence sujet. L'ANI est ensuite obtenue en moyennant les identités nucléotidiques de chaque alignement local. Nous comptons ainsi deux valeurs d'ANI pour un couple de SAGs.

Bien que l'ANI soit une métrique intéressante pour la comparaison de paires de génomes, elle ne permet pas de mettre en lumière des différences de similarité au niveau de régions génomiques particulières. Une analyse de la synténie des génomes a donc été effectuée entre le génome de référence MIT9312 et un SAG représentatif de chaque sous-population, le plus important en termes de taille d'assemblage. L'algorithme progressiveMauve, (Darling *et al.*, 2010), implémenté dans l'outil Mauve v2.4 (Darling *et al.*, 2004), a été utilisé pour l'alignement des séquences génomiques et la détection des régions conservées. L'algorithme peut être résumé comme suit :

- identification de régions similaires sans *gap*, en copie unique entre deux ou plusieurs génomes – notés LMA (*Local Multiple Alignments*) ;
- calcul d'une matrice de distances et construction d'un arbre-guide sur la base des LMA ;
- définition des LCBs. Il s'agit de définir récursivement des alignements locaux par paires de génomes à partir d'un ensemble de LMA ordonnés et de même orientation, sur lesquels est appliquée une analyse de points de cassure afin de définir des partitions minimales communes constituant les LCBs ;
- sur la base du calcul d'un score d'ancrage de type somme par paires (*SP anchor*), sélection des LCBs ancres le long de l'arbre-guide précédemment défini, puis raffinement itératif des arbres-guides et LCBs ancres suivi de l'ancrage récursif pour l'identification des ancres supplémentaires entre et à l'intérieur des LCBs ;
- alignements multiples sur les profils des LCBs ancres. Afin de capturer la totalité des régions homologues au voisinage des LCBs, les séquences des régions en dehors de ceux-ci sont assignées aléatoirement à leur LCB voisin et sont alignées conjointement avec eux ;
- raffinement itératif des alignements sur fenêtre glissante ;
- détection des zones alignées non-homologues par l'application d'un modèle HMM de décodage *a posteriori*. Les régions détectées comme non-homologues sont supprimées de l'alignement final.

2.3. Données taxonomiques et enrichissements fonctionnels

Des affiliations taxonomiques et des annotations fonctionnelles des gènes au sein de chaque COG *core* et flexibles ont été réalisées par recherche de similarités au moyen d'un BLASTP contre la base de données EggNOG v4.5 (*Evolutionary genealogy of genes: Non-supervised Orthologous Groups* ; Huerta-Cepas et al., 2016; Jensen et al., 2008). Cette base de données correspond à un ensemble de groupes de gènes orthologues, construite à partir de l'analyse de milliers de génomes eucaryotes, procaryotes et viraux pour établir, sur la base d'analyses phylogénétiques, les relations d'homologie. Elle permet également l'annotation fonctionnelle des COGs établis.
Dans cette étude, seules les séquences protéiques d'une longueur minimale de 60 acides aminés et pour lesquelles les résultats BLAST ont une *e-value* inférieure à 1e-5, une couverture d'alignement représentant au moins 50 % de la plus petite des deux protéines alignées et un pourcentage d'identité supérieur à 30 % ont été conservées.

2.3.1. Caractérisations de l'origine taxonomique des COGs

Dans la mesure où les gènes peuvent être issus d'événements évolutifs distincts, et notamment de HGTs, les affiliations ont été réalisées à l'échelle des gènes, puis généralisées à l'échelle du COG. Ces affiliations ont été caractérisées à différents niveaux taxonomiques, depuis le genre pour les gènes affiliés à *Prochlorococcus* et *Synechococcus*, jusqu'à des niveaux taxonomiques plus élevés (*i.e.*, classe voire phylum). Certains COGs présentant une hétérogénéité des affiliations de leurs gènes – du fait d'homologies lointaines – une catégorie dite « incertaine » a été définie. Celle-ci regroupe tous les COGs qui contiennent des gènes affiliés *a minima* à *Prochlorococcus* et/ou *Synechococcus*, mais aussi à d'autres groupes taxonomiques.

2.3.2. Catégorisation des gènes et enrichissements fonctionnels

Suite à l'analyse des résultats du BLASTP, les gènes ont été classifiés par catégories fonctionnelles telles que décrites par Tatusov *et al.* (1997, 2000), classification

également appliquée à la base de données EggNOG. Il existe quatre grandes catégories qui sont relatives (i) au « Stockage et traitement de l'information », (ii) aux « Processus cellulaires et signalisation », (iii) aux « Métabolismes » et (iv) aux gènes de fonctions peu ou mal caractérisées ; chacune d'entre elles est subdivisée en catégories fonctionnelles (Tableau 2.3).

Les gènes codant des protéines dont les fonctions sont mal caractérisées (*poorly characterized*) – *i.e.,* dont la fonction est seulement une prédiction générale (*general function prediction only*), voire inconnue (*unknown function*) – ont été écartés de l'analyse. À partir des annotations, les enrichissements fonctionnels ont pu être évalués en calculant les rapports observés/théoriques (O/E) pour chaque catégorie fonctionnelle en fonction des caractéristiques des gènes (*core*, flexibles), de leur localisation génomique (*backbone*, ISL, ambiguë) ou de leur affiliation taxonomique (*Prochlorococcus*, *Synechoccocus*, autres taxa, incertaine). Les valeurs théoriques ont été obtenues en multipliant le nombre de gènes (*core* ou flexibles en fonction de leur localisation génomique ou affiliation taxonomique) par le pourcentage de gènes totaux dans chaque catégorie fonctionnelle. Enfin, la significativité des enrichissements a été testée par des tests du χ^2.

Tableau 2.3. Catégories fonctionnelles telles que décrites par la classification de Tatusov *et al.* (2000).

Catégories fonctionnelles		
STOCKAGE ET TRAITEMENT DE L'INFORMATION	A	Traitement et modification de l'ARN
	B	Structure et dynamique de la chromatine
	J	Traduction, structure ribosomique et biogenèse
	K	Transcription
	L	Réplication, recombinaison et réparation
PROCESSUS CELLULAIRES ET SIGNALISATION	D	Contrôle du cycle cellulaire, division cellulaire et partitionnement des chromosomes
	M	Biogenèse des paroi cellulaire/membrane/enveloppe
	N	Motilité cellulaire
	O	Modification post-traductionnelle, turnover protéiques, chaperonnes
	T	Mécanismes de transduction du signal
	U	Trafic intracellulaire, sécrétion et transport vésiculaire
	V	Mécanismes de défense
	W	Structures extracellulaires
	Y	Structure nucléaire
	Z	Cytosquelette
MÉTABOLISME	C	Production et conversion d'énergie
	E	Transport et métabolisme des acides aminés
	F	Transport et métabolisme des nucléotides
	G	Transport et métabolisme des glucides
	H	Transport et métabolisme des coenzyme
	I	Transport et métabolisme des lipides
	P	Transport et métabolisme des ions inorganiques
	Q	Biosynthèse, transport et catabolisme des métabolites secondaires
MAL CARACTÉRISÉ	R	Prédiction de fonctions générales
	S	Fonction inconnue

2.4. Analyses des trajectoires évolutives

2.4.1. Pressions de sélection

Du fait de la dégénérescence du code génétique, un même acide aminé peut être codé par plusieurs codons. Ces derniers sont alors appelés codons synonymes (Figure 2.12). En conséquence, les différentes bases constituant les codons ne sont pas soumises aux mêmes contraintes évolutives pour garantir la fonction de la protéine codée. Cela donne lieu à deux types de substitutions. Celles dites synonymes n'entraînent aucune modification du signal à l'échelle protéique et surviennent majoritairement sur la troisième position du codon (position dégénérée). Elles ne sont pas soumises à sélection, et reflètent un processus neutre d'évolution, qui peut alors être caractérisé par un taux de substitutions synonymes par site synonyme noté *dS*. À l'inverse, les substitutions dites non synonymes entraînent une modification du signal en changeant la nature de l'acide aminé codé. Elles surviennent essentiellement sur les positions 1 et 2 des codons (positions non dégénérées) et sont soumises à sélection. Elles reflètent ainsi un processus de sélection positive, ou négative, caractérisé par un taux de substitutions non synonymes par site non synonyme, noté *dN*.

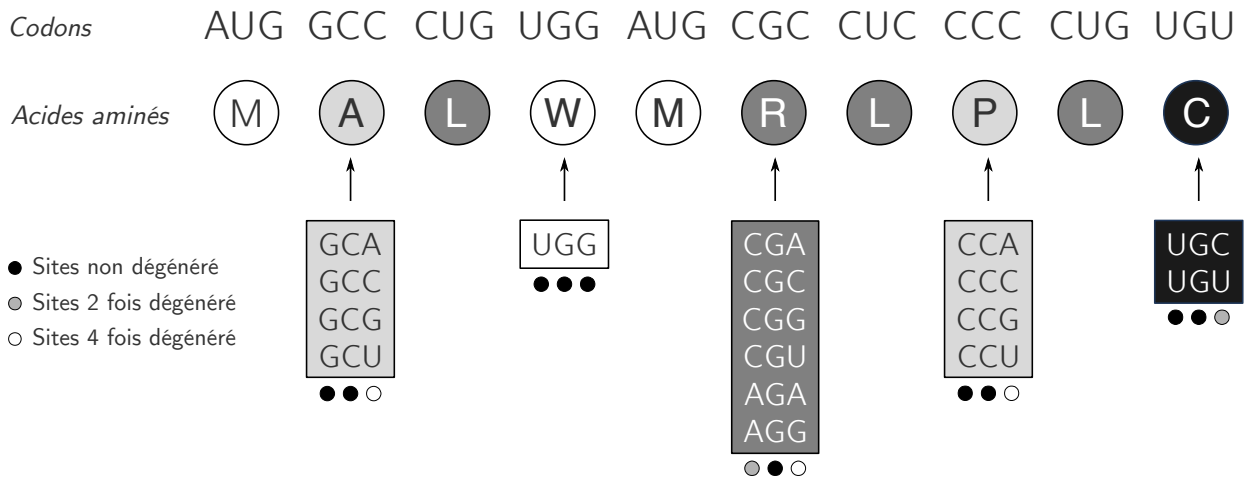


Figure 2.12. Représentation schématique de la dégénérescence du code génétique. Certains acides aminés sont codés par plusieurs codons alors qualifiés de synonymes, qui se différencient par un à deux nucléotides sur les positions 1 et 3 des codons. Ces positions sont dites dégénérées.

En analysant les substitutions sur les positions dégénérées et non dégénérées des codons, il est possible d'évaluer les pressions de sélection qui s'appliquent sur les séquences codantes des gènes (Ellegren, 2008). Celles-si sont estimées à partir du rapport entre le taux de substitutions synonymes et le taux de substitutions non synonymes, appelé valeur sélective et noté *dN/dS*. Un gène pour lequel *dN* est égal à *dS,* soit *dN/dS* = 1 (Figure 2.13A), n'est pas soumis à pression de sélection et les substitutions qu'il porte seront fixées aléatoirement dans la population selon le principe de dérive génétique. Lorsque *dN* > *dS* (Figure 2.13B), soit *dN/dS* > 1, il est possible d'inférer une pression de sélection positive à la séquence codante du gène. Dans ce cas, les substitutions non synonymes présentent un avantage sélectif et sont rapidement fixées dans la population. Enfin, un ratio *dN/dS* < 1 (*dN* < *dS*) (Figure 2.13C) traduit une pression de sélection négative. Les substitutions non synonymes désavantageuses sont alors éliminées de la population.

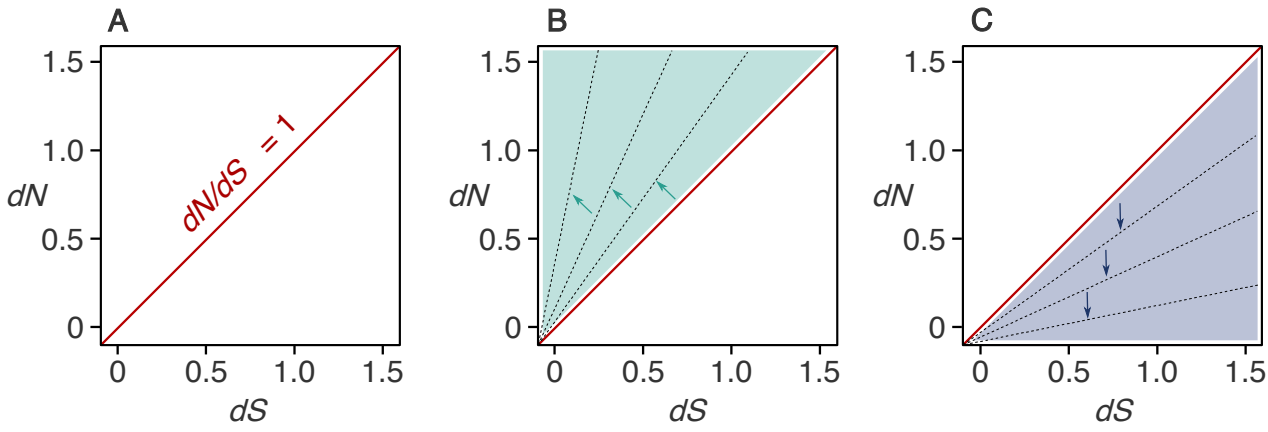


Figure 2.13. Inférence des processus évolutifs à partir des séquences codantes des gènes : divergence non synonyme (*dN*) *versus* divergence synonyme (*dS*). (A) Le rapport des taux de substitutions non synonymes sur les substitutions synonymes *dN/dS*=1 traduit une évolution neutre ; **(B)** un rapport *dN/dS* > 1 traduit une sélection positive ; **(C)** un rapport *dN/dS* < 1 traduit une sélection négative.

Les valeurs de *dN*, *dS* et leurs ratios *dN/dS* ont été estimés pour l'ensemble des COGs en copie unique communs au génome de référence MIT9312 (*core* et flexibles) et les COGs non partagés par MIT9312 mais communs, *a minima*, à deux sous-populations. Les arbres phylogénétiques au maximum de vraisemblance, générés à partir des alignements multiples des séquences nucléotidiques pour l'ensemble des COGs considérés, ont été inférés avec PhyML v3.0 (Guindon *et al.*, 2010) et le modèle d'évolution GTR+G. Les *dN, dS* et *dN/dS* ont ensuite été calculés par la méthode du maximum de vraisemblance

telle qu'implémentée dans *codeml* au sein du logiciel PAML v4.8a (Yang and Nielsen, 2000; Yang, 2007).

Les COGs ont ensuite été regroupés en fonction des valeurs de *dN*, *dS* et *dN/dS* en utilisant la méthode des *k*-moyennes (Hartigan and Wong, 1979). Le nombre optimal de groupements de COGs, au nombre de 5, a été défini par la méthode du coude (méthode *Elbow)* (Figure 2.14).

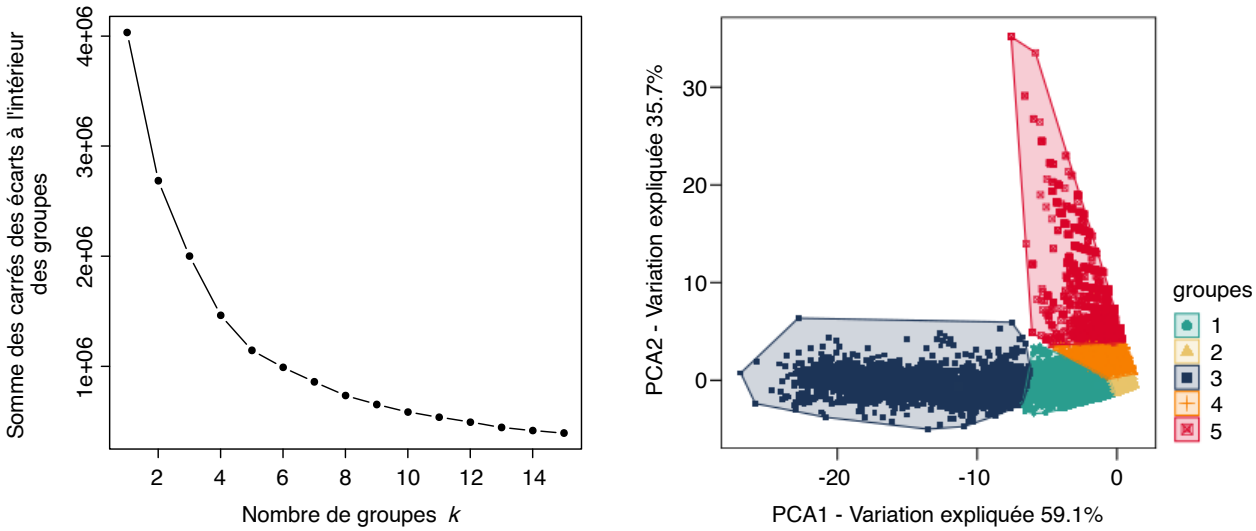


Figure 2.14. Analyse du regroupement de l'ensemble des COGs en copie unique en fonction des valeurs de *dN*, *dS* et *dN/dS* par la méthode des *k*-moyennes (*k-means*). La méthode de coude (méthode *Elbow*) permet de définir le nombre optimal de groupes *k* (*cf.* partie gauche) en effectuant un regroupement des *k*-moyennes pour un ensemble de valeurs *k* (*i.e.*, entre 2 et 15). Dans la mesure où trois variables sont considérées (*dN*, *dS* et *dN/dS*), une analyse en composantes principales (ACP) a été réalisée (*cf.* partie droite). Le pourcentage de variation expliquée par chaque axe est indiqué. Les deux premières composantes expliquent plus de 94 % de la variance.

2.4.2. Détection et quantification des événements de recombinaison homologue

Deux approches ont été mises en place pour détecter et évaluer la recombinaison dans la population de *Prochlorococcus,* écotype HLII, constituée de 87 SAGs. La première approche a consisté à ré-évaluer la structure de la population en autorisant de possibles événements de recombinaison au cours de son histoire évolutive et à caractériser

des points chauds de recombinaison à l'échelle des génomes. La seconde a permis d'identifier les catégories de COGs (*i.e.*, *core*, flexibles communs à MIT9312 et flexibles SAG-spécifiques) et les compartiments (*i.e.*, *backbone* et ISLs) susceptibles d'être concernés par les événements de recombinaison en détectant les signatures de recombinaison à l'échelle des COGs.

2.4.2.1. Structure de la population et recombinaison

Les relations de parenté entre individus sont habituellement représentées à l'aide d'arbres phylogénétiques, tenant compte de modèles évolutifs décrits principalement par des événements de mutations. D'autres modèles peuvent cependant être envisagés, pour lesquels sont considérés, entre autres, les événements de recombinaisons homologues ou hétérologues, visualisables par le biais de la construction de réseaux phylogénétiques (Sneath, 1975; Huson and Scornavacca, 2011).

Dans un premier temps, les distances phylogénétiques entre paires de SAGs ont été estimées à l'aide de l'outil RAxML v8.2.12 (Stamatakis, 2014) à partir de l'alignement multiple des séquences génomiques complètes produit par Kashtan et collaborateurs (2014). Les paramètres du modèle ont été choisis à partir de la phylogénie de génome *core* inférée au maximum de vraisemblance (*cf.* partie 2.2.2). Le modèle d'évolution GTR a été utilisé, l'hétérogénéité des taux de substitutions étant modélisée par une distribution *gamma*. Dans un second temps, les relations génétiques entre SAGs, exprimées par les distances ainsi calculées, ont été visualisées *via* la reconstruction d'un réseau phylogénétique par l'algorithme *Neighbor-Net* (Bryant and Moulton, 2004) implémenté dans le logiciel SplitsTree4 v4.15.1 (Huson and Bryant, 2006).

À partir de ce même alignement, les points chauds de recombinaison à l'échelle des génomes ont été détectés et quantifiés en utilisant ClonalFrameML v1.11-3 (Didelot *et al.*, 2015). La phylogénie du génome *core* a été utilisée comme phylogénie clonale de référence pour la population donnée. Brièvement, ClonalFrameML infère les événements de recombinaison de la manière suivante :

- les séquences ancestrales au niveau des nœuds internes de la phylogénie initiale sont reconstruites au maximum de vraisemblance ;

- les différents paramètres de recombinaison ainsi que les longueurs des branches de l'arbre sont alors estimés par l'algorithme espérance-maximisation de Baum–Welch ;
- un algorithme de Viterbi permet de déduire pour chaque site nucléotidique un statut « d'importation » – *i.e.*, qualifiant le fait qu'il soit issu ou non d'un événement de recombinaison.

Les différents paramètres ont été définis par défaut, *a priori* – *i.e.,* un rapport des taux de recombinaison sur mutation $R/\theta = 0{,}1$, l'inverse de la longueur moyenne des événements de recombinaison $1/\delta = 0{,}001$ et une distance moyenne entre les événements $v = 0{,}1$. Le ratio des transitions sur transversions $ts/tv = 1{,}70$ a été estimé en moyennant les ratios calculés pour chaque COG *core* avec le package R PopGenome v2.7.5 (Pfeifer *et al.*, 2014). L'impact relatif de la recombinaison par rapport aux mutations (r/m) a été calculé comme suit : $r/m = R/\theta \times \delta \times v$.

2.4.2.2. Détection des COGs recombinants

Chez les bactéries, la recombinaison homologue est synonyme de conversion génique, *i.e.*, transfert unidirectionnel de matériel génétique, depuis un donneur vers un receveur, impliquant des régions homologues. Afin de mesurer le potentiel de recombinaison des COGs *core*, flexibles communs à MIT9312 et flexibles SAG-spécifiques, leurs alignements multiples – avant suppression des *gaps* pour ne pas biaiser la détection des signaux de recombinaison – ont été soumis à quatre méthodes de détection des signatures de recombinaison :

- GENECONV (Sawyer, 1989). Cette méthode permet de détecter des événements de recombinaison en évaluant la significativité de la longueur de fragments identiques entre paires de séquences, au sein d'un alignement. Pour chaque paire de séquences, GENECONV identifie soit (i) les fragments les plus longs qu'elles partagent, soit (ii) un score d'alignement anormalement élevé pour cette paire de séquences (lorsque plusieurs longs fragments sont détectés). Les fragments sont délimités par deux sites polymorphes, ou un site polymorphe et une extrémité de l'alignement ; un fragment ne peut en aucun cas correspondre à la longueur totale de l'alignement des deux séquences. Sous réserve de la présence de sites variables dans les autres sé-

quences, ces fragments traduisent un potentiel événement de recombinaison entre les ancêtres des deux séquences. La significativité de ces inférences est estimée par permutations aléatoires des sites polymorphes. Les *p-values* globales (*i.e.*, estimées pour tous les fragments pour toutes les paires possibles) et locales (*i.e.*, estimées pour un fragment donné, pour une paire de séquences donnée) montrent respectivement l'existence de fragments recombinants à l'échelle des COGs et une information qualitative concernant les fragments recombinants à l'échelle des paires de gènes ;

- le test de Max χ^2 (Smith, 1992). Celui-ci permet de détecter le mosaïcisme – l'association de différents blocs au sein de séquences nucléotidiques – qui résulterait d'événements de recombinaison. Le principe de ce test repose sur l'analyse d'une paire de séquences alignées de longueur N contenant D sites polymorphes. Lorsqu'un point de rupture est défini à la position t, ceci détermine un bloc à gauche du point de rupture constitué de t sites, dont $x_1(t)$ polymorphes, et un bloc de $(N - t)$ sites à sa droite, dont $x_2(t)$ polymorphes (Figures 2.15A et 2.15B). Sous l'hypothèse nulle d'absence de recombinaison, les nombres attendus de sites polymorphes sont respectivement pour chaque bloc :

$$e_1(t)=\frac{D}{N}t \quad \text{et} \quad e_2(t)=\frac{D}{N}(N-t)$$

Afin de mesurer la déviation du nombre observé de sites polymorphes par rapport au nombre attendu, le χ^2 est calculé pour chaque point de rupture possible :

$$\chi 2(t)=\frac{[x_1(t)-e_1(t)]^2}{e_1(t)}+\frac{[x_2(t)-e_2(t)]^2}{e_2(t)}$$

Le point de rupture t pour lequel le χ^2 est maximal (Figure 2.15C) est alors le meilleur candidat à partir duquel est inféré un événement de recombinaison. *In fine*, la probabilité de l'hypothèse nulle d'absence de recombinaison est estimée comme la fréquence à laquelle le Max χ^2 est inférieur aux Max χ^2 obtenus suite à des permutations aléatoires des positions relatives des sites nucléotidiques ;

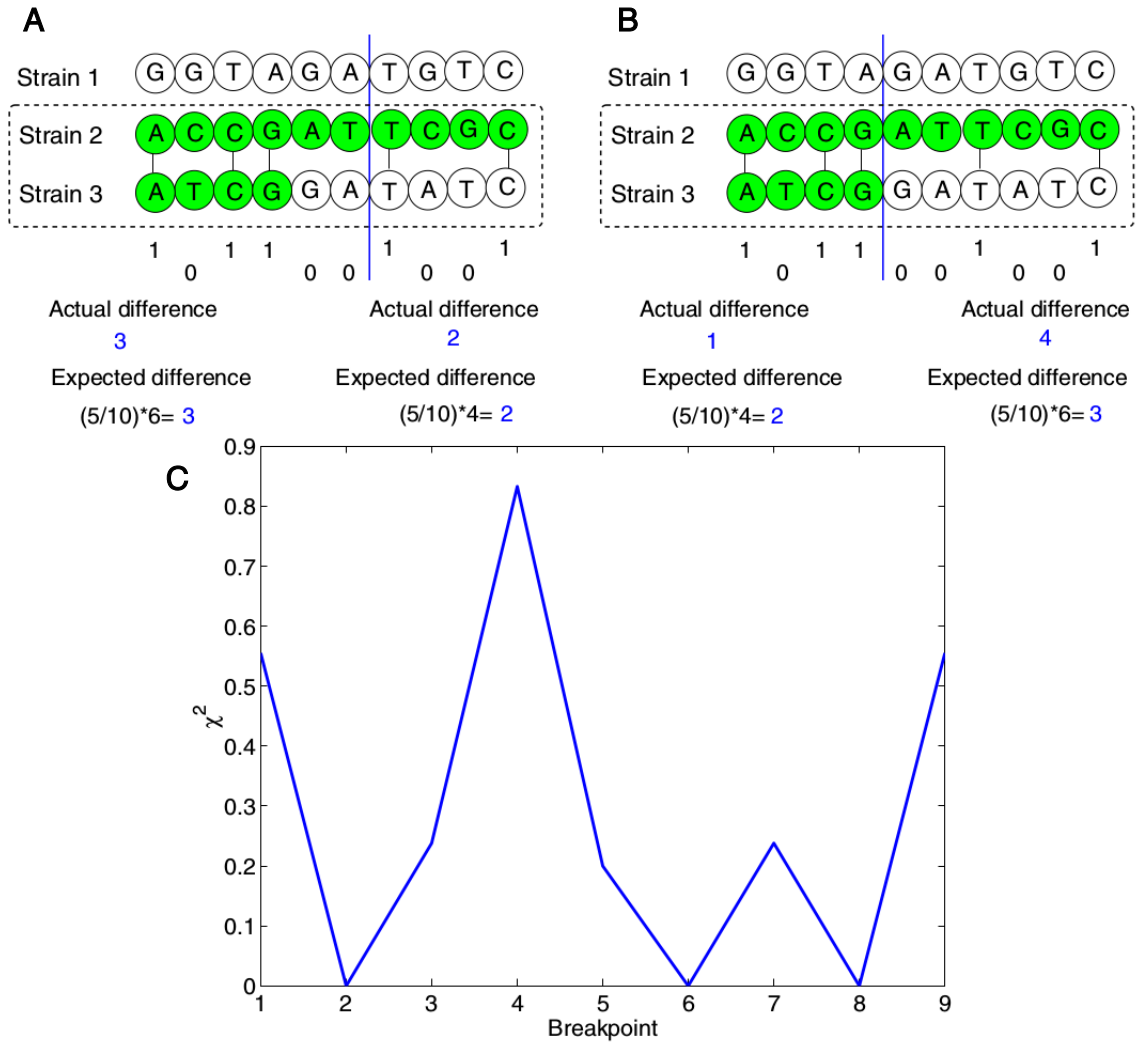


Figure 2.15. Illustration de la méthode du Max χ^2 (Husmeier *et al.*, 2006). Les partie **(A)** et **(B)** représentent l'alignement des séquences de trois souches. Les souches 1 et 2, représentées de couleurs différentes, dérivent d'un ancêtre commun, leur histoire évolutive n'intégrant pas de recombinaison. La souche 3 résulte d'un événement de recombinaison entre les ancêtres respectifs des souches 1 et 2 et est par conséquent une séquence mosaïque des deux autres. Si l'on compare les souches 2 et 3, la distribution des sites nucléotidiques est symbolisée par 0 lorsque les sites sont différents – *i.e.*, sites polymorphes – et par 1 lorsqu'ils sont identiques. La partie **(B)** montre la véritable localisation du point de rupture puisque la valeur du χ^2 est la plus élevée pour t=4, *i.e.*, de 0,83. **(C)** Graphique illustrant toutes les valeurs de χ^2 pour les points de rupture analysés. Le χ^2 est maximal lorsque le point de rupture est égal à 4.

- le test NSS (*Neighbor Similarity Score*) (Jakobsen and Easteal, 1996). Celui-ci repose sur le principe de parcimonie et le postulat selon lequel deux sites

informatifs (sites présentant *a minima* deux allèles différents, ces allèles étant chacun observés dans deux séquences au moins) (Figure 2.16A) sont compatibles s'il existe une unique histoire évolutive des séquences pour laquelle la répartition des allèles sur chaque site s'explique par une unique substitution nucléotidique au site considéré (Figure 2.16B). Le nombre de substitutions (c) est alors inférieur au nombre de nucléotides distincts (n) pour un site donné : $c = n - 1$ (Figure 2.16B). Les sites sont incompatibles lorsque pour toutes les histoires évolutives considérées $c > n - 1$, ce qui suggère qu'au moins l'un des deux sites serait issu d'un événement de recombinaison.

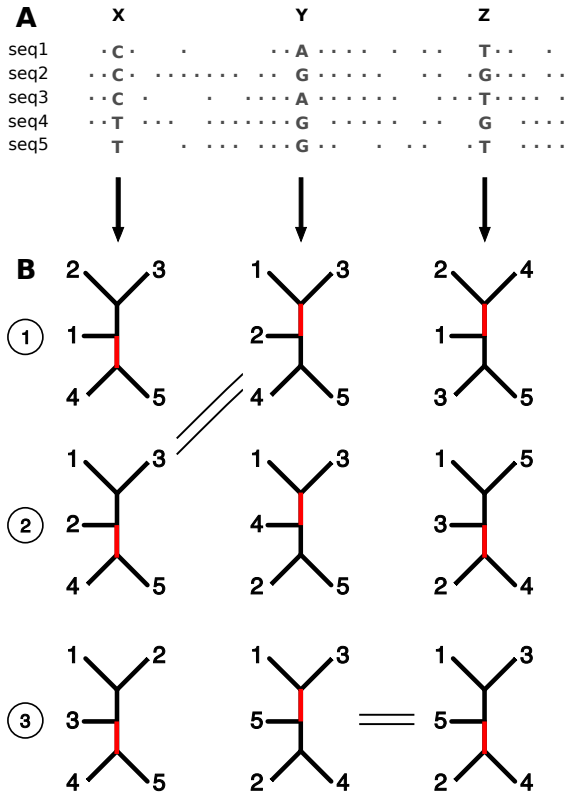


Figure 2.16. Détermination de la compatibilité entre deux sites informatifs (adapté de Jakobsen & Easteal, 1996). (A) Alignement de cinq séquences, au sein duquel trois sites informatifs – X, Y et Z – sont utilisés pour illustrer la détermination de la compatibilité. **(B)** Représentation des arbres pour lesquels $c = n - 1$, correspondant à chaque site informatif. La localisation des événements de substitution est indiquée en rouge. Les sites X et Y sont compatibles puisqu'ils ont en commun un arbre (respectivement les arbres 2 et 1) pour lequel $c = n - 1$. Il en est de même pour les sites Y et Z (les deux arbres 3). Cependant, les sites X et Z sont incompatibles dans la mesure où ils n'ont en commun aucun arbre pour lequel $c = n - 1$.

La compatibilité de chaque comparaison par paire de sites informatifs est mesurée et présentée sous la forme d'une matrice, où chaque entrée correspond à un site informatif selon sa localisation dans l'alignement. À partir de cette matrice, la proportion des regroupements de sites compatibles / incompatibles le long des séquences est estimée par le NSS. La significativité du test est mesurée en générant des matrices aléatoires, par permutations de l'ordre rela-

tif des sites informatifs, et en déterminant la proportion des scores aléatoires supérieurs ou égaux au score observé ;

- de la même manière, le test PHI (*Pairwise Homoplasy Index*) (Bruen *et al.*, 2006) détermine la compatibilité entre sites informatifs en intégrant cependant une notion de distance entre sites voisins. La probabilité d'un événement de recombinaison dépend ici de la distance physique entre les sites. Ainsi il est plus vraisemblable que les sites les plus éloignés soient incompatibles (Hudson and Kaplan, 1985; Hagenblad and Nordborg, 2002). Les séquences sont segmentées en fenêtres de longueur fixe pour lesquelles est calculée la statistique PHI pour *k* sites informatifs, *k* étant proportionnel au nombre total de sites informatifs dans l'alignement et à la longueur de la fenêtre. Cette statistique mesure donc directement la compatibilité entre sites proches plutôt qu'un regroupement de sites au sein d'une matrice de compatibilité (NSS). La significativité de la statistique est également obtenue par des permutations de l'ordre relatif des sites informatifs.

Le programme GENECONV v1.81 (Sawyer, 1999) a été utilisé avec les options /w123 pour l'initialisation du générateur de nombres aléatoires interne au programme, /lp permettant d'estimer les *p-values* locales, *gscale* = 2 autorisant les mésappariements entre les fragments. Les *p-values* ont été estimées à partir de 10 000 permutations. Seuls les COGs ayant une *p-value* globale inférieure à 0,05 ont été considérés comme recombinants.
Les statistiques associées aux tests de Max χ^2, NSS et PHI ont été calculées telles qu'implémentées dans le logiciel PhiPack (Bruen and Bruen, 2005). Concernant le test PHI, la longueur de la fenêtre a été fixée à 100 pb. Pour chaque test, les COGs sont considérés comme recombinants lorsque la *p-value* calculée sur la base de 1 000 permutations est inférieure à 0,05.

2.4.2.3. Identification des événements de recombinaison et de leur ancestralité

Un signal de recombinaison pour un COG recombinant peut être la conséquence d'un à plusieurs événements de recombinaison, et peut impliquer plusieurs gènes le constituant. Il apparaît donc nécessaire de compléter la caractérisation des COGs recombinants par l'identification des événements de recombinaison homologue pour l'ensemble

des gènes le constituant. Ceci a été réalisé à l'aide de l'outil fastGEAR (Mostowy *et al.*, 2017). La méthode développée dans cet outil permet de caractériser des lignées génétiques au sein d'une population à partir d'alignements de séquences et de détecter les événements de recombinaison probables, qu'ils se soient produits entre les lignées inférées ou qu'ils soient d'origine externe (Figure 2.17).

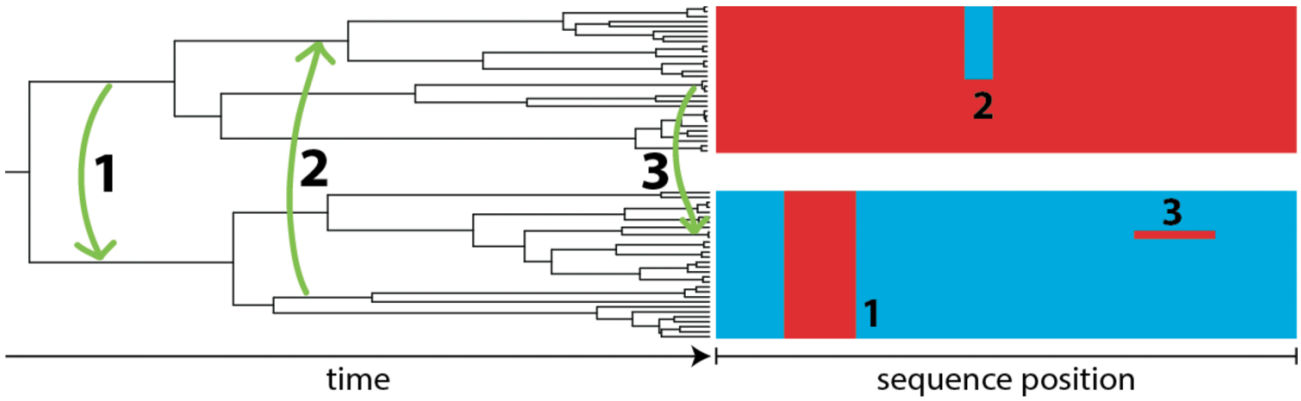


Figure 2.17. Définition des événements de recombinaison et de leur ancestralité dans fastGEAR (Mostowy *et al.*, 2017). La partie gauche représente un arbre phylogénétique caractérisant l'évolution de deux populations (ou lignées). La partie droite représente les alignements génétiques de l'ensemble des populations (ou lignées), les couleurs rouge et bleu représentant les modèles de séquences propres à chaque lignée. Dans un contexte de recombinaison, ces couleurs représentent les schémas ancestraux des résidus considérés. Trois événements de recombinaison représentés par des flèches vertes sur l'arbre sont à l'origine de trois blocs de brassage génétique. L'analyse des schémas de résidus sur les alignements permet de caractériser l'origine probable des ces résidus et donc d'inférer la localisation des événements de recombinaison dans l'arbre. Le schéma ancestral (1) affecte tous les isolats de la lignée bleu et suggère un événement de recombinaison ancestral, alors que les schémas ancestraux (2) et (3) n'affectant que certains isolats dans chacune des lignées, suggèrent des événements de recombinaison récents. Il est à noter que ces événements sont dits « récents ou ancestraux » uniquement par rapport aux lignées définies.

L'approche développée dans fastGEAR procède en quatre étapes :

- Identification des lignées génétiques : la délimitation des sous-populations, correspondant ici aux lignées génétiques, a été faite sur la base des travaux de Kashtan *et al.* (2014) et de la phylogénie des gènes *core* (*cf.* Chapitre 2, paragraphe 2.2.1) ;
- Identification des événements de recombinaison récents (Figure 2.18B) : ceux-ci sont détectés à l'aide d'un modèle HMM à l'échelle des souches au sein des

lignées. Il s'agit ici de comparer l'information portée par chaque site de chaque séquence au sein d'une lignée à l'ensemble des autres informations disponibles (*i.e.*, les autres souches de la même lignée ou les autres lignées). Ainsi, si un fragment provient d'une autre souche ou d'une autre lignée, le signal propre à cette souche / lignée sera détecté et l'origine du fragment lui sera attribuée. Si le fragment analysé ne provient pas d'une autre lignée et qu'il est par ailleurs très différent des autres isolats de la lignée à laquelle appartient les séquences qui le portent, son origine sera considérée comme externe ;

- Identification des événements de recombinaison ancestraux (Figure 2.18A) : ceux-ci sont identifiés avec une approche similaire à la détection des événements de recombinaison récents, à la différence que les investigations se font à l'échelle des lignées (et non plus des souches ou des isolats au sein des lignées), en faisant abstraction des sites associés à des événements de recombinaison récents. Il n'est pas possible ici de déterminer la direction des recombinaisons ancestrales ni si elles proviennent de sources externes ;
- Détection des faux-positifs : il s'agit ici de supprimer des événements inférés par erreur. Ceci est fait à travers l'analyse de la localisation de SNPs entre la souche cible et la souche / lignée ancestrale au sein et entre les segments détectés comme recombinants et la réalisation d'un test binomial pour calculer un facteur de Bayes (*Bayes factor* ou BF) (Bernardo and Smith, 2001), qui mesure la force du support des changements de densité des SNPs à l'inférence des événements de recombinaison.

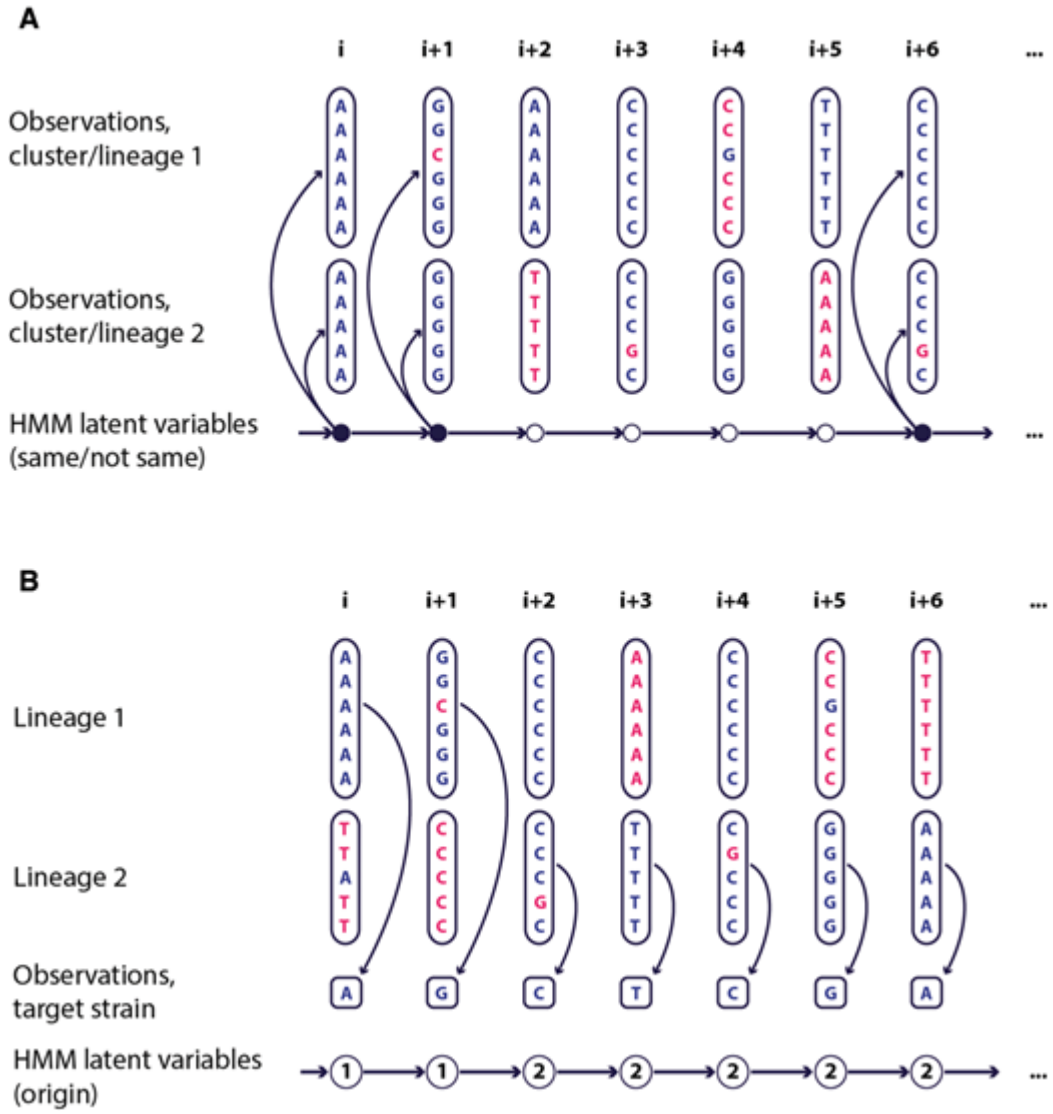


Figure 2.18. Structure des modèles de Markov cachés (HMM) pour la détection et la caractérisation d'événements de recombinaison dans fastGEAR (Mostowy *et al.*, 2017). (A) HMM utilisé pour identifier les lignées et déduire les recombinaisons ancestrales. Chaque colonne représente un site polymorphe dans l'alignement et les lignes représentent des souches. Les états observés de la chaîne sont les fréquences alléliques au sein de chaque groupe (dans le cas de l'identification de lignées) ou lignée (dans le cas de l'identification de recombinaisons ancestrales). Les états cachés de la chaîne représentent l'identité (ou l'absence d'identité) des fréquences des allèles dans les deux lignées au niveau des sites polymorphes. **(B)** HMM utilisé pour l'identification des recombinaisons récentes. Les états observés sont des valeurs de nucléotides observées dans la souche cible et les états cachés sont des origines possibles des nucléotides. Les origines possibles comprennent toutes les lignées observées plus une origine inconnue.

3. Analyse pangénomique d'une population bactérienne environnementale

Au cours des dernières années, les progrès réalisés dans le domaine de la génomique « cellule-unique » (SCG) ont amélioré l'échantillonnage et l'isolement de génomes de micro-organismes appartenant à une même population et coexistants dans un même environnement. Il devient donc possible d'étudier, à une échelle fine, les facteurs qui régissent la diversification de la structure et de l'organisation des génomes microbiens. Grâce au développement de ces approches, nous disposons maintenant de données génomiques à l'échelle populationnelle pour des espèces majeures de l'environnement comme, par exemple, *Prochlorococcus marinus*, une des espèces photosynthétiques les plus abondantes de la zone euphotique océanique, puisqu'elle est responsable de près de 10% de la productivité primaire marine. Sa diversité génétique se compose d'au moins 12 écotypes, parmi lesquels des écotypes de groupes haute lumière (HL) et de groupes basse lumière (LL) (*cf.* Chapitre 1.4). Il a été démontré que ces écotypes contiennent des ensembles de gènes fonctionnels variables ce qui leur permet de s'adapter, par exemple, à des changements de disponibilité de la lumière et suggère une répartition stable des groupes écologiquement distincts en niches.

Sur la base d'une approche SCG 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 a été mise en évidence pour ces sous-populations coexistantes avec notamment la caractérisation de compartiments génomiques incluant un *backbone* principalement composé de gènes *core* – hautement conservés entre tous les individus – et des ISLs principalement composés de gènes flexibles – partie variable de ces génomes.

Afin de répondre à l'un des objectifs de cette thèse qui était de mieux comprendre les fondements évolutifs de la différenciation des populations bactériennes dans l'environnement, une ré-évaluation du contenu génétique, du potentiel fonctionnel et de la proximité taxonomique du pangénome à l'échelle populationnelle 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) et telles qu'elles ont été proposées par Kashtan *et al.* (2014).

3.1. Phylogénie, différentiation des sous-populations et organisation des génomes

Le travail réalisé au cours de cette étude repose sur l'analyse de 87 SAGs (Tableau 2.1, Chapitre 2) de *Prochlorococcus* appartenant à l'écotype HLII. Ces SAGs sont distribués au sein de trois groupes (*i.e.*, cN2, c9301 et cN1) préalablement définis par Kashtan *et al.* (2014) sur la base d'une analyse de l'ITS (*cf.* Figure 2.1, Chapitre 2). Ces mêmes auteurs ont également défini l'existence de sept sous-populations majeures (C1 à C5, C8 et C9) en se basant cette fois-ci sur une analyse phylogénétique réalisée à partir de l'alignement des génomes complets de 96 SAGs, incluant ceux étudiés (Figure 3.1).

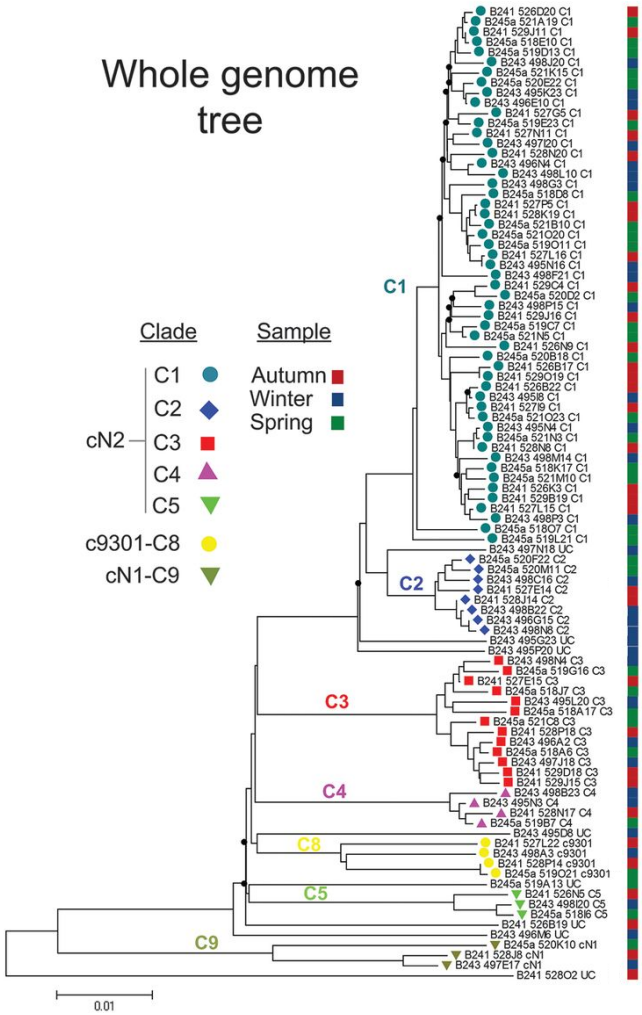


Figure 3.1. Arbre phylogénétique basé sur l'alignement des séquences génomiques de 96 SAGs de *Prochlorococcus* écotype HLII (Kashtan *et al.*, 2014). La saison de prélève-

ment est indiquée pour chaque SAG sur la droite de l'arbre grâce à un code couleur (automne – rouge ; hiver – bleu ; printemps – vert). Les cercles noirs au niveau des nœuds mettent en évidence des valeurs de *bootstrap* inférieures à 80 %. Chaque sous-population (C1 à C5, C8 et C9) est représentée par une couleur et une forme qui lui est propre. La divergence a été estimée par la distance *p* (fonction du nombre de substitutions observées au niveau des sites homologues). L'arbre a été produit avec la méthode du *Neighbor-Joining*.

Malgré leur congruence (*i.e.*, même délimitation des sous-populations), la phylogénie basée sur les séquences génomiques (Figure 3.1) ne soutient pas la monophylie des trois groupes (cN2, c9301 et cN1) fondée sur la phylogénie des ITS (Figure 2.1). En effet, le groupe c9301, constitué des SAGs du clade C8, forme une ramification au sein du groupe cN2.

Afin de compléter les travaux de Kashtan *et al.* (2014), une phylogénie basée sur l'analyse de l'histoire évolutive des sous-populations à l'échelle du génome *core*, a été réalisée. Elle permet d'exclure ainsi les informations portées par le génome flexible et les événements de recombinaison hétérologue qui y sont majoritairement associés. Après concaténation des alignements des séquences nucléotidiques des gènes *core* (*i.e.,* gènes partagés par les 13 génomes de souches cultivées de l'écotype HLII, soit 1 202 gènes en copie unique au total), une phylogénie au maximum de vraisemblance a été estimée, en utilisant le génome de référence MIT9312 comme groupe externe (Figure 3.2). Cette phylogénie confirme la distinction des sous-populations en sept clades (valeurs de *bootstrap* supérieures à 80%). Cependant, bien qu'elle souligne également la paraphylie du groupe cN2, elle l'explique par le partage d'un ancêtre commun entre les clades C3 et C8 (valeur de *bootstrap* égale à 100%).

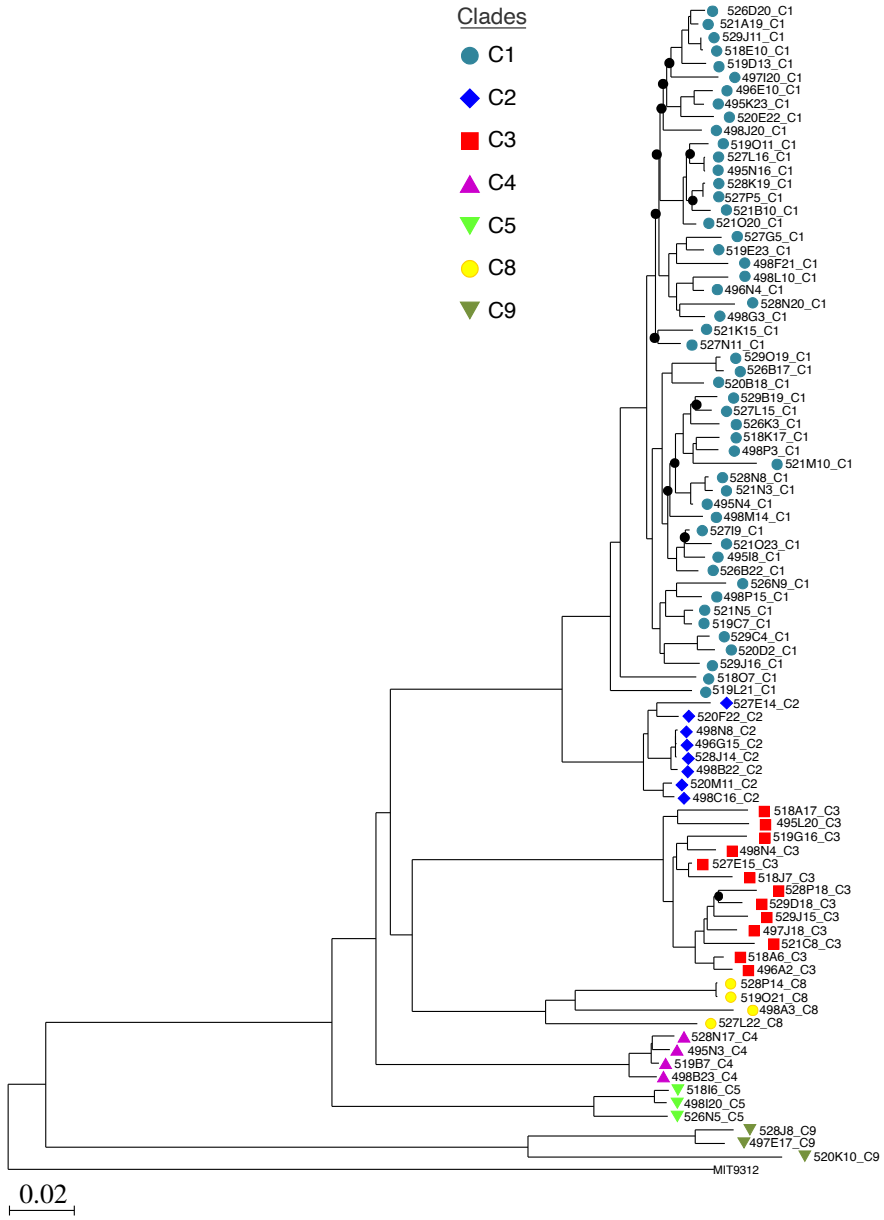


Figure 3.2. Arbre phylogénétique au maximum de vraisemblance basé sur la concaténation des alignements des gènes *core* présents en copie unique. Les gènes *core* (1 202 au total) sont partagés par 13 génomes de souches cultivées de l'écotype HLII de *Prochlorococcus*. Le génome de référence MIT9312 a été utilisé pour enraciner l'arbre. Les valeurs de *bootstrap* inférieures à 80 % sont signalées au niveau des nœuds par un cercle noir. Les différents clades sont représentés par des formes et des couleurs qui leurs sont propres et qui sont positionnées sur chaque feuille de l'arbre. Le modèle d'évolution utilisé est le GTR (*General Time Reversible, Tavaré, 1986)*.

Les ANI, reflet de la divergence entre les séquences génomiques, ont été évaluées par la comparaison des SAGs deux à deux. En effet, le pourcentage d'ANI reflète la pro-

portion des régions génomiques qui s'alignent entre deux génomes. Les résultats obtenus par cette méthode soutiennent bien la délimitation des sous-populations (Figure 3.3), assimilées ici à des OTUs.

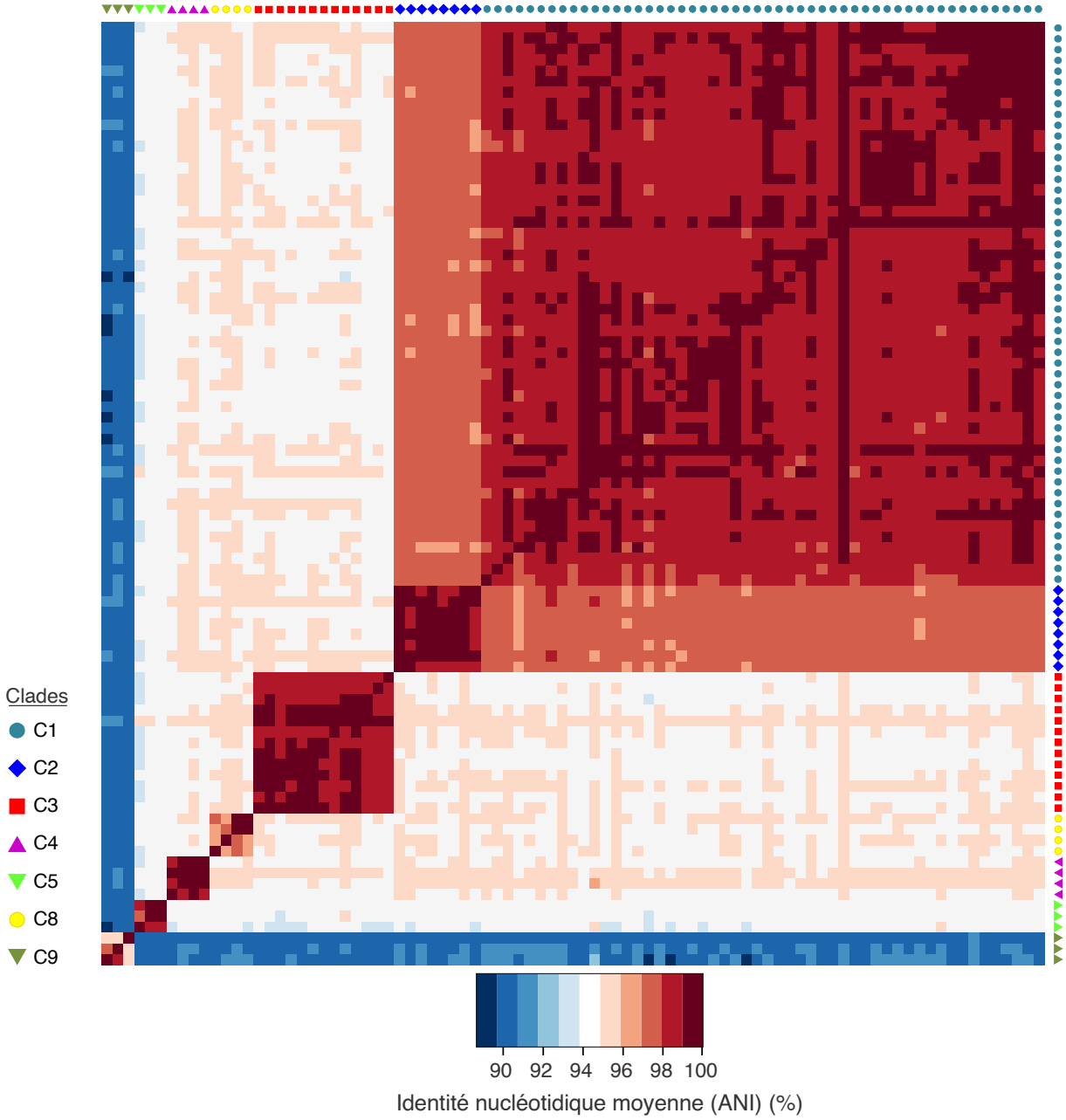


Figure 3.3. *Heatmap* montrant les pourcentages d'identité nucléotidique moyenne (ANI) calculés entre paires de SAGs pour l'ensemble des clades (C1 à C5, C8, C9). L'ordre des SAGs tels qu'ils apparaissent sur les lignes et les colonnes de la *heatmap* est identique à celui de l'arbre phylogénétique présenté en Figure 3.2. Chaque clade est caractérisé par une forme et un couleur qui lui est propre.

Les comparaisons par paires de SAGs entre clades mettent en évidence des pourcentages d'ANI homogènes de l'ordre de 94 % en moyenne pour les sous-populations C1 à C8 et de 90 % lorsque les comparaisons sont faites avec le clade C9 (Figure 3.3 ; Tableau 3.1). Ce dernier est le clade le plus divergent, ce qui est en accord avec sa position phylogénétique et son émergence comme branche la plus basale dans l'arbre (Figure 3.2). *A contrario*, les identités les plus élevées sont observées pour les plus proches parents, à savoir les clades C1 et C2 (ANI égale à 97 % en moyenne).

Lorsque les comparaisons sont réalisées entre les SAGs d'un même clade, les valeurs d'ANI sont supérieures à 98 %, exception faite pour les clades C8 (97%) et C9 (96%) (Tableau 3.1). Par conséquent, ces résultats reflètent non seulement le faible polymorphisme intra-clade observé par Kashtan *et al.* (2014), mais aussi la différenciation des allèles entre clade mise en évidence par ces mêmes auteurs.

Tableau 3.1. Pourcentages d'identité nucléotidique moyenne (ANI) obtenus après comparaison par paires de SAGs. Les ANI calculées entre paires de SAGs ont été moyennées en fonction de la comparaison inter- ou intra-clade considérée.

	C1	C2	C3	C4	C5	C8	C9
C1	98,68						
C2	97,3	99,18					
C3	94,66	94,8	98,83				
C4	94,78	94,93	94,65	99,15			
C5	94,16	94,17	94,03	94,16	98,53		
C8	94,75	94,8	94,8	94,72	94,24	97,33	
C9	90,31	90,37	90,25	90,44	90,08	90,36	96,25

Pour compléter les données en lien avec la différenciation des populations, le contenu génétique des sous-populations, ainsi que les segments génomiques partagés, ont été mis en évidence par la comparaison de leurs génomes. Pour cela, un SAG représentatif de chaque sous-population a été sélectionné sur la base de la taille de son assemblage (la plus grande), sa complétude (la plus importante) et sa contamination (la plus faible) (Tableau 3.2). Les séquences génomiques des sept SAGs retenus, ainsi que celle de la souche cultivée MIT9312 (choisie du fait de son égale distance évolutive avec les SAGs

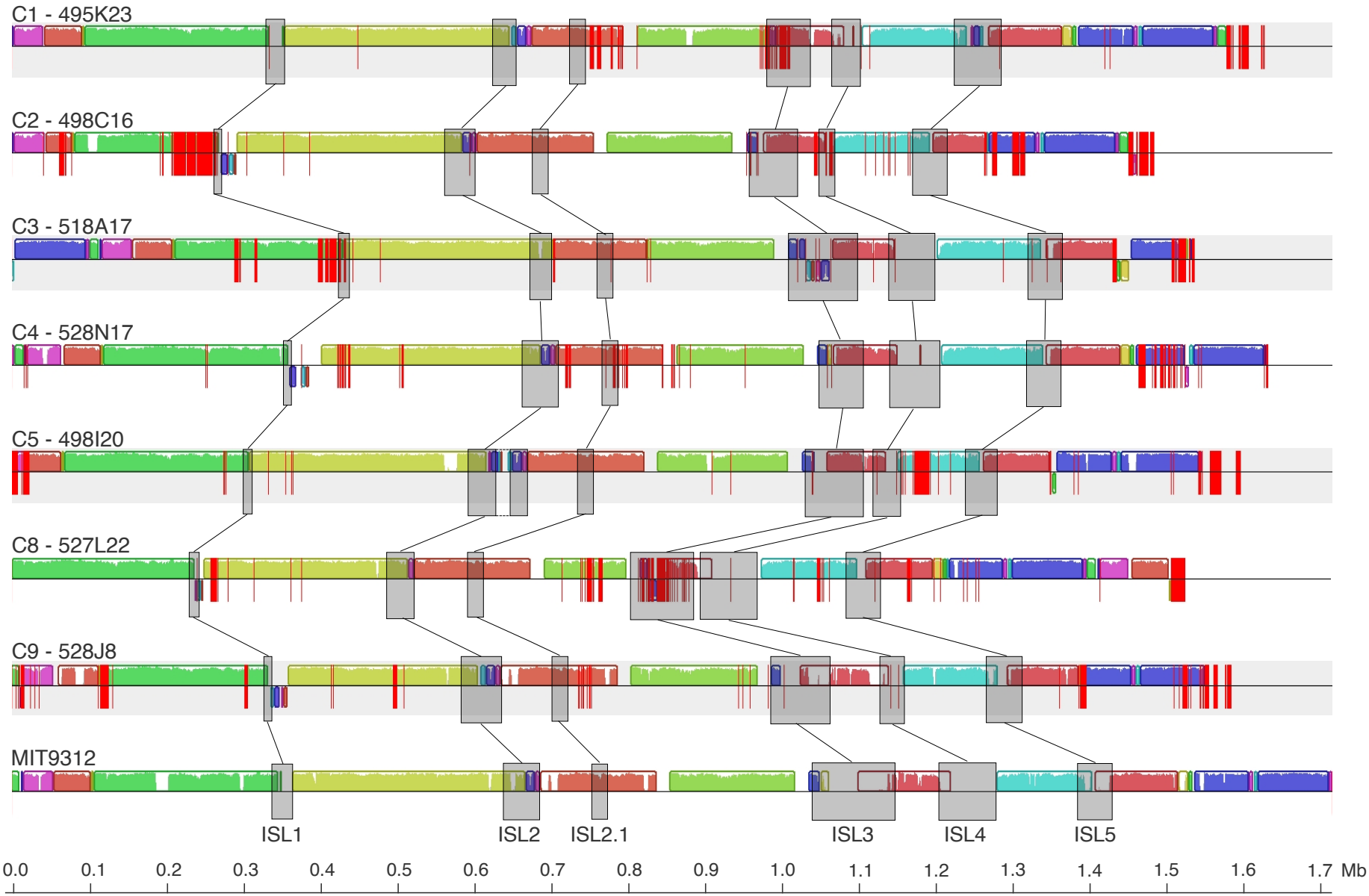
analysés) ont été alignées à l'aide de l'algorithme progressiveMauve (Darling *et al.*, 2010). Cette approche a permis de déterminer les LCBs entre les SAGs. Les segments communs à l'ensemble des SAGs comparés totalisent 1,33 Mpb soit 76 % de la taille du génome de MIT9312 (Figure 3.4). Prise individuellement, cette taille varie entre 1,35 (SAG 498C16 appartenant au clade C1) et 1,47 Mpb (SAG 518A17, clade C3), soit entre 87 et 92 % de la longueur des assemblages des SAGs et entre 79 et 86 % de la taille du génome de MIT9312 (Tableau 3.2).

Tableau 3.2. Caractéristiques des SAGs représentatifs de chaque sous-population et proportion des segments conservés (LCBs) avec le génome de référence MIT9312. La taille du génome de MIT9312 est de 1,71 Mpb.

Sous-population	SAG	Taille d'assemblage (Mpb)	Complétude SAGs (%)	Segments conservés (Mpb)	Couverture SAGs (%)	Couverture MIT9312 (%)
C1	495K23	1,62	96,92	1,42	87,65	83,08
C2	498C16	1,47	92,45	1,35	91,83	78,98
C3	518A17	1,62	97,39	1,47	90,74	86,00
C4	528N17	1,53	93,43	1,38	90,19	80,74
C5	498I20	1,58	94,63	1,41	89,24	82,49
C8	527L22	1,51	92,44	1,37	90,73	80,15
C9	528J8	1,57	92,62	1,38	87,90	80,74

Il existe, par conséquent, une forte synténie entre les SAGs, en adéquation avec ce qui a été observé entre les souches composant l'écotype HLII de *Prochlorococcus* (Yan *et al.*, 2018). Une diminution des similarités est toutefois observée au niveau des régions correspondant aux ISLs tels qu'ils ont été caractérisés dans le génome de MIT9312 (Avrani *et al.*, 2011) (Figure 3.4).

Figure 3.4. Alignement des génomes des SAGs représentatifs de chaque clade et de la souche cultivée de référence MIT9312 à l'aide de l'outil progressiveMauve (Darling *et al.*, 2010). Les segments conservés (LCBs) apparaissent en couleur tandis que les domaines uniques / variables pour un génome donné apparaissent en blanc. Les zones grisées correspondent aux six îlots génomiques (ISL1, ISL2, ISL2.1, ISL3, ISL4 et ISL5), tels qu'observés dans le génome de MIT9312. Les ISLs des SAGs et de MIT9312 sont connectés entre eux par des lignes grises. Les lignes verticales rouges indiquent les extrémités des *contigs*. Pour chaque SAG, son nom ainsi que celui de la sous-population à laquelle il est rattaché apparaissent en début de ligne.


C1 - 495K23
C2 - 498C16
C3 - 518A17
C4 - 528N17
C5 - 498I20
C8 - 527L22
C9 - 528J8
MIT9312
ISL1
ISL2
ISL2.1
ISL3
ISL4
ISL5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
Mb

3.2. Variabilité du contenu en gènes

Bien que l'organisation génomique soit conservée, il existe une réelle différentiation des sous-populations, telle que décrite par la phylogénie des gènes *core* et l'analyse des ANI. Par ailleurs, cette différenciation pourrait être, en partie, accentuée par une variabilité de leur contenu en gènes. Pour répondre à cette interrogation, le pangénome – collection de gènes *core* et flexibles caractéristiques d'un ensemble de génomes microbiens phylogénétiquement proches – associé aux 87 SAGs a été étudié. Ce pangénome est constitué de 7 125 COGs. Il est qualifié d'ouvert, dans la mesure où la courbe d'accumulation des COGs qui le décrit n'atteint pas un plateau (Figure 3.5). En outre, étant donné le nombre de COGs supplémentaires apporté par chaque nouveau génome séquencé, y compris à cette échelle de diversité, il est impossible, en l'état, d'estimer la taille du pangénome chez *Prochlorococcus*.

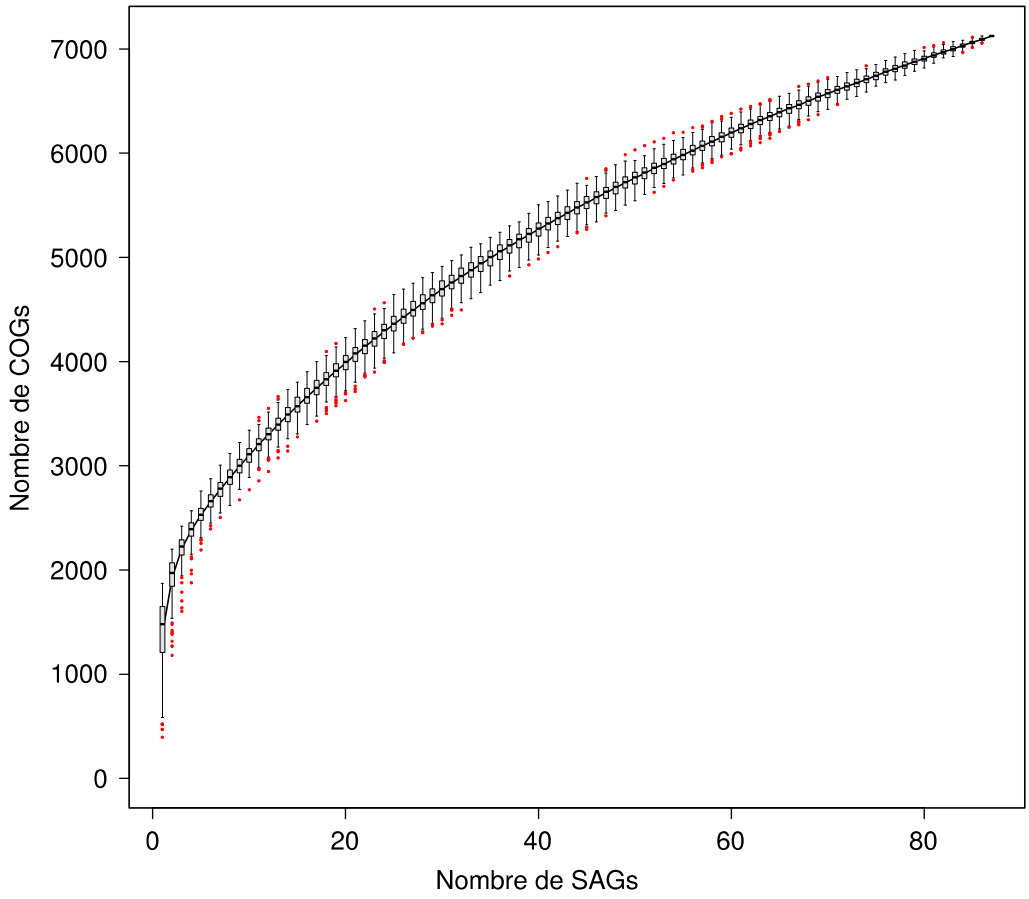


Figure 3.5. Pangénome des 87 SAGs analysés. La taille du pangénome est dépendante du nombre de génomes étudiés, et le nombre de clusters de gènes orthologues (COGs) augmente

à mesure qu'un génome est ajouté. Pour un ensemble de 87 SAGs, et pour l'estimation d'un pangénome constitué de *k* génomes, il existe 87!/(k!(87-k)!) combinaisons possibles de SAGs (Tettelin *et al.*, 2005). Ainsi, la distribution du nombre de COGs pour *k* génomes est représentée par un diagramme en boîte et les *outliers* par des points rouges.

Parmi les 7 125 COGs analysés, 1 410 sont catégorisés comme *core*, puisque communs aux 13 génomes de souches cultivées de *Prochlorococcus,* écotype HLII. Les COGs *core* sont présents *a minima* dans 27 SAGs, et en moyenne dans 60 SAGs (Figure 3.6). Bien que, par définition, le génome *core* soit commun à l'ensemble des individus étudiés, cette règle simple n'est pas respectée en raison de l'incomplétude des génomes obtenus par la méthode de séquençage cellule-unique. Parmi les COGs restants, dits flexibles (5 715 au total), 83 % sont présents dans 10 SAGs au moins, et 70 % *a minima* dans 50 SAGs. Seuls 382 sont partagés avec le génome de référence MIT9312.

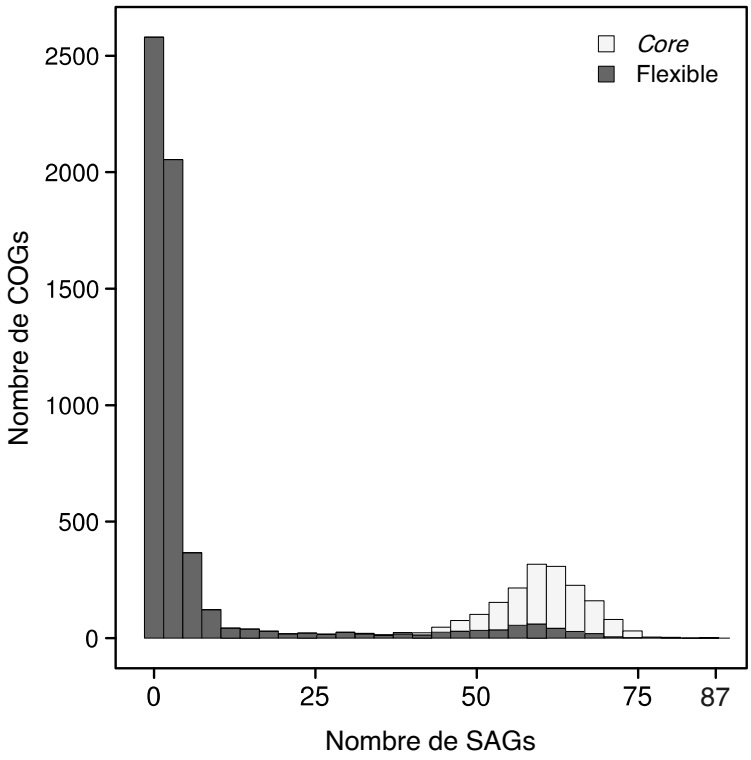


Figure 3.6. Profil de distribution des clusters de gènes orthologues (COGs) dans les SAGs en fonction des catégories *core* et flexible. Les COGs *core* (gris clair) représentent l'ensemble des COGs (1 410 au total) présents dans les 13 génomes des souches cultivées pour *Prochlorococcus,* écotype HLII (*i.e.,* MIT9311, MIT9314, MIT9401, MIT9301, MIT9312, MIT9107, MIT9201, MIT9321, MIT9202, MIT9215, SB, GP2 et AS9601). Les COGs flexibles (gris foncé ; 5 715 au total) correspondent à ceux retrouvés dans le génome de certaines souches ou spécifiques des sous-populations (C1 à C5, C8 et C9).

Les gènes flexibles spécifiques des SAGs étudiés (donc absents de la souche de référence MIT9312 ; 5 333 au total) montrent, en revanche, une distribution plus resserrée, puisque 94 % d'entre eux sont présents dans moins de 10 SAGs et seulement 2 % dans plus de 50 SAGs (Figure 3.6). La courbe d'accumulation décrivant l'évolution du nombre de COGs flexibles en fonction du nombre de SAGs (Figure 3.7) explique l'expansion du pangénome qui est observée puisqu'elle montre que plus il y a de SAGs plus le nombre de COGs est important.

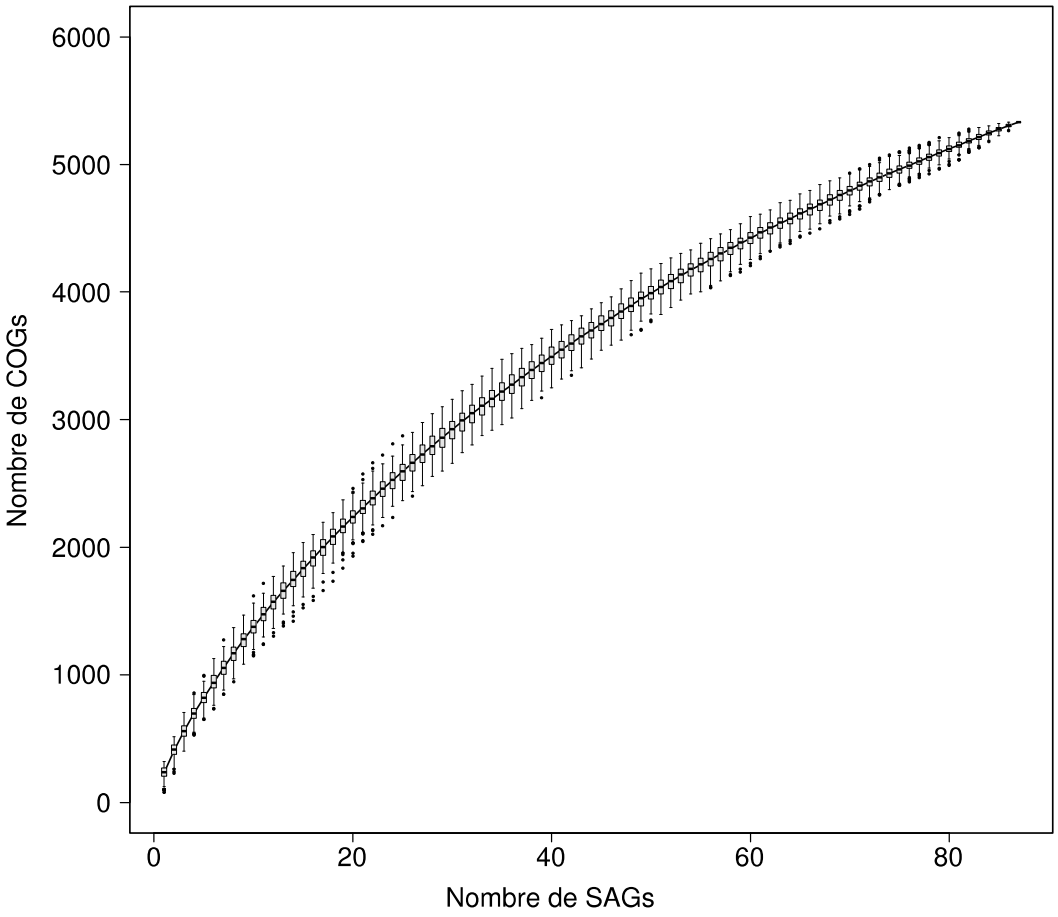


Figure 3.7. Évolution du nombre de COGs flexibles spécifiques des SAGs en fonction du nombre de SAGs. Dans cette étude, les COGs flexibles non partagés par le génome de référence MIT9312 sont au nombre de 5 333. Pour un ensemble de 87 SAGs, et pour l'estimation d'un pangénome constitué de *k* génomes, il existe 87!/(k!(87-k)!) combinaisons possibles de SAGs (Tettelin *et al.*, 2005). Ainsi, la distribution du nombre de COGs pour *k* génomes est représentée par un diagramme en boîte et les *outliers* par des points.

La prépondérance des COGs flexibles relativement rares pourrait s'expliquer par une spécificité à l'échelle des clades. Pour approfondir cette hypothèse, leur distribution à l'échelle des clades a donc été étudiée. Ceci a conduit à la définition de sept catégories, à savoir de la catégorie 1 (COGs observés dans un seul clade) à la catégorie 7 (COGs communs à l'ensemble des clades) (Figure 3.8 ; Tableau 3.3).

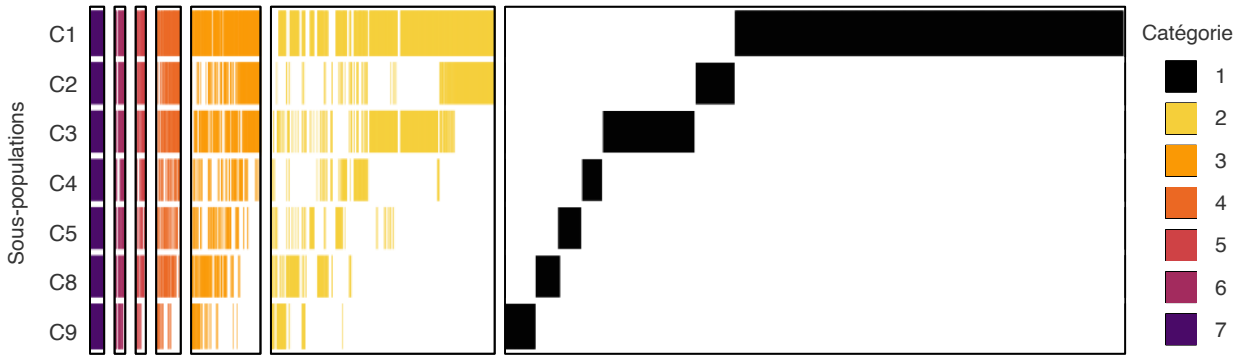


Figure 3.8. Représentation de la distribution des COGs SAG-spécifiques en fonction des sous-populations. Chaque COG est attribué à une catégorie selon sa représentativité au sein des sous-populations (catégories 1 à 7, *i.e.,* COGs spécifiques d'une sous-population en noir à communs à toutes en violet). Les catégories sont caractérisées par des couleurs différentes. Chaque barre verticale correspond à un COG.

Les proportions les plus importantes de COGs flexibles spécifiques des SAGs sont observées dans les deux premières catégories, avec 3 547 COGs dans la catégorie 1 et 1 135 COGs dans la catégorie 2 (Tableau 3.3). Par ailleurs, les COGs communs à l'ensemble des clades et appartenant donc à la catégorie 7 (81 au total) pourraient être considérés comme fixés dans les populations étudiées et donc assimilés à des gènes *core* à cette échelle de diversité.

Tableau 3.3. Nombre de COGs flexibles SAG-spécifiques en fonction de leur distribution au sein des clades. Sept catégories ont été définies selon la présence des COGs dans un unique clade (catégorie 1) ou plus (catégories 2 à 7). La catégorie 7 regroupe les COGs communs à l'ensemble des sous-populations étudiées.

		Nombre de COGs par catégorie						
		7	**6**	**5**	**4**	**3**	**2**	**1**
Sous-populations	**C1**	81	48	42	115	343	1 000	2 248
	C2	81	43	36	80	178	337	222
	C3	81	45	34	89	225	486	528
	C4	81	36	31	49	86	126	112
	C5	81	39	22	48	83	94	127
	C8	81	44	31	69	115	150	137
	C9	81	33	14	26	53	77	173
	Total	81	48	42	119	361	1 135	3 547

3.3. Relation entre complétude des SAGs et taille du pangénome – choix du jeu de données

Les SAGs étudiés montrent une grande disparité au niveau de la taille de leur assemblage et par conséquent, de leur complétude. Ce paramètre pouvant impacter cette étude, il est nécessaire de l'évaluer. En effet, bien qu'il existe une forte corrélation positive entre complétude et nombre total de COGs par SAG (ρ=0,94 ; p<0,001, corrélation de Spearman) (Figure 3.9A), celle-ci est quasi nulle (ρ=0,03 ; p=0,80, corrélation de Spearman) lorsque l'on compare, à l'échelle des SAGs, la complétude et le nombre de COGs spécifiques d'une unique sous-population (Figure 3.9B). Cette analyse montre donc qu'il est difficile, du fait des données partielles, d'évaluer la quantité d'information manquante, celle-ci étant variable d'un individu à l'autre, mais aussi de statuer sur la représentativité d'un individu (quelle que soit sa complétude, y compris pour des complétudes élevées), la part qui lui est spécifique étant également variable.

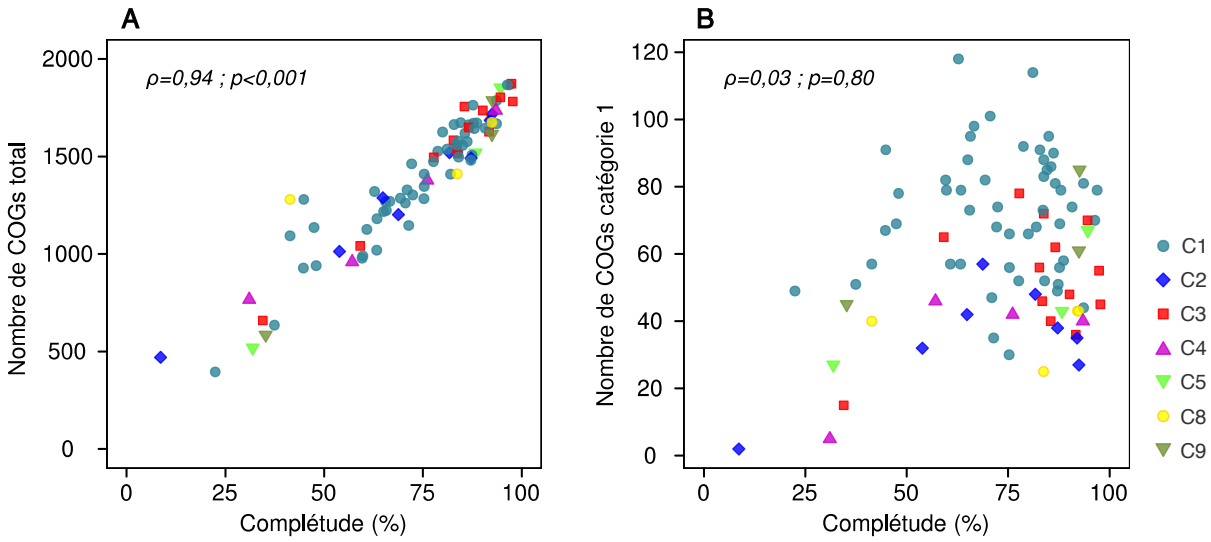


Figure 3.9. Corrélation entre complétude des SAGs (exprimée en %) et nombre de COGs. (A) Corrélation entre complétude et nombre de COGs flexibles total. **(B)** Corrélation entre complétude et COGs flexibles spécifiques d'une sous-population. Chaque point correspond à un SAG. Les SAGs d'une sous-population 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-values* associées sont indiqués.

Afin d'évaluer l'impact potentiel d'une complétude insuffisante sur les résultats de l'étude, une analyse ne tenant compte que de l'information portée par les SAGs quasi-complets a également été réalisée. Le jeu de données a ainsi été réduit aux 18 SAGs pour lesquels la complétude est supérieure à 90 %. Ces SAGs sont répartis dans les sept clades (*i.e.,* 5 SAGs du clade C1, 2 de C2, 5 de C3, 1 de C4, 1 de C5, 2 de C8 et 2 de C9 ; *cf.* Tableau 2.1), incluant donc ceux caractérisés par un faible nombre de SAGs (*i.e.,* C4, C5, C8 et C9). Cela réduit le jeu de données de COGs flexibles SAG-spécifiques à 3 286 sur les 5 333 identifiés avec les 87 SAGs. L'ensemble des clades étant représenté, la balance entre complétude, nombre de SAGs et représentativité des COGs flexibles spécifiques a pu être considérée dans chaque sous-population. Les résultats obtenus à partir des données réduites (Figure 3.10) montrent une distribution des COGs différente de celle observée avec le jeu de données total (Figure 3.8). En effet, par comparaison avec l'analyse globale, de nombreux COGs se trouvent partagés par un nombre plus restreint de clades et changent de catégorie (Figure 3.10). Certains, initialement partagés par au moins deux clades deviennent même spécifiques d'un unique clade. L'utilisation d'un jeu de données réduit, bien qu'intégrant des génomes dont la complétude est optimale, induirait des biais d'interprétation dans l'analyse des pangénomes, voire une sur-interprétation de la spécificité de COGs pour certaines sous-populations (même lorsque celles-ci ne sont représentées que par un faible nombre de SAGs).

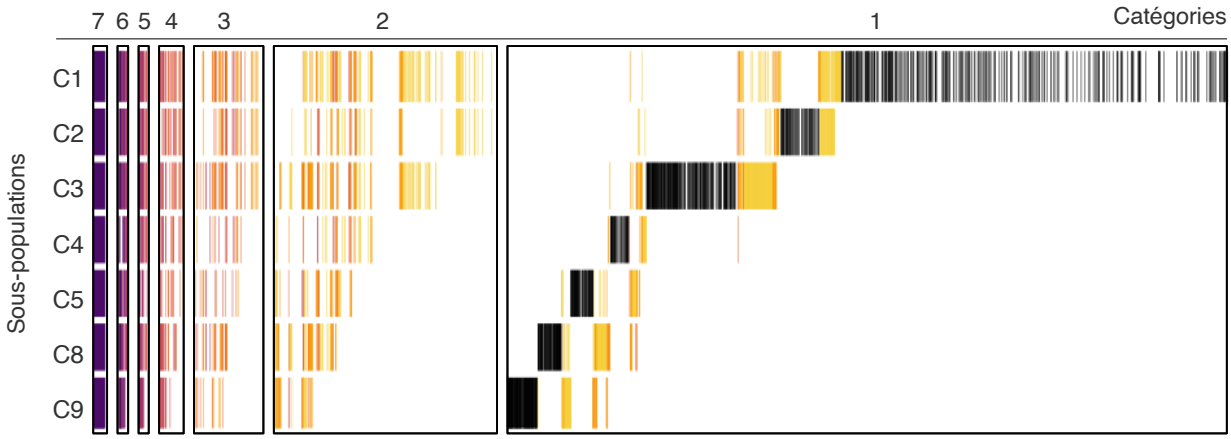


Figure 3.10. Représentation de la distribution des COGs SAG-spécifiques des 18 SAGs quasi-complets (3 286 au total) en fonction des sous-populations (C1 à C5, C8 et C9) et de leur attribution à une catégorie (catégories 1 à 7, *i.e.,* COGs spécifiques d'une sous-population à droite à COGs communs à toutes à gauche). Contrairement à

la Figure 3.8, 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.

Pour compléter cette analyse, la possibilité d'une répartition aberrante des COGs dans les sous-populations en fonction de la taille du jeu de données a été évaluée. Celle-ci a été appréciée par le biais d'un tirage aléatoire des COGs pour un nombre croissant de SAGs, à l'image de l'approximation de la taille d'un pangénome. La répartition des COGs dans les différentes sous-populations pour un nombre de SAGs donné a donc été comparée à celle effectivement observée quand l'ensemble des SAGs est considéré (Figure 3.11). Ainsi, si, après réduction du jeu de données, un COG devient spécifique d'une sous-population (catégorie 1) alors qu'il était observé dans au moins deux sous-populations lors de l'analyse globale, il sera considéré comme faux-positif. Une sur-estimation du nombre de COGs spécifiques d'une sous-population (catégorie 1) est constatée pour une partie des jeux de données réduits (Figure 3.11). En effet, la fréquence moyenne des faux-positifs ne cesse de croître jusqu'à atteindre un maximum pour 27 SAGs (en moyenne ±σ, 2 140 COGs ±90 COGs flexibles attribués à la catégorie 1, soit à 746 ±39 faux-positifs) (Figure 3.11). Ainsi, il est préférable d'intégrer l'ensemble des génomes, même partiels, pour proposer une image plus objective d'un pangénome, en particulier lorsque sont étudiées des populations structurées.

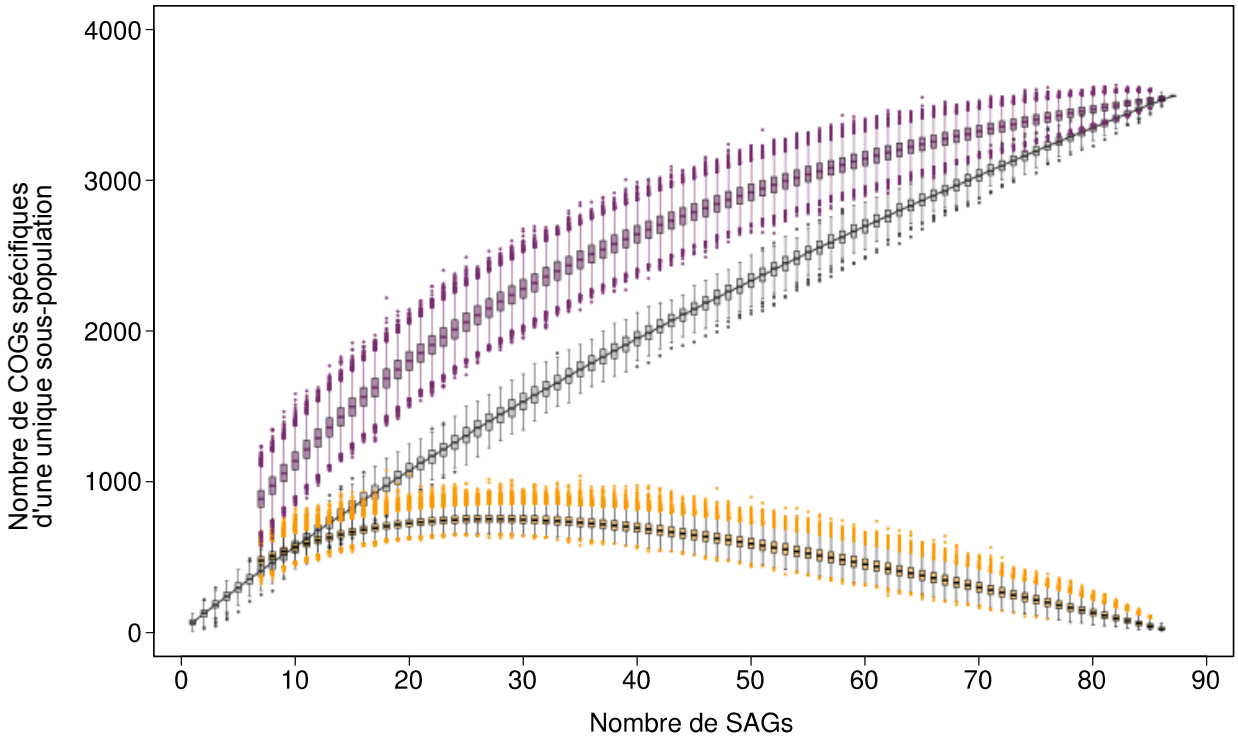


Figure 3.11. Evolution du nombre de COGs flexibles en fonction de nombre de SAGs. Pour une meilleure lisibilité des résultats, seuls ceux spécifiques d'une unique sous-population (catégorie 1) sont présentés. La courbe d'accumulation grise correspond au nombre de COGs de catégorie 1 caractérisés à partir de l'ensemble des 87 SAGs (3 547 au total). La courbe violette correspond au nombre de COGs de catégorie 1 inféré après le tirage aléatoire d'un nombre croissant de SAGs. La courbe orange restitue l'évolution du nombre de COGs affectés par erreur à la catégorie 1 (faux-positifs) en fonction du nombre de SAGs échantillonnés. Le nombre maximal de faux positifs est observé pour 27 SAGs.

Puisque l'information portée par l'ensemble des SAGs est plus pertinente que celle provenant uniquement de SAGs complets ou quasi-complets – et ce malgré une proportion importante de SAGs partiels – le choix a été fait de réaliser les analyses à partir du jeu de données global.

3.4. L'ébauche d'un paysage génomique

3.4.1. Assignation des COGs flexibles non partagés par MIT9312 à un compartiment génomique

À la lumière de la synténie à l'échelle des génomes (*cf.* Figure 3.4), un compartiment chromosomique – *i.e.*, *backbone* ou ISL – a été attribué à l'ensemble des COGs flexibles spécifiques des SAGs présents en copie unique (soit 5 290 COGs sur les 5 333). L'assignation à un compartiment a, dans un premier temps, été réalisée pour chaque gène de chaque COG par l'analyse de leur contexte génomique, puis extrapolée à l'échelle du COG *via* l'estimation de l'entropie de Shannon (notée H). En effet, cette méthode permet l'évaluation du nombre de compartiments retrouvés au sein d'un COG et de leur proportion (Figure 3.12).

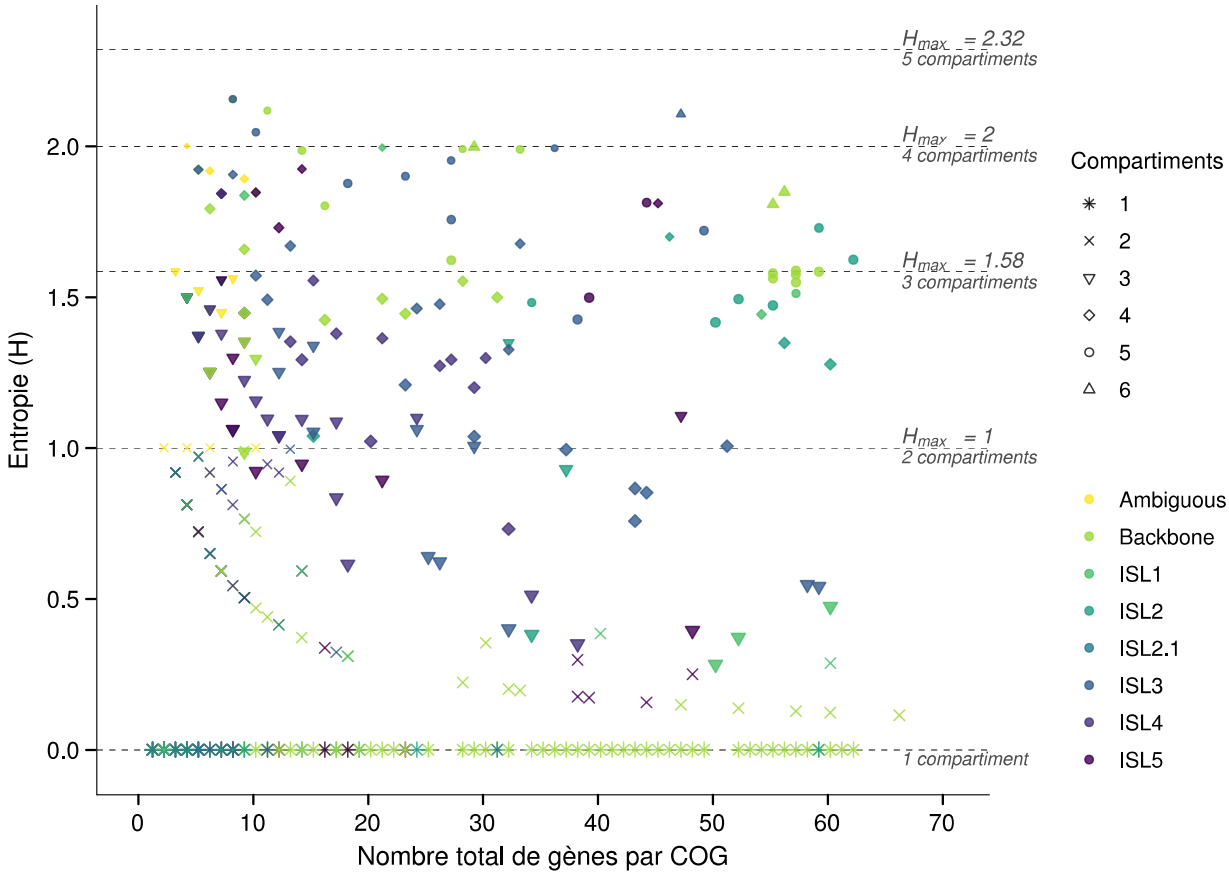


Figure 3.12. Valeurs d'entropie de Shannon (H) associées aux COGs flexibles non partagés avec MIT9312 en fonction du nombre de gènes par COG pour l'attribution d'un compartiment génomique. L'entropie est maximale lorsque les compartiments attribués aux gènes d'un même COG sont équiprobables. Elle est calculée selon la formule $H_{max} = \log_2(n)$, où n est égal au nombre de compartiments considérés. Les valeurs d'entropie maximales sont données selon le nombre de compartiments (*cf.* partie droite du graphique). Chaque COG est représenté par un symbole et une couleur. Les symboles représentent le nombre de compartiments différents retrouvés au sein d'un COG donné. Les couleurs correspondent au compartiment génomique attribué à l'échelle du COG.

L'entropie est maximale lorsque tous les compartiments associés à un COG sont équiprobables. Elle est par exemple égale à 1 si deux compartiments sont retrouvés en même proportion. L'assignation aux compartiments est considérée comme particulièrement robuste dans la mesure où moins de 8,5 % des COGs ont une entropie de Shannon supérieure ou égale à 1, alors que plus de 85 % ont une entropie égale à 0 – *i.e.,* un seul compartiment pour l'ensemble des gènes du COG.

Au final, les résultats obtenus montrent une assignation de 63,1 % des COGs analysés au *backbone* contre 31,6 % aux ISLs. Pour ces derniers, la distribution est comme suit : 8,2 % assignés à l'ISL3, 14,5 % à l'ISL4 alors que les derniers 8,9 % sont répartis entre ISL1, ISL2, ISL2.1 et ISL5. Pour finir, les COGs non assignés à un compartiment génomique (5,3%) du fait de l'équiprobabilité des attributions, ont été qualifiés d'ambigus (Tableau 3.4).

Tableau 3.4. Nombre et pourcentage de COGs assignés aux différents compartiments génomiques sur la base de l'estimation de leur entropie de Shannon (H).

Compartiment	Nombre de COGs	COGs (%)
Backbone	3 339	63
ISL1	72	1
ISL2	126	2
ISL2.1	39	1
ISL3	435	8
ISL4	766	14
ISL5	231	4
Ambigu	282	5

3.4.2. Distribution des COGs *core* et flexibles le long du génome

Les distributions des différentes catégories de COGs le long du génome (*i.e., core* et flexibles partagés et non partagés par le génome de référence MIT9312) ont été comparées (Figure 3.13A). Dans un premier temps, pour les COGs flexibles SAG-spécifiques, seuls ceux présents *a minima* dans deux sous-populations (catégories 2 à 7 ; 1 626 COGs au total ; Figure 3.8) ont été considérés pour éviter que le signal ne soit saturé par l'information apportée par ceux présents dans une unique sous-population (catégorie 1 ; 3 547 COGs).

Les distributions montrent une différence significative entre les compartiments (*backbone versus* ISLs ; $p<0{,}001$, test du χ^2). En effet, une plus forte densité de COGs *core* (comprise entre 0,38 et 0,54) est observée dans le *backbone* comparée aux ISLs (comprise entre 0,03 et 0,40), ce qui était attendu du fait de la définition même des compartiments *backbone* et ISLs (Kashtan *et al.*, 2014). Les COGs flexibles partagés par MIT9312, en revanche, sont principalement retrouvés dans les ISLs (densités de 0,21 à 0,45), exception faite pour l'ISL2.1 où leur densité est inférieure à 0,10. Les COGs flexibles SAG-spécifiques sont, quant à eux, principalement enrichis dans l'ISL3, l'ISL4 et l'ISL5 avec des densités comprises entre 0,30 et 0,76 (Figure 3.13B). La part des COGs flexibles partagés par MIT9312 illustre une échelle de diversité qui va au-delà des sous-populations de l'écotype HLII étudiées ici et laisse sous-entendre l'idée que cette flexibilité peut être partagée à une échelle de diversité large au sein de cet écotype. *A contrario*, l'enrichissement de l'ISL3, l'ISL4 et de l'ISL5 en COGs flexibles SAG-spécifiques laisse supposer une plus grande plasticité de ces îlots et de la région génomique qui les porte, par rapport à la partie du génome faisant suite à l'origine de réplication.

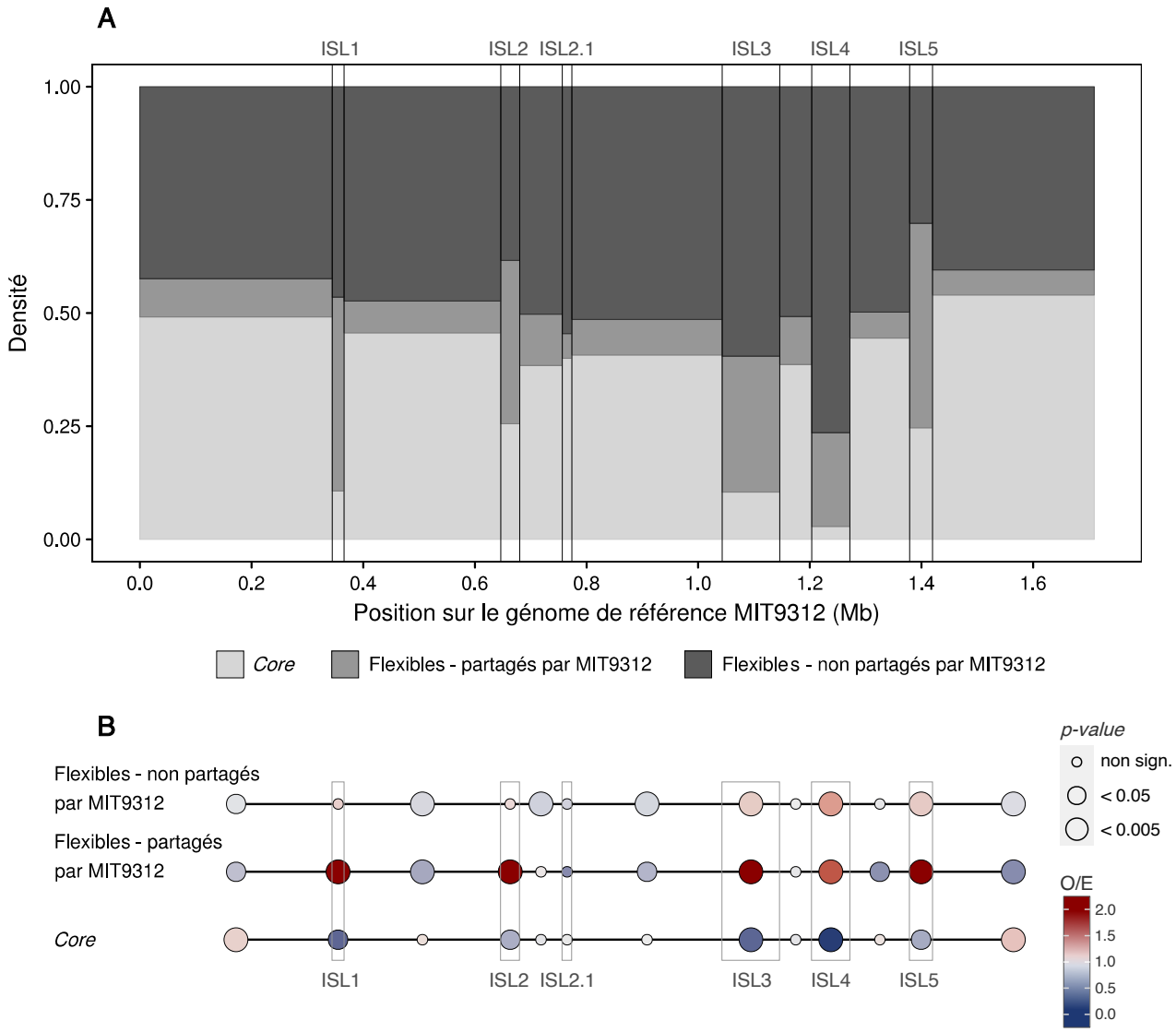


Figure 3.13. Densité de distribution des COGs le long du génome. (A) Densité de distribution des COGs exprimée pour chaque compartiment (*core versus* flexibles partagés *versus* non partagés par MIT9312). Concernant les COGs flexibles SAG-spécifiques, seuls ceux présents *a minima* dans deux sous-populations sont considérés. Le point 0.0 sur le graphique représente l'origine de réplication. **(B)** Densité de distribution 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* – flexible ; *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 associé.

La détermination de la distribution des COGs flexibles SAG-spécifiques au sein des clades (catégories 1 à 7) a mis en évidence une diversité inter-populationnelle notable (Figure 3.8). L'hypothèse selon laquelle cette distribution pourrait être associée à la répartition de ces COGs le long du génome a donc également été testée (Figure 3.14).

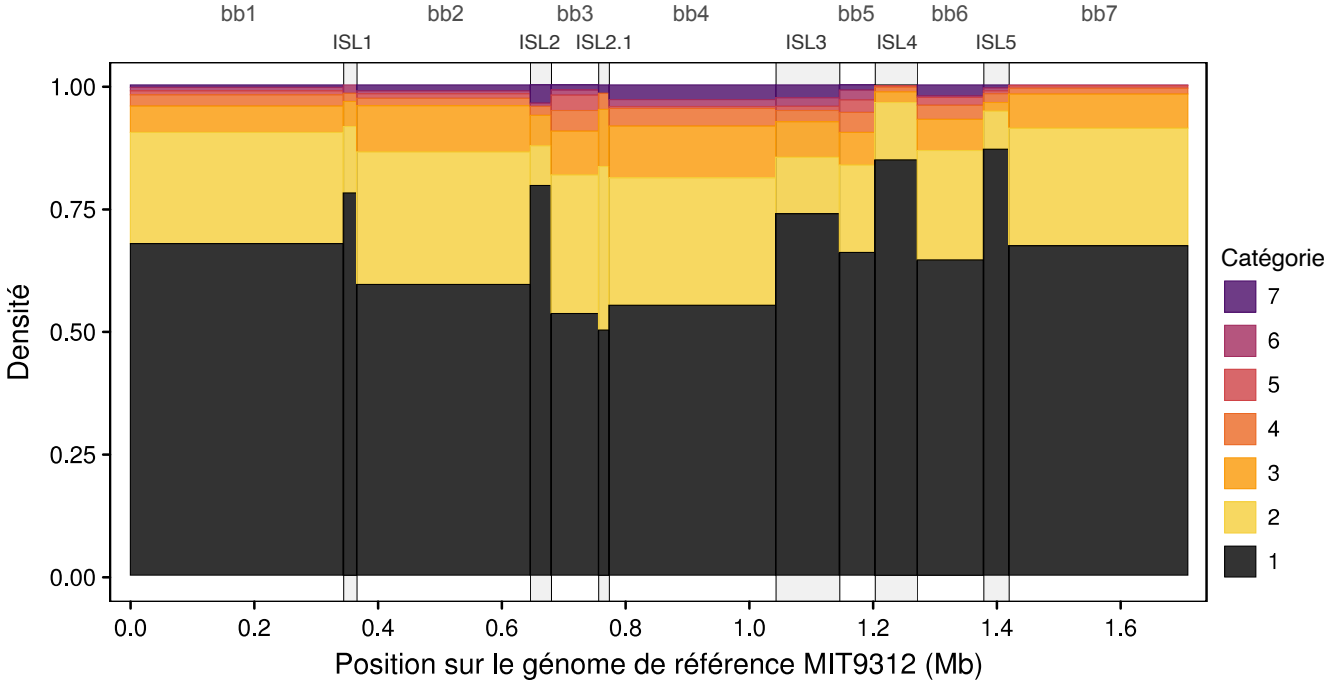


Figure 3.14. Densité de distribution des COGs flexibles SAG-spécifiques le long du génome en fonction de leur représentativité au sein des sous-populations (catégories 1 à 7, *i.e.*, COGs spécifiques d'une sous-population en noir à communs à toutes en violet).

Les résultats obtenus montrent des différences significatives de la répartition des COGs SAG-spécifiques le long du génome en fonction de leur distribution au sein des sous-populations et des compartiments génomiques (*p-value* <0,05, test du χ^2). Ainsi, les COGs composant la catégorie 1 sont principalement enrichis dans les ISL2, ISL3, ISL4 et ISL5 alors qu'ils sont appauvris dans le *backbone* (essentiellement la région située entre l'ISL1 et l'ISL2 ainsi que celle comprise entre l'ISL2 et l'ISL3, incluant l'ISL2.1). Les COGs de la catégorie 2 présentent des profils d'enrichissement inversés dans ces mêmes régions. Enfin les COGs de la catégorie 3 sont appauvris dans les ISL4 et ISL5 et enrichis dans les région du *backbone* situées entre ISL1 et ISL2 et ISL2.1 et ISL3.

Les catégories 4 à 7 présentent un enrichissement au niveau du *backbone* entre ISL2 et ISL4, ainsi que dans ISL3. Les COGs de ces catégories sont appauvris au niveau de ISL4

et de la région du *backbone* située après ISL5. Ces résultats suggèrent que la plus grande plasticité supposée des ISLs du fait des COGs flexibles SAG-spécifiques pourrait essentiellement être supportée par ceux de la catégorie 1.

3.5. Bilan

Ces dernières années, l'essor des techniques de NGS et de génomique « cellule-unique » a élargi le champ des possibles dans la recherche que ce soit en écologie, en sciences de l'environnement ou en évolution. Ceci a notamment permis d'accéder à une fraction encore peu connue de la diversité des micro-organismes de l'environnement, celle de la population. En effet, il est désormais possible d'accéder aux informations génétiques portées par des micro-organismes issus d'un même environnement et isolés individuellement. Cependant, malgré la levée de verrous technologiques (*i.e.*, extraction et séquençage du génome d'un unique micro-organisme), les génomes obtenus par la méthode SCG, *i.e.*, les SAGs, sont pour la plupart partiels et donc potentiellement inadaptés pour l'étude des pangénomes. Ainsi, ce chapitre a pour objectif de décrire et d'évaluer la pertinence du jeu de données dans le contexte des questions posées par cette thèse.

Les différentes phylogénies réalisées au cours de ce travail sur un jeu de données composé de 87 SAGs (dont 18 de complétude > à 90%) montrent une même structuration de la population de *Prochlorococcus,* écotype HLII, en sept clades. Cependant, bien que l'organisation génomique de ces clades soit conservée, ceux-ci se distinguent les uns des autres du fait, notamment, de leur contenu en gènes flexibles. En effet, une augmentation du nombre de SAGs se traduit par un accroissement du nombre de COGs, pour la plupart, SAG-spécifiques. Ceci est caractéristique d'un pangénome ouvert, en accord avec la nature de celui de *Prochlorococcus* tel que décrit à des échelles de diversité plus large (Kettler *et al.*, 2007; Biller *et al.*, 2014b; Delmont and Eren, 2018). De plus, il apparaît qu'une réduction du jeu de données aux seuls SAGs de forte complétude se traduit par l'apparition d'une proportion non négligeable de COGs SAG-spécifiques faux positifs. Ainsi, des COGs, partagés par plusieurs SAGs dans le jeu de données complet, voient leur distribution au sein des clades réduite allant, pour certains, jusqu'à une assignation à un unique clade (COGs de la catégorie 1). La conservation de l'information génétique portée par l'ensemble des SAGs, quelle que soit leur complétude, est donc essentielle afin d'estimer au mieux la part de la fraction la moins partagée du pangénome, d'autant qu'elle représente une part importante de ce dernier chez les sous-populations de *Prochlorococcus* étudiées. En effet, il se compose de près de 20 % de COGs *core*, et pour le génome

flexible de 32 % de COGs appartenant aux catégories 2 à 7 et de 48 % de COGs de la catégorie 1.

Si le pangénome reflète de fait la plasticité et le potentiel génétique des individus d'une population, celui-ci permet par ailleurs de caractériser un paysage génomique incluant des régions conservées (*backbone*) et variables (ISLs) le long du génome. D'un point de vue structurel, les approches développées dans ce chapitre confirment que les SAGs étudiés ont une organisation de leur génome qui est similaire à celle de la souche de référence MIT9312, à savoir contenant six ISLs dispersés au sein du *backbone* (*cf*. Figure 3.4). Comme attendu, leur génome flexible est essentiellement localisé au niveau des ISLs, en particulier ISL2, ISL3, ISL4 et ISL5 qui sont par ailleurs majoritairement constitués de COGs appartenant à la catégorie 1. En revanche, lorsque les COGs flexibles sont portés par au moins deux populations, ceux-ci sont préférentiellement retrouvés dans le *backbone* (*cf*. Figure 3.14).

On peut s'interroger dès lors sur les raisons d'une telle organisation le long du génome des COGs plus ou moins partagés par les différentes populations, tant d'un point de vue fonctionnel qu'évolutif. En effet, il est admis que les gènes flexibles portés par les ISLs sont issus de transferts horizontaux et confèrent un caractère adaptatif de par la fonction qu'ils codent. Ainsi, ces gènes devraient être soumis à des pressions de sélection dont il convient de préciser la nature. Ces différents éléments sont abordés dans le Chapitre 4.

4. Évaluation de la dynamique évolutive différentielle le long des compartiments génomiques

Les génomes bactériens sont majoritairement constitués de gènes flexibles (Lapierre and Gogarten, 2009) parmi lesquels certains permettent aux bactéries de s'adapter à des niches environnementales spécifiques. Ainsi, des îlots de gènes accessoires peuvent, par exemple, conférer une pathogénicité ou une résistance aux antibiotiques (Dobrindt *et al.*, 2004; Schmidt and Hensel, 2004). Pour *Prochlorococcus*, plus particulièrement l'écotype HL, des gènes codant des fonctions liées à des stress physiologiques et à l'assimilation de nutriments ont été mis en évidence au niveau des ISLs (Coleman *et al.*, 2006; Coleman and Chisholm, 2010). Des gènes non conservés conférant une résistance à différents virus ont également été identifiés au sein des ISL3 et ISL4, par exemple (Avrani *et al.*, 2011). Il est couramment admis que ces gènes flexibles sont principalement acquis par des HGTs (Lindell *et al.*, 2004; Coleman *et al.*, 2006) qui, par ailleurs, sont facilités par une grande diversité de micro-organismes partageant un même espace.

Afin de décrypter les fondements de la différenciation des sous-populations mise en évidence dans le chapitre précédent, des caractérisations taxonomiques, fonctionnelles et évolutives des COGs flexibles ont été réalisées. Ces analyses ont porté, d'une part, sur la comparaison des sous populations entre elles et, d'autre part, sur celle des compartiments génomiques.

4.1. Compartimentations taxonomique et fonctionnelle le long des génomes

4.1.1. Analyses taxonomiques et origines phylogénétiques des COGs

Afin d'évaluer l'origine phylogénétique des COGs, ainsi que leur intégrité (*i.e.,* l'homogénéité de l'affiliation à l'échelle des gènes), des analyses taxonomiques ont été réalisées. Celles-ci ont révélé que 97,4 % des COGs appartenant au génome *core* sont affiliés à l'écotype HLII, bien que 2,9 % d'entre eux contiennent des gènes affiliés à d'autres écotypes. Les 2,6 % restants, non affiliés à HLII, sont considérés comme incertains car contenant des gènes affiliés à divers groupes taxonomiques, dont *Prochlorococcus* et/ou *Synechococcus* (Figure 4.1).

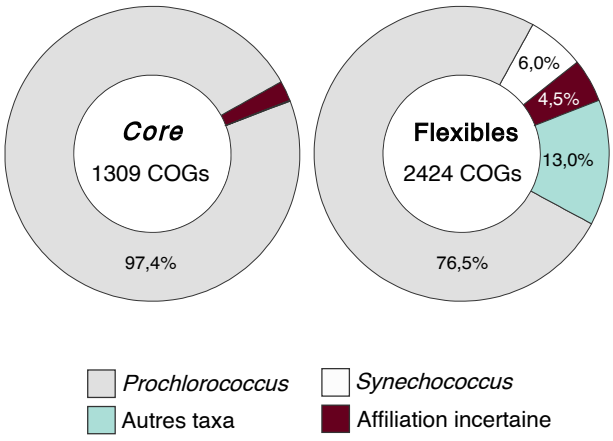


Figure 4.1. Affiliations taxonomiques des COGs *core* et flexibles. Quatre catégories de COGs ont été mises en évidence : ceux contenant des gènes majoritairement affiliés à (i) *Prochlorococcus*, (ii) *Synechococcus*, (iii) ceux comprenant des gènes attribués à divers groupes taxonomiques, incluant *Prochlorococcus* et/ou *Synechococcus* (affiliation incertaine) et (iv) ceux dont les gènes sont affiliés à des taxa autres que *Prochlorococcus* et *Synechococcus* (autres taxa).

En revanche, l'origine phylogénétique des 5 715 COGs flexibles semble moins évidente puisque près de la moitié d'entre eux contiennent des gènes sans équivalent dans la base de données EggNOG. Concernant les COGs ayant une affiliation taxonomique (*i.e.,* 2 424 au total), 76,5% sont affiliés à *Prochlorococcus*, avec une très grande majorité

à l'écotype HLII (94,4%), suivi de l'écotype HLI (3,3%) et des écotypes LL (2,3%). Parmi les COGs restants, 6 % sont affiliés à *Synechococcus*, 13 % à des taxa bactériens autres que *Prochlorococcus* et *Synechococcus*, et 4,5 % sont considérés comme incertains (Figure 4.1). Lorsqu'ils sont affiliés à d'autres taxa, les COGs appartenant aux protéobactéries prédominent (51,6%), suivis des cyanobactéries (14,4%), du groupe des bactéroïdetes / chlorobi (10,9%), des firmicutes (5,3%), des actinobactéries (2,3%) et des spirochètes (2,3%) (Figure 4.2A).

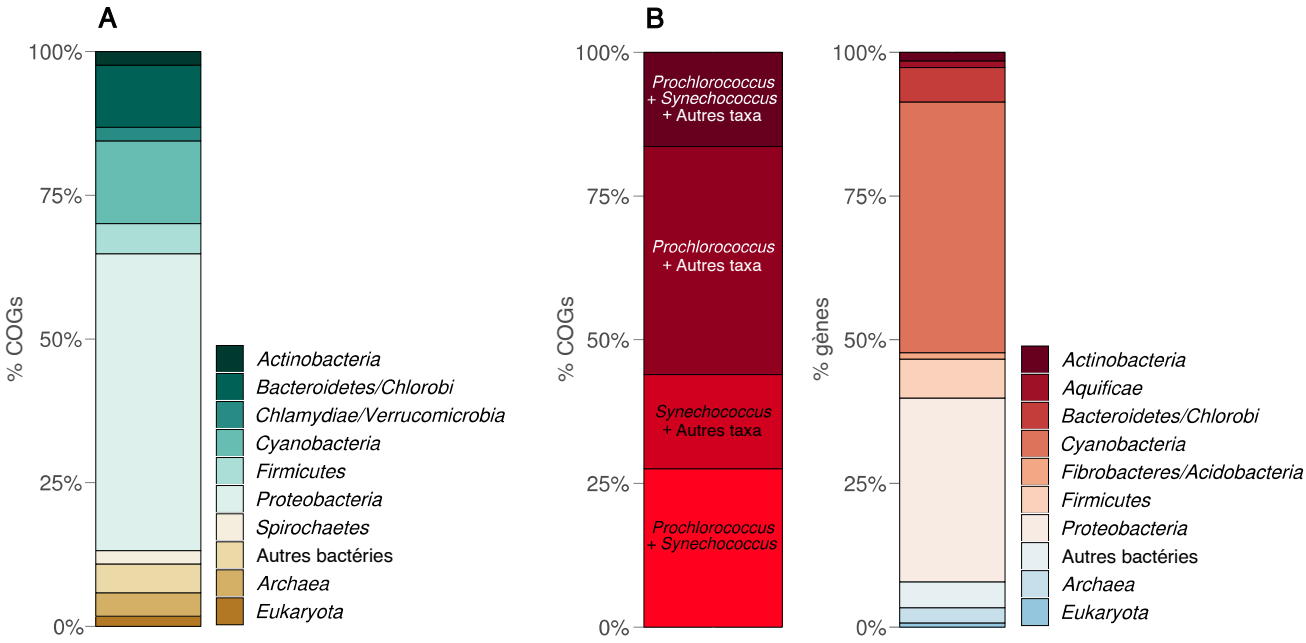


Figure 4.2. Distributions taxonomiques observées pour les COGs flexibles non affiliés à *Prochlorococcus* ou *Synechococcus*. (A) Distributions taxonomiques des COGs flexibles contenant des gènes affiliés à des taxa autres que *Prochlorococcus* et *Synechococcus*. **(B)** Distributions taxonomiques des COGs dont l'affiliation est incertaine. Pour ces COGs, sont indiquées, d'une part, les distributions en fonction de la présence de gènes affiliés à *Prochlorococcus* et/ou *Synechococcus* (*cf.* partie gauche) et d'autre part, la composition et l'abondance des gènes au sein des autres taxa quand mis en évidence (*cf.* partie droite). La catégorie « Autres bactéries » regroupe les taxa bactériens dont l'abondance est inférieure à 1 %.

Les COGs qualifiés d'incertains contiennent des gènes affiliés à *Prochlorococcus* (39,7 %), *Synechococcus* (16,4 %) voire à ces deux genres, associés (16,4%) ou non à d'autres taxa (27,5%). Les taxa autres que *Prochlorococcus* et *Synechococcus* sont principalement

des cyanobactéries, suivies des protéobactéries, des bactéroïdetes / chlorobi et des firmicutes (Figure 4.2B).

4.1.2. Enrichissements fonctionnels

Nous avons vu précédemment qu'il existait, d'une part, une diversité de COGs inter-populationnelle, et d'autre part, une répartition différentielle des COGs – *core versus* flexibles – dans les compartiments génomiques (*cf.* Chapitre 3). Ceci pourrait impliquer une spécialisation fonctionnelle à l'échelle des sous-populations, voire même des compartiments.

4.1.2.1. Spécialisation fonctionnelle des sous-populations

Le potentiel fonctionnel des COGs a, dans un premier temps, été évalué pour les sous-populations. Seuls les gènes ayant un homologue dans la base de données EggNOG et une fonction connue, ont été pris en compte, soit 47 % de l'ensemble des gènes *core* et flexibles. Excepté une sur-représentation de la catégorie fonctionnelle associée à la motilité cellulaire dans les clades C1 et C9, aucune différence dans la distribution des catégories fonctionnelles à l'échelle de la sous-population n'a été observée (p=0,39, test du χ^2) (Figure 4.3), appuyant l'hypothèse de travail selon laquelle la différentiation des sous-populations ne résulterait pas d'une dynamique évolutive propre à chacune d'elles. Par conséquent, les sous-populations ont été groupées lors de l'analyse des enrichissements fonctionnels à l'échelle des compartiments génomiques.

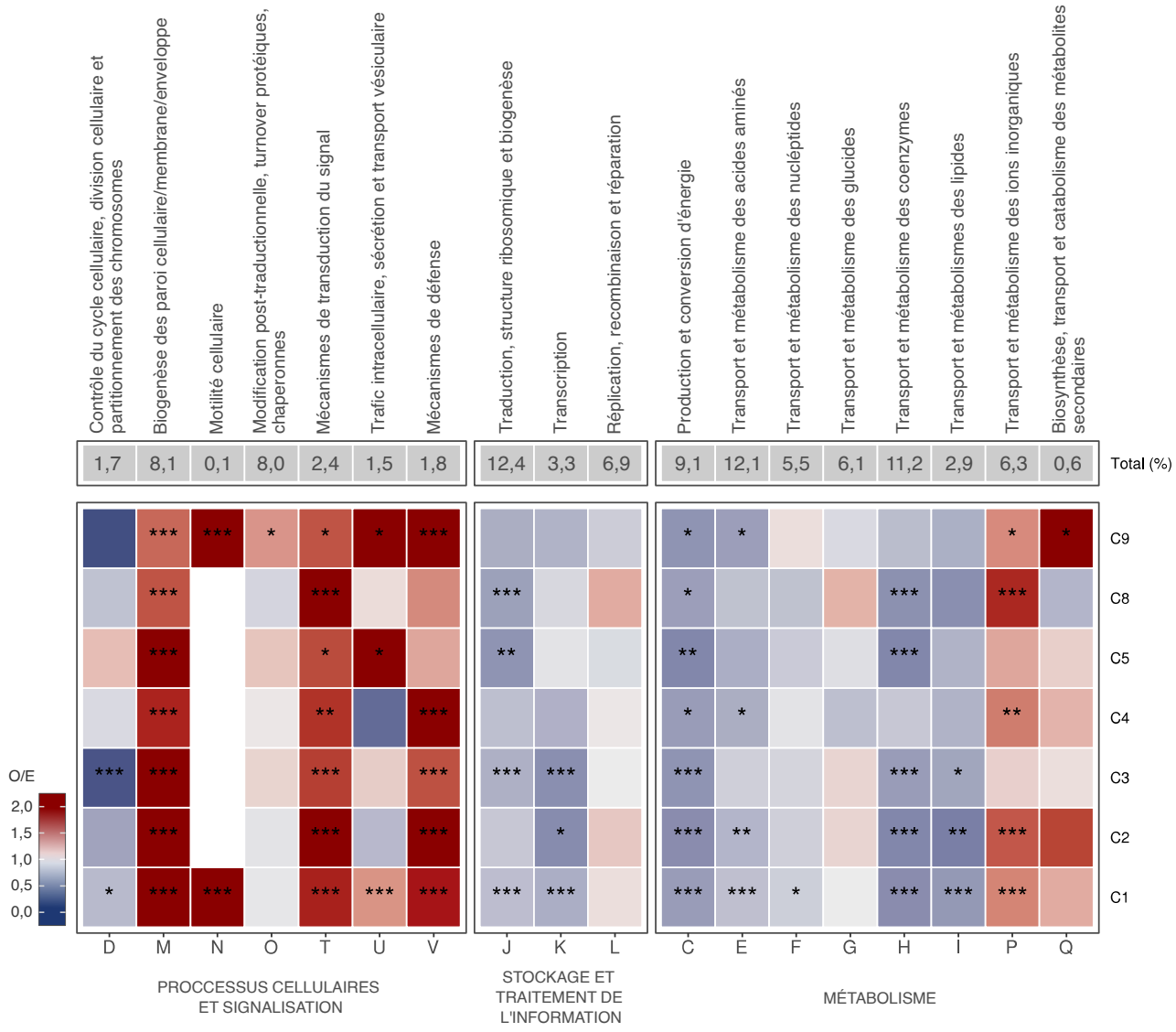


Figure 4.3. Représentation des enrichissements fonctionnels à l'échelle des sous-populations (C1 à C5, C8 et C9). Les catégories fonctionnelles ont été définies par comparaison des gènes à la base de données EggNOG. Les enrichissements sont illustrés par les rapports observés/théoriques (O/E) pour chaque catégorie fonctionnelle symbolisée par une lettre majuscule. Les pourcentages des gènes totaux attribués à chaque catégorie fonctionnelle sont indiqués sous les entêtes des catégories fonctionnelles correspondantes. Les valeurs observées correspondent au nombre de gènes attribués à chaque catégorie fonctionnelle dans chaque ensemble de données. Les valeurs théoriques (E) ont été obtenues en multipliant le nombre de gènes (fonction de la sous-population considérée) par le pourcentage de gènes de chaque catégorie fonctionnelle. Les différences de distribution des catégories fonctionnelles sur l'ensemble des sous-populations ont été testées par un test du χ^2 (p=0,39). 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 une sous-population donnée. Test du χ^2 : *, p-$value < 0,05$; **, p-$value < 0,01$; ***, p-$value < 0,001$.

4.1.2.2. Spécialisation fonctionnelle des compartiments

Les enrichissements fonctionnels ont été évalués en fonction du type de gènes (*i.e., core* et flexibles), de la localisation génomique (*i.e., backbone*, ISL et ambigu) et des affiliations taxonomiques (Figure 4.4). La comparaison de la distribution des catégories fonctionnelles entre les gènes *core* et les gènes flexibles a montré une sous-représentation de la quasi-totalité des catégories fonctionnelles associées aux hiérarchies « Métabolisme » (catégories J, K et L) et « Stockage et traitement de l'information » (catégories C, E, F, G, H, I, P et Q) au sein des gènes flexibles (Figure 4.4A). Les principales fonctions touchées sont celles liées aux mécanismes de transcription et traduction, la production d'énergie mais aussi le transport et le métabolisme des nucléotides, des acides aminés, des coenzymes et des lipides. Les gènes associés à ces catégories fonctionnelles sont majoritairement portés par le génome *core*, ce qui peut sans doute s'expliquer par le fait qu'elles regroupent principalement des gènes de ménage, donc essentiels au bon fonctionnement de toute cellule. C'est une raison également qui pourrait expliquer pourquoi, lorsqu'elles sont associées à des gènes flexibles, ceux-ci sont préférentiellement associés au *backbone* (Figure 4.4B). En effet, le *backbone* est considéré comme le squelette génomique, tandis que les ISLs sont des régions hypervariables connues pour porter de nombreux gènes du métabolisme secondaire (Rocap *et al.*, 2003; Coleman *et al.*, 2006; Martiny *et al.*, 2006; Kettler *et al.*, 2007; Avrani *et al.*, 2011).

Parallèlement, quelques catégories fonctionnelles montrent un enrichissement dans les gènes flexibles du fait de leur présence surnuméraire dans certains ISLs (Figure 4.4B). Deux catégories sont principalement concernées avec, d'une part, celle correspondant à la biosynthèse, au transport et au catabolisme des métabolites secondaires (Q ; Figure 4.4) qui est enrichie dans l'ISL3, l'ISL5 et le compartiment défini comme ambigu, et d'autre part, celle du transport et métabolisme des ions inorganiques (P ; Figure 4.4) enrichie dans l'ISL2 et l'ISL3. La première catégorie se caractérise par des gènes majoritairement annotés comme méthyltransférases, connues pour leur potentielle implication dans la réparation de l'ADN. En ce qui concerne la seconde catégorie, les gènes identifiés codent principalement des transporteurs de phosphates organiques et inorganiques. Au niveau taxonomique, les séquences les plus proches des gènes sur-représentés sont principalement affiliés aux protéobactéries et aux *Archaea* (Figure 4.4C). Il est également intéressant de noter que, bien que les gènes associés au transport et au métabolisme des glucides (G ; Fi-

gure 4.4) aient un rapport O/E proche de 1, leur enrichissement est extrêmement variable d'un compartiment à l'autre, avec un appauvrissement dans les ISLs et une sur-représentation dans le compartiment ambigu. Les gènes concernés codent des protéines telles que des transcétolases et des transaldolases, protéines jouant notamment un rôle dans le cycle de Calvin (Raines, 2003) (Figure 4.4B).

Pour la troisième catégorie hiérarchique, *i.e.,* « Processus cellulaires et signalisation », un enrichissement des gènes flexibles est observé dans la plupart des catégories fonctionnelles, à l'exception de celle impliquée dans le contrôle du cycle cellulaire, de la division cellulaire et du partitionnement des chromosomes (D ; Figure 4.4A). Ces enrichissements sont plus marqués au niveau de l'ISL3 et du *backbone*. Au niveau du *backbone*, cela concerne essentiellement les mécanismes de transduction du signal, avec des gènes codant des protéines telles que des histidines kinases impliquées dans la réponse au stress nutritionnel, et majoritairement affiliés à *Prochlorococcus* (T ; Figure 4.4C).

Deux catégories fonctionnelles, correspondant à la biogenèse de la paroi cellulaire (M) et aux mécanismes de défense (V) et sur-représentées au sein des gènes flexibles, ont également la particularité d'être associées à des COGs spécifiques à une unique sous-population (Figure 4.4D). Les gènes concernés par cet enrichissement codent principalement, pour la catégorie M, des glycosyltransférases et GDP-mannose 4,6-déshydratase, connues pour leur rôle dans la biosynthèse de lipopolysaccharides de la membrane externe et pour la catégorie V, des endonucléases et des transporteurs. La majorité des gènes constituant ces COGs (95,33%) sont localisés au niveau de l'ISL3, l'ISL4 et du compartiment ambigu. Ils sont, de plus, affiliés à une grande diversité de taxa (Figure 4.4C).

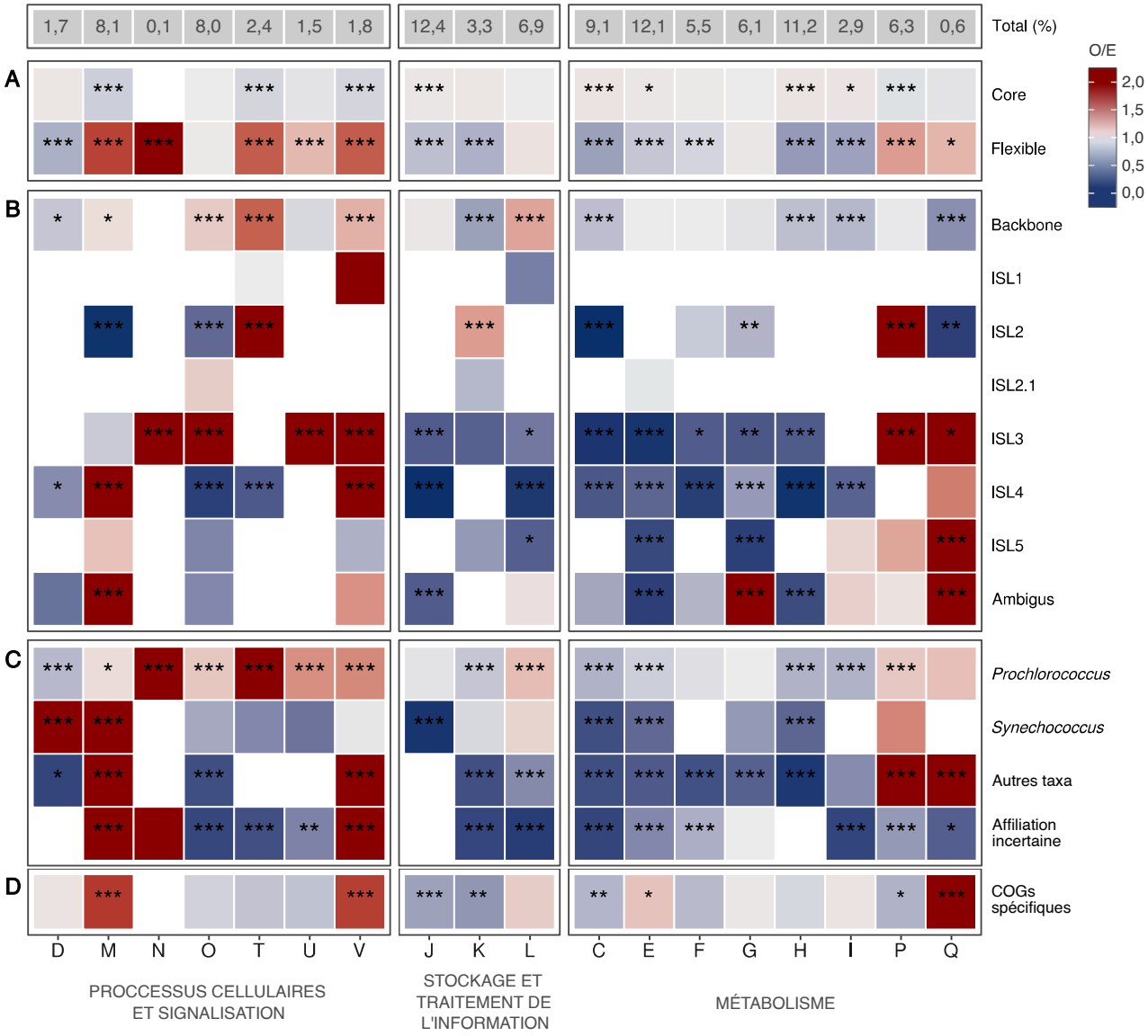


Figure 4.4. Représentation des enrichissements fonctionnels des COGs *core* et flexibles, obtenus par comparaison à la base de données EggNOG, selon leur catégorie (A), leur position sur le génome (B), leur affiliation (C), et leur appartenance à une unique sous-population (D). Les enrichissements sont illustrés par les rapports observés/théoriques (O/E) pour chaque catégorie fonctionnelle de la base de données EggNOG. Les catégories fonctionnelles, symbolisées par une lettre majuscule, sont telles que décrites dans la Figure 4.3. Les pourcentages des gènes totaux attribués à chaque catégorie fonctionnelle sont indiqués. Les valeurs observées correspondent au nombre de gènes attribués à chaque catégorie fonctionnelle dans chaque ensemble de données. Les valeurs théoriques (E) ont été obtenues en multipliant le nombre de gènes (*core* ou flexibles fonction de leur localisation génomique, de leur affiliation taxonomique ou de leur appartenance à une unique sous-population) par le pourcentage total de gènes de chaque catégorie fonctionnelle. Les différences de distribution des catégories fonctionnelles entre gènes *core* et flexibles **(A)**, entre

compartiments génomiques **(B)** et affiliations taxonomiques **(C)** ont été testées par des tests du 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* $< 0{,}001$.

4.2. Processus d'évolution différentielle le long des génomes

4.2.1. Hétérogénéité des pressions de sélection entre compartiments génomiques

Sur la base d'une l'analyse des indices de fixation (F_{ST}) comparant les fréquences alléliques des gènes *core* et flexibles entre sous-populations, Kashtan *et al.* (2014) ont suggérés que différents allèles pouvaient être fixés au sein même de différentes sous-populations, conduisant, *in fine*, à la différenciation de ces sous-populations. Pour identifier les processus évolutifs à l'origine de cette différenciation, l'approche développée dans cette étude s'appuie sur le calcul des rapports de *dN/dS* pour les gènes constituant les différents COGs.

4.2.1.1. Évaluation des biais associés aux estimations des ratios *dN/dS*

Le calcul des ratios *dN/dS* pour caractériser les pressions de sélection à l'échelle des gènes est une approche reconnue, mais dont la fiabilité est limitée lorsque les distances évolutives entre les gènes comparés sont faibles (Rocha *et al.*, 2006; Wolf *et al.*, 2009; dos Reis and Yang, 2013), ou lorsque les populations comparées ne sont pas différenciées (Kryazhimskiy and Plotkin, 2008). La comparaison des taux de substitutions synonymes (*dS*) entre SAGs au sein de chaque clade ou entre clades permet dans un premier temps d'évaluer ce possible biais dans notre jeu de données. Les valeurs de *dS* calculées varient de 0,02 ± σ 0,08 pour les comparaisons intra-clades à 0,12 ± σ 0,11 pour les comparaisons inter-clades (Figure 4.5). Ainsi, les très faibles valeurs de *dS* calculées au sein des clades pourraient induire des biais dans l'estimation des ratios *dN/dS*. C'est pourquoi seuls les gènes présents *a minima* dans deux clades (excluant toutes comparaisons intra-clades) et pour lesquels les valeurs de *dS* permettent une estimation fiable de ces ratios ont été retenus dans cette étude. Par ailleurs, les faibles valeurs de *dS* obtenues pour les comparaisons des clades C1 et C2, de l'ordre de celles des comparaisons intra-clades (0,05 ± σ 0,07) peuvent également introduire un biais dans l'estimation des *dN/dS* (Kryazhimskiy and Plotkin, 2008). De même, les comparaisons C1 – C2 n'ont pas été prises en compte.

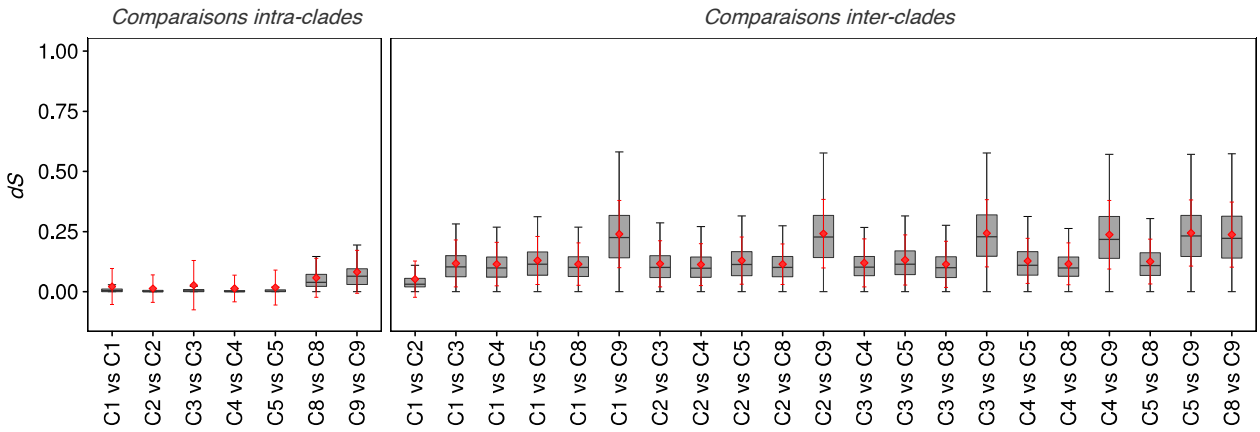


Figure 4.5. Distributions des valeurs de *dS* pour les comparaisons intra-clades (*cf.* partie gauche) et inter-clades (*cf.* partie droite). Pour une meilleure lisibilité du graphique, les *outliers* ne sont pas représentés ; moyennes et écart-types apparaissent en rouge.

Dans la mesure où l'estimation des ratios *dN/dS* dépend des temps de divergence séparant les gènes comparés, il apparaît pertinent d'envisager d'abord ces estimations par rapport à un génome de référence, extérieur aux sous-populations considérées et équidistant de tous les SAGs étudiés. Le génome de référence MIT9312 est à égale distance phylogénétique de tous les clades analysés dans cette étude (*cf.* Chapitre 2), ce qui implique une « homogénéité de temps » dans l'estimation des ratios de *dN/dS* entre MIT9312 et chacun des clades. D'un autre côté, s'affranchir d'un génome de référence permet d'inclure l'ensemble des COGs spécifiques des sous-populations, *i.e.*, ceux absents du génome de MIT9312. Cependant, comme cela a été précisé précédemment, les rapports *dN/dS* peuvent être largement influencés par les valeurs de *dS*, en particulier lorsque l'on compare des groupes taxonomiques très proches phylogénétiquement, tels que deux sous-populations d'un même écotype. Dans cet esprit, il est nécessaire de s'assurer que les ratios *dN/dS* calculés entre clades sont équivalents à ceux calculés entre MIT9312 et les différents clades. Ainsi, les comparaisons des rapports de *dN/dS* ont été effectuées à différentes échelles (MIT9312 *versus* clades et clades *versus* clades) pour les COGs localisés dans le *backbone*, qu'ils soient *core* (Figure 4.6) ou flexibles (Figure 4.7). Les ratios de *dN/dS* pour les 1 139 COGs *core* issus des comparaisons MIT9312 *versus* clades ne sont pas significativement différents de ceux issus des comparaisons clades *versus* clades (*dN/dS* moyen $\pm\sigma$ = 0,21 $\pm$0,14 et 0,22 $\pm$0,19, respectivement ; p=0,35, test de Wilcoxon). Par ailleurs, la grande similitude entre les graphiques de densité (Figure 4.6A), associée à

une forte corrélation positive entre les valeurs des *dN/dS* moyens par COG (ρ=0,83, p<0,001, corrélation de Spearman) (Figure 4.6B) ne révèlent aucun biais dans l'évaluation des pressions de sélection par l'approche inter-clade.

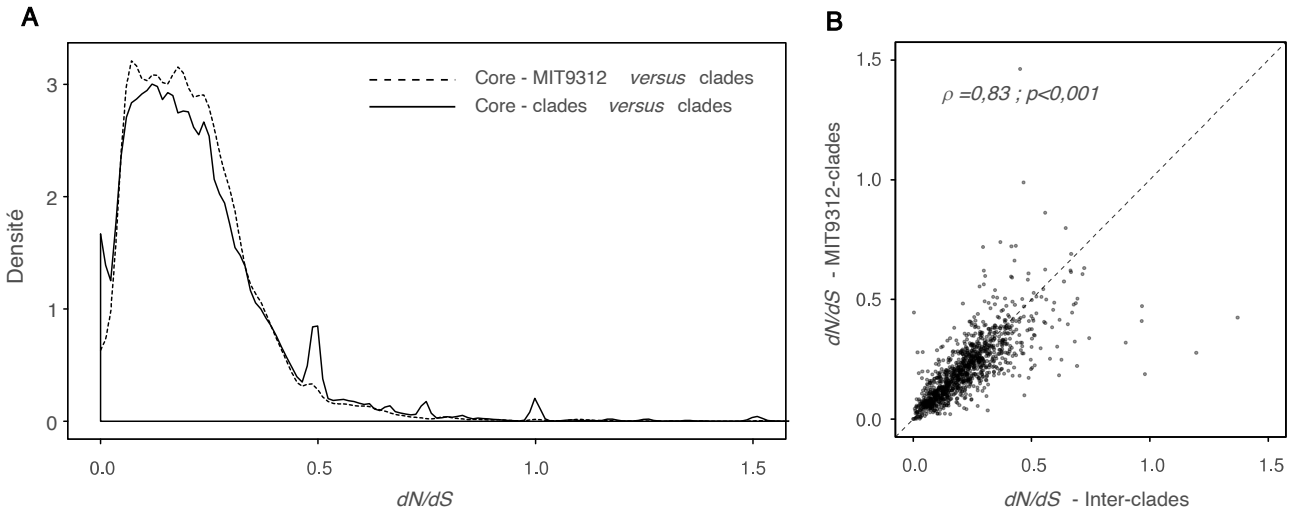


Figure 4.6. Comparaisons des rapports *dN/dS* pour les COGs *core* retrouvés dans le *backbone*. **(A)** Densité des rapports *dN/dS* obtenus par comparaison du génome de référence MIT9312 à chacun des clades (ligne pointillée) et par comparaison des clades deux à deux (ligne continue). **(B)** Corrélation entre les moyennes des ratios *dN/dS* estimées pour les comparaisons MIT9312-clades et inter-clades. Chaque point représente une valeur moyenne de *dN/dS* pour un COG donné. Le résultat du test de corrélation de Spearman (ρ) et la *p-value* associée sont indiqués. La ligne pointillée symbolise la diagonale.

Des résultats similaires ont été obtenus pour les distributions des ratios *dN/dS* entre les comparaisons MIT9312 *versus* clades et clades *versus* clades pour les 205 COGs flexibles retrouvés dans le *backbone* (Figure 4.7).

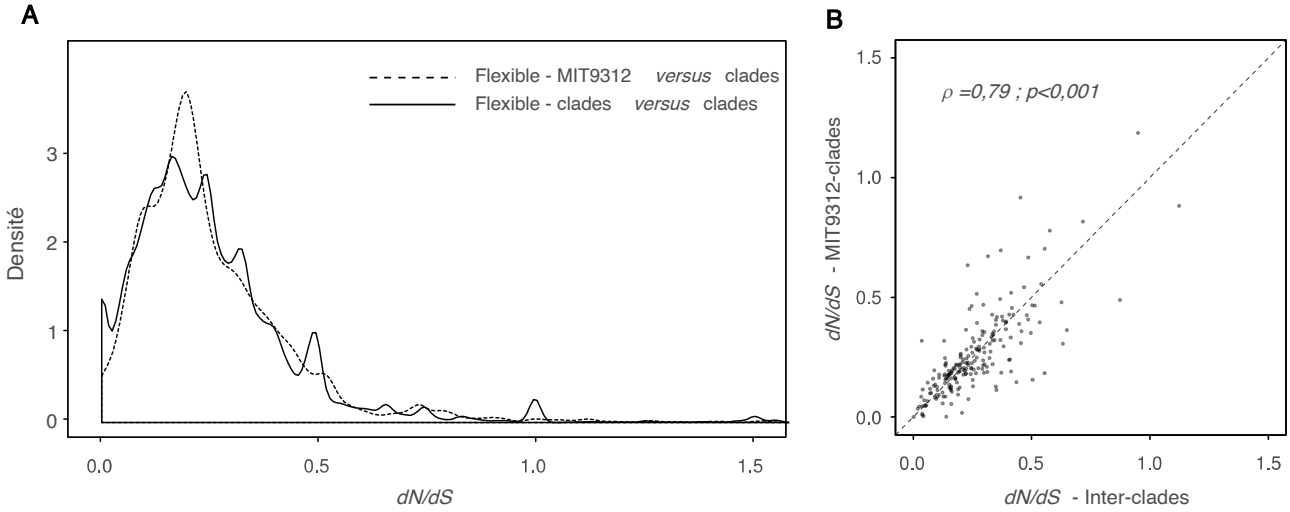


Figure 4.7. Comparaisons des rapports *dN/dS* pour les COGs flexibles retrouvés dans le *backbone*. **(A)** Densité des rapports *dN/dS* obtenus par comparaison du génome de référence MIT9312 à chacun des clades (ligne pointillée) et par comparaison des clades deux à deux (ligne continue). **(B)** Corrélation entre les moyennes des ratios *dN/dS* estimées pour les comparaisons MIT9312-clades et les comparaisons inter-clades. Chaque point représente une valeur moyenne de *dN/dS* pour un COG donné. Le résultat du test de corrélation de Spearman (ρ) et la *p-value* associée sont indiqués. La ligne pointillée symbolise la diagonale.

Les approches basées sur les comparaisons MIT9312 *versus* clades et clades *versus* clades donnant des résultats comparables en termes d'estimation des pressions de sélection, l'analyse des ratios *dN/dS* basée sur les comparaisons inter-clades a été privilégiée. Celle-ci permet de prendre en compte l'ensemble du génome flexible – notamment les COGs absents de MIT9312 – et de mener une étude plus exhaustive des trajectoires évolutives des ISLs, constitués majoritairement de COGs spécifiques des clades étudiés ici.

4.2.1.2. Des contraintes sélectives propres aux génomes *core* et flexible

Les estimations des contraintes sélectives réalisées à l'échelle des sous-populations montrent que quels que soient les clades comparés, génomes *core* et flexible confondus, les distributions des valeurs de *dN/dS* sont similaires (Figure 4.8). Ceci permet de faire l'hypothèse que, d'une part, le nombre de SAGs représentant chaque clade n'impacte pas ces estimations (*e.g.,* nombre de SAGs plus important pour le clade C1), et d'autre part, qu'il existe une homogénéité des contraintes sélectives à l'échelle des clades. Par conséquent, il n'y aurait pas de dynamique évolutive propre à chacune des sous-populations étudiées.

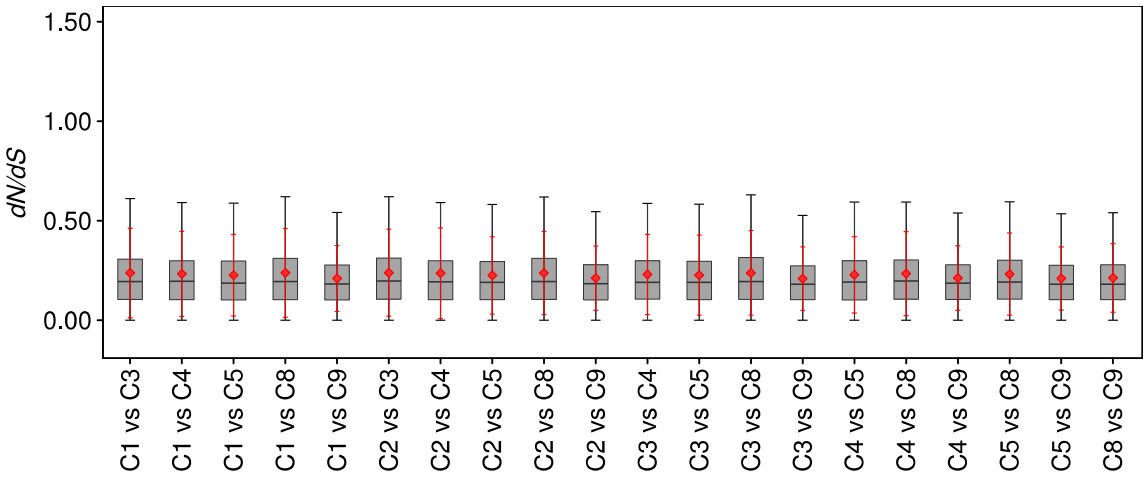


Figure 4.8. Distributions des valeurs de *dN/dS* issues des comparaisons inter-clades. Pour une meilleure lisibilité du graphique, les *outliers* ne sont pas représentés ; moyennes et écart-types apparaissent en rouge.

Les rapports de *dN/dS* sont majoritairement inférieurs à 1 (Figure 4.9), quelle que soit la nature des COGs considérés (*i.e., core* ou flexibles), traduisant des contraintes sélectives essentiellement négatives. Cependant, des différences significatives entre les *dN/dS* moyens sont observées en fonction de la nature des COGs, avec des valeurs plus faibles pour les COGs *core* (1 202 COGs) comparées à celles estimées pour les COGs flexibles partagés (310 COGs) ou non partagés (1 033 COGs) par MIT9312 (p<0,001, test de Kruskal-Wallis) (Figure 4.9). Les COGs *core* sont ainsi soumis à une plus forte pression de sélection négative.

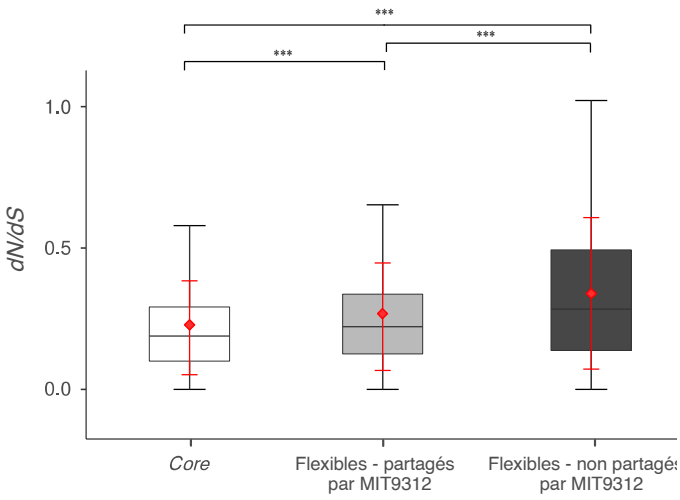


Figure 4.9. Distributions des valeurs de *dN/dS* estimées pour les COGs *core*, flexibles partagés ou non par le génome de référence MIT9312 (test de Kruskal-Wallis et test post-hoc des rangs signés de Wilcoxon avec correction de Bonferroni ; ***, *p-value* < 0,001). Seuls les ratios *dN/dS* > 1,5 sont représentés ; moyennes et écart-types sont indiqués en rouge.

4.2.1.3. Variations des contraintes sélectives à l'échelle des compartiments

Les estimations des contraintes sélectives le long des génomes montrent que des différences significatives s'observent à l'échelle des compartiments génomiques (p<0,001, test de Kruskal-Wallis) (Figure 4.10). De fortes contraintes sélectives, homogènes, sont observées pour les COGs *core* et flexibles partagés par MIT9312 dans le *backbone* (*dN/dS* ±σ moyens allant de 0,19 ±0,22 à 0,23 ±0,20 et de 0,21 ±0,19 à 0,29 ±0,26, respectivement). À l'inverse, ces contraintes sélectives sont variables pour les COGs flexibles non partagés par MIT9312 et inégalement réparties le long du *backbone*. À titre d'exemple, le ratio *dN/dS* moyen le plus faible (0,29 ±0,17) est observé dans la région située entre l'ISL2 et l'ISL2.1, alors que le plus élevé (0,73 ±0,82) concerne la région située entre l'ISL4 et l'ISL5 (Figure 4.11).

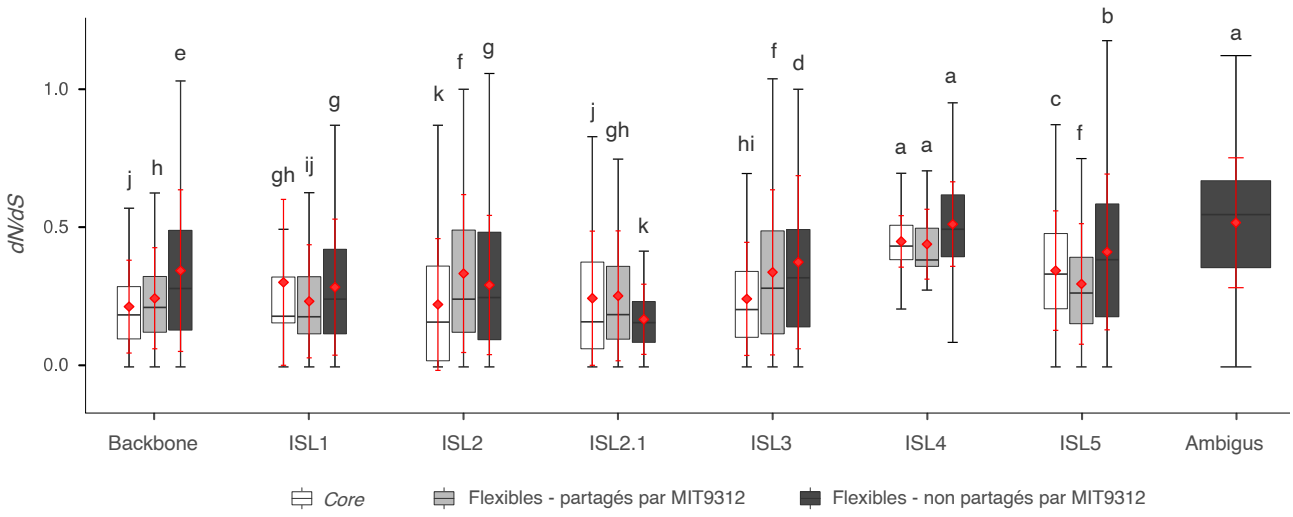


Figure 4.10. Distributions des valeurs de *dN/dS* estimées pour les COGs *core*, flexibles partagés ou non par le génome de référence MIT9312 en fonction des compartiments génomiques auxquels ils appartiennent (*backbone*, ISL, ambigus). Les différences significatives (catégorisant ainsi les compartiments ayant des patterns similaires) sont signalées par des lettres minuscules, avec a > b > c > d > e > f > g > h > i > j > k (test de Kruskal-Wallis et test post-hoc des rangs signés de Wilcoxon avec correction de Bonferroni, *p-value* < 0,05). Les *outliers* ne sont pas représentés ; moyennes et écart-types sont indiqués en rouge.

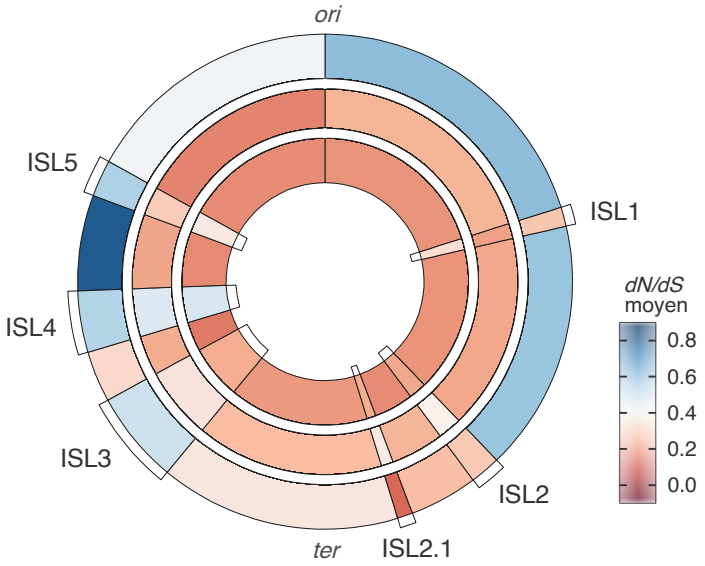


Figure 4.11. Représentation des ratios *dN/dS* moyens le long du chromosome (tel qu'organisé pour le génome de référence MIT9312). Le cercle interne représente les ratios estimés pour les COGs *core*, le cercle intermédiaire, les COGs flexibles partagés et le cercle externe, les COGs flexibles non partagés par MIT9312. L'origine (*ori*) et le terminus (*ter*) de réplication sont indiqués. ISL : îlots génomiques.

L'analyse des ISLs a, en outre, révélé des tendances contrastées en termes de contraintes sélectives, notamment entre l'ISL1, l'ISL2, et l'ISL2.1 d'une part, et l'ISL3, l'ISL4, l'ISL5 et ambigu d'autre part. Les pressions de sélection exercées sur l'ISL1, l'ISL2 et l'ISL2.1 sont globalement équivalentes à celles exercées sur les COGs *core* dans le *backbone*, *i.e.,* négatives, à l'exception des COGs flexibles de l'ISL2 partagés par MIT9312 (Figures 4.10 et 4.11). Quelques variations peuvent tout de même être notées. Le profil des COGs flexibles de l'ISL2.1 non partagés par MIT9312 (*dN/dS* moyen $\pm\sigma$ = 0,17 $\pm$0,13) montre une pression de sélection négative plus forte que celle observée dans le *backbone* (Figure 4.10). À l'inverse, les COGs *core* de l'ISL1 et flexibles partagés par MIT9312 de l'ISL2.1 présentent des contraintes sélectives légèrement moindres (Figure 4.10). Ces valeurs de *dN/dS* plus élevées (respectivement 0,35 $\pm$0,37 et 0,38 $\pm$0,62 en moyenne $\pm\sigma$) pourraient néanmoins être le résultat d'un biais d'échantillonnage dans la mesure où les effectifs associés à ces ISLs sont relativement faibles (un COG pour l'ISL1 et trois pour l'ISL2).

A contrario, une réduction des contraintes sélectives est observée pour l'ISL3, l'ISL4 et l'ISL5 (*dN/dS* moyens ±σ compris entre 0,36 ±0,35 pour l'ISL3 et 0,52 ±0,62 pour l'ISL5), exception faite des COGs *core* de l'ISL3 (*dN/dS* moyen ±σ = 0,26 ±0,25) (Figures 4.10 et 4.11). Les COGs de l'ISL4 et du compartiment ambigu présentent de loin les pressions sélectives les moins contraignantes (*dN/dS* ±σ moyens de 0,44 ±0,13 à 0,52 ±0,23, respectivement). Ces résultats sont tout de même à nuancer, puisque soutenus par un nombre réduit de COGs *core* (trois COGs) et flexibles partagés par MIT9312 (14 COGs) dans l'ISL4.

4.2.2. Signatures des taux de substitution spécifiques à chaque compartiment

Nous venons de voir qu'il existe une variation des ratios *dN/dS* dans les différents compartiments du génome chez *Prochlorococcus*, écotype HLII, traduisant des fluctuations des pressions de sélection locales le long du génome. Cependant, les valeurs des ratios *dN/dS* peuvent avoir plusieurs origines. Considérant une valeur de base pour le ratio *dN/dS* – par exemple, sa valeur moyenne au sein du génome – un abaissement de la valeur de ce ratio dans un compartiment peut s'expliquer :

- soit par une réduction du taux de substitutions non synonymes (*dN*), traduisant alors une augmentation des pressions de sélection négatives associées aux gènes considérés ;
- soit par une augmentation du taux de substitutions synonymes (*dS*) pour un *dN* constant, traduisant le fait que les gènes comparés ont un degré de divergence supérieur à la divergence moyenne du génome.

Il apparaît dès lors intéressant d'analyser la source des fluctuations du ratio *dN/dS*, par rapport aux valeurs de *dN* et de *dS*. Dans la suite de cette étude, nous parlerons de signature évolutive.

4.2.2.1. Définition de clusters de gènes en liens avec les valeurs de *dN*, *dS* et *dN/dS*

Les signatures des taux de substitution en fonction des compartiments génomiques ont été évaluées à partir des liens existant entre *dN*, *dS* et *dN/dS* estimés entre les gènes composant chacun des COGs *core* ou flexibles partagés ou non par le génome de référence MIT9312. La représentation graphique des ratios *dN/dS* estimés pour le jeu de données total en fonction des *dN* et *dS* laisse entrevoir *a minima* deux signatures évolutives. Afin de donner un poids statistique à cette observation (Figure 4.12), la méthode des *k*-moyennes (k-*means*) a été appliquée (*cf.* Chapitre 2). Sur la base de cette méthode de partitionnement des données, cinq clusters de gènes ont été définis, parmi lesquels :

- trois sont caractérisés par de faibles valeurs de *dS* (allant de 0,001 à 0,268), accompagnées de faibles valeurs de *dN* (inférieures à 0,340) et des rapports *dN/dS* variant de 0 à plus de 1,5 (cluster jaune : valeurs inférieures à 0,314 ; orange : de 0,296 à 1,146 ; rouge : de 1,133 à plus de 1,5). Les ratios *dN/dS* élevés sont associés à des *dS* faibles plutôt qu'à des *dN* élevés, suggérant ainsi une tendance générale pour des pressions de sélection négatives et une sélection d'arrière-plan (*i.e.*, diminution de la diversité génétique liée à la réduction de la fréquence d'allèles désavantageux) ;
- un (de couleur verte) est caractérisé par des valeurs de *dS* intermédiaires (comprises entre 0,173 et 1,076) et des valeurs de *dN* faibles (inférieures à 0,326), synonyme de séquences plus divergentes et de contraintes sélectives toujours négatives ;
- un dernier (de couleur bleu foncé) est caractérisé par des valeurs de *dS* allant de 0,252 à plus de 1,5 et des ratios *dN/dS* inférieurs à 1. Ces ratios semblent ici régis par les valeurs de *dN* (comprises entre 0,172 et 1,277), comme l'illustrent les points linéairement répartis autour de la diagonale sur la Figure 4.12 ce qui montre les liens entre *dN/dS* et *dN* (panel central).

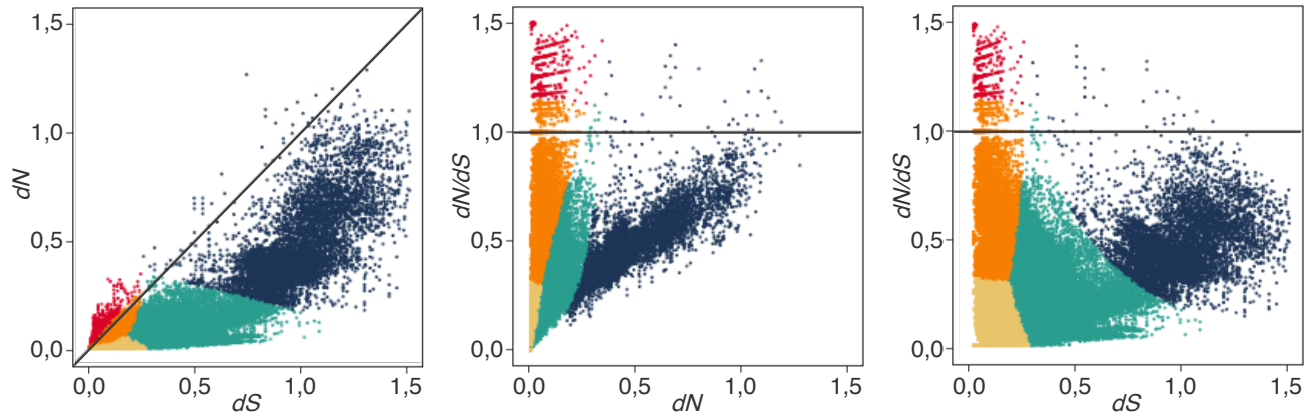


Figure 4.12. Relations entre les estimations des *dN*, *dS* et *dN/dS* pour les gènes des COGs *core* et flexibles partagés ou non par le génome de référence MIT9312, quel que soit le compartiment génomique considéré. Chaque point représente une comparaison deux à deux de gènes appartenant à un même COG et issus de deux sous-populations distinctes. Les clusters de gènes (jaune, orange, rouge, vert et bleu foncé) ont été définis par le partitionnement des données en *k*-moyennes (*k-means*). Le ratio *dN/dS* égal à 1 est représenté par une ligne grise. Par souci de lisibilité, seules les valeurs de *dN*, *dS* et *dN/dS* inférieures à 1,5 sont représentées.

4.2.4.2. Spécificités à l'échelle des compartiments

Il apparaît que les gènes au sein du génome peuvent être distingués selon trois grands types de signatures évolutives. La question est alors de savoir si ces signatures se distribuent de manière homogène le long du génome.

Concernant le *backbone*, la distribution tant des valeurs de *dN* que de *dS* varient considérablement d'un COG à l'autre, qu'il soit associé au génome *core* ou au génome flexible. Bien que quelques valeurs de *dS* puissent être supérieures à 1,5, la grand majorité (> 95 %) sont inférieures à 0,35 (moyenne *dS* $\pm\sigma$ = 0,100 $\pm$0,091). En comparaison, les *dN* affichent des valeurs et des variations bien moindres (moyenne *dN* $\pm\sigma$ = 0,026 $\pm$0,003). De plus, la relation des ratios *dN/dS*, par rapport aux valeurs de *dN* ou de *dS*, est comparable entre COGs *core* et flexibles (Figures 4.13 et 4.14).

Une analyse des ISLs montre que les patterns de variations des valeurs de *dN* et de *dS* sont contrastés pour les COGs *core* (Figure 4.13) et flexibles (Figure 4.14). Les gènes répartis sur l'ISL1, l'ISL2 et l'ISL2.1 sont principalement caractérisés pas des *dN* et des *dS* faibles et homogènes (clusters jaune, orange, rouge et vert), traduisant ainsi des pressions de sélection négatives importantes. Bien que leurs rapports *dN/dS* soient globalement

faibles, quelques gènes ont des valeurs proches de 1, voire supérieures (Figure 4.14). En comparaison, les gènes de l'ISL4 et ceux qualifiés d'ambigus se distinguent par des valeurs de *dN* et de *dS* plus élevées, des ratios *dN/dS* linéairement corrélés aux *dN*, ainsi que des *dS* proches d'une saturation (cluster bleu foncé). Les gènes situés dans l'ISL3 et l'ISL5, présentent, quant à eux, un profil mixte (ensemble des clusters de gènes).
Ainsi, les différents compartiments le long du génome semblent avoir des signatures évolutives contrastées, depuis des compartiments très contraints (ISL2.1) jusqu'à des compartiments très divergents (ISL4), tant pour le génome *core* que pour le génome flexible.

Figure 4.13. Relations entre les estimations des *dN*, *dS* et *dN/dS* pour les gènes des COGs *core* dans chaque compartiment génomique (*backbone* et ISLs). Chaque point représente une comparaison deux à deux de gènes appartenant à un même COG et issus de deux sous-populations distinctes. Les clusters de gènes (jaune, orange, rouge, vert et bleu foncé) ont été définis par le partitionnement des données en *k*-moyennes (*k-means*). Le ratio *dN/dS* égal à 1 est représenté par une ligne continue. Par souci de lisibilité, seules les valeurs de *dN*, *dS* et *dN/dS* inférieures à 1,5 sont représentées. ISL : îlot génomique.

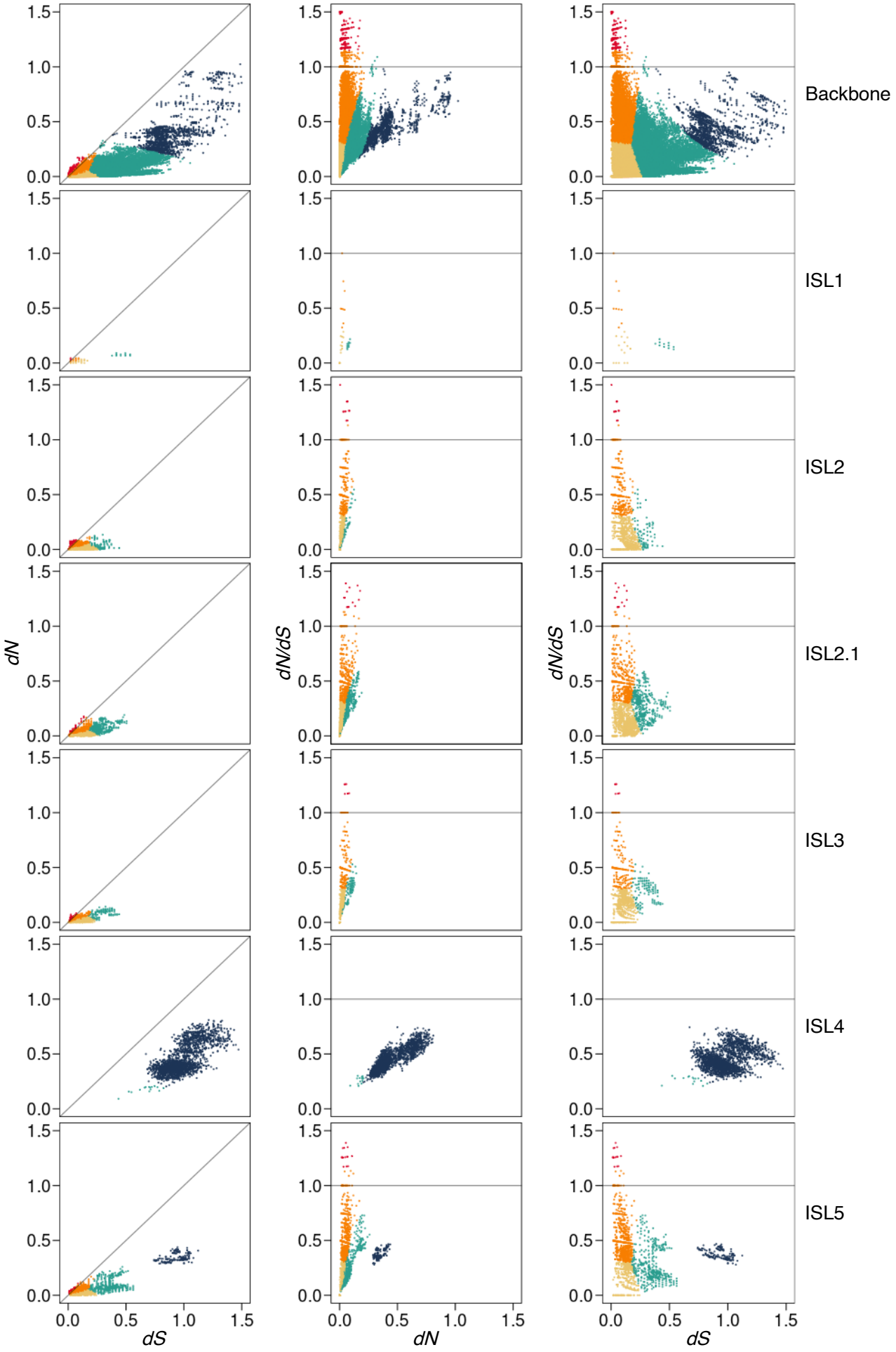
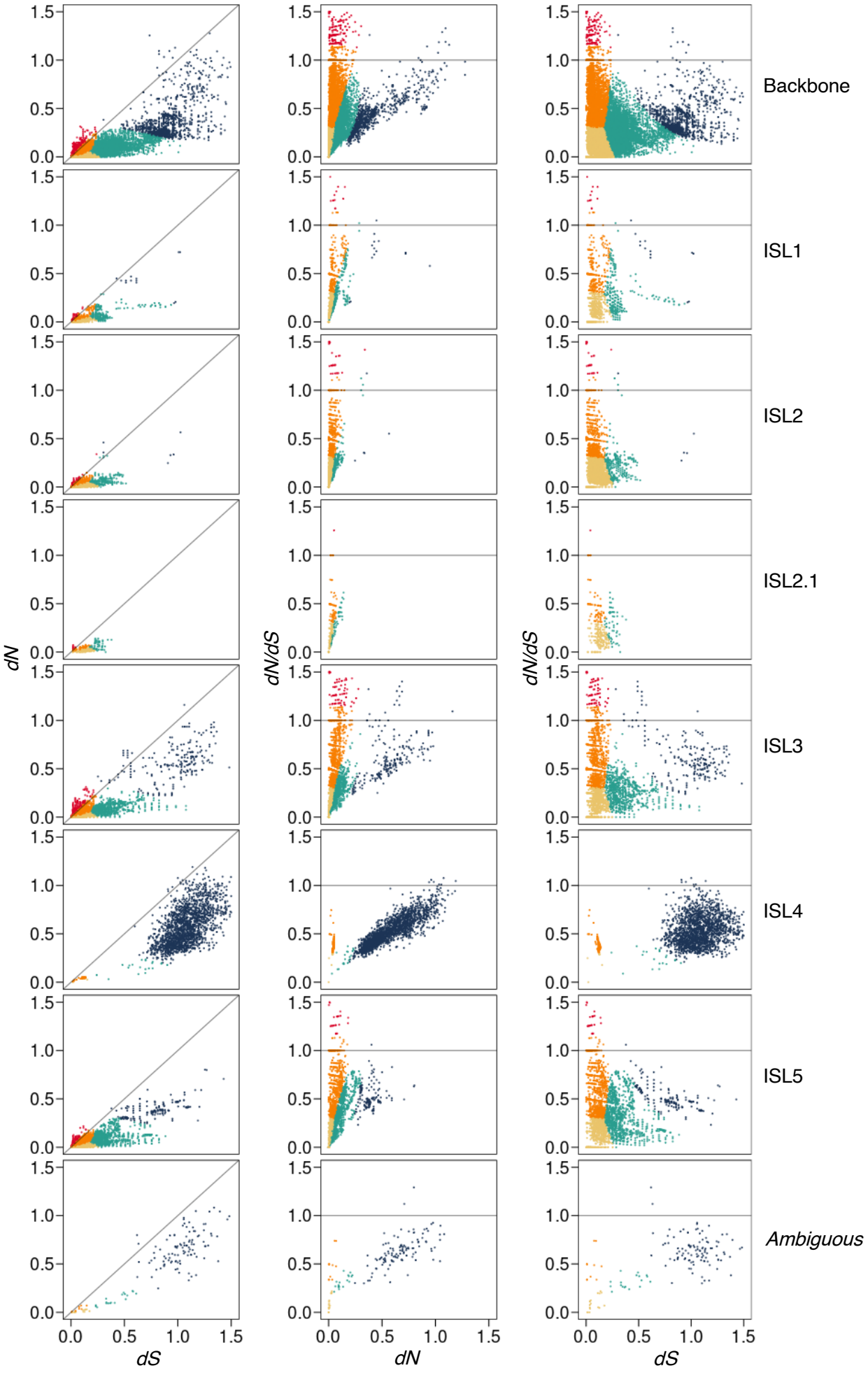

dN
dN/dS
dN/dS
dS
dN
dS
Backbone
ISL1
ISL2
ISL2.1
ISL3
ISL4
ISL5
0.0
0.5
1.0
1.5

Figure 4.14. Relations entre les estimations des *dN*, *dS* et *dN/dS* pour les gènes des COGs flexibles, partagés ou non par le génome de référence MIT9312, dans chaque compartiment génomique (*backbone*, ISLs et ambigus). Chaque point représente une comparaison deux à deux de gènes appartenant à un même COG et issus de deux sous-populations distinctes. Les clusters de gènes (jaune, orange, rouge, vert et bleu foncé) ont été définis par le partitionnement des données en *k*-moyennes (*k-means*). Le ratio *dN/dS* égal à 1 est représenté par une ligne continue. Par souci de lisibilité, seules les valeurs de *dN*, *dS* et *dN/dS* inférieures à 1,5 sont représentées. ISL : îlot génomique.


dN
dN/dS
dN/dS
Backbone
ISL1
ISL2
ISL2.1
ISL3
ISL4
ISL5
Ambiguous
dS
dN
dS
0.0
0.5
1.0
1.5

4.2.3. Distribution des COGs flexibles SAG-spécifiques dans les sous-populations et signatures évolutives

Par définition, les COGs flexibles ne sont pas partagés par l'ensemble des génomes analysés (*cf.* Chapitre 2), et peuvent être restreints à certains clades. La distribution d'un COG donné peut être plus ou moins importante au sein de ces clades. Ceci pourrait potentiellement avoir une incidence sur sa trajectoire évolutive, ce qui se traduira par une variation des valeurs de *dN* et *dS,* et donc des ratios *dN/dS*. C'est pourquoi nous nous intéresserons maintenant à la relation entre signature évolutive et distribution des COGs flexibles non partagés par MIT9312 dans les différents clades (Figure 4.15).

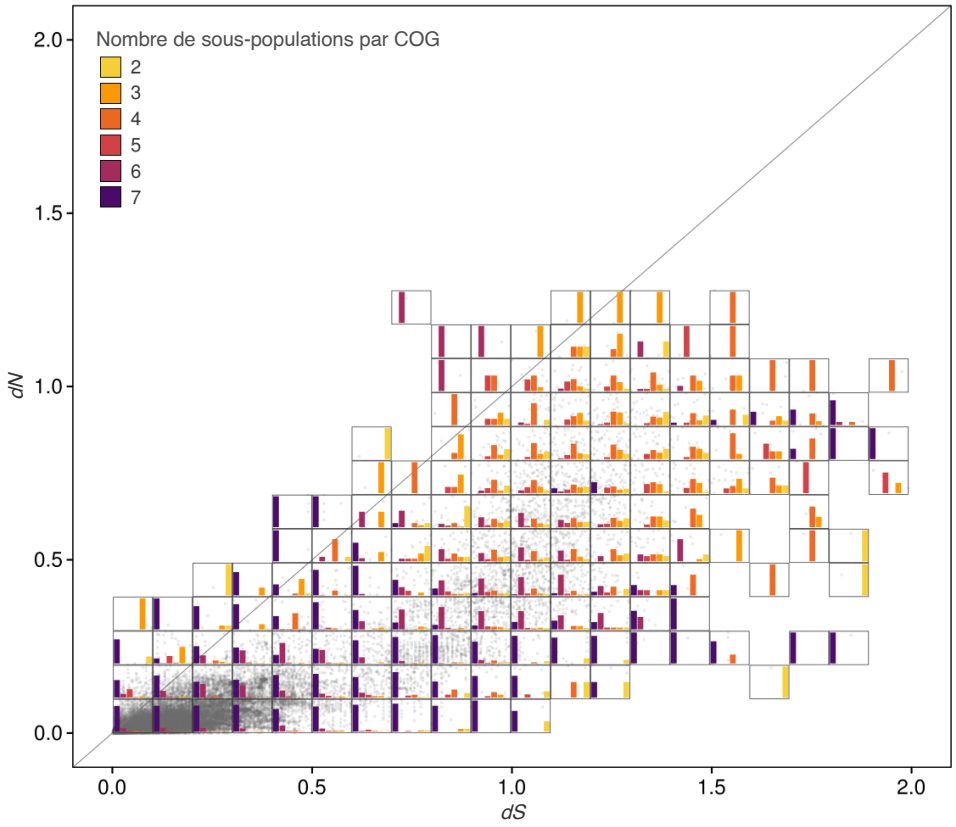


Figure 4.15. Relations entre les taux de substitution estimés pour les COGs flexibles non partagés par MIT9312 et leur distribution dans les différentes sous-populations. Le nombre de sous-populations varie entre 2 et 6 lorsqu'un COG n'est retrouvé que chez certaines d'entre elles ; il est égal à 7 si un COG est présent chez toutes les sous-populations. Le nombre de sous-populations par COG a donc été reporté sur les estimations des valeurs de *dN* et de *dS* (points gris). Chaque histogramme décrit la distribution du nombre de sous-po-

pulations 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.

Globalement, les COGs flexibles non partagés par MIT9312 sont caractérisés par des gènes dont les valeurs de *dS* peuvent être relativement variables, à l'exception des COGs partagés par les sept clades. Dans ce cas, les valeurs de *dS* sont rarement à saturation (Figure 4.15). Concernant les estimations des ratios de *dN/dS*, ils diffèrent sensiblement selon le nombre de clades dans lesquels ils sont retrouvés (p<0,001, test de Kruskal-Wallis). En effet, les COGs partagés par un grand nombre de clades ont tendance à afficher des valeurs de *dN/dS* plus faibles.

Près de 65,5 % des COGs flexibles SAG-spécifiques analysés ont une affiliation taxonomique, avec une proportion plus élevée pour ceux appartenant à l'ISL4 (78,7 % affiliés) par rapport aux autres ISLs (53,8 % affiliés en moyenne). De manière intéressante, des variations sont observées dans les affiliations taxonomiques en fonction des variations des signatures évolutives. Les COGs dont les *dS* sont faibles sont essentiellement affiliés à *Prochlorococcus* (Figure 4.16) et s'avèrent majoritairement associés à l'ISL3 et l'ISL5.

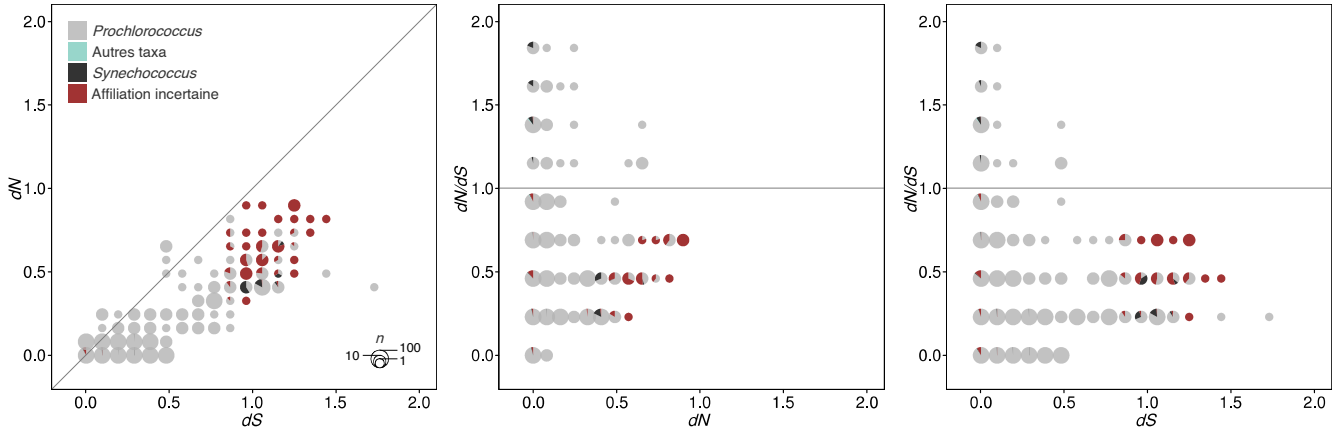


Figure 4.16. Relations entre taux de substitution estimés pour les COGs flexibles SAG-spécifiques appartenant à l'ISL3 et l'ISL5 et leurs affiliations taxonomiques. Les affiliations taxonomiques – à savoir *Prochlorococcus*, *Synechococcus*, autres groupes taxonomiques et affiliation incertaine – ont été reportées sur les estimations des *dN*, *dS* et *dN/dS*. Chaque diagramme circulaire illustre la distribution des affiliations dans une région 2D 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.

En revanche, les COGs ayant des *dS* saturés sont essentiellement classés comme incertains (*i.e.*, multiples affiliations, dont *Prochlorococcus* et/ou *Synechococcus*) (Figure 4.17) et se trouvent dans l'ISL4 ou sont considérés comme ambigus. Les COGs affiliés aux autres phyla bactériens (*i.e.*, ni *Prochlorococcus,* ni *Synechococcus)* sont caractérisés par de faibles valeurs de *dN* et de *dS,* suggérant une origine phylogénétique proche des gènes les constituant et sont enrichis dans l'ISL3 et l'ISL4 (p<0,005, test du χ^2).

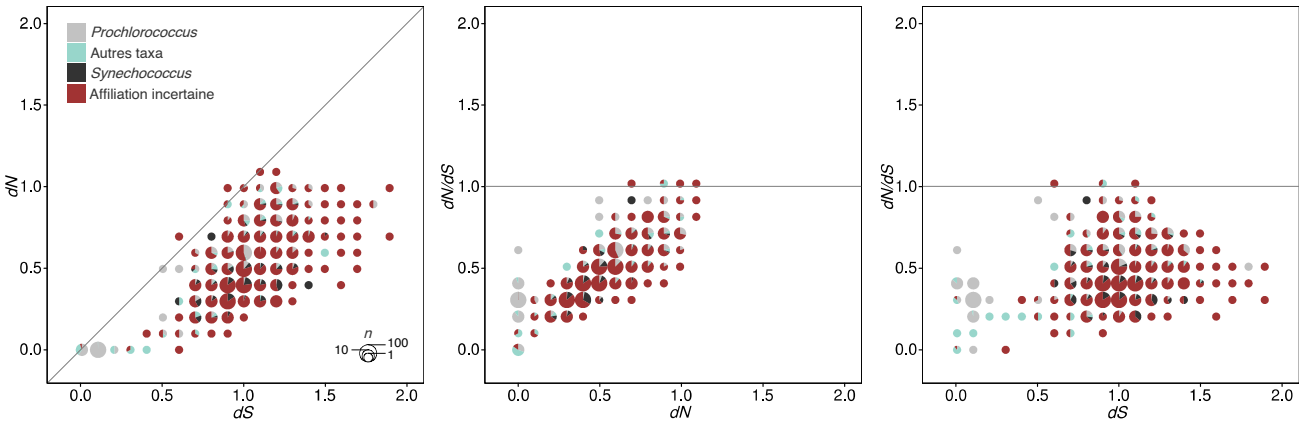


Figure 4.17. Relations entre les taux de substitution estimés pour les COGs flexibles SAG-spécifiques appartenant à l'ISL4 et considérés comme ambigus et leurs affiliations taxonomiques. Les affiliations taxonomiques – à savoir *Prochlorococcus*, *Synechococcus*, autres groupes taxonomiques et affiliation incertaine – ont été reportées sur les estimations des *dN, dS* et *dN/dS*. Chaque diagramme circulaire illustre la distribution des affiliations dans une région 2D 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.

4.3. Bilan

Ce quatrième chapitre présente les résultats de l'analyse du potentiel fonctionnel et des trajectoires évolutives des COGs constituant les génomes *core* et flexible de la population de *Prochlorococcus,* écotype HLII.

Au cours de cette étude, le potentiel fonctionnel du génome *core* a été utilisé comme point de comparaison pour la caractérisation des enrichissements fonctionnels au sein du génome flexible. L'analyse de la distribution des catégories attribuées aux gènes des COGs flexibles, comparativement au potentiel fonctionnel des SAGs, n'a pas permis de mettre en évidence de différences entre les sous-populations. Par contre, il existe une compartimentation fonctionnelle du génome flexible avec :

- une sous-représentation de fonctions associées au « Métabolisme » et au « Stockage de l'information » au sein de ce génome ;
- et une sur-représentation de celles associées aux « Processus cellulaires et de signalisation » dans l'ISL3 et le *backbone* mais aussi de celles associées à la « Biogenèse de la paroi cellulaire » et aux « Mécanismes de défense » dans l'ISL3, l'ISL4 et le compartiment ambigu. Les COGs associés à ces deux dernières catégories sont par ailleurs affiliés à une grande diversité de taxa et sont, pour l'essentiel, spécifiques à une sous-population, suggérant une acquisition *via* des HGTs.

Les principales fonctions sur-représentées dans le génome flexible identifiées ici, et leur distribution le long du génome, sont congruentes avec celles mises en évidence à une échelle de diversité plus large dans des études antérieures (Coleman *et al.*, 2006). De plus, les approches décrites dans ce paragraphe ont montré que parmi ces fonctions, certaines associées aux « Processus cellulaires et de signalisation » et impliquées dans la structuration de la membrane et les mécanismes de défense ne sont retrouvées que dans les COGs appartenant à un unique clade. Ceci pourrait signifier que les différences de distribution et d'abondance observées au niveau des écotypes pourraient s'appliquer à l'échelle des sous-populations.

D'un point de vue évolutif, des contraintes sélectives négatives inégales ont été détectées le long du *backbone* et en fonction des ISLs.

Ces signatures, comparables entre COGs *core* et flexibles, restreints ici à ceux partagés par au moins deux populations, varient en fonction :

- de l'intensité de la pression de sélection négative qui opère sur les gènes, associée ou non à une sélection d'arrière-plan ;
- du degré de divergence des séquences au sein des COGs.

Les résultats montrent, par ailleurs, que les COGs des ISL1, ISL2, et ISL2.1 sont principalement composés de gènes communs à tous les clades, affiliés au genre *Prochlorococcus* et soumis à une sélection négative forte, tandis que ceux des ISL3, ISL4 et ISL5 auraient tendance à être affiliés à d'autres taxa que les cyanobactéries et à être caractérisés par un relâchement de la pression de sélection négative. Ceci suggère que les COGs des ISL1, ISL2 et ISL2.1 seraient en cours de fixation dans les populations étudiées. Il est également à noter que les gènes composant les COGs de ces ISLs ont un *dS* très inférieur à celui des gènes des COGs *core*. Cette réduction de diversité pourrait s'expliquer par une sélection d'arrière-plan ou refléter des événements de recombinaison homologue entre les différents clades. Inversement, les gènes des COGs positionnés depuis ISL3 jusqu'à ISL4 ont tendance à avoir des *dS* élevés et pourraient être issus de HGTs.

En conclusion, la caractérisation des COGs constituant les génomes *core* et flexible considérés ici, suggère une répartition non aléatoire des fonctions portées par les gènes flexibles le long des génomes, associée à une dynamique évolutive contrastée. Ceci pourrait traduire, entre autres, des implications différentes des mécanismes de recombinaison homologue ou de HGTs le long des génomes, au sein même des populations bactériennes. Il apparaît dès lors intéressant d'évaluer l'importance de ces mécanismes dans le contexte de cette étude.

5. Impact des processus de recombinaison homologue et transferts horizontaux de gènes

Comme montré précédemment, les compartiments génomiques sont caractérisés par des signatures évolutives distinctes. Celles-ci ont révélé l'existence de deux ensembles de COGs montrant des profils d'évolution différents (*cf.* Figure 4.12). Un premier ensemble présente des valeurs de taux de substitution qui sont conformes soit à un schéma de sélection d'arrière-plan, soit à une possible homogénéisation des séquences géniques par le biais de recombinaison homologue (HR) entre clades. Ceux du deuxième ensemble présentent un relâchement des contraintes sélectives associé à un *dS* élevé et des affiliations taxonomiques incertaines, suggérant l'existence de flux de gènes avec des micro-organismes plus distants.

Comme les HR existent chez la majorité des espèces procaryotes, y compris *Prochlorococcus* (Vos and Didelot, 2009; Bobay and Ochman, 2017), il était pertinent de s'interroger sur la présence et les modalités de celles-ci au sein des sous-populations étudiées.

5.1. Balance entre mutations et recombinaison

5.1.1. Évidences de recombinaison homologue

La recombinaison est à l'origine d'échanges de matériel génétique et occasionne des irrégularités dans les phylogénies. Par conséquent, un arbre phylogénétique reconstruit pour un gène donné pourra être différent de celui inféré pour l'espèce étudiée. Ces histoires évolutives particulières peuvent être visualisées par des réseaux phylogénétiques, prenant en compte les événements de recombinaison et intégrant ainsi les réticulations dans les relations phylogénétiques (Huson, 1998).

Un réseau phylogénétique – établissant les liens entre les séquences génomiques des 87 SAGs de *Prochlorococcus*, écotype HLII, et matérialisant les événements de recombinaison – a été reconstruit avec l'algorithme *Neighbor–Net* (Bryant and Moulton, 2004) implémenté dans l'outil SplitsTree4 v4.15.1 (Huson and Bryant, 2006) (Figure 5.1). Le réseau obtenu montre la structuration de la population de *Prochlorococcus* avec une démarcation des différents clades telle que définie dans cette étude par la phylogénie reconstruite à partir du génome *core* d'une part, et par l'estimation des ANI, d'autre part (*cf.* Chapitre 4). Ainsi, les événements de recombinaison suggérés dans cette analyse ne portent probablement que sur une petite partie des génomes.

Le réseau phylogénétique met en évidence que tous les clades sont soumis à des événements de HR – tant au sein des clades qu'entre clades – sans toutefois que ceux-ci ne brouillent le signal génétique distinguant les sous-populations. Les sous-populations C1 et C2 présentent une diversité moindre, du fait de leur divergence récente, tandis que C9 est le clade le plus divergent. En outre, bien que les sous-populations C3, C4, C5 et C8 soient bien délimitées, les relations phylogénétiques entre elles sont moins évidentes. Les événements de recombinaison entre ces sous-populations pourraient ainsi expliquer la paraphylie du groupe cN2 mise en évidence dans les phylogénies reconstruites précédemment (*cf.* Chapitres 2 et 3). Comme on peut le voir sur la Figure 5.1 (encadré), les échanges de matériel génétique à l'intérieur du clade C1, qui est représenté par un plus grand nombre de SAGs, sont nombreux.

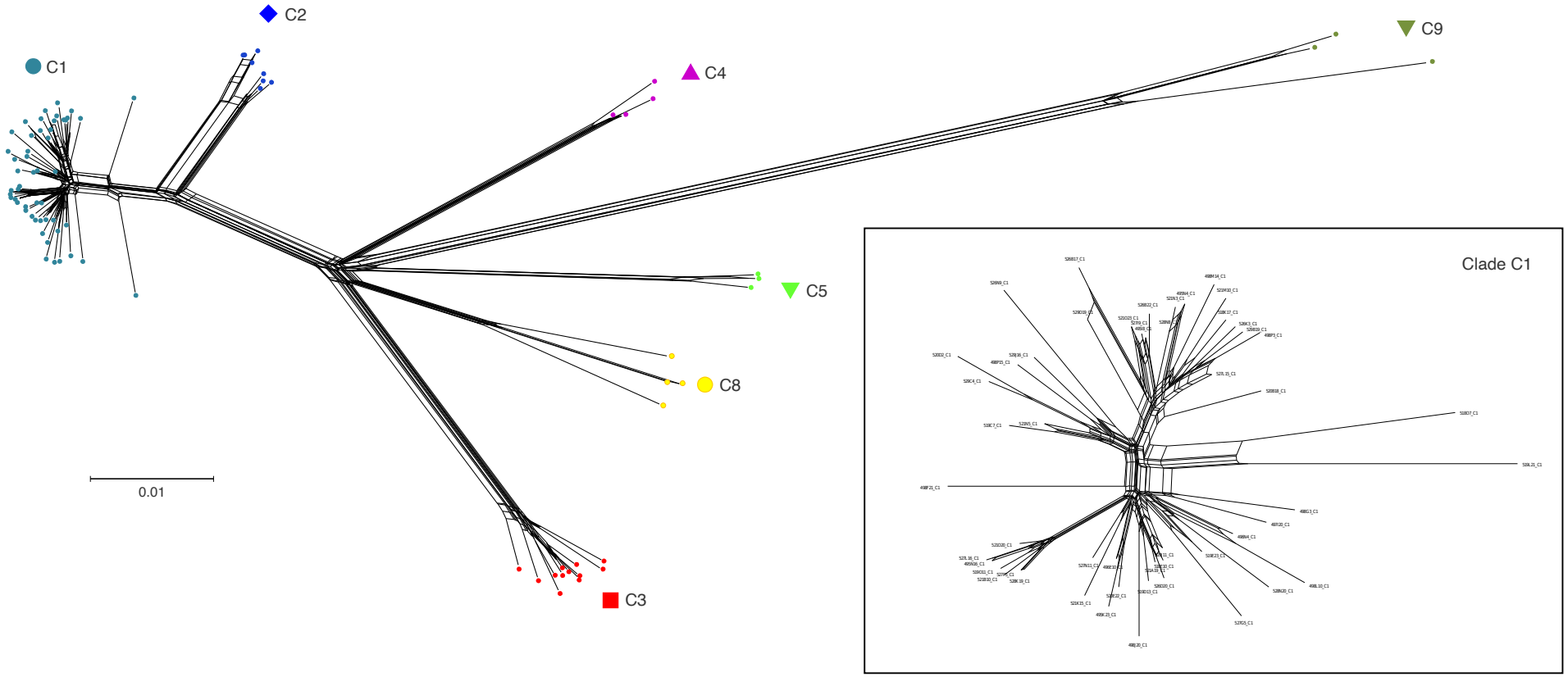


Figure 5.1. Réseau phylogénétique des 87 SAGs de *Prochlorococcus* analysés dans cette étude. Le réseau a été reconstruit avec l'algorithme *Neighbor-Net* (Bryant and Moulton, 2004) implémenté dans le logiciel SplitsTree4 v4.15.1 (Huson and Bryant, 2006). Au préalable, les distances par paires de séquences de l'alignement multiple des génomes complets ont été estimées au maximum de vraisemblance à l'aide de l'outil RAxML v8 (Stamatakis, 2014). Chaque clade (C1 à C5, C8, C9) est représenté par une forme et une couleur. Cette approche a également été appliquée au clade C1 afin de voir plus en détails les réticulations, indiquant des événements de recombinaison à l'intérieur de celui-ci.

5.1.2. Évaluation des taux de recombinaison à l'échelle des génomes

Afin d'avoir une image détaillée des événements de recombinaison qui s'opèrent au sein de la population de *Prochlorococcus* étudiée, leur détection et quantification ont été réalisées à l'échelle des génomes avec l'outil ClonalFrameML (Didelot *et al.*, 2015) (Figure 5.2). Cet outil permet notamment d'estimer, dans un contexte phylogénétique, la contribution relative de la recombinaison par rapport à la mutation dans la génération du polymorphisme observé au niveau des sous-populations. L'arbre phylogénétique présenté en Figure 5.2 correspond à la phylogénie clonale déduite par ClonalFrameML tout en tenant compte de l'influence de la recombinaison. La topologie est ici encore équivalente à celle obtenue par une estimation au maximum de vraisemblance sur le génome *core*, c'est-à-dire une démarcation en sept sous-populations. Les moyennes des différents paramètres estimés – à savoir le ratio des taux de recombinaison sur mutation, la longueur moyenne des fragments recombinés et la distance du matériel génétique importé – sont respectivement égales à $R/\theta = 0{,}105$; $\delta = 554$ pb et $v = 0{,}042$. Compte tenu de ces différents paramètres, il est possible d'évaluer l'impact relatif de la recombinaison comparativement aux mutations sur la divergence des séquences génomiques. Ainsi, bien que la fréquence des événements de recombinaison soit près de 10 fois moindre que les événements mutationnels ($R/\theta = 0{,}105$), chaque événement de recombinaison introduit en moyenne $\delta v = 23$ substitutions. La recombinaison cause donc deux fois plus de substitutions que les mutations (ratio $r/m = R/\theta \times \delta \times v = 2{,}45$). Ceci suggère que la recombinaison joue un rôle certainement non négligeable dans la diversification des sous-populations de *Prochlorococcus,* écotype HLII, à l'échelle du génome. Par ailleurs, 2 956 événements de recombinaison ont été détectés, soit en moyenne 33,97 événements par SAG. Sachant que la longueur moyenne des événements de recombinaison $\delta = 554$ pb, il est estimé que, pour un SAG donné, 18 819 nucléotides seraient associés à la recombinaison, contre 13 737 ± 14 000 nucléotides d'après Kashtan *et al.* (2014). Bien qu'obtenues à l'aide d'approches différentes, ces estimations sont du même ordre de grandeur.

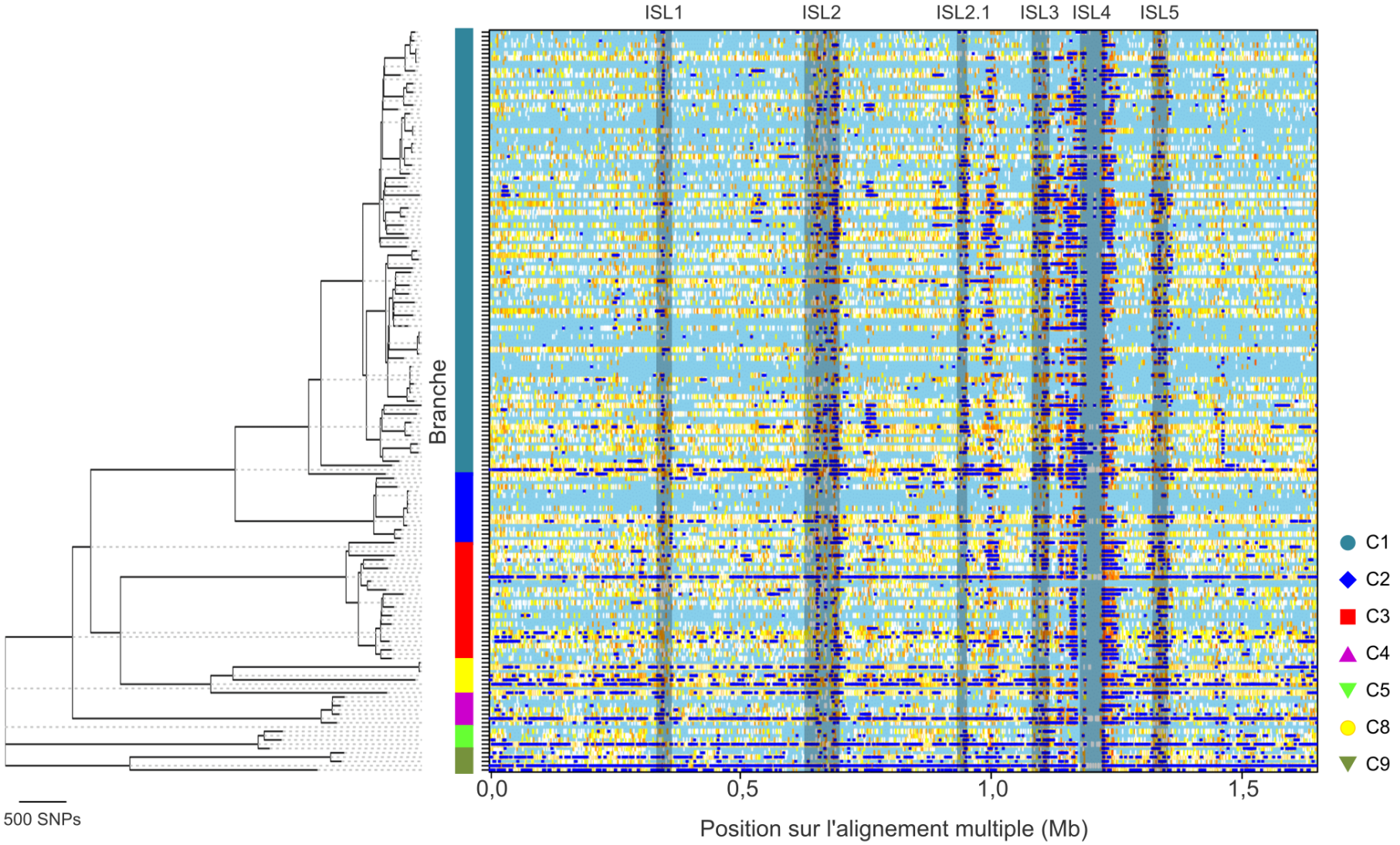

ISL1
ISL2
ISL2.1
ISL3
ISL4
ISL5
Branche
500 SNPs
0,0
0,5
1,0
1,5
Position sur l'alignement multiple (Mb)
C1
C2
C3
C4
C5
C8
C9

Figure 5.2. *Heatmap* montrant la détection d'événements de recombinaison le long du génome des 87 SAGs de *Prochlorococcus* (clades C1 à C5, C8 et C9). L'arbre phylogénétique est inféré au maximum de vraisemblance par ClonalFrameML (Didelot and Wilson, 2015) en tenant compte des événements de recombinaison. À chaque branche de l'arbre correspond une ligne de la *heatmap* et différentes couleurs qualifient le polymorphisme retrouvé entre les séquences ; bleu clair : absence de polymorphisme ; blanc : substitution compatible avec la phylogénie; du jaune au rouge : sites homoplasiques, le rouge représentant le plus fort degré d'incompatibilité avec la phylogénie. Les barres horizontales représentées en bleu foncé indiquent les événements de recombinaison détectés. Chaque clade est représenté par une forme et une couleur. Les îlots génomiques (ISLs) sont positionnés sur le génome.

Une même analyse, réalisée à partir des alignements multiples des gènes *core,* montre également l'existence d'événements de recombinaison à l'échelle du génome *core*, mais avec quelques différences par rapport au génome total – *i.e.*, $R/\theta = 0{,}108$; $\delta = 388$ pb et $v = 0{,}051$. Comme souligné précédemment pour les génomes complets, les événements de recombinaison à l'échelle du génome *core* introduisent deux fois plus de substitutions que les mutations ($r/m = 2{,}16$), avec pour chaque événement une moyenne de $\delta v = 20$ substitutions. La principale différence entre les deux conditions s'observe au niveau de la longueur moyenne des fragments recombinés qui est 30 % inférieure pour le génome *core* (majorité des fragments recombinés de taille inférieure à 1 Kpb ; Figure 5.3), ce qui est cohérent avec la taille des fragments considérés dans ces deux analyses. Une plus forte hétérogénéité dans la distribution de la longueur des fragments, une forte proportion de fragments courts (< 500 pb) et une longue queue de distribution sont observées pour le génome total. Les événements impliquant de très grands fragments (≥ 10 Kpb) sont rares et principalement observés pour le génome total, ce qui expliquerait la différence des δ entre les deux analyses.

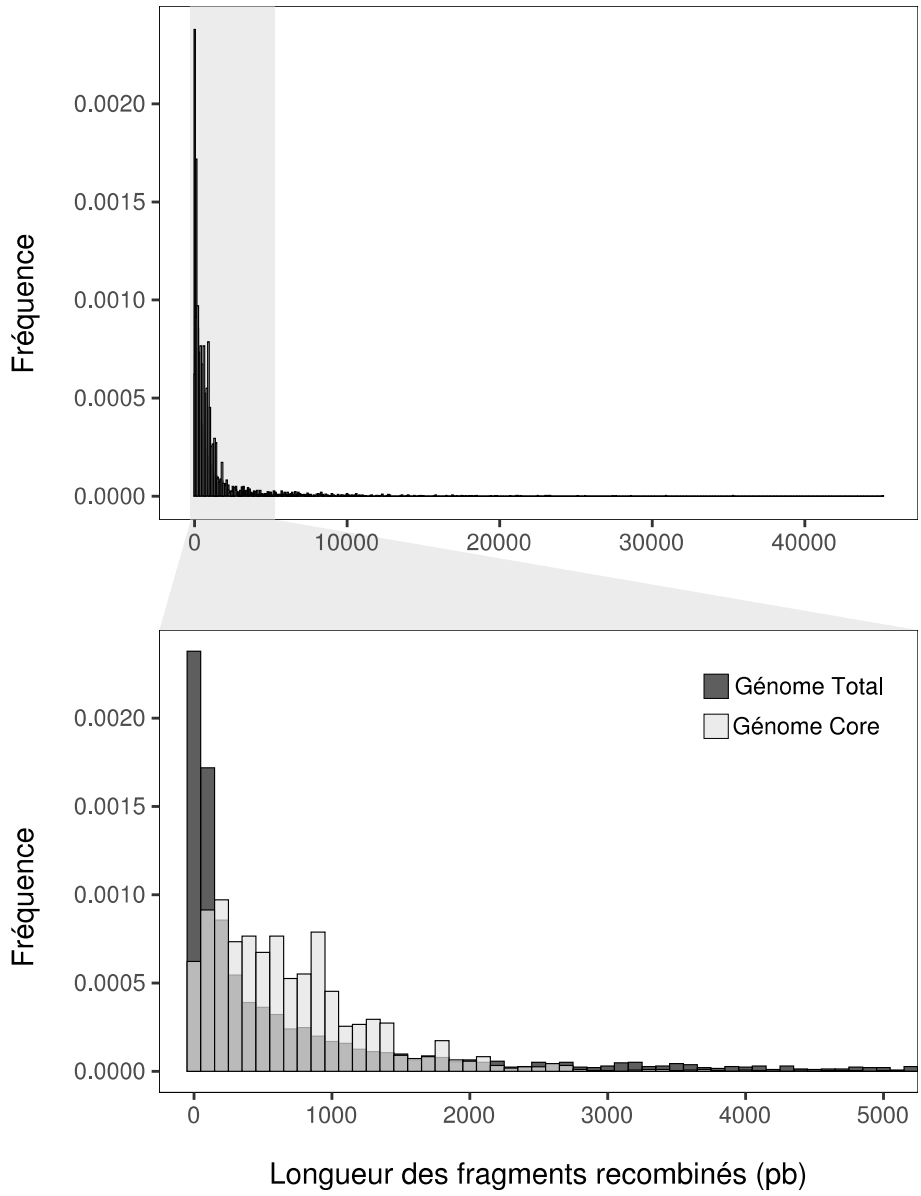


Figure 5.3. Caractérisation de la fréquence de recombinaison et de la longueur moyenne des fragments échangés pour le génome total et le génome *core* pour la population de *Prochlorococcus* étudiée.

5.1.3. Points chauds de recombinaison

De nombreux événements de recombinaison ont été détectés au sein des génomes des différents clades, y compris au niveau des gènes *core,* et semblent se concentrer au niveau de certaines régions spécifiques (Figure 5.2). Ainsi, la recombinaison pourrait être préférentiellement associée à certains compartiments génomiques.

Les pourcentages moyens de sites recombinants ont été estimés sur des fenêtres glissantes de 100 pb le long des génomes. Des points chauds, ou *hotspots*, de recombinaison, caractérisés par une nette augmentation des sites présentant des empreintes de recombinaison, ont ainsi pu être identifiés (Figure 5.4). Ceci est plus marqué pour les ISL2, ISL3, ISL5 et le *backbone* situé entre les ISL2.1 et ISL3 où le pourcentage de sites recombinants peut atteindre 20 %. En revanche, très peu d'événements de recombinaison ont été détectés pour l'ISL4 – ceci pourrait s'expliquer par sa forte variabilité entre les génomes, que ce soit en termes de similarité de séquence ou de contenu en gènes. Les régions du *backbone* encadrant cet ISL sont, à l'inverse, particulièrement soumises à la recombinaison, contenant près de 60 % des sites associés aux événements de recombinaison.

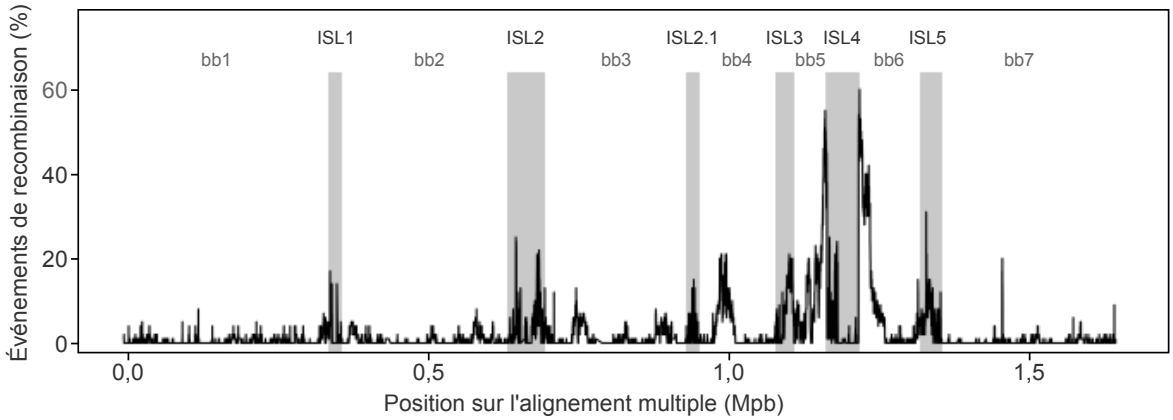


Figure 5.4. Représentation de la distribution de la fréquence des *hotspots* de recombinaison le long du génome des 87 SAGs de *Prochlorococcus*. Le pourcentage d'événements de recombinaison, détectés par ClonalFrameML, a été estimé sur des fenêtres glissantes de 100 pb. ISL : îlot génomique ; bb : *backbone*

5.2. Répartition spatiale des événements de recombinaison homologue

Des points chauds de HR ont été détectés le long des génomes, notamment en lien avec certains ISLs. Par conséquent, l'incidence de la recombinaison sur les COGs – en fonction de leur localisation génomique – a été étudiée.

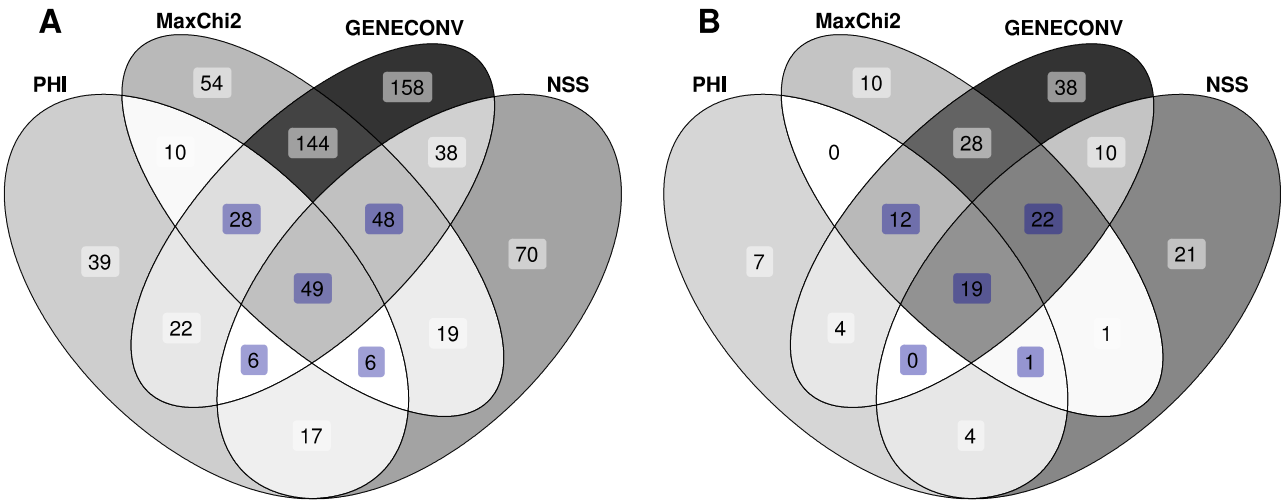
5.2.1. Quantification des gènes présentant un signal de recombinaison homologue

Dans un premier temps, la proportion de gènes présentant des traces d'événements de recombinaison a été quantifiée pour chaque catégorie de COGs – *i.e.*, *core*, flexibles communs à MIT9312 et flexibles SAG-spécifiques. Compte-tenu de l'histoire évolutive de chaque gène, relative d'une part aux événements mutationnels ayant eu lieu avant ou après la recombinaison (entraînant l'accumulation plus ou moins importante de polymorphisme) et d'autre part, à la réciprocité potentielle de la recombinaison (homologue *versus* hétérologue), il peut être difficile de quantifier les événements de recombinaison avec précision (Chan *et al.*, 2006). Afin d'assurer au mieux leur quantification, quatre outils détectant différents types de signaux ont été utilisés (*cf.* Chapitre 2). Ceci a permis de mettre en évidence entre 271 et 899 COGs avec un signal de recombinaison, selon l'outil utilisé, sur les 4 264 COGs constitués de gènes présents en copie unique et retrouvés *a minima* chez trois individus (1 202 COGs *core*, 335 flexibles partagés et 2 727 non partagés par le génome de référence MIT9312), pour un total de 1 224 COGs détectés comme recombinants toutes méthodes confondues (Tableau 5.1).

Tableau 5.1. Nombre de COGs présentant un signal de recombinaison pour chaque outil utilisé (Max χ^2, NSS, PHI et GENECONV) et chaque catégorie (*core*, flexibles partagés et non partagés par le génome de référence MIT9312).

		Flexibles		
	Core	partagés par MIT9312	non partagés par MIT9312	Total
Max χ^2	358	93	118	569
NSS	253	78	108	439
Phi	177	47	47	271
GENECONV	493	133	273	899
Total	708	177	339	1 224

Les différentes méthodes donnant des résultats hétérogènes, seuls les COGs présentant un signal de recombinaison avec au moins trois des quatre outils ont été considérés comme recombinants par la suite, soit 241 COGs au total (137 *core*, 54 flexibles partagés et 50 flexibles non partagés par le génome de référence MIT9312 ; Figure 5.5).



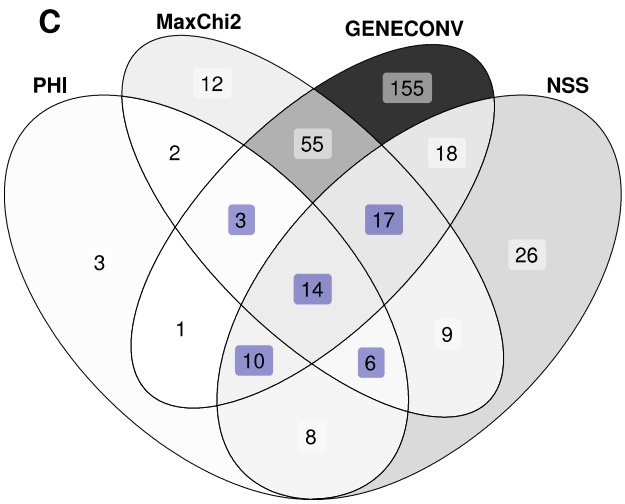


Figure 5.5. Diagrammes de Venn représentant la répartition du nombre de COGs recombinants en fonction des outils utilisés (Max χ^2, NSS, GENECONV, PHI). Nombre d'événements pour : **(A)** les COGs *core* ; **(B)** flexibles partagés et **(C)** non partagés par le génome de référence MIT9312. Pour chaque catégorie, le nombre de COGs détectés comme recombinants par au moins trois outils est mis en évidence par un encadré bleu.

Les résultats obtenus montrent que les événements de recombinaison impactent toutes les catégories de COGs mais de façon inégale. En effet, 16,12 % des COGs flexibles communs à MIT9312 présentent des signatures de recombinaison contre 11,40 % des COGs *core* et 1,83% des COGs flexibles non partagés par le génome de référence MIT9312. La répartition des COGs recombinants en fonction du compartiment génomique (*backbone versus* ISL) est également hétérogène. Ainsi, alors qu'aucun COG recombinant n'est détecté dans l'ISL1, cela concernerait 1 % des COGs de l'ISL5 et jusqu'à 15 % de ceux de l'ISL2 (Tableau 5.2). Au niveau du *backbone*, une proportion significative de COGs recombinants est observée dans les régions bordant l'ISL4. Ces résultats sont cohérents avec les points chauds de recombinaison identifiés à l'échelle des génomes (Figure 5.4).

Tableau 5.2. Caractéristiques des COGs recombinants présents dans les différents compartiments génomiques.

	COGs recombinants (%)	Nb. moyen de gènes / COG recomb.[a]	Nb. moyen de clades / COG recomb.[b]	Identité nucléotidique (%)
Core	11,40	60,83	7,96	96,20
Flexibles – communs à MIT9312	16,12	42,83	6,66	90,82
Flexibles – clade-spécifiques	1,83	6,38	2,12	90,17
bb1	5,34	28,48	4,42	95,75
ISL1	0,00	0,00	0,00	0,00
bb2	8,11	26,75	4,34	95,05
ISL2	14,85	30,13	4,56	93,47
bb3	6,17	27,75	4,25	92,41
ISL2.1	4,41	24,71	4,15	94,26
bb4	5,47	24,76	4,19	91,56
ISL3	6,70	13,57	3,25	83,14
bb5	16,40	21,45	3,84	88,08
ISL4	6,42	4,81	1,88	73,46
bb6	13,64	24,85	4,31	92,61
ISL5	1,00	18,67	4,10	93,73
bb7	4,02	29,49	4,58	96,03

[a] Nombre moyen de gènes par COG recombinant ;
[b] Nombre moyen de clades par COG recombinant.

La détection des événements de HR peut être affectée par différents paramètres dont :

- la taille des COGs, qui reflète leur occurrence dans les SAGs. La probabilité d'identifier un événement de recombinaison dépend ainsi de la taille de l'échantillon que constitue le nombre de gènes par COG ;
- le nombre de clades dans lesquels sont retrouvés les COGs ;
- les similarités de séquence des gènes au sein d'un COG, synonyme de distance évolutive (Tableau 5.2).

Le poids de chacune de ces variables dans la détection des COGs recombinants a été évalué au travers d'une analyse itérative de tous les modèles de régression pouvant être construits à partir de celles-ci. Les modèles obtenus ont ensuite été comparés entre eux afin de retenir celui qui décrit le mieux les variations observées. D'après le modèle retenu, le nombre de COGs détectés comme recombinants dépend principalement du nombre de gènes présents dans les COGs, indépendamment de la catégorie (*core versus* flexible) ou du compartiment (*backbone versus* ISLs) (R^2 = 0,43, *p–value* < 0,05). Les autres variables n'ont, en revanche, qu'un impact négligeable sur la détection de COGs recombinants.

5.2.2. COGs recombinants *versus* événements de recombinaison

Les analyses précédentes indiquent qu'environ 11 % des COGs *core* et 16 % des COGs flexibles partagés par MIT9312 sont concernés par la HR, contre moins de 2 % des COGs SAG-spécifiques. Lorsqu'un COG présente un signal de recombinaison, il est envisageable que plusieurs événements – impliquant plusieurs gènes le constituant – lui soient associés.

Le nombre d'événements de recombinaison homologue pour chaque COG recombinant a été estimé à l'aide de l'outil fastGEAR (Mostowy *et al.*, 2017). Cette analyse a permis de détecter des événements de recombinaison pour 212 des 241 COGs recombinants. Leur nombre est plus important pour les COGs *core* et flexibles communs à MIT9312, avec respectivement, 9,21 et 10,39 événements détectés par COG en moyenne, contre 4,32 en moyenne pour les COGs flexibles SAG-spécifiques (Tableau 5.3). Cette variabilité s'observe également pour les différents compartiments génomiques. À cette échelle, le nombre moyen d'événements de recombinaison par COG recombinant varie de 2,00 pour les COGs associés à ISL2.1 et ISL5 à 35,93 pour ceux localisés dans l'ISL4, avec une moyenne de 8,60 événements par COG recombinant et par compartiment (6,67 si l'ISL4 est exclu).

Tableau 5.3. COGs recombinants et événements de recombinaison inférés avec fastGEAR (Mostowy *et al.*, 2017) en fonction des compartiments génomiques.

	COGs recombinants (%)	Nb. moyen d'év. / COG recomb.[a]
Core	10,40	9,21
Flexibles – Communs à MIT9312	14,63	10,39
Flexibles – clade-spécifiques	1,39	4,32
bb1	3,42	6,68
ISL1	0,00	0,00
bb2	5,16	6,51
ISL2	8,91	7,22
bb3	4,85	4,73
ISL2.1	2,94	2,00
bb4	4,01	4,70
ISL3	6,70	3,79
bb5	13,76	8,50
ISL4	5,28	35,93
bb6	9,79	10,89
ISL5	1,00	2,00
bb7	2,56	4,93

[a] Nombre moyen d'événements de recombinaison par COG recombinant.

Dans la mesure où la détection de signaux de recombinaison est expliquée en partie par le nombre de gènes constituant les COGs, le nombre d'événements de recombinaison a été pondéré par le nombre total de gènes des COGs recombinants pour une catégorie donnée (*i.e., core*, flexibles partagés ou non partagés par MIT9312) ou un compartiment génomique (*i.e., backbone* et ISLs). Ceci a permis de mettre en évidence que, dès lors qu'ils sont recombinants, les COGs SAG-spécifiques sont touchés par un nombre important

d'événements de recombinaison (Figure 5.6). Ce constat est également valable pour ISL3, ISL4 et les régions du *backbone* situées de part et d'autre de ISL4 (*i.e.,* bb5 et bb6).

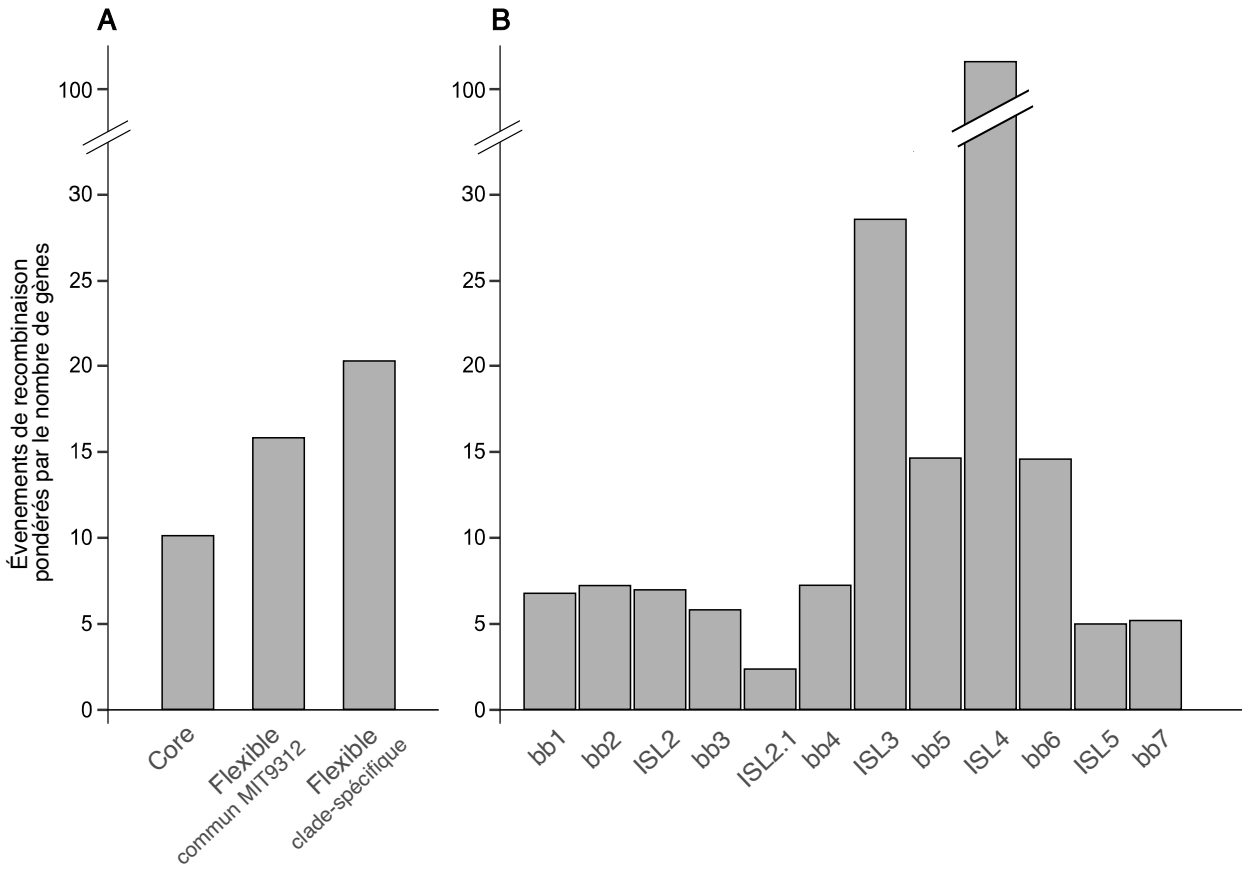


Figure 5.6. Représentation du nombre d'événements de recombinaison par COG recombinant, pondéré par leur taille (en nombre de gènes). (A) Pour les catégories de COGs. **(B)** Pour les compartiments génomiques.

5.2.3. Évaluation de l'ancestralité des événements de recombinaison homologue

La recombinaison implique un échange de matériel entre bactéries donneuses et receveuses. Il peut donc être intéressant, d'une part, d'analyser la direction des événements de recombinaison (existe-t-il des sous-populations dont les individus sont préférentiellement donneurs et/ou receveurs ?) et d'autre part, de définir l'ancestralité de ces évé-

nements (les gènes touchés par des événements inférés comme anciens seraient plus susceptibles d'être fixés dans les sous-populations).

La direction des événements de recombinaison ainsi que leur ancestralité ont été analysées avec l'outil fastGEAR (Mostowy *et al.*, 2017). Une recombinaison ancestrale affecte l'ensemble des individus d'une sous-population. L'événement qui lui est associé est survenu entre les ancêtres communs des sous-populations donneuse et receveuse concernées. Les événements dits récents n'affectent pour leur part qu'une partie des individus de la sous-population (*cf.* Figure 2.17, Chapitre 2). Il devient dans ce cas possible d'inférer une direction aux transferts de matériel génétique. Les résultats de fastGEAR montrent que les COGs *core* et flexibles ont expérimentés, en moyenne, bien plus d'événements de recombinaison récents qu'ancestraux (Tableau 5.4). Ceci peut s'expliquer soit par l'élimination du matériel génétique non adaptatif (sans avantage sélectif) échangé lors des événements ancestraux soit, au contraire, par le transfert de ce matériel génétique à l'ensemble de la population, conduisant alors à son homogénéisation.

Tableau 5.4. Nombre d'événements de recombinaison récents et ancestraux inférés pour les COGs détectés comme recombinants par fastGEAR (Mostowy *et al.*, 2017).

		Flexibles		
	Core	**Communs MIT9312**	**Clade-spécifiques**	
Nombre d'événements	888	328	157	**Récents**
Nombre de COGs[a]	117	41	37	
Moyenne par COG[b]	7,59	8,00	4,24	
Nombre d'événements	263	181	7	**Ancestraux**
Nombre de COGs[a]	71	22	7	
Moyenne par COG[b]	3,70	8,23	1,00	

[a] Nombre de COGs impliqués dans des événements récents et/ou ancestraux ;
[b] Nombre moyen d'événements de recombinaison récents et/ou ancestraux par COG recombinant.

La distribution des recombinaisons récentes et ancestrales dépend de la catégorie des COGs (*core,* flexibles partagés et non partagés). Alors que les deux sont retrouvées au sein des COGs *core* et flexibles communs à MIT9312, très peu d'événements ancestraux ont été détectés au sein des COGs flexibles SAG-spécifiques (Figure 5.7).

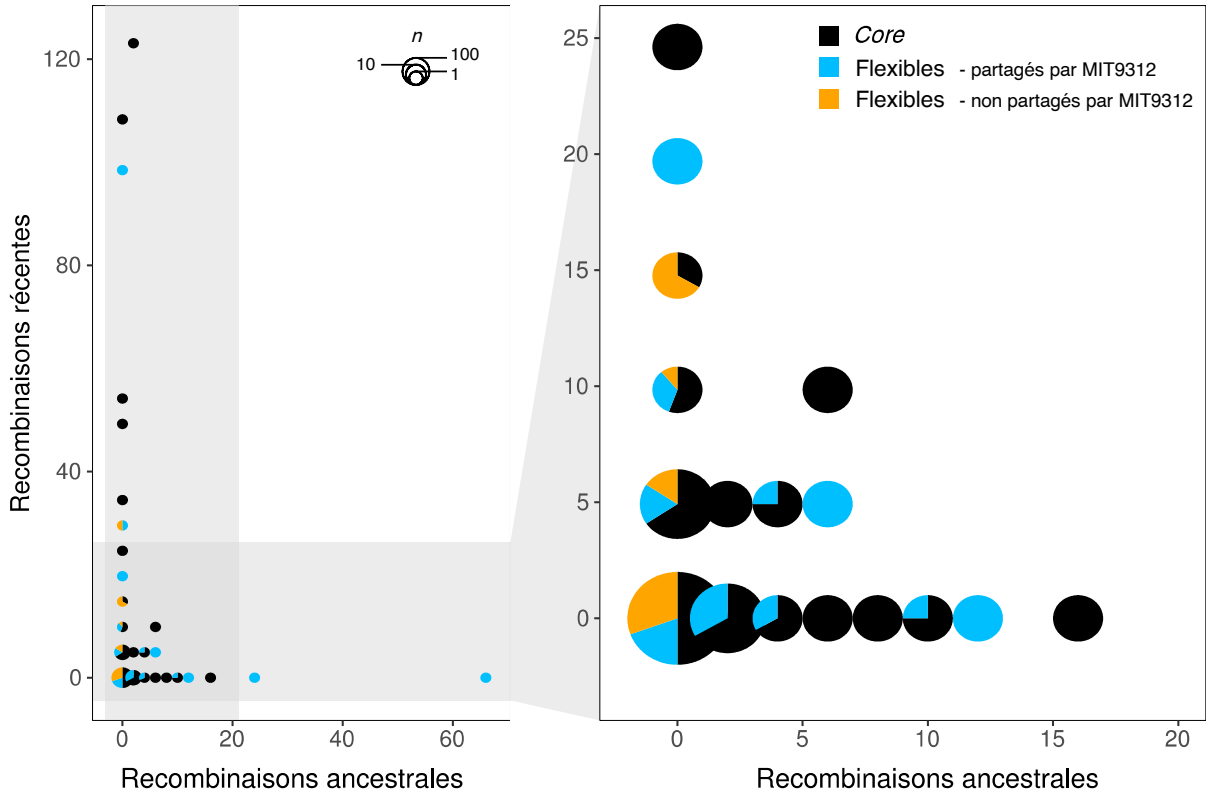


Figure 5.7. Représentation de le fréquence du nombre d'événements ancestraux (abscisse) et récents (ordonnée) détectés dans les COGs recombinants en fonction de leur catégorie (*core*, flexibles partagés et non partagés par MIT9312). Les diagrammes circulaires illustrent la proportion de COGs associés aux catégories *core* (noir), flexibles partagés par MIT9312 (bleu) et non partagés par MIT9312 (jaune). La taille des diagrammes est proportionnelle au nombre *n* de COGs (au moins une observation ; de 10 à 100 observations ; plus de 100 observations).

Pour l'ensemble des sous-populations étudiées, les couples donneur / receveur impliqués dans les recombinaisons récentes ont été identifiés. Ils s'apparentent respectivement à une sous-population donnée et à un ou plusieurs SAGs au sein d'une sous population. Si la part des sous-populations donneuses peut être estimée directement à partir de leur dénombrement dans les événements de recombinaison récents, la part des différentes

sous-populations au sein des SAGs receveurs dépend du nombre de SAGs constituant chacune des populations analysées et est proportionnelle à la complétude de ces derniers (Figure 5.8).

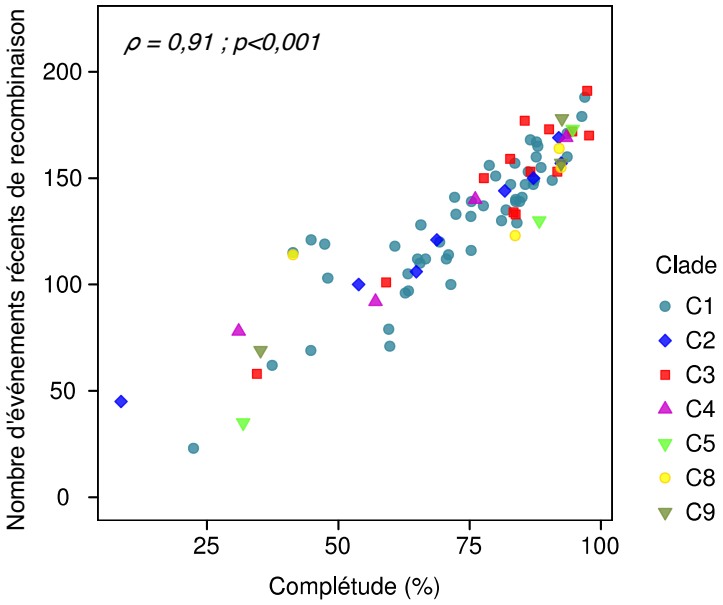


Figure 5.8. Corrélation entre complétude des SAGs (exprimée en %) et nombre d'événements de recombinaison récents. Chaque point correspond à un SAG. Tous les SAGs d'une sous-population donnée sont représentés par une couleur et une forme qui leurs sont propres. Le résultat de la corrélation de Spearman (ρ) et la *p-value* associée sont indiqués.

On constate que la part des sous-populations en tant que clades donneurs ou receveurs est différente et cette part est également différente en fonction des catégories de COGs. Quelles que soient les catégories de COGs, les clades C1 et C3 (qui représentent respectivement 60 et 15 % du jeu de données en termes de nombre de SAGs, mais aussi 30 et 13 % du jeu de données en termes d'assemblage) sont les récipiendaires principaux de ces événements de recombinaison récents (Figure 5.9). Pour les COGs *core*, les résultats montrent que le clade C8 (4,6 % des SAGs, 4,1 % de l'assemblage) ainsi que, dans une moindre mesure, les clades C2 (9,2 % des SAGs, 12,8 % de l'assemblage) et C9, reçoivent beaucoup de matériel. Les clades C4 et C5, en revanche, seraient faiblement représentés parmi les receveurs. Pour les COGs flexibles communs à MIT9312, tous les clades sont récipiendaires d'événements récents alors que seuls les clades C1 à C4 et C9 le sont pour les COGs flexibles SAG-spécifiques.

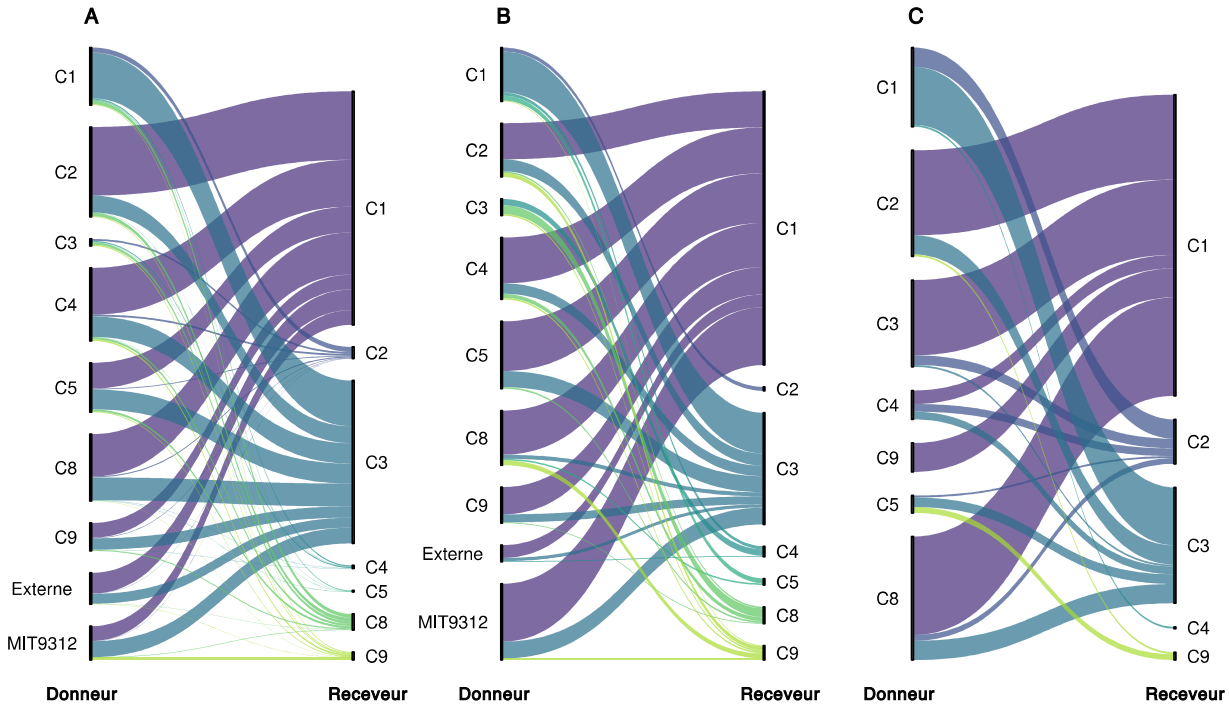


Figure 5.9. Représentation des recombinaisons récentes montrant les interactions donneur / receveur entre clades. L'ancestralité des événements de recombinaison a été définie à l'aide de l'outil fastGEAR (Mostowy *et al.*, 2017). Les couples donneur / receveur ont été identifiés pour les COGs recombinants des catégories **(A)** *core*, **(B)** flexibles communs à MIT9312 et **(C)** SAG-spécifiques. Certains donneurs correspondent au génome de référence MIT9312 alors que d'autres sont extérieurs au jeu de données analysé et regroupés sous le terme *externe*.

Concernant les donneurs, il n'est possible de considérer que les clades auxquels ils appartiennent. La fréquence à laquelle ces clades donneurs participent effectivement à des événements de recombinaison ne semble pas dépendre de la taille des clades (Figure 5.9). Ainsi, à titre d'exemple, aucun couple n'est observé avec pour donneur le clade C3 et pour receveur le clade C1 (Figures 5.9A et 5.8B), alors que ces deux clades sont les plus représentés au sein de notre jeu de données (52 SAGs pour C1 et 13 pour C3). De même, les clades C4 à C9, faiblement représentés en termes de nombre de SAGs, prennent part à autant d'événements, en tant que donneurs, que les clades C1, C2 ou C3. Par ailleurs, la proximité phylogénétique ne semble pas affecter la proportion de clades donneurs. Bien que le couple C1 / C2 – les deux clades les plus proches phylogénétiquement – soit le plus représenté pour les COGs *core* (Figure 5.9A), ceci n'est pas vérifié pour les COGs flexibles partagés ou non par le génome de référence MIT9312 (Figures 5.9A et B).

Toutes ces disparités sont plus marquées pour les COGs flexibles SAG-spécifiques (Figure 5.9C), sans doute du fait du plus faible nombre d'événements de recombinaison qui leurs sont associés.

Concernant les recombinaisons ancestrales, il est possible d'identifier un couple donneur / receveur sans pour autant inférer une direction aux événements de recombinaison, ni connaître avec exactitude l'origine du fragment transféré. Ainsi, de par la nature même des événements et de leur ancestralité, un couple s'apparente à deux sous-populations. Seuls sept événements ancestraux de recombinaison ont été détectés dans les COGs clade-spécifiques, ils ne sont donc pas montrés. Les profils de recombinaison ancestrale des COGs *core* et flexibles communs à MIT9312 sont quant à eux relativement similaires (Figure 5.10).

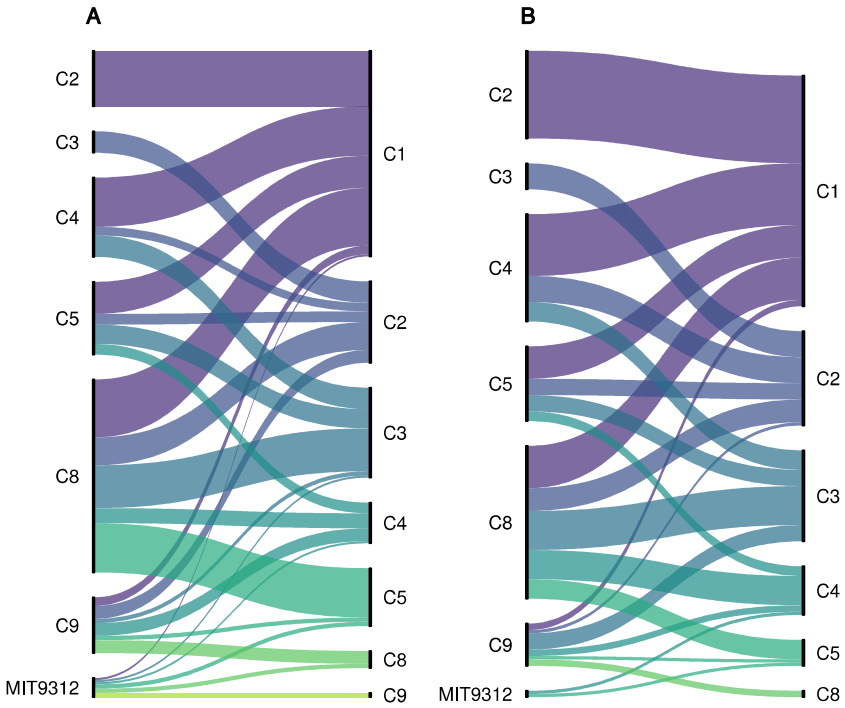


Figure 5.10. Représentation des événements de recombinaisons ancestrales montrant les interactions entre clades. L'ancestralité des événements de recombinaison a été définie à l'aide de l'outil fastGEAR (Mostowy *et al.*, 2017). Les couples de clades ont été identifiés pour les COGs recombinants des catégories **(A)** *core* et **(B)** flexibles communs à MIT9312.

Les sous-populations C1 et C8 sont majoritairement représentées. Ceci suggère que la détection des événements ancestraux serait indépendante de la taille des clades – le clade C8 n'étant constitué que de quatre SAGs. Elle serait également indépendante de la proximité phylogénétique – les couples C4 / C1, C8 / C1 ou C8 / C5 ont une occurrence équivalente à celle du couple constitué des sous-populations C1 / C2 ayant récemment divergé (Figure 5.11).

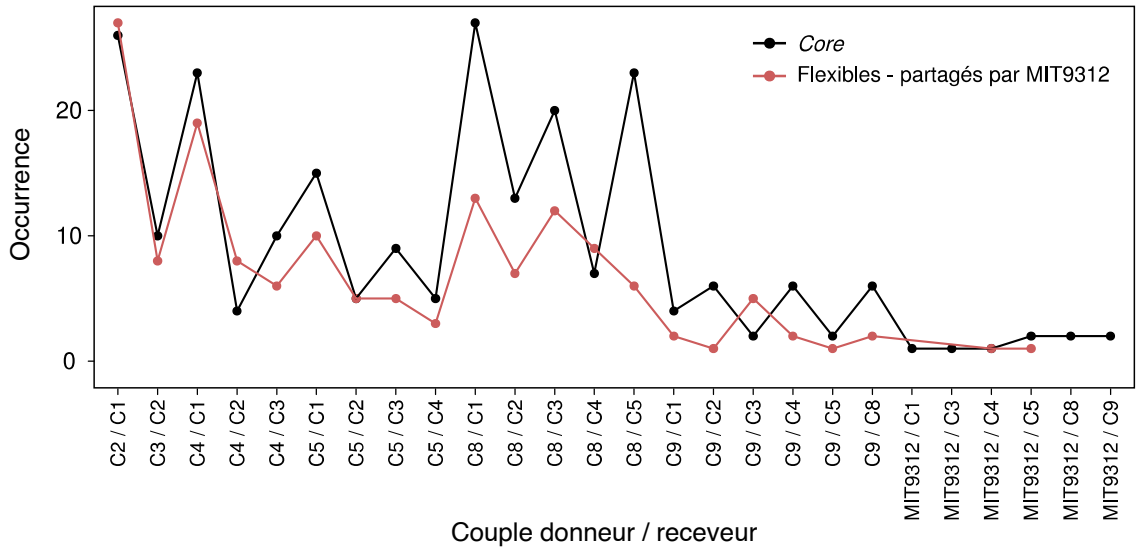


Figure 5.11. Représentation de l'occurrence des couples donneurs / receveurs impliqués dans des événements de recombinaisons ancestrales.

Il semblerait que l'implication des clades C1 et C8 soit majoritairement dictée par l'abondance relative des cellules apparentées à ces deux clades dans l'environnement (Table 5.5).

Tableau 5.5. Abondance relative des *clusters* d'ITS obtenue pour chaque clade à partir des données de SCG (Kashtan *et al.*, 2014). Au total, 1 381 séquences d'ITS, amplifiées et séquencées suite au tri des cellules par cytométrie en flux, ont été obtenues pour trois campagnes d'échantillonnage (automne, hiver, été) et correspondent à *Prochlorococcus*, écotype HL. Pour chaque sous-population, l'abondance relative (± se) a été estimée, et reflète un pourcentage de la population totale.

	Automne	**Hiver**	**Été**
C1	14,4 ± 1,6	3,3 ± 0,8	17,8 ± 1,6
C2	1,5 ± 0,9	0,9 ± 0,3	1,3 ± 0,4
C3	2,9 ± 0,8	2,8 ± 0,7	4,3 ± 1,1
C4	0,8 ± 0,5	1,5 ± 0,2	0,3 ± 0,2
C5	0,4 ± 0,4	0,4 ± 0,2	0,2 ± 0,2
C8	20,6 ± 2,2	17,5 ± 1,9	13,9 ± 1,6
C9	6,6 ± 0,2	9,7 ± 2,4	9,2 ± 0,3

6. Discussion et Perspectives

6.1. Diversification des populations et adaptation de niche

Les modèles d'évolution associés aux concepts de différenciation des espèces mettent en avant des processus de balayage sélectif à l'échelle des gènes (BSC) ou à l'échelle des génomes (ESC). La spéciation de populations bactériennes pour leur adaptation à des niches spécifiques est par ailleurs documentée non seulement pour des individus dans un contexte expérimental (Wiser *et al.*, 2013), mais également pour des individus dans un contexte environnemental (Shapiro *et al.*, 2012; Kent *et al.*, 2016; Larkin *et al.*, 2016).

Dans le cadre de cette thèse, l'analyse phylogénétique basée sur la concaténation des COGs *core* présents en copie unique et l'analyse des ANI pour les sous-populations de l'écotype HLII du genre *Prochlorococcus* provenant des BATS ont donné des résultats qui corroborent ceux générés par l'analyse des séquences de l'ITS (Kashtan *et al.*, 2014) et des segments génomiques strictement identiques entre paires de génomes (Arevalo *et al.*, 2019). En effet, quelle que soit l'approche utilisée, les résultats obtenus montrent une structuration en clades, ce qui pourrait refléter une partition ancienne de ces sous-populations, en lien avec une adaptation à différentes niches écologiques. Cette hypothèse est supportée, d'un point de vue génétique, par la prédominance d'allèles différents pour les gènes *core*, fixés au sein de sous-populations, et associés à des ensembles de gènes flexibles spécifiques des sous-populations (Kashtan *et al.*, 2014). Elle l'est également d'un point de vue écologique par l'observation de variations de l'abondance relative de ces sous-populations en fonction des saisons, ce qui reflète une réponse potentiellement adaptative aux conditions environnementales, bien que les moteurs de cette différenciation restent incertains (Larkin *et al.*, 2016). En revanche, d'un point de vue évolutif, aucune adaptation de niche ne peut être avancée, dans la mesure où aucune empreinte de sélection positive n'a été mise en évidence à l'échelle des COGs, les ratios *dN/dS* estimés soutenant globalement des pressions de sélections négatives à la fois inter- et intra-clades.

Sur la base d'un modèle d'évolution des génomes, Marttinen et collaborateurs (2015) montrent que la recombinaison peut tout ou partie expliquer la structuration des populations bactériennes dont la diversité génétique est limitée, sans implication d'adaptation de niches. Ainsi, il pourrait ne pas être nécessaire d'invoquer des processus adaptatifs pour expliquer la structuration en clades de cette population HLII. Cependant,

lors de l'étude des espèces dominantes de l'intestin humain, des résultats similaires ont été obtenus et ont conduit les auteurs de ces travaux à suggérer que la sélection positive, si elle est présente, ne peut l'emporter sur le signal de sélection négative (Garud *et al.*, 2019).

À l'échelle des SAGs étudiés, l'analyse des signatures de sélection a mis en lumière une pression de sélection négative (estimée ici par une analyse de *dN/dS*), dont l'intensité varie selon la nature des COGs, *i.e.*, *core* ou flexibles et leur appartenance au *backbone* ou aux ISLs (Chapitre 4, Figures 4.10 et 4.11). Pour exemple, la concentration dans des îlots spécifiques (ISL2 et ISL2.1) de COGs communs à l'ensemble des sous-populations et soumis aux contraintes sélectives les plus fortes suggère que certains décrits comme flexibles à l'échelle du genre sont en passe d'être fixés (*i.e.*, devenir *core*) dans celles-ci. Ceci rejoint les conclusions faites par Avrani et collaborateurs (2011) pour ISL2.1 à l'échelle de la diversité globale de l'écotype HLII. Les très fortes valeurs de F_{ST} dans ces îlots révèlent par ailleurs une diversité allélique au sein de ces COGs (Figure 6.1) qui pourrait soutenir la séparation de niches écologiques associées aux différents clades. Cette hypothèse rejoint les propositions de Kashtan et collaborateurs (2014) faites sur la base d'une analyse des F_{ST} sur les gènes *core*.

Ceci est également soutenu par les caractéristiques évolutives et fonctionnelles de certains COGs présents dans ces îlots. À titre d'exemple, deux COGs présents dans ISL2 codent des protéines liées à l'utilisation des phosphonates, généralement retrouvées chez des organismes évoluant dans des environnements pauvres en phosphore (Feingersch *et al.*, 2012). Les individus appartenant à l'écotype HLII sont globalement retrouvés dans les eaux de surface dont la concentration en phosphore est limitée, et ainsi la diversité des gènes et des allèles associés au métabolisme du phosphore serait directement liée à la disponibilité de cet élément (Martiny *et al.*, 2006, 2009; Coleman and Chisholm, 2010).

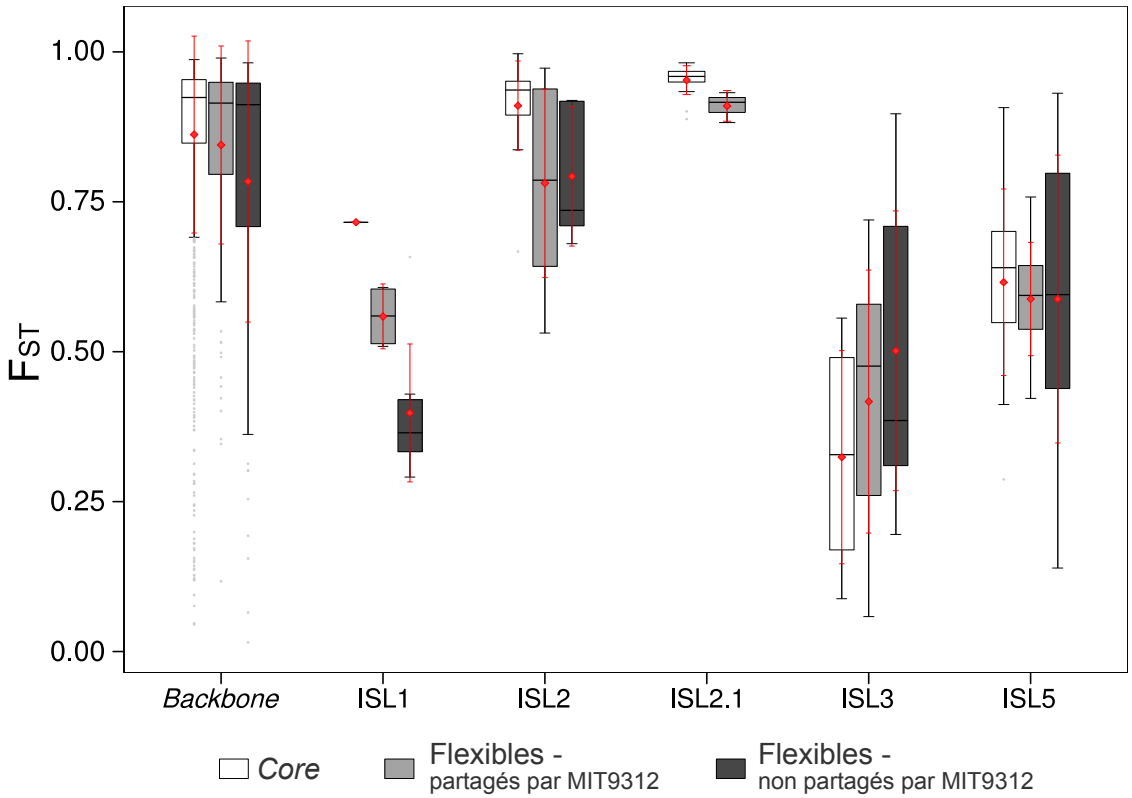


Figure 6.1. Distributions des valeurs de F_{ST} dans les COGs *core* et flexibles partagés ou non par le génome de référence MIT9312 en fonction des compartiments génomiques (*backbone* et ISLs). Les F_{ST} ont été calculés par Kashtan *et al.* (2014) afin d'évaluer la différenciation génomique entre les sous-populations. Rouge : moyenne ± écart-type (sd) ; blanc : gènes *core*, gris clair : gènes flexibles partagés par MIT9312 ; gris foncé : gènes flexibles non partagés par MIT9312.

6.2. Organisation du génome et fluidité du pangénome

Au cours de leurs travaux, López-Pérez et collaborateurs (2014) ont défini deux types d'îlots génomiques, des îlots dits « de remplacement », caractérisés par la présence de gènes non-homologues mais codant pour des fonctions semblables, et des îlots dits « additifs » qui affichent une variabilité de contenu en gènes du fait de HGTs *via* des éléments génétiques mobiles (Figure 6.2).

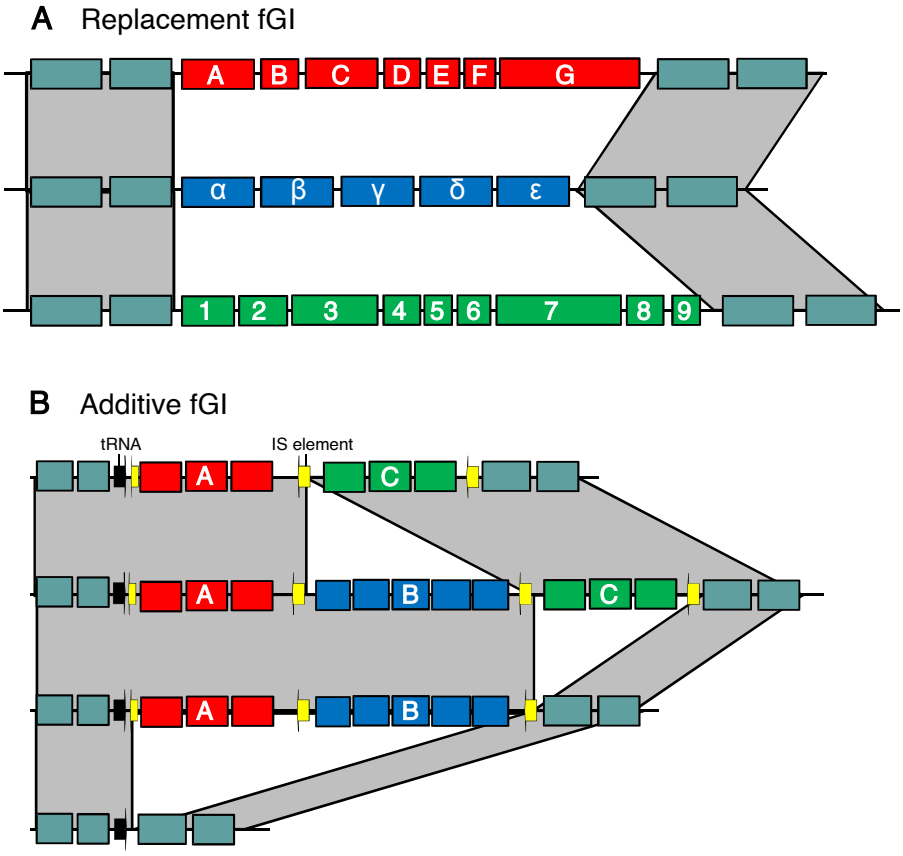


Figure 6.2. Représentation schématique de l'organisation des îlots de remplacement (A) et additifs (B) (Rodriguez-Valera *et al.*, 2016). Pour les îlots de remplacement **(A)**, les lettres, symboles et chiffres indiquent des gènes non-homologues alors que les régions grisées représentent des régions synthéniques et homologues. Pour les îlots additifs **(B)**, plusieurs « cassettes » sont identifiées par différentes lettres et couleurs. Celles-ci sont encadrées par des ARNt et des éléments génétiques mobiles, comme des séquences d'insertion (IS), les deux jouant un rôle dans les transferts horizontaux. fGI : îlots génomiques flexibles.

Sur la base de cette définition, l'intervalle génomique compris entre ISL3 et ISL5, avec une pression de sélection négative relâchée et des COGs peu partagés entre sous-populations, pourrait s'apparenter à un îlot génomique additif. En revanche, l'assimilation de ISL1, ISL2, et ISL2.1 à des îlots de remplacement semble plus osée, malgré la présence de COGs flexibles sous forte pression de sélection, largement partagés entre sous-populations et associés aux catégories fonctionnelles « Processus cellulaires et signalisation », « Transport et métabolisme des ions inorganiques » et « Mécanismes de défense ». Contrairement aux deux autres ISLs, ISL2 aussi affiche un pourcentage de recombinaison parmi les plus élevés le long du génome (Figures 5.4 et 5.6 ; Tableau 5.2), une caractéristique observée pour des îlots de remplacement ayant subi un événement récent (López-Pérez *et al.*, 2014).

Par ailleurs, López-Pérez *et al.* (2014) ont également montré que les îlots de remplacement sont la cible d'événements de recombinaison fréquents impliquant des ensembles de gènes nécessaires à la réalisation d'une fonction plutôt que des gènes isolés. Ces événements sont facilités par la forte conservation de séquences maintenue au voisinage des îlots. Ceci rejoint la notion d'îlots de variabilité introduite par Oliveira et collaborateurs (2017), pour qui le mécanisme de recombinaison homologue repose sur la similarité de séquences entre un ADN exogène et les gènes *core* flanquant les îlots. Au cours de cette thèse, une hétérogénéité de la longueur des fragments recombinants entre SAGs de *Prochlorococcus* a été mise en évidence, associée à une distribution de valeurs étalée vers des tailles de fragments longs. Il serait intéressant de cartographier plus finement ces fragments en fonction de leur longueur afin d'évaluer plus précisément les propriétés des régions qui les bornent (*e.g.*, nature des COGs, degré de conservation des séquences) et les modalités de remplacement des gènes flexibles au sein des îlots.

Les îlots ISL2 et ISL2.1 (et plus généralement les îlots de remplacement), du fait de leur concentration en gènes flexibles essentiels pour l'ensemble des sous-populations, pourraient tenir lieu de «méga-loci » tels qu'ils ont été définis par Schmutzer and Barraclough (2019). En effet, ces auteurs suggèrent qu'en présence de flux de gènes entre des populations divergentes, une forte proportion de gènes localement adaptatifs dans un nombre réduit de loci pourrait être favorisée. Cela permettrait (i) de réduire l'impact négatif des insertions des gènes transférés horizontalement le long du génome et (ii) d'augmenter l'efficacité relative de la sélection sur quelques « méga-loci » par rapport à une dispersion

de nombreux loci à effet réduit. Ce point de vue est en accord avec l'idée d'une coévolution entre les génomes hôtes vis-à-vis de l'intégration d'éléments génétiques mobiles sur des sites cibles (*via* des intégrases par exemple), de manière à ce que le coût lié à la disruption de fonctions suite à leur intégration soit minimal en termes de fitness (Oliveira *et al.*, 2017).

Les îlots ISL3 et ISL4, composés de COGs affiliés à une grande diversité de taxa et essentiellement spécifiques à une sous-population, pourraient être associés à la catégorie des îlots génomiques additifs. Ces caractéristiques suggèrent que ces COGs ont été acquis par HGTs. Ils sont enrichis dans les catégories fonctionnelles « Biogenèse de la paroi cellulaire » et « Mécanismes de défense », catégories également mises en évidence à l'échelle des écotypes HLI et HLII par Coleman et collaborateurs (2006).

L'analyse des événements de recombinaison à l'échelle des sous-populations a par ailleurs montré un pourcentage élevé de gènes *core* en copie unique impliqués dans des processus de recombinaison (>10%) et de COGs recombinants dans les régions génomiques au voisinage de ISL3 et ISL4. Ces résultats soulignent l'importance des régions conservées du *backbone* au voisinage des îlots additifs dans la dynamique évolutive de ces régions. Cependant, ces COGs *core* affichent peu d'événements de recombinaison récents comparativement aux COGs flexibles. Ceci laisse penser que des processus autres que la recombinaison homologue interviennent dans la dynamique de ces régions variables à l'échelle des sous-populations. En outre, il apparaît qu'un nombre d'événements de recombinaison plus important est détecté pour les COGs flexibles clade-spécifiques que pour les COGs *core* ou partagés par MIT9312. Ainsi, les gènes flexibles moins conservés à l'échelle évolutive pourraient prendre un part importante dans la fluidité des génomes dans ces zones de variabilité à des échelles de temps courts au sein des populations.

Les îlots génomiques additifs sont typiquement associés à des éléments génétiques mobiles et ont des contenus en « gènes cassettes » variables et généralement flanqués d'intégrases ou de transposases (Figure 6.2) (López-Pérez *et al.*, 2014). Bien que de tels éléments aient longtemps été difficiles à caractériser chez *Prochlorococcus*, Hackl et collaborateurs (2020) ont récemment identifiés des éléments mobiles apparentés à des transposons de type cargo nommés *tycheposons*. Ceux-ci contiennent des gènes codant des intégrases ciblant des sites spécifiques (comme des gènes d'ARNt et ARNtm) ainsi que des

protéines impliquées dans les processus de réplication autonome (*e.g.*, polymérases, primases). Ils embarquent par ailleurs de longs fragments d'ADN (>10 Kpb), porteurs de gènes de fonctions généralement inconnues, mais aussi dans certains cas, des gènes de *packaging* viraux ou d'interférence avec la machinerie virale (*tycheposons* PICI-like). Lorsqu'ils sont annotés, ces gènes révèlent un large spectre de fonctions à caractère écologique, majoritairement en lien avec l'assimilation de nutriments tels que l'azote, le phosphore ou le fer. Du fait de l'intégration de ces *tycheposons* dans les îlots génomiques *via* les gènes cibles de type ARNt situés à proximité, un « gradient d'insertion temporel » serait induit avec un matériel génétique d'autant plus récent qu'il serait proche du gène ciblé lors de son insertion (Figure 6.3).

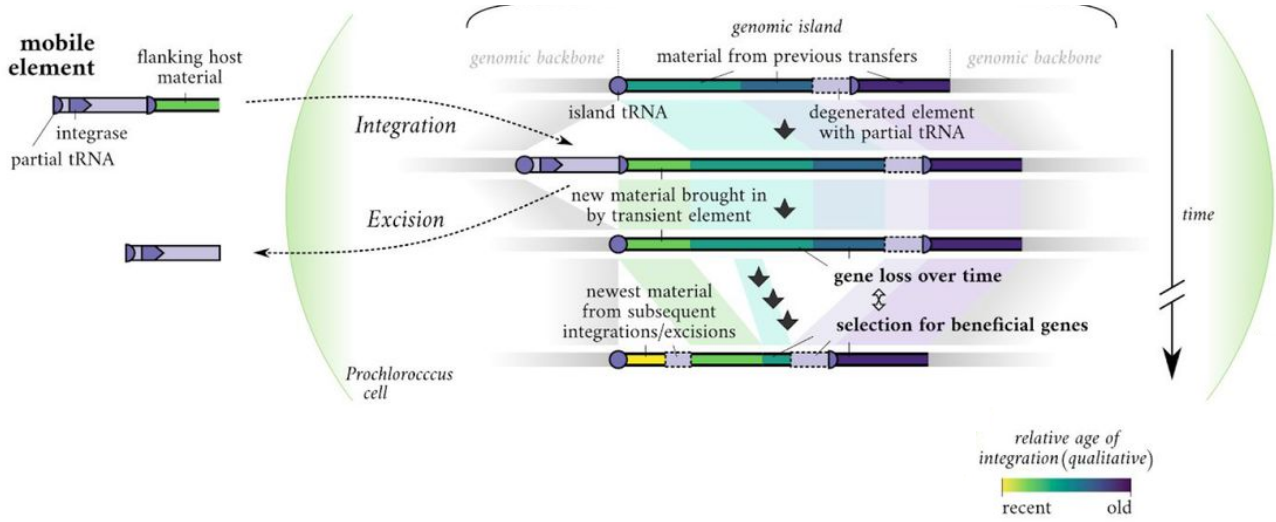


Figure 6.3. Modèle proposé pour la formation d'îlots génomiques *via* l'activité d'éléments génétiques mobiles (Hackl *et al.*, 2020). Les éléments mobiles s'intègrent et s'excisent au niveau du gène de l'ARNt à l'extrémité proximale de l'îlot. Le matériel génétique introduit mais non excisé par la suite, tel que l'ADN flanquant provenant d'autres hôtes et les éléments dégénérés, s'accumule à côté de l'ARNt et entraîne un gain de gènes et la croissance de l'îlot. Le gain de gènes est équilibré par la perte de gènes et la sélection, ne préservant que les acquisitions avantageuses qui peuvent à leur tour être fixées dans la descendance. En raison du « gradient d'insertion temporel » du processus, l'hétérogénéité intra-clade observée est plus élevée à l'extrémité proximale de l'îlot, juste à côté de l'ARNt, et diminue vers l'extrémité distale de l'îlot.

Il est également proposé que les *tycheposons* PICI-like sont en mesure de détourner la machinerie virale lors d'une infection, leur permettant de se disséminer dans la population

tout en réduisant l'efficacité de dissémination du phage. Cependant la majorité des *tyche-posons* détectés ne semblent pas en mesure d'interférer avec les virus. Ils pourraient par contre se propager dans les populations *via* la dispersion de vésicules extracellulaires, phénomène récemment appelé vésiduction (Soler and Forterre, 2020). La production de vésicules a en effet été décrite dans le genre *Prochlorococcus* (Biller *et al.*, 2014c).

6.3. Évolution du pangénome

L'analyse des signatures de sélection dans le génome flexible a révélé deux ensembles de COGs. Le premier ensemble est caractérisé par des valeurs de *dN* faibles et de *dS* inférieures à la moyenne de celles observées pour les *dS* des gènes *core*. Les COGs de cet ensemble sont retrouvés dans tous les compartiments du génome à l'exception de ISL4, et sont les composants exclusifs du génome flexible de ISL1, ISL2 et ISL2.1. Bien que cela soit cohérent avec une sélection d'arrière plan (Price and Arkin, 2015), les faibles valeurs de *dS*, entraînant des ratios *dN/dS* particulièrement élevés, pourraient également refléter une sélection négative sur les substitutions synonymes ou une homogénéisation de la diversité des séquences entre les clades par RH (Hanage, 2016). L'existence d'une faible pression de sélection négative au niveau des positions synonymes a été montrée à de nombreuses reprises, en lien avec (i) le choix d'usage de codons préférentiel induisant une meilleure fidélité et efficacité de la traduction (Gouy and Gautier, 1982; Comeron *et al.*, 1999; Sharp *et al.*, 2005), (ii) la régulation transcriptionnelle (Parmley and Hurst, 2007; Thorpe *et al.*, 2017) ou (iii) la composition nucléotidique (GC %) des génomes (Hildebrand *et al.*, 2010; Rocha and Feil, 2010; Raghavan *et al.*, 2012). Cette dernière hypothèse est proposée pour expliquer la faible teneur en nucléotides GC des génomes de l'écotype HLII (autour de 30%), traduisant une adaptation à un environnement oligotrophe (*i.e.*, faible teneur en azote et phosphore) (Grzymski and Dussaq, 2011; Giovannoni *et al.*, 2014). Cette hypothèse s'applique cependant à l'ensemble du génome, et ne permet pas d'expliquer les profils de *dN* et *dS* associés aux COGs des ISL1, ISL2 et ISL2.1. Aucune sélection transcriptionnelle/traductionnelle n'a en revanche été détectée (Yu *et al.*, 2012; Batut *et al.*, 2014).

D'après des analyses évolutives réalisées sur les génomes, les variations du génome flexible seraient liées aux HGTs (*i.e.*, intégration de gènes non-homologues), tandis que la RH contribuerait à l'homogénéisation du génome *core* (*i.e.*, conversion génomique) (Polz *et al.*, 2013). La forte similitude des séquences associées aux COGs caractérisés par des valeurs de *dN* faibles, notamment ceux retrouvés dans ISL2 et ISL2.1 en passe de devenir *core* dans les sous-populations, pourrait dès lors résulter de RH. De plus, celle-ci pourrait également augmenter l'efficacité de la sélection en réduisant l'effet Hill-Robertson (perte

de diversité génétique du fait d'une réduction de l'efficacité de la sélection dans les populations de petite taille) (Hill and Robertson, 1966). Dans un contexte de populations structurées, la RH inter-clade pourrait également augmenter la taille efficace des populations (*Ne*) pour les gènes dont la circulation s'étendrait au-delà des sous-populations, tout en limitant la différenciation des clades dans une gamme de divergence limitée (Marttinen *et al.*, 2015). Ceci est conforme à la nature combinatoire des différents allèles portés par les gènes *core* du *backbone* et des gènes flexibles, comme le proposent Kashtan et collaborateurs (2014) pour les gènes liés à l'adaptation aux fortes intensités lumineuses ou aux phosphonates.

Le deuxième ensemble de COGs, principalement retrouvé dans ISL3, ISL4 et ISL5, est caractérisé par un relâchement des contraintes sélectives associées à des valeurs de *dS* élevées, suggérant des événements de HGTs (Castillo-Ramírez et al., 2011). En outre, l'analyse taxonomique de ces COGs a montré qu'ils étaient majoritairement affiliés à des taxa autres que les cyanobactéries. Les HGTs sont reconnus comme étant un moteur de l'évolution, contribuant à l'adaptation des organismes aux environnements changeants par l'expansion et la conversion de familles de gènes (Ochman *et al.*, 2000; Gogarten *et al.*, 2002; Wiedenbeck and Cohan, 2011). Par conséquent, la surreprésentation des gènes impliqués dans les « Mécanismes de défense » ou la « Biogénèse de la paroi cellulaire » dans ISL3 et ISL4 refléterait l'acquisition de gènes qui pourraient être transitoirement adaptatifs, comme par exemple, lors des périodes d'infection par des phages (Coleman *et al.*, 2006; Kettler *et al.*, 2007; Avrani *et al.*, 2011). Cependant, les HGTs adaptatifs sont principalement documentés pour les gènes présentant un avantage sélectif important, ce qui pourrait ne pas être vrai pour la plupart d'entre eux.

Dans le cas d'un HGT quasi neutre, la sélection dépend de *Ne* (Kuo *et al.*, 2009). Pour McInerney et collaborateurs (McInerney *et al.*, 2017), étant donné qu'une grande *Ne* augmente de fait l'efficacité de la sélection, le génome flexible est essentiellement adaptatif, les gènes légèrement délétères acquis étant rapidement éliminés. Cependant, il est également attendu que ces populations portent un nombre plus important d'allèles quasi neutres. Ainsi, l'hypothèse d'une évolution neutre du pangénome ne peut être rejetée, dans la mesure où celle-ci est alors plus parcimonieuse (Baumdicker *et al.*, 2012; Andreani *et al.*, 2017; Vos and Eyre-Walker, 2017). Dans un troisième modèle, celui de barrière à la dérive – ou *drift-barrier model* (Lynch, 2010; Bobay and Ochman, 2018), la perte de

gènes flexibles est aléatoire pour une population avec une petite *Ne*, alors qu'une grande *Ne* augmente :

- la proportion de gènes avec un coefficient de sélection $s < 0$, qui sont alors perçus comme délétères ;
- la probabilité de fixation des gènes légèrement avantageux ;
- le temps de fixation ou la perte de gènes quasi neutres, et donc la taille et la diversité du génome flexible.

Il a été proposé que la taille du pangénome chez *Prochlorococcus* pourrait provenir de leur grande *Ne* [estimé entre 10^6 (Price and Arkin, 2015) et 10^{13} (Kashtan *et al.*, 2014)]. Cependant, dans le cas de populations structurées, comme on l'observe ici, l'évolution d'un gène presque neutre acquis à partir d'une lignée éloignée pourrait être limitée au clade dans lequel il a été introduit. Ainsi, un HGT distant pourrait évoluer dans un contexte de *Ne* inférieure à la taille de la population totale, un schéma qui pourrait expliquer les caractéristiques de ISL4 (c'est-à-dire enrichi en COGs spécifiques à un clade, impliqués dans des mécanismes de défense et affiliés à des taxa autres que les cyanobactéries). Cependant, la *Ne* associée aux gènes transférés pourrait également être élargie, par exemple par des événements locaux de RH, favorisant une empreinte sélective et la persistance de gènes acquis avec un effet marginal. L'apparition à la fois de HGTs (*dS* élevés avec affiliation incertaine) et de gènes sélectionnés (*dN/dS* faibles avec affiliation à *Prochlorococcus*) dans les ISL3 et ISL5 pourrait être le résultat de tels processus.

6.4. Conclusions et Perspectives

6.4.1. Résumé des principaux résultats

Les travaux réalisés au cours de cette thèse visaient à mieux comprendre les mécanismes mis en jeu lors de la différenciation de populations bactériennes libres de l'environnement, avec pour modèle un ensemble de SAGs de sous-populations cooccurrentes de l'écotype HLII de *Prochlorococcus*.

Une partie de mes recherches a tout d'abord porté sur une analyse du pangénome en termes de contenu en gènes. Celle-ci a révélé la nature ouverte du pangénome à cette échelle de diversité, en accord avec une augmentation du nombre de SAGs se traduisant par une augmentation du nombre de COGs, pour la plupart SAG-spécifiques. Ce pangénome se caractérise, par ailleurs, par un paysage génomique comprenant des régions conservées (*backbone*) et variables (ISLs) – constituées majoritairement de gènes *core* et de gènes flexibles, respectivement. Dès lors, je me suis intéressée aux raisons de cette organisation des COGs des génomes *core* et flexible, tant d'un point de vue fonctionnel qu'évolutif. Une répartition non aléatoire des fonctions portées par les gènes flexibles, associée à une dynamique évolutive différente des compartiments génomiques, a été mise en évidence. Cette observation amène à la question de la relation entre les flux de gènes, *via* des événements de recombinaison, et cette compartimentation. Ainsi, des point chauds de recombinaison ont été détectés au sein des génomes de l'ensemble des clades, suggérant une répartition spatiale particulière de ces événements, dont il reste à déterminer si elle est la cause d'une occurrence non aléatoire de ces événements ou le fruit d'une rétention différentielle des séquences recombinantes le long du génome.

6.4.2. Limites associées aux données issues de séquençage SCG

Par le biais de cette étude, il a été montré que le choix du jeu de données est d'une grande importance. Un jeu de données réduit aux seuls SAGs les plus complets entraîne-

rait une interprétation erronée de la nature du pangénome, avec une sur-estimation du nombre de COGs clade-spécifiques.

Parallèlement aux problèmes de complétude, les projets de séquençage SCG restent encore relativement peu développés à grande échelle. Est-il alors envisageable d'étudier la dynamique des génomes, à l'échelle des populations, à partir de l'assemblage de métagénomes ? En effet, à ce jour, les données métagénomiques provenant d'une large sélection d'écosystèmes constituent la majorité des séquences disponibles dans les bases de données publiques (Qin *et al.*, 2010; Bork *et al.*, 2015). De plus, l'analyse des métagénomes à l'échelle populationnelle et l'étude de la diversité intra-spécifique sont devenues familières avec des outils tels que metaSNV (Costea *et al.*, 2017), ConStrains (Luo *et al.*, 2015), StrainPhlAn (Truong *et al.*, 2017), POGENOM (Sjöqvist *et al.*, 2021), inStrain (Olm *et al.*, 2021) ou STRONG (Quince *et al.*, 2021). Cependant, ces données contiennent des éléments génomiques provenant de nombreux organismes différents et en pratique, leur assemblage peut conduire à la reconstruction de séquences génomiques chimériques (Greenwald *et al.*, 2017). Celles-ci résultent par exemple (i) d'artefacts liés à l'expérimentation, (ii) de l'assemblage de régions très similaires entre des génomes coexistants dans une même communauté microbienne, ou encore (iii) du choix de l'assembleur (Sinha *et al.*, 2015). Dans ce contexte, il apparaît pertinent de proposer une nouvelle méthodologie pour évaluer le chimérisme à l'échelle populationnelle. Celle-ci pourrait reposer sur la simulation de métagénomes, de complexité croissante, à partir d'un sous-ensemble des SAGs de *Prochlorococcus*, répartis sur différents clades (C1, C2, C3 et C8) de l'écotype HLII et en association avec des métagénomes réels. Dans un premier temps, les données ainsi produites pourraient être assemblées à l'aide de différents assembleurs, sélectionnés sur la base de leur performance sur différentes communautés microbiennes (Greenwald *et al.*, 2017). Les *contigs* chimériques, provenant de l'assemblage d'au moins deux SAGs appartenant à un même clade ou à deux clades différents, pourront être détectés par l'alignement des lectures sur les *contigs* nouvellement reconstruits. La quantification et la mesure du chimérisme pourra alors se faire à l'aide de deux métriques :

- l'entropie de Shannon (Shannon, 1948) pour caractériser le chimérisme au sein d'un *contig*, *i.e.*, quels SAGs / clades sont présents et dans quelle proportion ? ;

- la divergence de Kullback-Leibler (Kullback and Leibler, 1951) pour caractériser le chimérisme à l'échelle du métagénome, *i.e.*, la distribution observée des SAGs / clades diffère-t-elle de la distribution théorique dans le métagénome ?

Dans un deuxième temps, les facteurs qui pourraient influencer le degré de chimérisme, tels que (i) l'abondance des SAGs / clades au sein des métagénomes simulés, (ii) la structure et la complexité de la communauté, (iii) la parenté phylogénétique et (iv) le contenu génétique, pourront être évalués.

6.4.3. Contraintes environnementales, flux de gènes et évolution des pangénomes

Dans une phylogénie génomique comprenant des SAGs du Pacifique et des BATS, Kashtan et collaborateurs (2017) identifient au moins deux sous-populations co-occurrentes du Pacifique phylogénétiquement distinctes de celles des BATS. Les populations de ces deux régions du globe sont par ailleurs caractérisées par des ensembles de gènes flexibles distincts. Une analyse comparative de l'occurrence et des processus de sélection qui s'appliquent sur les gènes flexibles contraints (comme ceux de ISL2 et ISL2.1) à l'échelle des sous-populations Atlantique et Pacifique pourrait permettre une évaluation plus fine du caractère essentiel et adaptatif des gènes flexibles sous forte contrainte sélective.
Parallèlement à l'évaluation de cette hypothèse sur différentes populations d'un même écotype, il serait intéressant d'étudier les populations d'autres écotypes, voire d'autres espèces dont on sait qu'ils répondent plus fortement aux contraintes de l'environnement. C'est par exemple le cas de l'écotype HLI, pour lequel une corrélation plus forte entre saisonnalité et facteurs environnementaux a été observée (Larkin *et al.*, 2016).

La répartition non aléatoire des gènes flexibles et l'analyse de la RH ont permis de mettre en évidence des compartiments génomiques particulièrement soumis aux flux de gènes. De plus, il existerait une relation entre la répartition des gènes transférés et l'ances-

tralité des événements qui leurs sont associés, apportant des éléments de réponse concernant la fluidité des génomes sur différentes périodes évolutives.

Au cours de ces travaux, seuls les gènes issus de RH, impliquant la conversion de gènes homologues, ont été identifiés. Il conviendrait donc d'étudier plus finement les processus de recombinaison hétérologue et confirmer l'origine exogène des gènes potentiellement issus de HGTs (*i.e.*, apports de matériel génétique nouveau), identifiés *via* les affiliations taxonomiques, par des approches paramétriques (*i.e.*, basées sur des signatures génomiques spécifiques telles que la composition en dinucléotides GC) et phylogénétiques.

L'évaluation des contraintes sélectives associées à l'ensemble des gènes transférés (*i.e.*, issus de RH et HGTs) pourrait nous éclairer sur leur trajectoire évolutive. Ceci permettrait dans un premier temps d'établir un lien entre les gains et les pertes de gènes transférés et leur répartition spatiale dans les différents compartiments génomiques. Plus spécifiquement, l'analyse longitudinale des pressions de sélection permettrait de tester l'impact relatif des *tycheposons* sur la dynamique évolutive des ISL1, ISL2 et ISL2.1 (potentiellement îlots de remplacement) *versus* ISL3, ISL4 et ISL5 (*a priori* additifs), dans la mesure où il a été proposé l'existence d'un « gradient d'insertion temporel » des gènes transférés par le biais de ces éléments mobiles (Hackl *et al.*, 2020).

Les travaux réalisés au cours de cette thèse ont permis d'avancer l'hypothèse que l'évolution du pangénome analysé est liée à une rétention différentielle des gènes transférés, conséquence d'une fluctuation de *Ne* le long du génome. Des variations intra-génomiques de *Ne* ont d'ores et déjà été constatées chez de nombreuses espèces eucaryotes (Gossmann *et al.*, 2011). Par ailleurs, le brassage de ces populations par le biais d'événements de recombinaison peut engendrer une variation de *Ne* sur certains loci le long du génome (Nordborg, 1997). Si les fluctuations de *Ne* ont été suggérées, elles ne sont cependant pas démontrées pour notre modèle d'étude. L'évaluation de *Ne* à l'échelle des gènes, le long du génome, constitue donc une perspective importante de ces travaux.

Bibliographie

Acinas, S. G., **Marcelino, L. A.**, **Klepac-Ceraj, V.** and **Polz, M. F.** (2004) 'Divergence and Redundancy of 16S rRNA Sequences in Genomes with Multiple rrn Operons', *Journal of Bacteriology*, 186(9), pp. 2629–2635. doi: 10.1128/JB.186.9.2629-2635.2004.

Ahlgren, N. A., **Rocap, G.** and **Chisholm, S. W.** (2006) 'Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies', *Environmental Microbiology*, 8(3), pp. 441–454. doi: 10.1111/j.1462-2920.2005.00910.x.

Altschul, S. F., **Gish, W.**, **Miller, W.**, **Myers, E. W.** and **Lipman, D. J.** (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Andersson, S. G. e. and **Kurland, C. G.** (1998) 'Reductive evolution of resident genomes', *Trends in Microbiology*. Elsevier Current Trends, pp. 263–268. doi: 10.1016/S0966-842X(98)01312-2.

Andreani, N. A., **Hesse, E.** and **Vos, M.** (2017) 'Prokaryote genome fluidity is dependent on effective population size', *The ISME Journal*, 11(7), pp. 1719–1721. doi: 10.1038/ismej.2017.36.

Arevalo, P., **VanInsberghe, D.**, **Elsherbini, J.**, **Gore, J.** and **Polz, M. F.** (2019) 'A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations', *Cell*, 178(4), pp. 820--834.e14. doi: 10.1016/J.CELL.2019.06.033.

Avery, O. T., **MacLeod, C. M.** and **McCarty, M.** (1944) 'Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III', *Journal of experimental medicine*, 79(2), pp. 137–158.

Avrani, S., **Wurtzel, O.**, **Sharon, I.**, **Sorek, R.** and **Lindell, D.** (2011) 'Genomic island variability facilitates *Prochlorococcus*–virus coexistence', *Nature*, 474(7353), pp. 604–608. doi: 10.1038/nature10172.

Bakker, H. C. den, **Cummings, C. A.**, **Ferreira, V.**, **Vatta, P.**, **Orsi, R. H.**, **Degoricija, L.**, **Barker, M.**, **Petrauskene, O.**, **Furtado, M. R.** and **Wiedmann, M.** (2010) 'Comparative genomics of the bacterial genus Listeria : Genome evolution is characterized by limited gene acquisition and limited gene loss', *BMC Genomics 2010 11:1*, 11(1), pp. 1–20. doi: 10.1186/1471-2164-11-688.

Bao, Y. J., **Shapiro, B. J.**, **Lee, S. W.**, **Ploplis, V. A.** and **Castellino, F. J.** (2016) 'Phenotypic differentiation of Streptococcus pyogenes populations is induced by recombination-driven gene-specific sweeps', *Scientific Reports*, 6(1), pp. 1–13. doi: 10.1038/srep36644.

Batut, B., **Knibbe, C.**, **Marais, G.** and **Daubin, V.** (2014) 'Reductive genome evolution at both ends of the bacterial population size spectrum', *Nature Reviews Microbiology*, 12(12), pp. 841–850. doi: 10.1038/nrmicro3331.

Baumdicker, F., **Hess, W. R.** and **Pfaffelhuber, P.** (2012) 'The infinitely many genes model for the distributed genome of bacteria.', *Genome biology and evolution*, 4(4), pp. 443–456. doi: 10.1093/gbe/evs016.

Bendall, M. L., Stevens, S. L. R., Chan, L.-K., Malfatti, S., Schwientek, P., Tremblay, J., Schackwitz, W., Martin, J., Pati, A., Bushnell, B., *et al.* (2016) 'Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations', *The ISME Journal*, 10(7), pp. 1589–1601. doi: 10.1038/ismej.2015.241.

Bernardo, J. M. and **Smith, A. F.** (2001) 'Bayesian theory, West Susex, England'. John Wiley \& Sons.

Berube, P. M., Biller, S. J., Hackl, T., Hogle, S. L., Satinsky, B. M., Becker, J. W., Braakman, R., Collins, S. B., Kelly, L., Berta-Thompson, J., *et al.* (2018) 'Single cell genomes of *Prochlorococcus*, *Synechococcus*, and sympatric microbes from diverse marine environments', *Scientific Data*, 5, p. 180154. doi: 10.1038/sdata.2018.154.

Biller, S. J., Berube, P. M., Berta-Thompson, J. W., Kelly, L., Roggensack, S. E., Awad, L., Roache-Johnson, K. H., Ding, H., Giovannoni, S. J., Rocap, G., *et al.* (2014a) 'Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*', *Scientific Data*, 1, p. 140034. doi: 10.1038/sdata.2014.34.

Biller, S. J., **Berube, P. M.**, **Lindell, D.** and **Chisholm, S. W.** (2014b) '*Prochlorococcus*: the structure and function of collective diversity', *Nature Reviews Microbiology*, 13(1), pp. 13–27. doi: 10.1038/nrmicro3378.

Biller, S. J., **Schubotz, F.**, **Roggensack, S. E.**, **Thompson, A. W.**, **Summons, R. E.** and **Chisholm, S. W.** (2014c) 'Bacterial Vesicles in Marine Ecosystems', *Science*, 343(6167), pp. 183–186. doi: 10.1126/SCIENCE.1243457.

Bobay, L.-M. (2020) 'The prokaryotic species concept and challenges', *The Pangenome*, pp. 21–49.

Bobay, L.-M. and **Ochman, H.** (2017) 'Biological Species Are Universal across Life's Domains', *Genome Biology and Evolution*, 9(3), pp. 491–501. doi: 10.1093/gbe/evx026.

Bobay, L.-M. and **Ochman, H.** (2018) 'Factors driving effective population size and pan-genome evolution in bacteria', *BMC Evolutionary Biology*, 18(1), p. 153. doi: 10.1186/s12862-018-1272-4.

Bolotin, E. and **Hershberg, R.** (2016) 'Bacterial intra-species gene loss occurs in a largely clocklike manner mostly within a pool of less conserved and constrained genes', *Scientific Reports 2016 6:1*, 6(1), pp. 1–9. doi: 10.1038/srep35168.

Bork, P., **Bowler, C.**, **De Vargas, C.**, **Gorsky, G.**, **Karsenti, E.** and **Wincker, P.** (2015) 'Tara Oceans studies plankton at Planetary scale', *Science*, 348(6237), p. 873. doi: 10.1126/SCIENCE.AAC5605.

Brockhurst, M. A., **Harrison, E.**, **Hall, J. P. J.**, **Richards, T.**, **McNally, A.** and **MacLean, C.** (2019) 'The Ecology and Evolution of Pangenomes', *Current Biology*, 29(20), pp. R1094--R1103. doi: 10.1016/J.CUB.2019.08.012.

Bruen, T. and **Bruen, T.** (2005) 'PhiPack: PHI test and other tests of recombination', *McGill University, Montreal, Quebec*.

Bruen, T. C., **Philippe, H.** and **Bryant, D.** (2006) 'A Simple and Robust Statistical Test for Detecting the Presence of Recombination', *Genetics*, 172(4).

Bryant, D. and **Moulton, V.** (2004) 'Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks', *Molecular Biology and Evolution*, 21(2), pp. 255–265. doi: 10.1093/molbev/msh018.

Cadillo-Quiroz, H., **Didelot, X.**, **Held, N. L.**, **Herrera, A.**, **Darling, A.**, **Reno, M. L.**, **Krause, D. J.** and **Whitaker, R. J.** (2012) 'Patterns of Gene Flow Define Species of Thermophilic Archaea', *PLoS Biology*. Edited by N. H. Barton, 10(2), p. e1001265. doi: 10.1371/journal.pbio.1001265.

Camacho, C., **Coulouris, G.**, **Avagyan, V.**, **Ma, N.**, **Papadopoulos, J.**, **Bealer, K.** and **Madden, T. L.** (2009) 'BLAST+: architecture and applications', *BMC Bioinformatics*, 10(1), p. 421. doi: 10.1186/1471-2105-10-421.

Castresana, J. (2000) 'Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis', *Molecular Biology and Evolution*, 17(4), pp. 540–552. doi: 10.1093/oxfordjournals.molbev.a026334.

Chan, C. X., **Beiko, R. G.** and **Ragan, M. A.** (2006) 'Detecting recombination in evolving nucleotide sequences', *BMC Bioinformatics*, 7(1), p. 412. doi: 10.1186/1471-2105-7-412.

Charlesworth, B., **Morgan, M. T.** and **Charlesworth, D.** (1993) 'The effect of deleterious mutations on neutral molecular variation.', *Genetics*, 134(4), pp. 1289–1303.

Cohan, F. M. (2001) 'Bacterial Species and Speciation', *Systematic Biology*. Edited by M. Kane, 50(4), pp. 513–524. doi: 10.1080/10635150118398.

Cohan, F. M. and **Perry, E. B.** (2007) 'A Systematics for Discovering the Fundamental Units of Bacterial Diversity', *Current Biology*. Cell Press, pp. R373--R386. doi: 10.1016/j.cub.2007.03.032.

Coleman, M. L. and **Chisholm, S. W.** (2010) 'Ecosystem-specific selection pressures revealed through comparative population genomics.', *Proceedings of the National*

Academy of Sciences of the United States of America, 107(43), pp. 18634–18639. doi: 10.1073/pnas.1009480107.

Coleman, M. L., **Sullivan, M. B.**, **Martiny, A. C.**, **Steglich, C.**, **Barry, K.**, **DeLong, E. F.** and **Chisholm, S. W.** (2006) 'Genomic Islands and the Ecology and Evolution of *Prochlorococcus*', *Science*, 311(5768).

Collins, R. E. and **Higgs, P. G.** (2012) 'Testing the Infinitely Many Genes Model for the Evolution of the Bacterial Core Genome and Pangenome', *Molecular Biology and Evolution*, 29(11), pp. 3413–3425. doi: 10.1093/molbev/mss163.

Comeron, J. M., **Kreitman, M.** and **Aguadé, M.** (1999) 'Natural Selection on Synonymous Sites Is Correlated With Gene Length and Recombination in Drosophila', *Genetics*, 151(1), pp. 239–249. doi: 10.1093/GENETICS/151.1.239.

Costea, P. I., **Munch, R.**, **Coelho, L. P.**, **Paoli, L.**, **Sunagawa, S.** and **Bork, P.** (2017) 'metaSNV: A tool for metagenomic strain level analysis', *PLOS ONE*. Edited by K. Wang, 12(7), p. e0182392. doi: 10.1371/journal.pone.0182392.

Croucher, N. J., Harris, S. R., Fraser, C., Quail, M. A., Burton, J., Van Der Linden, M., McGee, L., Von Gottberg, A., Song, J. H., Ko, K. S., *et al.* (2011) 'Rapid pneumococcal evolution in response to clinical interventions', *Science*, 331(6016), pp. 430–434. doi: 10.1126/science.1198545.

Darling, A. C. E., **Mau, B.**, **Blattner, F. R.** and **Perna, N. T.** (2004) 'Mauve: Multiple alignment of conserved genomic sequence with rearrangements', *Genome Research*, 14(7), pp. 1394–1403. doi: 10.1101/gr.2289704.

Darling, A. C. E., **Mau, B.** and **Perna, N. T.** (2010) 'progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement', *PLoS ONE*. Edited by J. E. Stajich, 5(6), p. e11147. doi: 10.1371/journal.pone.0011147.

Darriba, D., **Taboada, G. L.**, **Doallo, R.** and **Posada, D.** (2012) 'JModelTest 2: More models, new heuristics and parallel computing', *Nature Methods*. Nature Publishing Group, p. 772. doi: 10.1038/nmeth.2109.

Delmont, T. O. and **Eren, A. M.** (2018) 'Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome', *PeerJ*, 6, p. e4320. doi: 10.7717/peerj.4320.

Denef, V. J., **Kalnejais, L. H.**, **Mueller, R. S.**, **Wilmes, P.**, **Baker, B. J.**, **Thomas, B. C.**, **VerBerkmoes, N. C.**, **Hettich, R. L.** and **Banfield, J. F.** (2010) 'Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities', *Proceedings of the National Academy of Sciences*, 107(6), pp. 2383–2390. doi: 10.1073/PNAS.0907041107.

Didelot, X., **Méric, G.**, **Falush, D.** and **Darling, A. E.** (2012) 'Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli.', *BMC genomics*, 13, p. 256. doi: 10.1186/1471-2164-13-256.

Didelot, X., **Walker, A. S.**, **Peto, T. E.**, **Crook, D. W.** and **Wilson, D. J.** (2016) 'Within-host evolution of bacterial pathogens', *Nature Reviews Microbiology 2016 14:3*, 14(3), pp. 150–162. doi: 10.1038/nrmicro.2015.13.

Didelot, X., Wilson, D. J., Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A., Crook, D. W., Köser, C. U., Ellington, M. J., Cartwright, E. J. P., *et al.* (2015) 'ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes', *PLOS Computational Biology*. Edited by A. Prlic, 11(2), p. e1004041. doi: 10.1371/journal.pcbi.1004041.

Dobrindt, U., **Hochhut, B.**, **Hentschel, U.** and **Hacker, J.** (2004) 'Genomic islands in pathogenic and environmental microorganisms', *Nature Reviews Microbiology 2004 2:5*, 2(5), pp. 414–424. doi: 10.1038/nrmicro884.

Domingo-Sananes, M. R. and **McInerney, J. O.** (2021) 'Mechanisms That Shape Microbial Pangenomes', *Trends in Microbiology*. doi: 10.1016/j.tim.2020.12.004.

Dufresne, A., **Garczarek, L.** and **Partensky, F.** (2005) 'Accelerated evolution associated with genome reduction in a free-living prokaryote', *Genome Biology*, 6(2), p. R14. doi: 10.1186/gb-2005-6-2-r14.

Dufresne, A., Ostrowski, M., Scanlan, D. J., Garczarek, L., Mazard, S., Palenik, B. P., Paulsen, I. T., Tandeau de Marsac, N., Wincker, P., Dossat, C., *et al.* (2008) 'Unravelling the genomic mosaic of a ubiquitous genus of marine cyanobacteria', *Genome Biology*, 9(5), p. R90. doi: 10.1186/gb-2008-9-5-r90.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., *et al.* (2003) 'Genome sequence of the cyanobacterium *Prochlorococcus* marinus SS120, a nearly minimal oxyphototrophic genome', *Proceedings of the National Academy of Sciences of the United States of America*, 100(17), pp. 10020–10025. doi: 10.1073/pnas.1733211100.

Dutilh, B. E., **van Noort, V.**, **van der Heijden, R. T. J. M.**, **Boekhout, T.**, **Snel, B.** and **Huynen, M. A.** (2007) 'Assessment of phylogenomic and orthology approaches for phylogenetic inference', *Bioinformatics*, 23(7), pp. 815–824. doi: 10.1093/bioinformatics/btm015.

Eddy, S. R. (2009) 'A new generation of homology search tools based on probalistic inference', in *Genome Informatics 2009*. Published by Imperial College Press and distributed by Wolrd Scientific Publishing Co., pp. 205–211. doi: 10.1142/9781848165632_0019.

Ellegren, H. (2008) 'Comparative genomics and the study of evolution by natural selection', *Molecular Ecology*, 17(21), pp. 4586–4596. doi: 10.1111/j.1365-294X.2008.03954.x.

Feil, E. J. (2004) 'Small change: keeping pace with microevolution', *Nature Reviews Microbiology 2004 2:6*, 2(6), pp. 483–495. doi: 10.1038/nrmicro904.

Feingersch, R., **Philosof, A.**, **Mejuch, T.**, **Glaser, F.**, **Alalouf, O.**, **Shoham, Y.** and **Béjà, O.** (2012) 'Potential for phosphite and phosphonate utilization by *Prochlorococcus*', *ISME Journal*, 6(4), pp. 827–834. doi: 10.1038/ismej.2011.149.

Fitch, W. M. (1970) 'Distinguishing Homologous from Analogous Proteins', *Systematic Zoology*, 19(2), p. 99. doi: 10.2307/2412448.

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K. W., Lomas, M. W., Veneziano, D., *et al.* (2013) 'Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*.', *Proceedings of the National Academy of Sciences of the United States of America*, 110(24), pp. 9824–9829. doi: 10.1073/pnas.1307701110.

Fraser, C., **Hanage, W. P.** and **Spratt, B. G.** (2007) 'Recombination and the nature of bacterial speciation', *Science*. American Association for the Advancement of Science, pp. 476–480. doi: 10.1126/science.1127573.

Garud, N. R., **Good, B. H.**, **Hallatschek, O.** and **Pollard, K. S.** (2019) 'Evolutionary dynamics of bacteria in the gut microbiome within and across hosts', *PLoS biology*, 17(1), p. e3000102. doi: 10.1371/journal.pbio.3000102.

Gillings, M. R. (2016) 'Lateral gene transfer, bacterial genome evolution, and the Anthropocene', *Annals of the New York Academy of Sciences*. doi: 10.1111/nyas.13213.

Giovannoni, S. J., **Cameron Thrash, J.** and **Temperton, B.** (2014) 'Implications of streamlining theory for microbial ecology', *ISME Journal*. Nature Publishing Group, pp. 1553–1565. doi: 10.1038/ismej.2014.60.

Gogarten, J. P., **Doolittle, W. F.** and **Lawrence, J. G.** (2002) 'Prokaryotic Evolution in Light of Gene Transfer', *Molecular Biology and Evolution*, 19(12), pp. 2226–2238. doi: 10.1093/oxfordjournals.molbev.a004046.

Gogarten, J. P. and **Townsend, J. P.** (2005) 'Horizontal gene transfer, genome innovation and evolution', *Nature Reviews Microbiology*. Nature Publishing Group, pp. 679–687. doi: 10.1038/nrmicro1204.

Goris, J., **Konstantinidis, K. T.**, **Klappenbach, J. A.**, **Coenye, T.**, **Vandamme, P.** and **Tiedje, J. M.** (2007) 'DNA–DNA hybridization values and their relationship to whole-genome sequence similarities', *International Journal of Systematic and Evolutionary Microbiology*, 57(1), pp. 81–91. doi: 10.1099/ijs.0.64483-0.

Gossmann, T. I., **Woolfit, M.** and **Eyre-Walker, A.** (2011) 'Quantifying the variation in the effective population size within a genome.', *Genetics*, 189(4), pp. 1389–1402. doi: 10.1534/genetics.111.132654.

Gouy, M. and **Gautier, C.** (1982) 'Codon usage in bacteria: correlation with gene expressivity', *Nucleic Acids Research*, 10(22), pp. 7055–7074. doi: 10.1093/NAR/10.22.7055.

Greenwald, W. W., **Klitgord, N.**, **Seguritan, V.**, **Yooseph, S.**, **Venter, J. C.**, **Garner, C.**, **Nelson, K. E.** and **Li, W.** (2017) 'Utilization of defined microbial communities enables effective evaluation of meta-genomic assemblies', *BMC Genomics 2017 18:1*, 18(1), pp. 1–11. doi: 10.1186/S12864-017-3679-5.

Griffith, F. (1928) 'The Significance of Pneumococcal Types.', *The Journal of hygiene*, 27(2), pp. 113–159.

Grzymski, J. J. and **Dussaq, A. M.** (2011) 'The significance of nitrogen cost minimization in proteomes of marine microorganisms', *The ISME Journal 2012 6:1*, 6(1), pp. 71–80. doi: 10.1038/ismej.2011.72.

Guindon, S., **Dufayard, J.-F.**, **Lefort, V.**, **Anisimova, M.**, **Hordijk, W.** and **Gascuel, O.** (2010) 'New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0', *Systematic Biology*, 59(3), pp. 307–321. doi: 10.1093/sysbio/syq010.

Guttman, D. S. and **Dykhuizen, D. E.** (1994) 'Detecting selective sweeps in naturally occurring Escherichia coli.', *Genetics*, 138(4), pp. 993–1003. doi: 10.1093/GENETICS/138.4.993.

Hackl, T., Laurenceau, R., Ankenbrand, M. J., Bliem, C., Cariani, Z., Thomas, E., Dooley, K. D., Arellano, A. A., Hogle, S. L., Berube, P., *et al.* (2020) 'Novel integrative elements and genomic plasticity in ocean ecosystems', *bioRxiv*, p. 2020.12.28.424599. doi: 10.1101/2020.12.28.424599.

Hagenblad, J. and **Nordborg, M.** (2002) 'Sequence Variation and Haplotype Structure Surrounding the Flowering Time Locus FRI in Arabidopsis thaliana', *Genetics*, 161(1).

Hanage, W. P. (2016) 'Not so simple after all: Bacteria, their population genetics, and recombination', *Cold Spring Harbor Perspectives in Biology*, 8(7). doi: 10.1101/cshperspect.a018069.

Hao, W. and **Golding, G. B.** (2010) 'Inferring Bacterial Genome Flux While Considering Truncated Genes', *Genetics*, 186(1), pp. 411–426. doi: 10.1534/GENETICS.110.118448.

Hartigan, J. A. and **Wong, M. A.** (1979) 'Algorithm AS 136: A K-Means Clustering Algorithm', *Applied Statistics*, 28(1), p. 100. doi: 10.2307/2346830.

Hildebrand, F., **Meyer, A.** and **Eyre-Walker, A.** (2010) 'Evidence of Selection upon Genomic GC-Content in Bacteria', *PLoS Genetics*. Edited by M. W. Nachman, 6(9), p. e1001107. doi: 10.1371/journal.pgen.1001107.

Hill, W. G. and **Robertson, A.** (1966) 'The effect of linkage on limits to artificial selection', *Genetical Research*, 8(3), pp. 269–294. doi: 10.1017/S0016672300010156.

Huang, S., **Wilhelm, S. W.**, **Harvey, H. R.**, **Taylor, K.**, **Jiao, N.** and **Chen, F.** (2012) 'Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans', *ISME Journal*, 6(2), pp. 285–297. doi: 10.1038/ismej.2011.106.

Huddleston, J. R. (2014) 'Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes', *Infection and Drug Resistance*, 7, p. 167. doi: 10.2147/IDR.S48820.

Hudson, R. R. and **Kaplan, N. L.** (1985) 'Statistical properties of the number of recombination events in the history of a sample of DNA sequences.', *Genetics*, 111(1).

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., *et al.* (2016) 'eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences', *Nucleic Acids Research*, 44(D1), pp. D286--D293. doi: 10.1093/nar/gkv1248.

Hunt, D. E., **David, L. A.**, **Gevers, D.**, **Preheim, S. P.**, **Alm, E. J.** and **Polz, M. F.** (2008) 'Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton', *Science*, 320(5879), pp. 1081–1085. doi: 10.1126/SCIENCE.1157890.

Husmeier, D., **Dybowski, R.** and **Roberts, S.** (2006) *Probabilistic modeling in bioinformatics and medical informatics*. Springer Science \& Business Media.

Huson, D. H. (1998) 'SplitsTree: Analyzing and visualizing evolutionary data', *Bioinformatics*, 14(1), pp. 68–73. doi: 10.1093/bioinformatics/14.1.68.

Huson, D. H. and **Bryant, D.** (2006) 'Application of Phylogenetic Networks in Evolutionary Studies', *Molecular Biology and Evolution*, 23(2), pp. 254–267. doi: 10.1093/molbev/msj030.

Huson, D. H. and **Scornavacca, C.** (2011) 'A survey of combinatorial methods for phylogenetic networks', *Genome Biology and Evolution*. Oxford University Press, pp. 23–35. doi: 10.1093/gbe/evq077.

Jakobsen, I. B. and **Easteal, S.** (1996) 'A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences', *Bioinformatics*, 12(4), pp. 291–295.

Jensen, L. J., **Julien, P.**, **Kuhn, M.**, **von Mering, C.**, **Muller, J.**, **Doerks, T.** and **Bork, P.** (2008) 'eggNOG: Automated construction and annotation of orthologous groups of genes', *Nucleic Acids Research*, 36(SUPPL. 1), pp. D250--D254. doi: 10.1093/nar/gkm796.

Johnson, J. L. and **Whitman, W. B.** (2007) 'Similarity Analysis of DNAs', in *Methods for General and Molecular Microbiology*. ASM Press, pp. 624–652. doi: 10.1128/9781555817497.ch26.

Johnson, Z. I., **Zinser, E. R.**, **Coe, A.**, **McNulty, N. P.**, **Woodward, E. M. S.** and **Chisholm, S. W.** (2006) 'Niche Partitioning Among *Prochlorococcus* Ecotypes Along Ocean-Scale Environmental Gradients', *Science*, 311(5768).

Kashtan, N., **Roggensack, S. E.**, **Berta-Thompson, J. W.**, **Grinberg, M.**, **Stepanauskas, R.** and **Chisholm, S. W.** (2017) 'Fundamental differences in diversity and genomic population structure between Atlantic and Pacific *Prochlorococcus*', *The ISME Journal*, 11(9), pp. 1997–2011. doi: 10.1038/ismej.2017.64.

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. R., Stocker, R., *et al.* (2014) 'Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild *Prochlorococcus*', *Science*, 344(6182), pp. 416–420. doi: 10.1126/science.1248575.

Katoh, K. and **Standley, D. M.** (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30(4), pp. 772–780. doi: 10.1093/molbev/mst010.

Kelly, L., **Huang, K. H.**, **Ding, H.** and **Chisholm, S. W.** (2012) 'ProPortal: A resource for integrated systems biology of *Prochlorococcus* and its phage', *Nucleic Acids Research*, 40(D1), pp. D632--D640. doi: 10.1093/nar/gkr1022.

Kent, A. G., **Baer, S. E.**, **Mouginot, C.**, **Huang, J. S.**, **Larkin, A. A.**, **Lomas, M. W.** and **Martiny, A. C.** (2019) 'Parallel phylogeography of *Prochlorococcus* and *Synechococcus*', *ISME Journal*, 13(2), pp. 430–441. doi: 10.1038/s41396-018-0287-6.

Kent, A. G., **Dupont, C. L.**, **Yooseph, S.** and **Martiny, A. C.** (2016) 'Global biogeography of *Prochlorococcus* genome diversity in the surface ocean.', *The ISME Journal*, 10(8), pp. 1856–1865. doi: 10.1038/ismej.2015.265.

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., *et al.* (2007) 'Patterns and Implications of Gene Gain and Loss in the Evolution of *Prochlorococcus*', *PLoS Genetics*, 3(12), p. e231. doi: 10.1371/journal.pgen.0030231.

Kimura, M. (1969) 'The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations', *Genetics*, 61(4), p. 893.

Kislyuk, A. O., **Haegeman, B.**, **Bergman, N. H.** and **Weitz, J. S.** (2011) 'Genomic fluidity: an integrative view of gene diversity within microbial populations', *BMC Genomics*, 12(1), p. 32. doi: 10.1186/1471-2164-12-32.

Koeppel, A., Perry, E. B., Sikorski, J., Krizanc, D., Warner, A., Ward, D. M., Rooney, A. P., Brambilla, E., Connor, N., Ratcliff, R. M., *et al.* (2008) 'Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics', *Proceedings of the National Academy of Sciences*, 105(7), pp. 2504–2509. doi: 10.1073/PNAS.0712205105.

Konstantinidis, K. T. and **Tiedje, J. M.** (2005) 'Genomic insights that advance the species definition for prokaryotes', *Proceedings of the National Academy of Sciences of the United States of America*, 102(7), pp. 2567–2572. doi: 10.1073/pnas.0409727102.

Koonin, E. V (2011) *The logic of chance: the nature and origin of biological evolution*. FT press.

Koonin, E. V and **Wolf, Y. I.** (2008) 'Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world', *Nucleic Acids Research*, 36(21), pp. 6688–6719. doi: 10.1093/nar/gkn668.

Krause, D. J. and **Whitaker, R. J.** (2015) 'Inferring Speciation Processes from Patterns of Natural Variation in Microbial Genomes', *Systematic Biology*, 64(6), pp. 926–935. doi: 10.1093/sysbio/syv050.

Kryazhimskiy, S. and **Plotkin, J. B.** (2008) 'The Population Genetics of dN/dS', *PLoS Genetics*. Edited by T. Gojobori, 4(12), p. e1000304. doi: 10.1371/journal.pgen.1000304.

Kullback, S. and **Leibler, R. A.** (1951) 'On information and sufficiency', *The annals of mathematical statistics*, 22(1), pp. 79–86.

Kunin, V. and **Ouzounis, C. A.** (2003) 'The balance of driving forces during genome evolution in prokaryotes', *Genome Research*, 13(7), pp. 1589–1594. doi: 10.1101/gr.1092603.

Kuo, C.-H., **Moran, N. A.** and **Ochman, H.** (2009) 'The consequences of genetic drift for bacterial genome complexity.', *Genome research*, 19(8), pp. 1450–1454. doi: 10.1101/gr.091785.109.

Land, M., Hauser, L., Jun, S. R., Nookaew, I., Leuze, M. R., Ahn, T. H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., *et al.* (2015) 'Insights from 20 years of bacterial genome sequencing', *Functional and Integrative Genomics*. Springer Verlag, pp. 141–161. doi: 10.1007/s10142-015-0433-4.

Lapierre, P. and **Gogarten, J. P.** (2009) 'Estimating the size of the bacterial pan-genome', *Trends in genetics*, 25(3), pp. 107–110.

Larkin, A. A., **Blinebry, S. K.**, **Howes, C.**, **Lin, Y.**, **Loftus, S. E.**, **Schmaus, C. A.**, **Zinser, E. R.** and **Johnson, Z. I.** (2016) 'Niche partitioning and biogeography of high light adapted *Prochlorococcus* across taxonomic ranks in the North Pacific.', *The ISME journal*, 10(7), pp. 1555–1567. doi: 10.1038/ismej.2015.244.

Lindell, D., **Sullivan, M. B.**, **Johnson, Z. I.**, **Tolonen, A. C.**, **Rohwer, F.** and **Chisholm, S. W.** (2004) 'Transfer of photosynthesis genes to and from *Prochlorococcus* viruses', *Proceedings of the National Academy of Sciences*, 101(30), pp. 11013–11018. doi: 10.1073/PNAS.0401526101.

Lobkovsky, A. E., **Wolf, Y. I.** and **Koonin, E. V.** (2013) 'Gene Frequency Distributions Reject a Neutral Model of Genome Evolution', *Genome Biology and Evolution*, 5(1), pp. 233–242. doi: 10.1093/gbe/evt002.

López-Pérez, M., **Martin-Cuadrado, A.-B.** and **Rodriguez-Valera, F.** (2014) 'Homologous recombination is involved in the diversity of replacement flexible genomic islands in aquatic prokaryotes', *Frontiers in Genetics*, 5(MAY), pp. 1–1. doi: 10.3389/fgene.2014.00147.

Lorenz, M. G. and **Wackernagel, W.** (1994) 'Bacterial gene transfer by natural genetic transformation in the environment', *Microbiological Reviews*, 58(3), pp. 563–602. doi: 10.1128/MR.58.3.563-602.1994.

Luo, C., **Knight, R.**, **Siljander, H.**, **Knip, M.**, **Xavier, R. J.** and **Gevers, D.** (2015) 'ConStrains identifies microbial strains in metagenomic datasets', *Nature Biotechnology*, 33(10), pp. 1045–1052. doi: 10.1038/nbt.3319.

Lynch, M. (2010) 'Evolution of the mutation rate', *Trends in Genetics*, 26(8), pp. 345–352. doi: 10.1016/j.tig.2010.05.003.

Majewski, J. and **Cohan, F. M.** (1999) 'DNA Sequence Similarity Requirements for Interspecific Recombination in Bacillus', *Genetics*, 153(4).

Majewski, J., **Zawadzki, P.**, **Pickerill, P.**, **Cohan, F. M.** and **Dowson, C. G.** (2000) 'Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation', *Journal of Bacteriology*, 182(4), pp. 1016–1023. doi: 10.1128/JB.182.4.1016-1023.2000.

Makarova, K. S., **Sorokin, A. V.**, **Novichkov, P. S.**, **Wolf, Y. I.** and **Koonin, E. V.** (2007) 'Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea', *Biology Direct*. BioMed Central, pp. 1–20. doi: 10.1186/1745-6150-2-33.

Malmstrom, R. R., **Coe, A.**, **Kettler, G. C.**, **Martiny, A. C.**, **Frias-Lopez, J.**, **Zinser, E. R.** and **Chisholm, S. W.** (2010) 'Temporal dynamics of *Prochlorococcus* ecotypes in the

Atlantic and Pacific oceans', *The ISME Journal*, 4(10), pp. 1252–1264. doi: 10.1038/ismej.2010.60.

Martiny, A. C., **Coleman, M. L.** and **Chisholm, S. W.** (2006) 'Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation', *Proceedings of the National Academy of Sciences*, 103(33), pp. 12552–12557. doi: 10.1073/PNAS.0601301103.

Martiny, A. C., **Huang, Y.** and **Li, W.** (2009) 'Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions', *Environmental Microbiology*, 11(6), pp. 1340–1347.

Marttinen, P., **Croucher, N. J.**, **Gutmann, M. U.**, **Corander, J.** and **Hanage, W. P.** (2015) 'Recombination produces coherent bacterial species clusters in both core and accessory genomes.', *Microbial genomics*, 1(5), p. e000038. doi: 10.1099/mgen.0.000038.

Mayr, E. (1942) 'Systematics and the Origin of Species', *Annals of the Entomological Society of America*, 36(1), pp. 138–139. doi: 10.1093/aesa/36.1.138a.

McInerney, J. O., **McNally, A.** and **O'Connell, M. J.** (2017) 'Why prokaryotes have pangenomes', *Nature Microbiology*, 2(4), p. 17040. doi: 10.1038/nmicrobiol.2017.40.

Medini, D., **Donati, C.**, **Tettelin, H.**, **Masignani, V.** and **Rappuoli, R.** (2005) 'The microbial pan-genome', *Current Opinion in Genetics & Development*, 15(6), pp. 589–594. doi: 10.1016/J.GDE.2005.09.006.

Mira, A., **Ochman, H.** and **Moran, N. A.** (2001) 'Deletional bias and the evolution of bacterial genomes', *Trends in Genetics*, 17(10), pp. 589–596. doi: 10.1016/S0168-9525(01)02447-7.

Mirkin, B. G., **Fenner, T. I.**, **Galperin, M. Y.** and **Koonin, E. V.** (2003) 'Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes', *BMC Evolutionary Biology*, 3(1), pp. 1–34. doi: 10.1186/1471-2148-3-2.

Moore, L. R., **Rocap, G.** and **Chisholm, S. W.** (1998) 'Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes', *Nature*, 393(6684), pp. 464–467. doi: 10.1038/30965.

Moran, N. A. (1996) 'Accelerated evolution and Muller's rachet in endosymbiotic bacteria', *Proceedings of the National Academy of Sciences of the United States of America*, 93(7), pp. 2873–2878. doi: 10.1073/pnas.93.7.2873.

Mostowy, R., **Croucher, N. J.**, **Andam, C. P.**, **Corander, J.**, **Hanage, W. P.** and **Marttinen, P.** (2017) 'Efficient Inference of Recent and Ancestral Recombination within

Bacterial Populations', *Molecular Biology and Evolution*, 34(5), pp. 1167–1182. doi: 10.1093/molbev/msx066.

Nordborg, M. (1997) 'Structured Coalescent Processes on Different Time Scales', *Genetics*, 146(4).

Ochman, H., **Lawrence, J. G.** and **Groisman, E. a** (2000) 'Lateral gene transfer and the nature of bacterial innovation.', *Nature*, 405(6784), pp. 299–304. doi: 10.1038/35012500.

Oliveira, P. H., **Touchon, M.**, **Cury, J.** and **Rocha, E. P. C.** (2017) 'The chromosomal organization of horizontal gene transfer in bacteria', *Nature Communications*, 8(1), p. 841. doi: 10.1038/s41467-017-00808-w.

Olm, M. R., **Crits-Christoph, A.**, **Bouma-Gregson, K.**, **Firek, B. A.**, **Morowitz, M. J.** and **Banfield, J. F.** (2021) 'inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains', *Nature Biotechnology 2021 39:6*, 39(6), pp. 727–736. doi: 10.1038/s41587-020-00797-0.

Pachiadaki, M. G., Brown, J. M., Brown, J., Bezuidt, O., Berube, P. M., Biller, S. J., Poulton, N. J., Burkart, M. D., La Clair, J. J., Chisholm, S. W., *et al.* (2019) 'Charting the Complexity of the Marine Microbiome through Single-Cell Genomics', *Cell*, 179(7), pp. 1623-1635.e11. doi: 10.1016/J.CELL.2019.11.017.

Papke, R. T., **Zhaxybayeva, O.**, **Feil, E. J.**, **Sommerfeld, K.**, **Muise, D.** and **Doolittle, W. F.** (2007) 'Searching for species in haloarchaea', *Proceedings of the National Academy of Sciences*, 104(35), pp. 14092–14097. doi: 10.1073/PNAS.0706358104.

Parks, D. H., **Imelfort, M.**, **Skennerton, C. T.**, **Hugenholtz, P.** and **Tyson, G. W.** (2015) 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.', *Genome research*, 25(7), pp. 1043–1055. doi: 10.1101/gr.186072.114.

Parmley, J. L. and **Hurst, L. D.** (2007) 'How Common Are Intragene Windows with K A > K S Owing to Purifying Selection on Synonymous Mutations?', *Journal of Molecular Evolution*, 64(6), pp. 646–655. doi: 10.1007/s00239-006-0207-7.

Partensky, F., **Hess, W. R.** and **Vaulot, D.** (1999) '*Prochlorococcus*, a marine photosynthetic prokaryote of global significance.', *Microbiology and molecular biology reviews : MMBR*, 63(1), pp. 106–127.

Pelgrift, R. Y. and **Friedman, A. J.** (2013) 'Nanotechnology as a therapeutic tool to combat microbial resistance', *Advanced Drug Delivery Reviews*, 65(13–14), pp. 1803–1815. doi: 10.1016/J.ADDR.2013.07.011.

Perna, N. T., Plunkett, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., *et al.* (2001) 'Genome sequence of

enterohaemorrhagic Escherichia coli O157:H7', *Nature 2001 409:6819*, 409(6819), pp. 529–533. doi: 10.1038/35054089.

Pfeifer, B., **Wittelsbürger, U.**, **Ramos-Onsins, S. E.** and **Lercher, M. J.** (2014) 'PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R', *Molecular Biology and Evolution*, 31(7), pp. 1929–1936. doi: 10.1093/molbev/msu136.

Polz, M. F., **Alm, E. J.** and **Hanage, W. P.** (2013) 'Horizontal gene transfer and the evolution of bacterial and archaeal population structure', *Trends in Genetics*, 29(3), pp. 170–175. doi: 10.1016/j.tig.2012.12.006.

Preheim, S. P., **Timberlake, S.** and **Polz, M. F.** (2011) 'Merging taxonomy with ecological population prediction in a case study of Vibrionaceae', *Applied and Environmental Microbiology*, 77(20), pp. 7195–7206. doi: 10.1128/AEM.00665-11.

Price, M. N. and **Arkin, A. P.** (2015) 'Weakly Deleterious Mutations and Low Rates of Recombination Limit the Impact of Natural Selection on Bacterial Genomes.', *mBio*, 6(6), pp. e01302--15. doi: 10.1128/mBio.01302-15.

Pritchard, L., **Glover, R. H.**, **Humphris, S.**, **Elphinstone, J. G.** and **Toth, I. K.** (2016) 'Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens', *Analytical Methods*, 8(1), pp. 12–24. doi: 10.1039/C5AY02550H.

Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I., Koonin, E. V, Kolsto, A. B., Koonin, E. V, Galperin, M. Y., Casjens, S., Bellgard, M. I., *et al.* (2014) 'Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes', *BMC Biology*, 12(1), p. 66. doi: 10.1186/s12915-014-0066-4.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al.* (2010) 'A human gut microbial gene catalogue established by metagenomic sequencing', *Nature 2010 464:7285*, 464(7285), pp. 59–65. doi: 10.1038/nature08821.

Quince, C., **Nurk, S.**, **Raguideau, S.**, **James, R.**, **Soyer, O.**, **Summers, J.**, **Limasset, A.**, **Eren, A.**, **Chikhi, R.** and **Darling, A.** (2021) 'STRONG: metagenomics strain resolution on assembly graphs', *Genome biology*, 22(1). doi: 10.1186/S13059-021-02419-7.

Raghavan, R., **Kelkar, Y. D.** and **Ochman, H.** (2012) 'A selective force favoring increased G+C content in bacterial genes.', *Proceedings of the National Academy of Sciences of the United States of America*, 109(36), pp. 14504–14507. doi: 10.1073/pnas.1205683109.

Raines, C. A. (2003) 'The Calvin cycle revisited', *Photosynthesis Research 2003 75:1*, 75(1), pp. 1–10. doi: 10.1023/A:1022421515027.

dos Reis, M. and **Yang, Z.** (2013) 'The unbearable uncertainty of Bayesian divergence time estimation', *Journal of Systematics and Evolution*, 51(1), pp. 30–43. doi: 10.1111/j.1759-6831.2012.00236.x.

Rice, P., **Longden, I.** and **Bleasby, A.** (2000) 'EMBOSS: the European Molecular Biology Open Software Suite.', *Trends in genetics : TIG*, 16(6), pp. 276–277. doi: 10.1016/S0168-9525(00)02024-2.

Richter, M. and **Rosselló-Móra, R.** (2009) 'Shifting the genomic gold standard for the prokaryotic species definition', *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), pp. 19126–19131. doi: 10.1073/pnas.0906412106.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., *et al.* (2003) 'Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation', *Nature*, 424(6952), pp. 1042–1047. doi: 10.1038/nature01947.

Rocha, E. P. C. and **Feil, E. J.** (2010) 'Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria?', *PLOS Genetics*, 6(9), p. e1001104. doi: 10.1371/JOURNAL.PGEN.1001104.

Rocha, E. P. C., **Smith, J. M.**, **Hurst, L. D.**, **Holden, M. T. G.**, **Cooper, J. E.**, **Smith, N. H.** and **Feil, E. J.** (2006) 'Comparisons of dN/dS are time dependent for closely related bacterial genomes', *Journal of Theoretical Biology*, 239(2), pp. 226–235. doi: 10.1016/J.JTBI.2005.08.037.

Rodriguez-Valera, F., **Martin-Cuadrado, A. B.** and **López-Pérez, M.** (2016) 'Flexible genomic islands as drivers of genome evolution', *Current Opinion in Microbiology*. Elsevier Ltd, pp. 154–160. doi: 10.1016/j.mib.2016.03.014.

Roux, S., **Enault, F.**, **le Bronner, G.** and **Debroas, D.** (2011) 'Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems', *FEMS Microbiology Ecology*, 78(3), pp. 617–628. doi: 10.1111/J.1574-6941.2011.01190.X.

Rusch, D. B., **Martiny, A. C.**, **Dupont, C. L.**, **Halpern, A. L.** and **Venter, J. C.** (2010) 'Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions', *Proceedings of the National Academy of Sciences of the United States of America*, 107(37), pp. 16184–16189. doi: 10.1073/pnas.1009513107.

Santos, S. R. and **Ochman, H.** (2004) 'Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins', *Environmental Microbiology*, 6(7), pp. 754–759. doi: 10.1111/J.1462-2920.2004.00617.X.

Sawyer, S. (1989) 'Statistical tests for detecting gene conversion.', *Molecular biology and evolution*, 6(5), pp. 526–538.

Sawyer, S. A. (1999) 'GENECONV: a computer package for the statistical detection of gene conversion', *http://www.math.wustl.edu/~sawyer*.

Schluter, D. (2009) 'Evidence for ecological speciation and its alternative.', *Science (New York, N.Y.)*, 323(5915), pp. 737–741. doi: 10.1126/science.1160006.

Schmidt, H. and **Hensel, M.** (2004) 'Pathogenicity Islands in Bacterial Pathogenesis', *Clinical Microbiology Reviews*, 17(1), pp. 14–56. doi: 10.1128/CMR.17.1.14-56.2004.

Schmutzer, M. and **Barraclough, T. G.** (2019) 'The role of recombination, niche-specific gene pools and flexible genomes in the ecological speciation of bacteria', *Ecology and Evolution*, 9(8), pp. 4544–4556. doi: 10.1002/ece3.5052.

Sela, I., **Wolf, Y. I.** and **Koonin, E. V.** (2021) 'Assessment of assumptions underlying models of prokaryotic pangenome evolution', *BMC Biology*, 19(1), pp. 1–15. doi: 10.1186/s12915-021-00960-2.

Sela, I., **Wolf, Y. I.** and **Koonin, E. V** (2016) 'Theory of prokaryotic genome evolution.', *Proceedings of the National Academy of Sciences of the United States of America*, 113(41), pp. 11399–11407. doi: 10.1073/pnas.1614083113.

Shannon, C. E. (1948) 'A Mathematical Theory of Communication', *The Bell system technical journal*, 27, pp. 379–423.

Shapiro, B. J., **David, L. A.**, **Friedman, J.** and **Alm, E. J.** (2009) 'Looking for Darwin's footprints in the microbial world', *Trends in Microbiology*, 17(5), pp. 196–204. doi: 10.1016/J.TIM.2009.02.002.

Shapiro, B. J., **Friedman, J.**, **Cordero, O. X.**, **Preheim, S. P.**, **Timberlake, S. C.**, **Szabó, G.**, **Polz, M. F.** and **Alm, E. J.** (2012) 'Population genomics of early events in the ecological differentiation of bacteria.', *Science (New York, N.Y.)*, 336(6077), pp. 48–51. doi: 10.1126/science.1218198.

Sharp, P. M., **Bailes, E.**, **Grocock, R. J.**, **Peden, J. F.** and **Sockett, R. E.** (2005) 'Variation in the strength of selected codon usage bias among bacteria', *Nucleic Acids Research*, 33(4), pp. 1141–1153. doi: 10.1093/NAR/GKI242.

Sheppard, S. K., **McCarthy, N. D.**, **Falush, D.** and **Maiden, M. C. J.** (2008) 'Convergence of Campylobacter species: Implications for bacterial evolution', *Science*, 320(5873), pp. 237–239. doi: 10.1126/science.1155532.

Sinha, R., **Abnet, C. C.**, **White, O.**, **Knight, R.** and **Huttenhower, C.** (2015) 'The microbiome quality control project: baseline study design and future directions', *Genome Biology 2015 16:1*, 16(1), pp. 1–6. doi: 10.1186/S13059-015-0841-8.

Sjöqvist, C., **Delgado, L. F.**, **Alneberg, J.** and **Andersson, A. F.** (2021) 'Ecologically coherent population structure of uncultivated bacterioplankton', *The ISME Journal 2021 15:10*, 15(10), pp. 3034–3049. doi: 10.1038/s41396-021-00985-z.

Smith, G. R. (1991) 'Conjugational recombination in E. coli: myths and mechanisms', *Cell*, 64(1), pp. 19–27.

Smith, J. M. (1992) 'Analyzing the mosaic structure of genes', *Journal of molecular evolution*, 34(2), pp. 126–129.

Smith, J. M. and **Haigh, J.** (1974) 'The hitch-hiking effect of a favourable gene', *Genetics Research*, 23(1), pp. 23–35.

Smith, J. M., **Smith, N. H.**, **O'Rourke, M.** and **Spratt, B. G.** (1993) 'How clonal are bacteria?', *Proceedings of the National Academy of Sciences of the United States of America*, 90(10), pp. 4384–4388. doi: 10.1073/pnas.90.10.4384.

Sneath, P. H. A. (1975) 'Cladistic representation of reticulate evolution', *Systematic Zoology*, 24(3), pp. 360–368.

Snel, B., **Bork, P.** and **Huynen, M.** (2000) 'Genome evolutiongene fusion versus gene fission', *Trends in Genetics*, 16(1), pp. 9–11. doi: 10.1016/S0168-9525(99)01924-1.

Soler, N. and **Forterre, P.** (2020) 'Vesiduction: the fourth way of HGT', *Environmental Microbiology*, 22(7), pp. 2457–2460. doi: 10.1111/1462-2920.15056.

Stackebrandt, E. and **Goebel, B. M.** (1994) 'Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology', *International Journal of Systematic Bacteriology*. Microbiology Society, pp. 846–849. doi: 10.1099/00207713-44-4-846.

Stamatakis, A. (2014) 'RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30(9), pp. 1312–1313. doi: 10.1093/bioinformatics/btu033.

Sullivan, J. T., Trzebiatowski, J. R., Cruickshank, R. W., Gouzy, J., Brown, S. D., Elliot, R. M., Fleetwood, D. J., McCallum, N. G., Rossbach, U., Stuart, G. S., *et al.* (2002) 'Comparative sequence analysis of the symbiosis island of Mesorhizobium loti strain R7A', *Journal of Bacteriology*, 184(11), pp. 3086–3095. doi: 10.1128/JB.184.11.3086-3095.2002.

Sun, Z. and **Blanchard, J. L.** (2014) 'Strong Genome-Wide Selection Early in the Evolution of *Prochlorococcus* Resulted in a Reduced Genome through the Loss of a Large Number of Small Effect Genes', *PLoS ONE*, 9(3). doi: 10.1371/journal.pone.0088837.

Sung, W., **Ackerman, M. S.**, **Miller, S. F.**, **Doak, T. G.** and **Lynch, M.** (2012) 'Drift-barrier hypothesis and mutation-rate evolution.', *Proceedings of the National Academy of Sciences of the United States of America*, 109(45), pp. 18488–18492. doi: 10.1073/pnas.1216223109.

Talavera, G. and **Castresana, J.** (2007) 'Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments', *Systematic Biology*. Edited by K. Kjer, R. Page, and J. Sullivan, 56(4), pp. 564–577. doi: 10.1080/10635150701472164.

Tatusov, R. L., **Galperin, M. Y.**, **Natale, D. A.** and **Koonin, E. V** (2000) 'The COG database: A tool for genome-scale analysis of protein functions and evolution', *Nucleic Acids Research*. Oxford University Press, pp. 33–36. doi: 10.1093/nar/28.1.33.

Tatusov, R. L., **Koonin, E. V** and **Lipman, D. J.** (1997) 'A genomic perspective on protein families', *Science*, 278(5338), pp. 631–637. doi: 10.1126/science.278.5338.631.

Tavaré, S. (1986) 'Some probabilistic and statistical problems in the analysis of DNA sequences.', *American Mathematical Society : Lectures on Mathematics in the Life Sciences*, 17, pp. 57–86.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V, Crabtree, J., Jones, A. L., Durkin, A. S., *et al.* (2005) 'Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".', *Proceedings of the National Academy of Sciences of the United States of America*, 102(39), pp. 13950–13955. doi: 10.1073/pnas.0506758102.

Thompson, C. C., **Silva, G. G. Z.**, **Vieira, N. M.**, **Edwards, R.**, **Vicente, A. C. P.** and **Thompson, F. L.** (2013) 'Genomic taxonomy of the genus *Prochlorococcus*', *Microbial ecology*, 66(4), pp. 752–762.

Thorpe, H. A., **Bayliss, S. C.**, **Hurst, L. D.** and **Feil, E. J.** (2017) 'Comparative Analyses of Selection Operating on Nontranslated Intergenic Regions of Diverse Bacterial Species.', *Genetics*, 206(1), pp. 363–376. doi: 10.1534/genetics.116.195784.

Tindall, B. J., **Rosselló-Móra, R.**, **Busse, H. J.**, **Ludwig, W.** and **Kämpfer, P.** (2010) 'Notes on the characterization of prokaryote strains for taxonomic purposes', *International Journal of Systematic and Evolutionary Microbiology*, 60(1), pp. 249–266. doi: 10.1099/ijs.0.016949-0.

Toft, C. and **Andersson, S. G. E.** (2010) 'Evolutionary microbial genomics: insights into bacterial host adaptation', *Nature Reviews Genetics 2010 11:7*, 11(7), pp. 465–475. doi: 10.1038/nrg2798.

Touchon, M. and **Rocha, E. P. C.** (2016) 'Coevolution of the Organization and Structure of Prokaryotic Genomes.', *Cold Spring Harbor perspectives in biology*, 8(1), p. a018168. doi: 10.1101/cshperspect.a018168.

Treangen, T. J., **Ambur, O. H.**, **Tonjum, T.** and **Rocha, E. P.** (2008) 'The impact of the neisserial DNA uptake sequences on genome evolution and stability', *Genome Biology 2008 9:3*, 9(3), pp. 1–17. doi: 10.1186/GB-2008-9-3-R60.

Treangen, T. J. and **Rocha, E. P. C.** (2011) 'Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes', *PLoS Genetics*. Edited by N. A. Moran, 7(1), p. e1001284. doi: 10.1371/journal.pgen.1001284.

Truong, D. T., **Tett, A.**, **Pasolli, E.**, **Huttenhower, C.** and **Segata, N.** (2017) 'Microbial strain-level population structure and genetic diversity from metagenomes', *Genome Research*, 27(4), pp. 626–638. doi: 10.1101/gr.216242.116.

Varghese, N. J., **Mukherjee, S.**, **Ivanova, N.**, **Konstantinidis, K. T.**, **Mavrommatis, K.**, **Kyrpides, N. C.** and **Pati, A.** (2015) 'Microbial species delineation using whole genome sequences', *Nucleic Acids Research*, 43(14), pp. 6761–6771. doi: 10.1093/nar/gkv657.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., *et al.* (2004) 'Environmental Genome Shotgun Sequencing of the Sargasso Sea', *Science*, 304(5667), pp. 66–74. doi: 10.1126/SCIENCE.1093857.

Vernikos, G., **Medini, D.**, **Riley, D. R.** and **Tettelin, H.** (2015) 'Ten years of pan-genome analyses', *Current Opinion in Microbiology*, 23, pp. 148–154. doi: 10.1016/j.mib.2014.11.016.

Větrovský, T. and **Baldrian, P.** (2013) 'The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses', *PLoS ONE*. Edited by J. Neufeld, 8(2), p. e57923. doi: 10.1371/journal.pone.0057923.

Vos, M. and **Didelot, X.** (2009) 'A comparison of homologous recombination rates in bacteria and archaea', *The ISME Journal*, 3(2), pp. 199–208. doi: 10.1038/ismej.2008.93.

Vos, M. and **Eyre-Walker, A.** (2017) 'Are pangenomes adaptive or not?', *Nature Microbiology*, 2(12), p. 1576. doi: 10.1038/s41564-017-0067-5.

Vos, M., **Hesselman, M. C.**, **te Beek, T. A.**, **van Passel, M. W. J.** and **Eyre-Walker, A.** (2015) 'Rates of Lateral Gene Transfer in Prokaryotes: High but Why?', *Trends in Microbiology*, 23(10), pp. 598–605. doi: 10.1016/j.tim.2015.07.006.

Vrancianu, C. O., **Popa, L. I.**, **Bleotu, C.** and **Chifiriuc, M. C.** (2020) 'Targeting Plasmids to Limit Acquisition and Transmission of Antimicrobial Resistance', *Frontiers in Microbiology*, 0, p. 761. doi: 10.3389/FMICB.2020.00761.

Vulić, M., **Dionisio, F.**, **Taddei, F.** and **Radman, M.** (1997) 'Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria', *Proceedings of the National Academy of Sciences of the United States of America*, 94(18), pp. 9763–9767. doi: 10.1073/pnas.94.18.9763.

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler,', 0, Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E., Stackebrandt,', E., *et al.* (1987) *Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics*, *INTERNATIONAL JOURNAL OF SYSTEMATIC BACTERIOLOGY*.

Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S.-R., Boutin, A., Hackett, J., *et al.* (2002) 'Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli', *Proceedings of the National Academy of Sciences*, 99(26), pp. 17020–17024.

West, N. J., **Lebaron, P.**, **Strutton, P. G.** and **Suzuki, M. T.** (2011) 'A novel clade of *Prochlorococcus* found in high nutrient low chlorophyll waters in the South and Equatorial Pacific Ocean', *ISME Journal*, 5(6), pp. 933–944. doi: 10.1038/ismej.2010.186.

Wiedenbeck, J. and **Cohan, F. M.** (2011) 'Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches', *FEMS Microbiology Reviews*, 35(5), pp. 957–976. doi: 10.1111/j.1574-6976.2011.00292.x.

von Wintersdorff, C. J. H., **Penders, J.**, **van Niekerk, J. M.**, **Mills, N. D.**, **Majumder, S.**, **van Alphen, L. B.**, **Savelkoul, P. H. M.** and **Wolffs, P. F. G.** (2016) 'Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer', *Frontiers in Microbiology*, 0(FEB), p. 173. doi: 10.3389/FMICB.2016.00173.

Wiser, M. J., **Ribeck, N.** and **Lenski, R. E.** (2013) 'Long-term dynamics of adaptation in asexual populations', *Science*, 342(6164), pp. 1364–1367. doi: 10.1126/science.1243357.

Woese, C. R. and **Fox, G. E.** (1977) 'Phylogenetic structure of the prokaryotic domain: The primary kingdoms', *Proceedings of the National Academy of Sciences of the United States of America*, 74(11), pp. 5088–5090. doi: 10.1073/pnas.74.11.5088.

Wolf, J. B. W., **Künstner, A.**, **Nam, K.**, **Jakobsson, M.** and **Ellegren, H.** (2009) 'Nonlinear Dynamics of Nonsynonymous (dN) and Synonymous (dS) Substitution Rates Affects Inference of Selection', *Genome Biology and Evolution*, 1, pp. 308–319. doi: 10.1093/gbe/evp030.

Yan, W., **Wei, S.**, **Wang, Q.**, **Xiao, X.**, **Zeng, Q.**, **Jiao, N.** and **Zhang, R.** (2018) 'Genome Rearrangement Shapes *Prochlorococcus* Ecological Adaptation', *Appl. Environ. Microbiol.*, 84(17), pp. e01178--18. doi: 10.1128/AEM.01178-18.

Yang, Z. (2007) 'PAML 4: Phylogenetic Analysis by Maximum Likelihood', *Molecular Biology and Evolution*, 24(8), pp. 1586–1591. doi: 10.1093/molbev/msm088.

Yang, Z. and **Nielsen, R.** (2000) 'Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models', *Molecular Biology and Evolution*, 17(1), pp. 32–43. doi: 10.1093/oxfordjournals.molbev.a026236.

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen, J. A., Heidelberg, K. B., Manning, G., Li, W., *et al.* (2007) 'The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families', *PLOS Biology*, 5(3), p. e16. doi: 10.1371/JOURNAL.PBIO.0050016.

Yu, T., **Li, J.**, **Yang, Y.**, **Qi, L.**, **Chen, B.**, **Zhao, F.**, **Bao, Q.** and **Wu, J.** (2012) 'Codon usage patterns and adaptive evolution of marine unicellular cyanobacteria *Synechococcus* and *Prochlorococcus*', *Molecular Phylogenetics and Evolution*, 62(1), pp. 206–213. doi: 10.1016/J.YMPEV.2011.09.013.

Zawadzki, P., **Roberts, M. S.** and **Cohan, F. M.** (1995) 'The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust.', *Genetics*, 140(3).

Zhaxybayeva, O., **Doolittle, W. F.**, **Papke, R. T.** and **Gogarten, J. P.** (2009) 'Intertwined Evolutionary Histories of Marine *Synechococcus* and *Prochlorococcus* marinus', *Genome Biology and Evolution*, 1, pp. 325–339. doi: 10.1093/gbe/evp032.

Zinder, N. D. and **Lederberg, J.** (1952) 'Genetic exchange in Salmonella', *Journal of bacteriology*, 64(5), pp. 679–699.

Zinser, E. R., **Johnson, Z. I.**, **Coe, A.**, **Karaca, E.**, **Veneziano, D.** and **Chisholm, S. W.** (2007) 'Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean', *Limnology and Oceanography*, 52(5), pp. 2205–2220. doi: 10.4319/lo.2007.52.5.2205.

Annexes

Article 1 / Première autrice : Gardon *et al.*, 2020

Gardon, H., **Biderre-Petit, C.**, **Jouan-Dufournel, I.** and **Bronner, G.** (2020) 'A drift-barrier model drives the genomic landscape of a structured bacterial population', *Molecular Ecology*. doi: 10.1111/mec.15628.

Received: 5 March 2020 | Revised: 26 August 2020 | Accepted: 28 August 2020

DOI: 10.1111/mec.15628

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

A drift-barrier model drives the genomic landscape of a structured bacterial population

Hélène Gardon | Corinne Biderre-Petit | Isabelle Jouan-Dufournel | Gisèle Bronner

Laboratoire Microorganismes: Génome et Environnement, Université Clermont Auvergne, CNRS, Clermont-Ferrand, France

Correspondence
Hélène Gardon, LMGE, UMR CNRS 6023, Université Clermont Auvergne, Campus Universitaire des Cézeaux, 1 impasse Amélie Murat, Aubiere Cedex 63178, France.
Email: helene.gardon@uca.fr

Abstract

Bacterial populations differentiate over time and space to form distinct genetic units. The mechanisms governing this diversification are presumed to result from the ecological context of living units to adapt to specific niches. Recently, a model assuming the acquisition of advantageous genes among populations rather than whole genome sweeps has emerged to explain population differentiation. However, the characteristics of these exchanged, or flexible, genes and whether their evolution is driven by adaptive or neutral processes remain controversial. By analysing the flexible genome of single-amplified genomes of co-occurring populations of the marine *Prochlorococcus* HLII ecotype, we highlight that genomic compartments – rather than population units – are characterized by different evolutionary trajectories. The dynamics of gene fluxes vary across genomic compartments and therefore the effectiveness of selection depends on the fluctuation of the effective population size along the genome. Taken together, these results support the drift-barrier model of bacterial evolution.

KEYWORDS

bacterial genome diversity, evolutionary mechanisms, pangenome, *Prochlorococcus*, single-cell analyses

1 | INTRODUCTION

The diversification of free bacterial species in the environment is assumed to result from their adaptation to specific ecological niches. However, the full understanding of the forces driving these differentiations also relies on evaluating their genome dynamics, in light of populational mechanisms such as selection, genetic drift and recombination. Based on general species definition by Mayr (1942), populations result from gene flow discontinuities within a species, leading to genetically cohesive units that can be distinguished according to their genome characteristics. However, this definition of species hardly applies to bacteria as population boundaries remain elusive due to gene fluxes occurring even among distant relatives. Likewise, gene content variations of conspecific organisms, which gave rise to the concept of pangenome (Medini et al., 2005; Tettelin et al., 2005), blur the genetic cohesion of the microbial population. Yet analyses based on comparative genomics have also suggested that high recombination rates lead to the exchange of advantageous genes within a bacterial population (Cadillo-Quiroz et al., 2012; Shapiro et al., 2012). These genes, rather than genomes, would sweep through the evolving population, leading to both its genetic cohesion and ecological differentiation at the species level. Genes acquired by horizontal transfer are indeed frequently reported as adaptive (McInerney et al., 2017; Sela et al., 2016). However, it has also been suggested that the distribution of flexible genes could be neutral (Baumdicker et al., 2012). Furthermore, species with larger effective population size (N_e) have greater genetic diversity, and by extension a highly diverse pangenome (Andreani et al., 2017). As N_e affects the effectiveness of selection, it may impact the number of flexible genes that would be retained through selection (Bobay & Ochman, 2018).

In recent years, the tremendous progress of single-cell genomics (SCG) has greatly improved the sampling of coexisting subpopulations. This progress allowed the investigation of the factors

that govern the diversification of the microbial genome structure and organization at a finer scale, such as for *Prochlorococcus marinus*. This cyanobacterium is one of the most abundant photosynthetic species in the ocean euphotic zone, responsible for up to 10% of the marine primary productivity (Flombaum et al., 2013; Partensky et al., 1999). Its genetic diversity spans at least 12 distinct ecotypes (Biller et al., 2014; Kashtan et al., 2014; Malmstrom et al., 2010; Moore et al., 1998; Rocap et al., 2003), broadly separated into high-light (HL) and low-light (LL) ecotypes. All these ecotypes were shown to contain different sets of functional genes and to adjust differently to environmental changes, suggesting a stable niche partitioning of ecologically distinct groups (Kent et al., 2016; Larkin et al., 2016). From a large-scale SCG approach it was recently proposed that *Prochlorococcus* populations in the Atlantic Ocean are composed of hundreds of subpopulations resulting from an ancient niche partitioning (Kashtan et al., 2014) and that population differentiation was occurring among *Prochlorococcus* (Stolyar & Marx, 2019). Coexisting subpopulations showed a fine-scale sequence diversity, i.e., a "genomic backbone" comprised primarily of within subpopulations core genes with distinct fixed alleles and several genomic islands (ISLs) mostly composed of flexible genes in combination with specific core alleles or shared among different backbones (Kashtan et al., 2014). This, in addition to the large population size and open pangenome of the *Prochlorococcus* genus, makes it a valuable taxon to study pangenome evolution.

On the basis of the work of Kashtan et al. (2014), who suggest that variations in co-occurring subpopulations within the *Prochlorococcus* HLII ecotype are targeted on specific genome regions, we investigate the evolutionary underpinnings of bacterial genome differentiation among these subpopulations, with a focus on the flexible genome. By analysing synonymous versus nonsynonymous substitution rates (d*N*/d*S*) of single-amplified genomes (SAGs), we assess the nature and strength of selection on genomic compartments (core versus flexible, backbone versus ISLs). Overall, despite clear delineation of these subpopulations according to genome phylogeny, average nucleotide identity (ANI) analysis and content in flexible genes, we do not find significant differences in average d*N*/d*S* among clades. However, by analysing the evolutionary rates of clusters of orthologous genes (COGs), we demonstrate that ISLs are characterized by differences in selective pressures that shed light on different evolutionary trajectories. This variation in the efficacy of selection – associated with distinct sets of genes in specialized genomic compartments – could result from the fluctuating N_e along the genome.

2 | MATERIALS AND METHODS

2.1 | SAG data sets

In total, 87 SAGs of the marine cyanobacterium *Prochlorococcus* belonging to the HLII ecotype (Table S1) were examined to investigate the evolutionary dynamics of free-living bacteria. These SAGs were a subset of 96 SAGs collected at the Bermuda-Atlantic Times-series Study (BATS) site, during three samplings between November 2008 and April 2009 (Kashtan et al., 2014). The original set was reduced only to SAGs assigned to the seven phylogenetically delineated subpopulations distributed among three clusters defined at 98% identity of the ITS (Kashtan et al., 2014), i.e., C1 to C5 within the cluster cN2, C8 within the cluster c9301 and C9 within the cluster cN1, and also excluded the contaminated SAG 518D8. The SAG sequences were downloaded from the National Center for Biotechnology Information (NCBI) (Table S2). Their assembly size ranged from 0.37 to 1.62 Mb, with an average GC content of 31.3%. We used CheckM (Parks et al., 2015) to estimate their completeness and contamination (Figure S1). Their completeness approximated 8.6%–97.4%, with ~14% of SAGs being classified as partial (<50% of completeness; with an overrepresentation of SAGs from C1 clade [7 over 12]), ~45% as substantial (≥50%–70% of completeness), ~19% as moderate (≥70 to 90% of completeness) and ~21% as near-complete (≥90% of completeness; fairly distributed over all clades, including those with a small number of SAGs). They all had less than 2.3% contamination (Figure S1b). Because of synteny of the HLII *Prochlorococcus* ecotype genomes (Yan et al., 2018) and to limit the complexity of the information, we used a reference genome in all analyses performed. We needed a reference genome that (i) did not branch with any clades studied (that excludes two cultured strains, i.e., MIT9301, AS9601); (ii) with a close relatedness with all clades; and (iii) but not too much either (excluding the three most distant, i.e., MIT9107, MIT9123 and MIT9116, and those closest to MIT9301 and AS9601; [Kent et al., 2019]). Among the three remaining (i.e., MIT9302, MIT9311 and MIT9312), *P. marinus str.* MIT9312 (accession number ABB49062.1) was chosen because it historically denotes the eMIT9312/HLII ecotype (Biller et al., 2014). Its genome, 1.71 Mb in size and with an average GC content of 31%, contained 1,962 CDS and showed the presence of six ISLs scattered all along its genomic backbone (Avrani et al., 2011; Table S3).

2.2 | Genome scale comparisons

The ANI allowed for the delineation of operational units at the genome level (Varghese et al., 2015). ANI was calculated both within and among subpopulations using the pyani package (Pritchard et al., 2016). It was estimated by aligning fragments of 1,020 nt (Klappenbach et al., 2007) with BLASTN+ (Altschul et al., 1990; Camacho et al., 2009) and averaging the sequence identity between pairs of genomes.

Genome synteny analysis was performed on the MIT9312 reference genome and representative SAGs were selected for each subpopulation (the largest ones). The whole genome alignment and the detection of LCBs were generated using Mauve 2.4 with the progressiveMauve algorithm (Darling et al., 2010) and default settings.

2.3 | COG assignments

Overall, 7,125 COGs previously defined (Kashtan et al., 2015) were analysed. These COGs were determined by inferring pairwise homologous relationships using the method described by Kelly et al. (2012). Briefly, they first assigned orthology relationships between genes using reciprocal best BLASTP hits (*e*-value ≤ 1e-5; sequence identity > 35%; alignment length > 75% of the length of the shorter protein of the two compared) followed by transitively clustering orthologs together. They then built Hidden Markov Model (HMM) profiles (Eddy, 2009) of each cluster to integrate most divergent homologous genes missed by the BLAST approach. It is assumed that homologous relationships are transitive within a COG; thus, all genes from a cluster are homologous to any other gene in the cluster.

First of all, we looked at for the balance between completeness and number of SAGs, some of them having less than 50% completeness, which could affect the result of the analysis. As we showed that many SAGs, even incomplete, were more informative than a reduced set of complete SAGs (Figure S2), they were all considered in the subsequent analysis as well as all COGs. Of all the 7,125 COGs, 1,410 were identified as core (Table S4; Figures S3 and S4a), namely, common to the available genomes for cultured strains of the HLII ecotype (i.e., MIT9311, MIT9314, MIT9401, MIT9301, MIT9312, MIT9107, MIT9201, MIT9321, MIT9202, MIT9215, SB, GP2, and AS9601), among which 1,397 were composed of single-copy genes. The remaining COGs were considered as flexible (5,715 COGs in total) and were either shared by some but not all cultured strains (11.35%) or specific to SAGs. Flexible COGs detected in at least one SAG but absent from the MIT9312 reference genome were assigned to a genomic compartment (i.e., ISLs or genomic backbone) according to the location of the closest pair of genes referenced in MIT9312 that bounded these flexible COGs (Figure S3). We assumed that if two contiguous genes found in MIT9312 belonged to a unique compartment, the flexible genes between them also belonged to this compartment, otherwise they were classified as "ambiguous". The compartment assignment was subsequently inferred at the COG level on a majority rule basis. However, since flexible COGs might contain genes located in different compartments, the Shannon entropy was computed (a) to evaluate the variability of compartment assignments at the gene level within a COG (a higher Shannon entropy reflected a higher variability of the gene distribution in the different compartments for the concerned COGs); and (b) to assess the accuracy of compartment assignment at the COG level compared with its genes (a lower Shannon entropy reflected a more representative location at the COG level). For each COG, entropy was calculated as follows:

$$H(x) = -\sum_{i=1}^{n} P_i \log_2 P_i \quad (1)$$

where n is the number of genomic compartments (ISL1; ISL2; ISL2.1; ISL3; ISL4; ISL5; backbone; *ambiguous*) and P_i is the proportion of genes arising from genomic compartment i within the COG.

2.4 | Taxonomic affiliation and functional enrichments

For each core and flexible COG, functional and taxonomic annotations of the genes they contained were performed using BLASTP against the EggNOG v4.5 database (Huerta-Cepas et al., 2016). Only genes with a minimum length of 60 amino acids and hits with an *e*-value lower than 1e-5, a minimum alignment coverage of 50% and an identity of 30% were kept. Since all genes within a COG may not have a unique taxonomic affiliation, we defined a category called "uncertain", which standed for COGs encompassing genes at least affiliated with *Prochlorococcus* and/or *Synechococcus* and with other bacterial taxa. For each COG, a preliminary allocation of their genes to the different EggNOG functional categories was also performed. Genes from the "Poorly characterized" category were discarded from the functional analyses. The gene functional enrichment was subsequently assessed by computing observed/expected (O/E) ratios of functional categories according to the genomic location or taxonomic affiliation of corresponding genes. The expected values were obtained by multiplying the number of genes (core or flexible genes as a function of their genomic location or taxonomic affiliation) by the percent of total genes in each functional category. The enrichments were tested through chi-squared tests.

2.5 | Multiple sequence alignments and phylogenetic analysis

For each core and flexible COG, gene sequence alignments were performed at the amino acid level using MAFFT v7.271 (Katoh & Standley, 2013) (linsi option), and DNA sequences were imposed on the protein alignments (tranalign, EMBOSS v6.6.0.0) (Rice et al., 2000). Gaps were deleted with Gblocks (Castresana, 2000; allowing smaller final blocks with gap positions).

The trimmed alignments of single-copy core COGs found in both the MIT9312 reference genome and at least one SAG of each subpopulation were concatenated with missing sequences treated as gaps. A maximum likelihood tree was inferred with PhyML v3.0 (Guindon et al., 2010), using a GTR + I + G model of evolution, as determined by jModelTest v2.1.10 (Darriba et al., 2012), and a bootstrap threshold of 100.

2.6 | Substitution rates estimation

The nonsynonymous substitutions per nonsynonymous site (dN), the synonymous substitutions per synonymous site (dS) and their ratio (dN/dS) were estimated for all single-copy COGs common to MIT9312 (either core or flexible) and for flexible single-copy COGs not found in MIT9312 but common to at least two subpopulations. Nonsynonymous and synonymous substitution rates were calculated using the maximum likelihood method as implemented in codeml from PAML v4.8a (Yang, 2007; Yang & Nielsen, 2000).

Maximum likelihood phylogenetic trees were computed for each COG with the GTR + G model as implemented in PhyML v3.0 (Guindon et al., 2010).

The COGs data set was clustered according to d*N*, d*S* and d*N*/d*S* values using *k*-means clustering (Hartigan & Wong, 1979). The optimal number of clusters defined by the elbow method was five.

3 | RESULTS

3.1 | Prochlorococcus co-occurring subpopulation phylogenetics

We analysed 87 SAGs of the HLII ecotype (Table S1) spread over three clusters, i.e., cN2, c9301 and cN1 as defined by phylogenetic analysis of their ITS by Kashtan et al. (2014). Using a whole-genome sequence phylogeny, they showed that these SAGs were distributed over seven major subpopulations (C1 to C5, C8 and C9), also referred to as clades throughout the paper. Despite their congruency with both phylogenies supporting the same subpopulation delineation, the tree based on whole-genome sequences did not follow the monophyly of the three clusters defined with ITS sequences. To reinforce these data, we inferred a maximum likelihood phylogeny from the concatenated alignment of 1,202 core genes, using the MIT9312 strain as the outgroup (Figure 1a). Our results confirmed the robust delimitation of clades (bootstrap values > 80%) and the paraphyly of the cluster cN2 because of the C8 and C3 clustering (100% bootstrap support). Our genome-wide ANI analysis was in accordance with the phylogeny, depicting the same subpopulation demarcation (Figure 1b), with the interclade ANI being close to 94% on average when analysing all pairwise comparisons from C1 to C8. The highest identity was observed for the closest relatives C1 and C2 (97% ANI on average), whereas C9 was the most divergent with 90% ANI on average with other subpopulations (Figure 1b). This finding is consistent with its emergence as the most basal branch of our tree (Figure 1a). By comparison, intra-clade ANI was higher (>98%), except for C8 (97%) and C9 (96%) (Table S5). This is in accordance with

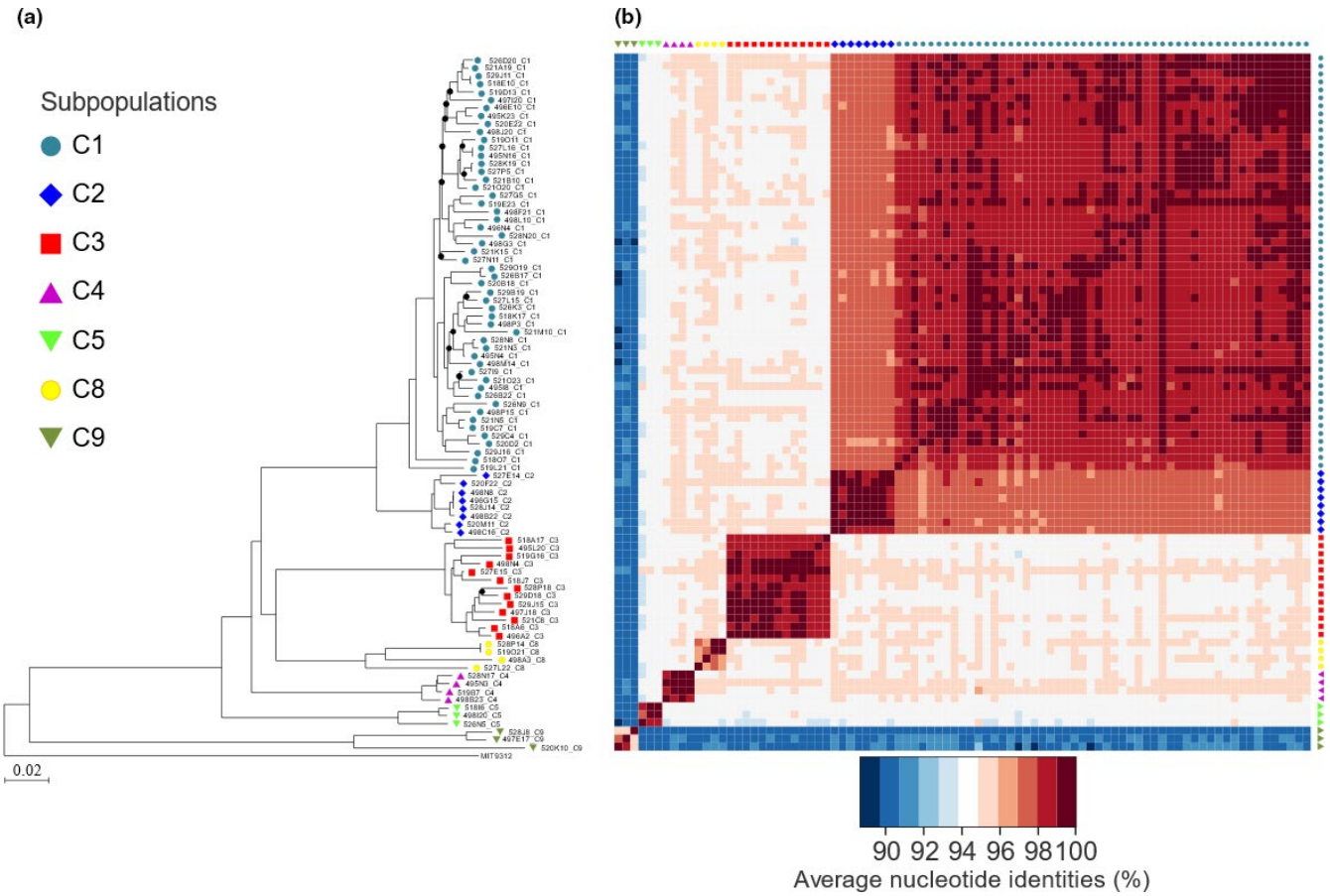


FIGURE 1 Phylogenetic relationships of 87 *Prochlorococcus* HLII ecotype single-amplified genomes (SAGs) distributed over seven major subpopulations (C1 to C5, C8 and C9). (a) The maximum likelihood phylogenetic tree inferred from the concatenated alignment of 1,202 single-copy core genes for all selected SAGs (within cN2 cluster: 52 C1, eight C2, 13 C3, four C4, three C5; within c9301 cluster: four C8 and within cN1 cluster: three C9). The reference genome MIT9312 was used to root the tree. Bootstrap supports < 80% are marked by black dots on the internal nodes. (b) Heatmap describing the pairwise average genome-wide nucleotide identity (ANI) (%) between SAG. Rows and columns are arranged according to the phylogenetic tree. Coloured dots used for both figures represent the different clades

a low intraclade polymorphism and allele differentiation between clades (Kashtan et al., 2014).

3.2 | Genome organization among subpopulations

The subpopulations were also investigated for their gene content and shared genomic regions, by aligning the genomic sequences of one representative SAG for each clade (the longest near complete SAG; Figure S1; Table S1), with the MIT9312 strain being used as the reference genome because of its equidistance to all SAGs investigated. Conserved segments of locally collinear blocks (LCBs) detected in at least seven genomes represented approximately 76% (1.33 Mb) of the MIT9312 genome length. SAG alignments covered between 79% and 86% of the MIT9312 genome length (the alignment fraction for each SAG against MIT9312 was as follow: C1 – 495K23: 1.42 Mb; C2 – 498C16: 1.35 Mb; C3 – 518A17: 1.47 Mb; C4 – 528N17: 1.38 Mb; C5 – 498I20: 1.41 Mb; C8 – 527L22: 1.37 Mb; C9 – 528J8: 1.38 Mb) (Figure S5). Therefore, genomes showed relatively high synteny, consistently to what was reported within *Prochlorococcus* ecotypes (Yan et al., 2018), however, with slight shifts in the locations of the six ISLs previously characterized in MIT9312 (Avrani et al., 2011; Table S3; Figure S5). In light of this collinearity, 5,290 single-copy COGs absent from MIT9312 were assigned to the chromosomal compartments (backbone or ISLs). A compartment was inferred for each gene within COGs, and the majority compartment was assigned at the COG level. These assignments were robust, as less than 8.5% of the COGs had a Shannon entropy equal to or higher than one (i.e., for which at least two compartments had substantial occurrence). However, as the compartment boundaries can be fuzzy, more specifically those of ISLs, the assignment of a few of these COGs should be taken with caution. Overall, 63.1% of the COGs were assigned to the backbone, 8.2% and 14.5% were allocated to ISL3 and ISL4, respectively, 8.9% were spread over ISL1, ISL2, ISL2.1 and ISL5, and the remaining 5.3% were tagged as "ambiguous". The relative density of the assigned COGs was 2.5-fold higher in ISLs (5.8 COGs per Mb) than in the backbone (2.3 per Mb) on average, except for ISL4 (11.2 per Mb) and ISL2.1 (2.3 per Mb). COGs shared by several subpopulations (≥5) were enriched in all ISLs except ISL4, whereas those found in a single subpopulation were enriched in ISL3 and ISL4.

3.3 | Taxonomic affiliation of COGs

Taxonomic analyses were performed to assess the phylogenetic origin of COGs as well as their integrity (i.e., homogeneity of their gene affiliations). Regarding the core COGs (1,309 in total; Figure 2a; Table S4; Figure S4b), 97.4% were clearly affiliated with the HLII

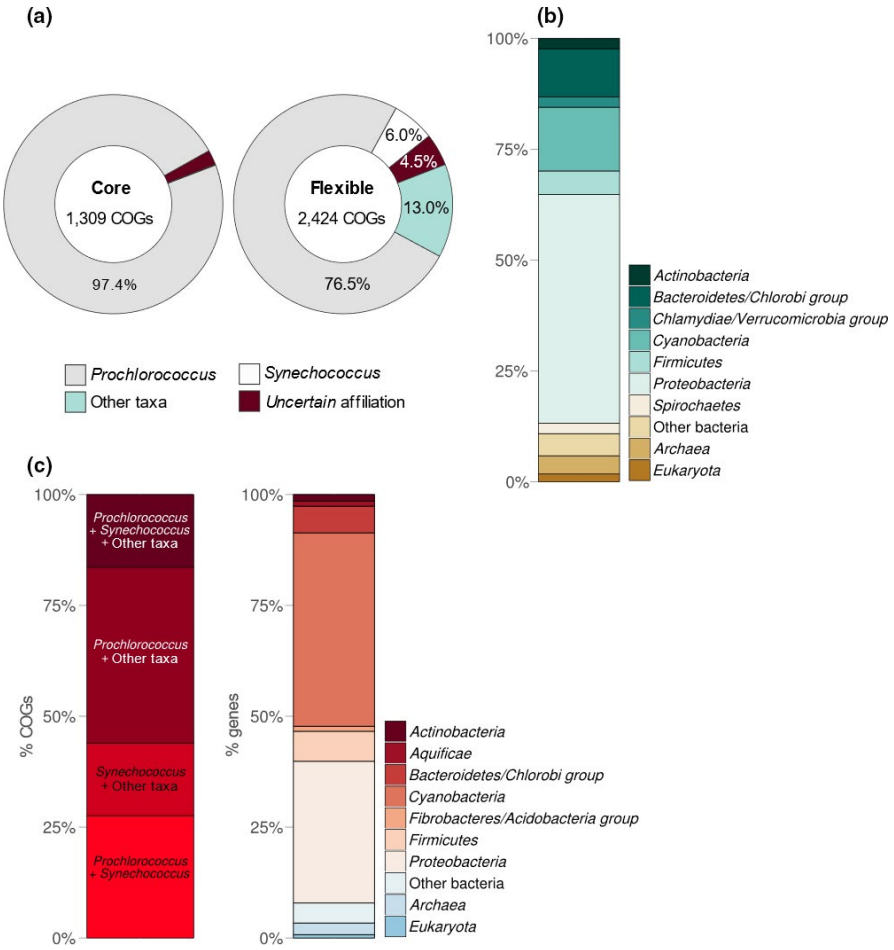


FIGURE 2 Taxonomic distributions of core and flexible clusters of orthologous genes (COGs) identified in SAGs. (a) Taxonomic affiliations of core and flexible COGs highlighting the proportion of those affiliated with the genera *Prochlorococcus* or *Synechococcus* or other taxonomic groups, or having an *uncertain* affiliation, i.e., containing genes assigned to various taxonomic groups, including *Prochlorococcus* and/or *Synechococcus*. (b) Taxonomic distributions of flexible COGs assigned to other taxa. (c) Taxonomic distributions of COGs tagged as *uncertain* with the proportion of those containing genes affiliated with *Prochlorococcus* and/or *Synechococcus* with or without other taxa (left) and the composition and abundance of genes within the category of other taxa (right). Bacterial taxa with less than 1% of abundance are grouped in the category "other bacteria" (b and c)

Prochlorococcus ecotype, with 2.9% of them containing genes affiliated with other ecotypes. The last 2.6% of COGs were tagged as *uncertain*, as they clustered genes with varying taxonomy, including *Prochlorococcus* and/or *Synechococcus* (Figure 2a).

In contrast, the phylogenetic origin of the 5,715 flexible COGs was less obvious since almost half of them contained genes with no counterpart in the EggNOG database. Regarding those with taxonomic affiliation (i.e., 2,424 in total; Figure 2a; Table S4; Figure S4c), 76.5% were related to *Prochlorococcus*, among which 94.4% consisted of genes affiliated with the HLII ecotype, 3.3% with the HLI ecotype and 2.3% with the LL ecotypes. Among the remaining COGs, 6% were related to *Synechococcus*, 13% to bacterial taxa other than *Prochlorococcus* and *Synechococcus*, and 4.5% were *uncertain* (Figure 2a). When affiliated with other bacterial taxa, COGs belonging to *Proteobacteria* overdominated (51.6%), followed by *Cyanobacteria* (14.4%), the *Bacteroidetes/Chlorobi* group (10.9%), *Firmicutes* (5.3%), *Actinobacteria* (2.3%) and *Spirochaetes* (2.3%) (Figure 2b). COGs tagged *uncertain* contained genes affiliated with *Prochlorococcus* (39.7%), *Synechococcus* (16.4%) or both, either associated with other taxa (16.4%) or not (27.6%) (Figure 2c). Taxa other than *Prochlorococcus* and *Synechococcus* were mostly *Cyanobacteria*, followed by *Proteobacteria*, *Bacteroidetes/Chlorobi* and *Firmicutes* (Figure 2c).

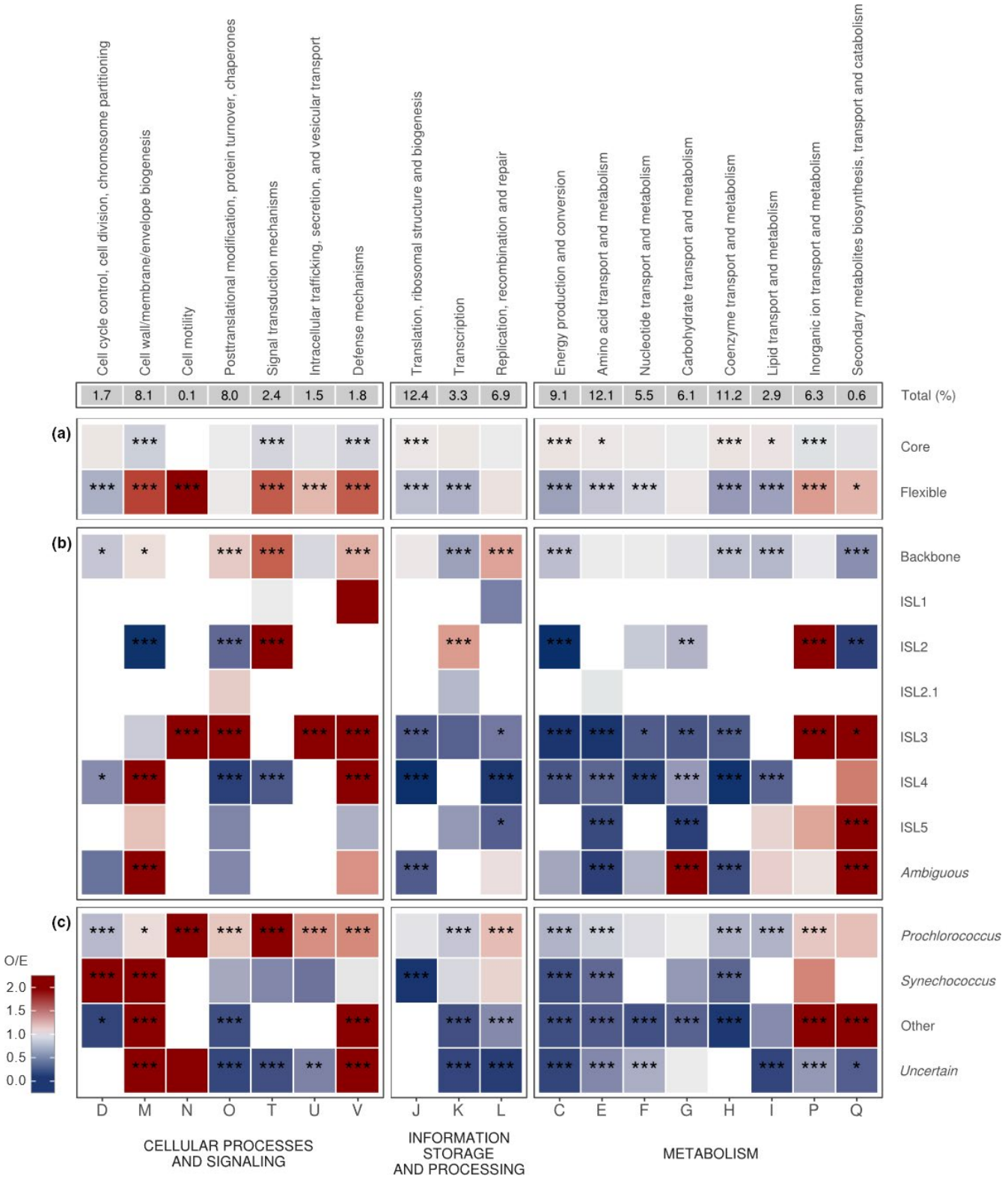
3.4 | Functional characterization of flexible COGs

The functional potential of COGs was investigated at the genomic compartment level. Only genes with known function were considered, which represented 47% of all core and flexible genes. Excepting the over-representation of the functional category "Cell motility" in clades C1 and C9, our results showed no difference in the distribution of functional categories at the subpopulation scale ($p = .39$, chi-squared test; Figure S6a). Therefore, subpopulations were considered as a whole in subsequent analyses.

The distribution of the functional categories assigned to genes from flexible COGs was compared to those from core COGs (Figure 3a) and was also analysed depending on the genomic compartments (Figure 3b) and taxonomic affiliations (Figure 3c). Our results highlighted an overall under-representation of flexible genes in the hierarchical categories "Information storage and processing" and "Metabolism", primarily impacting the functional categories involved in the mechanisms of transcription and translation as well as those in the energy production and the transport and metabolism of nucleotides, amino acids, coenzymes and lipids. This is in accordance with the fact that these categories mainly group housekeeping genes. When found in flexible COGs, these categories were preferentially located in the backbone (Figure 3b). Conversely, two functional categories were over-represented as a result of their enrichment in the ISLs (Figure 3b), namely the categories "Secondary metabolites biosynthesis, transport and catabolism", enriched in ISL3, ISL5 and in the "ambiguous" compartment (with most genes annotated as methyltransferase, thus possibly involved in DNA repair), and "Inorganic ion transport and metabolism", enriched in ISL2 and ISL3 (mainly transporters of inorganic and organic phosphate). The flexible genes encoding these functional categories mostly belonged to *Proteobacteria* and *Archaea* (Figure 3c). Interestingly, although exhibiting an O/E ratio close to 1, the category "Carbohydrate transport and metabolism" was under-represented in ISLs and over-represented in the "ambiguous" compartment (with genes encoding protein such as transketolase and transaldolase, thus possibly linked to the Calvin cycle; Figure 3b).

Regarding the hierarchical category "Cellular processes and signaling", genes associated with the flexible COGs were over-represented in most functional categories, except for "Cell cycle control, cell division and chromosome partitioning", where they were under-represented (Figure 3a). These over-representations were especially marked in the ISL3 (four out of seven functional categories) and the backbone, particularly the functional category "Signal transduction mechanisms" (with genes encoding proteins such as histidine kinases involved in response to nutrient stress), the associated genes being mainly affiliated with *Prochlorococcus* (Figure 3c). Finally, two functional categories over-represented in flexible COGs compared to core COGs, i.e., "Cell wall biogenesis" (mostly genes involved in the biosynthesis of outer membrane lipopolysaccharide protein such as glycosyltransferases and GDP-mannose 4,6-dehydratase) and "Defence mechanisms" (such as endonuclease and transporters), were found in COGs specific to a single subpopulation (Figure S6b). Most genes from these COGs (95.33%) belonged to ISL3, ISL4 and "ambiguous" compartments and were affiliated with a large variety of taxa (Figure 3c). We can notice that, despite the high proportion of SAGs in C1 compared to the one in other clades, which increased its weight in the pool of COGs specific to one clade, the same enrichments were observed in all clades (Figure S6a).

FIGURE 3 COGs functional annotations and distributions according to their category (a), location (b) and affiliation (c). The total percentages of genes (%) assigned to each EggNOG functional category (symbolized by a capital letter) are indicated. Observed/expected (O/E) ratios of core (a) and flexible (a–c) genes, according to their genomic location (backbone, genomic islands or "ambiguous") (b) and their taxonomic affiliation (*Prochlorococcus*, *Synechococcus*, other or "uncertain" taxonomic groups) (c). The observed values (O) correspond to the number of genes assigned to each functional category. The expected values (e) were obtained by multiplying the number of genes (core, flexible, or flexible genes as a function of their genomic location or taxonomic affiliation) by the total percentage of genes in each functional category. The white boxes indicate the lack of genes involved in the functional category considered. Differences in the distribution of functional categories between core and flexible genes (a), over all genomic locations (b) and over all taxonomic affiliations (c) were tested using chi-squared tests ($p < .005$ for all a, b and c groupings). Chi-squared tests were also performed for each line in the figure (each category against all others at once) to test the significance of enrichment for a given location or a given taxonomy. Chi-squared test: *, p-value < .05; **, p-value < .01; ***, p-value < .005. ISL: genomic island



3.5 | Heterogeneity of nonsynonymous (d*N*) to synonymous (d*S*) substitution rate ratios between genomic compartments

Since the F_{ST} analysis by Kashtan et al. (2014) suggested the fixation of different alleles among subpopulations that diverged at least few million years ago, it is therefore permissible to use the d*N*/d*S* approach to have deeper insights into evolution processes. In this context, to evaluate the selective pressure on COGs, we computed the ratios of d*N*/d*S* for genes in COGs recovered in at least two clades for which d*S* values could guarantee reliable estimates (Figure S7a). Thus, the C1 and C2 clade comparison was not considered because of

their low mean d*S* (±*SD* 0.05 ± 0.07; Kryazhimskiy & Plotkin, 2008). Moreover, to rule out possible bias in d*N*/d*S* ratios computed for interclade comparisons (dos Reis & Yang, 2013; Rocha et al., 2006; Wolf et al., 2009), we first compared the distribution of d*N*/d*S* ratios of core COGs from the backbone (1,139 in total) estimated from either the comparisons to MIT9312 or interclade analyses. Overall, the absence of significant difference of mean d*N*/d*S* between the two analyses (mean d*N*/d*S* ± *SD* = 0.21 ± 0.14 and 0.22 ± 0.19, respectively; $p = .35$, Wilcoxon test) associated with high similarity between the density plots (Figure 4a) and a strong positive correlation between per COGs mean d*N*/d*S* values ($\rho = 0.83$, $p < .001$, Spearman rank correlation; Figure 4b) support no bias in the evaluation of selective pressure. Similar results were obtained for the flexible COGs shared with MIT9312 (not shown). As these two approaches provided comparable results, we used the interclade analysis because it allowed taking into account the flexible genes not shared with MIT9312 and was not impacted by the over-representation of C1 members and the great number of COGs specific to this clade (Figure S7c). Then we investigated the evolutionary patterns of genomic compartments.

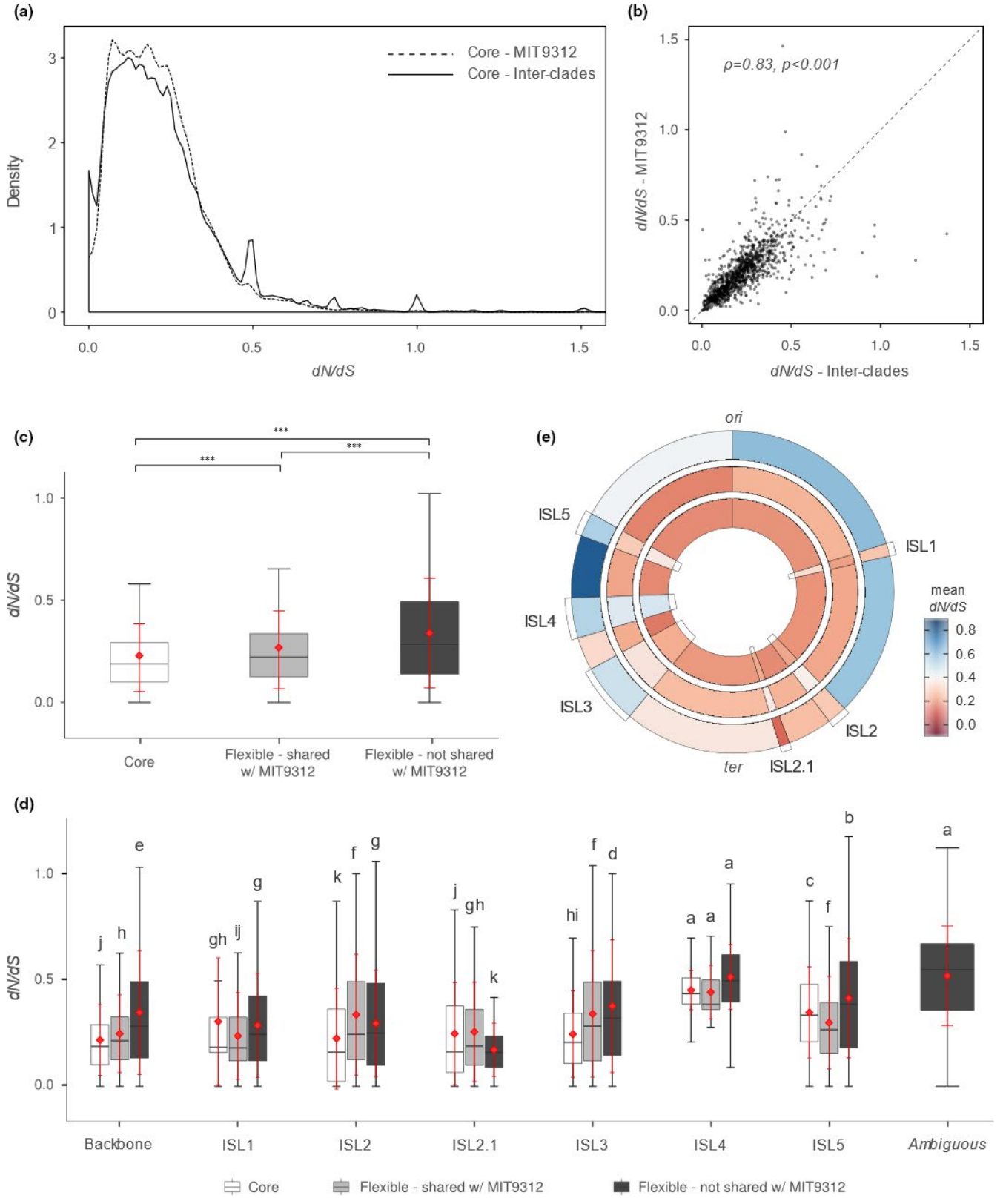
Overall, all pairwise comparisons between clades showed similar distributions of d*N*/d*S* values (Figure S7) suggesting that (i) these estimations are independent of clade abundances; and (ii) selective constraints are homogeneous among clades. Most d*N*/d*S* ratios were below 1, whatever the nature of the COGs (i.e., core or flexible), suggesting that the selective pressure was essentially negative as also found by Kashtan et al. (2014). However, the mean d*N*/d*S* ratios were significantly different between COGs assigned as core (1,202 COGs), flexible shared (310 COGs) and not shared with MIT9312 (1,033 COGs) ($p < .001$, Kruskal–Wallis test; Table S4; Figures S4c and d), the former experiencing the stronger negative selection (Figure 4c). At the compartment level, significant differences between the d*N*/d*S* ratios were also revealed ($p < .001$, Kruskal–Wallis test; Figure 4d). Thus, core and flexible genes shared with MIT9312 showed strong homogeneous selective constraints in the backbone (mean d*N*/d*S* ± *SD* ranging from 0.19 ± 0.22 to 0.23 ± 0.20 and from 0.21 ± 0.19 to 0.29 ± 0.26, respectively), while they were more variable in ISLs. Conversely, flexible genes not shared with MIT9312 exhibited variable selective constraints, unevenly scattered along the backbone. The lowest mean d*N*/d*S* ± *SD* ratio (0.29 ± 0.17) was observed in the region between ISL2 and ISL2.1, while the highest (0.73 ± 0.82) was between ISL4 and ISL5 (Figure 4e).

The analysis of ISLs also revealed contrasting patterns. ISL1, ISL2 and ISL2.1 experienced negative selective pressures similar to what was observed for core genes in the backbone except for ISL2 flexible COGs shared with MIT9312 (Figure 4d and e). Nonetheless, the profile of the ISL2.1 flexible genes not shared with MIT9312 (mean d*N*/d*S* ± *SD* = 0.17 ± 0.13) suggested a negative selective pressure stronger than the one observed in the backbone (Figure 4d). Conversely, ISL1 core genes and ISL2.1 flexible genes shared with MIT9312 showed weaker selective constraints (Figure 4d). However, these higher d*N*/d*S* values (0.35 ± 0.37 and 0.38 ± 0.62 on average ± *SD*, respectively) might be the result of a sampling bias because there was only one COG in ISL1 and three in ISL2.1. Furthermore, we observed reduced selective constraints in ISL3, ISL4 and ISL5 (ranging from [mean d*N*/d*S* ± *SD*] 0.36 ± 0.35 for ISL3 to 0.52 ± 0.62 for ISL5), except for core genes in ISL3 (mean d*N*/d*S* ± *SD* = 0.26 ± 0.25; Figure 4d and e). ISL4 and genes assigned as "ambiguous" exhibited by far the least constrained selective pressures (mean d*N*/d*S* ± *SD* from 0.44 ± 0.13 to 0.52 ± 0.23, respectively). Apparent reduced constraints on core and flexible COGs shared with MIT9312 in ISL4 might be hazardous to interpret as these values were sustained by few COGs (three core and 14 flexible).

3.6 | Substitution rate signatures of genomic compartments

To investigate substitution rate signatures depending on genomic compartments, we assessed the links between d*N*, d*S* and d*N*/d*S* estimated among genes within COGs. For the backbone, both d*N* and d*S* varied widely among COGs, whether core or flexible. Although the d*S* rates varied up to >1.5, more than 95% of estimated values were less than 0.35 (mean d*S* ± *SD* = 0.100 ± 0.091). By comparison, d*N* rates displayed lower values and variations (mean d*N* ± *SD* = 0.026 ± 0.003). Additionally, the relationship of d*N*/d*S* ratios versus d*N* or d*S* was similar when comparing core and flexible COGs (Figures S8 and S9). Five clusters of genes were distinguished based on *k*-means clustering (Figure S10). Among them, three were characterized by low d*S* values (ranging from 0.001 to 0.268), low d*N* values (<0.340) and d*N*/d*S* ratios varying from 0 to >1.5 (clusters yellow: <0.314; orange: 0.296–1.146; red: 1.133–>1.5; Figures S8

FIGURE 4 Selective pressure according to the genomic compartments. (a) Density of d*N*/d*S* ratios estimated for core COGs found in the backbone. Dashed line: values for MIT9312-clade pairwise comparisons; solid line: values for interclade pairwise comparisons. (b) Correlation between mean d*N*/d*S* of MIT9312-clade pairwise comparisons and mean d*N*/d*S* of interclade pairwise comparisons. The points represent mean d*N*/d*S* values and were averaged per COG. Spearman's rank correlation test (ρ) and the associated *p*-value are indicated. The dashed line symbolizes the diagonal. (c) Boxplot of d*N*/d*S* value distributions in the core genes and flexible genes shared or not shared with MIT9312 (Kruskal–Wallis test and post hoc Wilcoxon signed-rank test with Bonferroni correction; ***, *p*-value < .001). (d) Boxplot showing d*N*/d*S* value distributions over the genomic compartments (backbone, genomic islands and "ambiguous"). Significant differences (thus categorizing compartments with similar values) are indicated by lowercase letters, a > b > c > d > e > f > g > h > i > j > k (Kruskal–Wallis test and post hoc Wilcoxon signed-rank test with Bonferroni correction, *p*-value < .05). (c) and (d) For a better understanding, only d*N*/d*S* < 1.5 are shown; in red: mean ± the standard deviation (SD); white: core genes, light grey: flexible genes shared with MIT9312 and dark grey: flexible genes not shared with MIT9312. (e) Representation of mean d*N*/d*S* values along the chromosome (as organized in the MIT9312 reference genome) computed for genes within core COGs (inner circle) and flexible shared (middle circle) and not shared (outer circle) with MIT9312. The origin (*ori*) and terminus (*ter*) of replication are shown. ISL: genomic island



and S9). Here, high d*N*/d*S* ratios were associated with low d*S* rather than high d*N* rates, suggesting a general trend of negative selective pressure and background selection. The fourth cluster (green) was characterized by intermediary d*S* values (from 0.173 to 1.076) and low d*N* values (<0.326), reflecting more divergent sequences but still negative selective pressure. The last cluster (dark blue) showed d*S* values ranging from 0.252 to >1.5 and d*N*/d*S* ratios < 1. Here, the d*N*/d*S* ratios seemed governed by d*N* values (from 0.172 to 1.277), as illustrated by the dots linearly spread around the major diagonal of d*N*/d*S* versus d*N* plot in Figures S8 and S9.

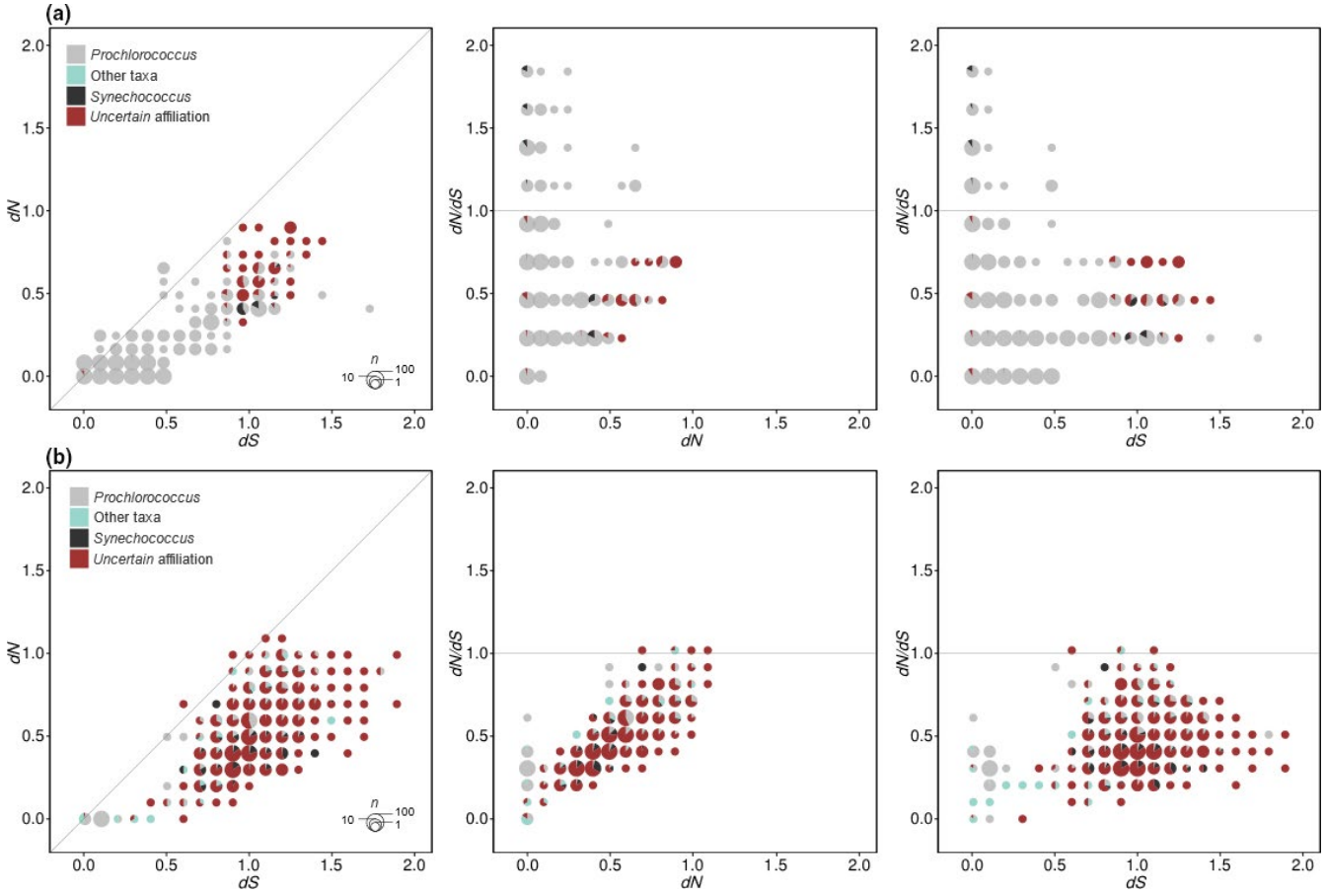


FIGURE 5 Relationships between substitution rates in flexible COGs and their taxonomic affiliations. Taxonomic affiliation – i.e., *Prochlorococcus*, *Synechococcus*, other taxonomic groups or "uncertain" – of flexible COGs in ISL3 and ISL5 (a) and ISL4 and tagged as "ambiguous"; and (b) plotted against d*N*, d*S* and d*N*/d*S* estimates. The pie charts depict the distribution of taxonomic affiliation in a 2D region of the graph corresponding to d*N*, d*S* and d*N*/d*S* values within a range of 0.1. The size of each pie chart is proportional to the number of observations *n* for the 2D region considered (at least one observation, from 10 to 100 observations, more than 100 observations). Horizontal and diagonal lines in each panel represent a d*N*/d*S* ratio equal to 1. Left: d*N* versus d*S*, Middle: d*N*/d*S* versus d*N*, Right: d*N*/d*S* versus d*S*

Regarding ISLs, the patterns of d*N* and d*S* variations among genes were contrasted, whether for core or flexible COGs. Genes distributed over the ISL1, ISL2 and ISL2.1 mostly had low and homogeneous d*N* and d*S* values (yellow, orange, red and green clusters), suggesting substantial negative selective pressure. Though their d*N*/d*S* ratios were essentially low, a few genes had values close to or greater than 1 (Figure S9). By comparison, genes in ISL4 or tagged as "ambiguous" were characterized by higher d*N* and d*S* values, with d*N*/d*S* ratios linearly linked with d*N*, and d*S* close to or at saturation (dark blue cluster). Genes located in ISL3 and ISL5, in contrast, displayed a mixed profile (all clusters).

3.7 | Evolutionary signature of COGs depending on their distribution in co-occurring subpopulations

Because flexible COGs are not recovered in all clades, we investigated the behaviours of d*N*, d*S* and d*N*/d*S* ratios according to the COGs distribution among clades. Overall, COGs were characterized by genes of low to high d*S* values, except those shared by all clades, which were depleted in genes with saturated d*S* (Figure S11). d*N*/d*S* estimates differed significantly depending on the number of clades where COGs were found ($p < .001$, Kruskal–Wallis test), i.e., those shared by a substantial number of clades tended to display lower d*N*/d*S* values.

Up to 65.5% of flexible COGs analysed for d*N*/d*S* ratios were taxonomically assigned, with a higher proportion for those found in ISL4 (78.7% affiliated) compared to other ISLs (53.8% affiliated on average). COGs with low d*S* values were essentially affiliated with *Prochlorococcus* (Figure 5a) and were mostly found in ISL3 and ISL5. COGs with saturated d*S* were assigned as "uncertain" (i.e., with multiple affiliations including *Prochlorococcus* or *Synechococcus*; Figure 5b) and were located in ISL4 and the *ambiguous* compartment. COGs affiliated with "other bacterial phyla" (neither *Prochlorococcus* nor *Synechococcus*) were characterized by both low d*N* and d*S* suggesting the close origin of their shared genes and were enriched in ISL3 and ISL4 ($p < .005$, chi-squared test).

4 | DISCUSSION

Bacterial species diversification to adapt to specific niches is documented not only for micro-organisms in the context of experimental conditions (Wiser et al., 2013) but also for some environmental bacteria (Kent et al., 2016; Larkin et al., 2016; Shapiro et al., 2012). In this study, we consider co-occurring SAGs of the *P. marinus* HLII ecotype, for which phylogenetic analyses of ITS (Kashtan et al., 2014) or concatenated single-copy core COGs (this study) depicted a structuring into clades that might reflect ancient niche partitioning. This finding is also supported by the predominance of different alleles in core genes that are fixed within subpopulations and distinct sets of flexible genes among them (Kashtan et al., 2014). Moreover, these subpopulations were characterized by variations in their relative abundance with seasonality (Kashtan et al., 2014) suggesting an adaptation; however, the drivers of this differentiation remain elusive (Larkin et al., 2016). Here, d*N*/d*S* values supported general negative selection among clades, in the same way as what was observed at the intraclade level (Kashtan et al., 2014), while no positive selection was found. This finding is congruent with the analysis of prevalent species in the human gut, which suggested that positive selection, if present, may not overpower the signal of negative selection (Garud et al., 2019). Furthermore, Marttinen et al. (2015) showed that a parsimonious model without niche or diversifying selection, but including recombination, induced structured populations within a stable range of genetic diversity. Thus, adaptation might not be necessary to explain the structuring of this HLII population into clades. To decipher between masked or absence of positive selection, similar analysis of d*N*/d*S* ratio could be performed on HLI ecotype subpopulations, for which a stronger correlation between seasonality and environmental factors was observed (Larkin et al., 2016).

Uneven negative selective constraints were detected on flexible genes along the backbone and among ISLs (Figure 4d and e), supporting allelic variations along the genome (Kashtan et al., 2014). Overall, ISL1, ISL2 and ISL2.1 showed COGs under strongest purifying selection, whereas ISL3, ISL4 and ISL5 tended to concentrate COGs with relaxed selection. ISL1, ISL2 and ISL2.1 were predominantly characterized by COGs shared by all clades and affiliated with *Prochlorococcus* (Figures S8 and S9), suggesting their ongoing "fixation". Moreover, the high F_{ST} values observed for COGs in ISL2 and ISL2.1 (Figure S12) suggest their potential role in stable niche partitioning. For instance, in ISL2, two COGs shared by all subpopulations and subject to strong selective constraints (means d*N*/d*S* < 0.19) were related to phosphonate utilization, which could suggest adaptation to low-phosphate environments (Feingersch et al., 2012). Recently, Schmutzer and Barraclough (2019) suggested that, in the presence of gene fluxes among diverging populations, the concentration of locally adapted genes in a reduced number of loci could be favoured, as it would (a) reduce the negative impact of insertions along the genome of horizontally transferred genes; and (b) increase the relative efficacy of selection on a few "mega" loci compared to many dispersed loci of reduced effect. Thus, ISL1, ISL2 and ISL2.1 might concentrate "flexible" genes that are essential for all these clades. They might, therefore, be considered as "core" in these clades, as proposed for ISL2.1 regarding the HLII ecotype (Avrani et al., 2011).

Selective signatures in the flexible genome also revealed two sets of COGs. A first set (consisting of the yellow, orange and red clusters), found in all genomic compartments (i.e., backbone and ISLs), was characterized by low d*N* values and d*S* below the mean d*S* of core genes (Figures S8 and S9). Although this is consistent with general background selection (Price & Arkin, 2015), the low d*S*, driving an unusually high d*N*/d*S* (red cluster), could also reflect strong negative selection on synonymous substitutions (Parmley & Hurst, 2007) or homogenization of sequence diversity among clades through homologous recombination (HR) (Hanage, 2016). HR could increase selection efficacy by reducing the Hill-Robertson effect (Hill & Robertson, 1966). In a context of structured populations, interclade HR could also increase N_e for genes whose circulation would spread beyond subpopulations, while constraining clade differentiation within a divergence mode (Marttinen et al., 2015). This is in line with the combinatorial nature of backbone and flexible genes as proposed by Kashtan et al. (2014) for the high-light or phosphonate related genes. COGs in the second set (dark blue cluster), primarily found in ISL3, ISL4 and ISL5, displayed relaxation of selective constraints associated with high d*S*, suggesting horizontal gene transfer (HGT) (Castillo-Ramírez et al., 2011). Moreover, they were not affiliated with *Cyanobacteria* (Figure 5). HGT is recognized as a driver of evolution, contributing to the adaptation to changing environments through the expansion and conversion of gene families (Gogarten et al., 2002; Ochman et al., 2000; Wiedenbeck & Cohan, 2011). Therefore, the over-representation of genes involved in defence mechanisms or cell wall biogenesis in ISL3 and ISL4 (Figure 3) could reflect the acquisition of genes that may be transiently adaptive during phage infection periods (Avrani et al., 2011; Coleman et al., 2006; Kettler et al., 2007). However, adaptive HGTs are primarily documented for genes with large selective advantage, which might not be true for most of them. In case of near neutral HGT, selection depends on N_e (Kuo & Ochman, 2009), with potentially different outcomes on the flexible genome. For McInerney et al. (2017), as large N_e enhances selection efficacy, the flexible genome is essentially adaptive, with slightly deleterious acquired genes being eliminated. However, neutral evolution of the pangenome could not be rejected (Andreani et al., 2017; Baumdicker et al., 2012; Vos & Eyre-Walker, 2017). Additionally, in the drift-barrier model (Bobay & Ochman, 2018; Lynch, 2010), the loss of flexible genes is random for small N_e, while larger N_e would increase (a) the proportion of genes with selection coefficient $s < 0$ that would be perceived as deleterious; (b) the fixation probability of slightly advantageous genes; and (iii) the fixation time or loss of quasi neutral genes, and thus the size and diversity of the flexible genome. It was proposed that the large pangenome of *Prochlorococcus* species might originate from their large N_e (estimated between 10^6 [Price & Arkin, 2015] and 10^{13} [Kashtan et al., 2014]). However, in the case of structured populations, as observed here, the evolution of a near neutral gene acquired from distant lineage might be restricted to the

clade in which it was introduced. Thus, distant HGT might evolve in a context of lower N_e, a pattern that might explain the ISL4 characteristics (i.e., enriched in COGs specific to one clade, involved in defence mechanisms and not affiliated to *Cyanobacteria*) (Figure 5b). However, the N_e of HGT could also be enlarged through, e.g., local HR, favouring a selective footprint and the persistence of acquired genes with marginal effect. The occurrence of both HGT (high d*S* with uncertain affiliation) and selected genes (low d*N*/d*S* with *Prochlorococcus* affiliation) in ISL3 and ISL5 (Figure 5a) could be the result of such processes. Overall, our results highlighting two sets of genes with distinct evolutionary trajectories (i.e., strong negative selection versus selection relaxation) in a structured population are in accordance with the drift-barrier model. Furthermore, the structuring of the genetic information along the genome might depend on the dynamics of gene fluxes among clades within a structured population, especially for flexible genes. Rather than a nonrandom acquisition of genes with regard to their genomic location, we may consider the differential retention probability of transferred genes as a consequence of fluctuating N_e along the genome.

ACKNOWLEDGEMENTS

The work of H.G. was supported by a PhD fellowship funded by the Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. We are grateful to Cécile Lepère with the improvement of the manuscript.

AUTHOR CONTRIBUTIONS

G.B., and C.B.-P. conceived and coordinated this study. G.B., H.G., and C.B.-P. designed the study. H.G., G.B., C.B.-P., and I.J.-D. performed the analysis and analysed the data. H.G., G.B., C.B.-P., and I.J.-D. contributed to the writing and the reviewing of the manuscript.

DATA AVAILABILITY STATEMENT

All data used for this manuscript are available and open access. Genomic sequences are available on the NCBI genome assembly database, BioProject PRJNA239833, PRJNA239872 and PRJNA239873. FTP addresses of these genomic sequences downloaded from the NCBI are provided in Table S2. Sequences of genes classified into clusters of orthologous genes (COGs) that were used for the analyses in this manuscript are available on DRYAD https://datadryad.org/stash/dataset/doi:10.5061/dryad.9r0p6.

ORCID

Hélène Gardon https://orcid.org/0000-0001-8976-1372

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andreani, N. A., Hesse, E., & Vos, M. (2017). Prokaryote genome fluidity is dependent on effective population size. *The ISME Journal*, *11*(7), 1719–1721. https://doi.org/10.1038/ismej.2017.36

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., & Lindell, D. (2011). Genomic island variability facilitates *Prochlorococcus*–virus coexistence. *Nature*, *474*(7353), 604–608. https://doi.org/10.1038/nature10172

Baumdicker, F., Hess, W. R., & Pfaffelhuber, P. (2012). The infinitely many genes model for the distributed genome of bacteria. *Genome Biology and Evolution*, *4*(4), 443–456. https://doi.org/10.1093/gbe/evs016

Biller, S. J., Berube, P. M., Berta-Thompson, J. W., Kelly, L., Roggensack, S. E., Awad, L., Roache-Johnson, K. H., Ding, H., Giovannoni, S. J., Rocap, G., Moore, L. R., & Chisholm, S. W. (2014). Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Scientific Data*, *1*, 140034. https://doi.org/10.1038/sdata.2014.34

Bobay, L.-M., & Ochman, H. (2018). Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evolutionary Biology*, *18*(1), 153. https://doi.org/10.1186/s12862-018-1272-4

Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., Krause, D. J., & Whitaker, R. J. (2012). Patterns of gene flow define species of Thermophilic Archaea. *PLOS Biology*, *10*(2), e1001265. https://doi.org/10.1371/journal.pbio.1001265

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Castillo-Ramírez, S., Harris, S. R., Holden, M. T. G., He, M., Parkhill, J., Bentley, S. D., & Feil, E. J. (2011). The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Path*, *7*(7), e1002129. https://doi.org/10.1371/journal.ppat.1002129

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Coleman, M. L., Sullivan, M., Martiny, A., Steglich, C., Barry, K., DeLong, E., & Chisholm, S. (2006). Genomic Islands and the ecology and evolution of *Prochlorococcus*. *Science*, *311*(5768), 1768–1770. https://doi.org/10.1126/science.1122050

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: Multiple genome alignment with gene gain, Loss and Rearrangement. *PLoS One*, *5*(6), e11147. https://doi.org/10.1371/journal.pone.0011147

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012, August 1). JModelTest 2: More models, new heuristics and parallel computing. *Nature Methods*, *9*, 772. https://doi.org/10.1038/nmeth.2109

dos Reis, M., & Yang, Z. (2013). The unbearable uncertainty of Bayesian divergence time estimation. *Journal of Systematics and Evolution*, *51*(1), 30–43. https://doi.org/10.1111/j.1759-6831.2012.00236.x

Eddy, S. R. (2009). A new generation of homology search tools based on probalistic inference. *Genome Informatics*, *2009*, 205–211. https://doi.org/10.1142/9781848165632_0019

Feingersch, R., Philosof, A., Mejuch, T., Glaser, F., Alalouf, O., Shoham, Y., & Béjà, O. (2012). Potential for phosphite and phosphonate utilization by *Prochlorococcus*. *ISME Journal*, *6*(4), 827–834. https://doi.org/10.1038/ismej.2011.149

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., & Martiny, A. C. (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(24), 9824–9829. https://doi.org/10.1073/pnas.1307701110

Garud, N. R., Good, B. H., Hallatschek, O., & Pollard, K. S. (2019). Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLOS Biology*, *17*(1), e3000102. https://doi.org/10.1371/journal.pbio.3000102

Gogarten, J. P., Doolittle, W. F., & Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*, *19*(12), 2226–2238. https://doi.org/10.1093/oxfordjournals.molbev.a004046

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to

estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Hanage, W. P. (2016). Not so simple after all: Bacteria, their population genetics, and recombination. *Cold Spring Harbor Perspectives in Biology*, *8*(7), a018069. https://doi.org/10.1101/cshperspect.a018069

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, *8*(3), 269–294. https://doi.org/10.1017/S0016672300010156

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C., & Bork, P. (2016). eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–D293. https://doi.org/10.1093/nar/gkv1248

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. r, Stocker, R., Follows, M. j, Stepanauskas, R., & Chisholm, S. W. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science*, *344*(6182), 416–420. https://doi.org/10.1126/science.1248575

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. R., Stocker, R., Follows, M. J., Stepanauskas, R., & Chisholm, S. W. (2015). Data from: Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.9r0p6

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kelly, L., Huang, K. H., Ding, H., & Chisholm, S. W. (2012). ProPortal: A resource for integrated systems biology of *Prochlorococcus* and its phage. *Nucleic Acids Research*, *40*(D1), D632–D640. https://doi.org/10.1093/nar/gkr1022

Kent, A. G., Baer, S. E., Mouginot, C., Huang, J. S., Larkin, A. A., Lomas, M. W., & Martiny, A. C. (2019). Parallel phylogeography of *Prochlorococcus* and *Synechococcus*. *ISME Journal*, *13*(2), 430–441. https://doi.org/10.1038/s41396-018-0287-6

Kent, A. G., Dupont, C. L., Yooseph, S., & Martiny, A. C. (2016). Global biogeography of *Prochlorococcus* genome diversity in the surface ocean. *ISME Journal*, *10*(8), 1856–1865. https://doi.org/10.1038/ismej.2015.265

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G. M., Richardson, P., & Chisholm, S. W. (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLOS Genetics*, *3*(12), e231. https://doi.org/10.1371/journal.pgen.0030231

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

Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. *PLOS Genetics*, *4*(12), e1000304. https://doi.org/10.1371/journal.pgen.1000304

Kuo, C.-H., & Ochman, H. (2009). The fate of new bacterial genes. *FEMS Microbiology Reviews*, *33*(1), 38–43. https://doi.org/10.1111/j.1574-6976.2008.00140.x

Larkin, A. A., Blinebry, S. K., Howes, C., Lin, Y., Loftus, S. E., Schmaus, C. A., Zinser, E. R., & Johnson, Z. I. (2016). Niche partitioning and biogeography of high light adapted *Prochlorococcus* across taxonomic ranks in the North Pacific. *ISME Journal*, *10*(7), 1555–1567. https://doi.org/10.1038/ismej.2015.244

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*(8), 345–352. https://doi.org/10.1016/j.tig.2010.05.003

Malmstrom, R. R., Coe, A., Kettler, G. C., Martiny, A. C., Frias-Lopez, J., Zinser, E. R., & Chisholm, S. W. (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *The ISME Journal*, *4*(10), 1252–1264. https://doi.org/10.1038/ismej.2010.60

Marttinen, P., Croucher, N. J., Gutmann, M. U., Corander, J., & Hanage, W. P. (2015). Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*, *1*(5), e000038. https://doi.org/10.1099/mgen.0.000038.

Mayr, E. (1942). Systematics and the origin of species. *Annals of the Entomological Society of America*, *36*(1), 138–139. https://doi.org/10.1093/aesa/36.1.138a

McInerney, J. O., McNally, A., & O'Connell, M. J. (2017). Why prokaryotes have pangenomes. *Nature Microbiology*, *2*(4), 17040. https://doi.org/10.1038/nmicrobiol.2017.40

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, *15*(6), 589–594. https://doi.org/10.1016/J.GDE.2005.09.006

Moore, L. R., Rocap, G., & Chisholm, S. W. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature*, *393*(6684), 464–467. https://doi.org/10.1038/30965

Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, *405*(6784), 299–304. https://doi.org/10.1038/35012500

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

Parmley, J. L., & Hurst, L. D. (2007). How common are intragene windows with K A > K S owing to purifying selection on synonymous mutations? *Journal of Molecular Evolution*, *64*(6), 646–655. https://doi.org/10.1007/s00239-006-0207-7

Partensky, F., Hess, W. R., & Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiology and Molecular Biology Reviews*, *63*(1), 106–127. https://doi.org/10.1128/MMBR.63.1.106-127.1999

Price, M. N., & Arkin, A. P. (2015). Weakly deleterious mutations and low rates of recombination limit the impact of natural selection on bacterial genomes. *Mbio*, *6*(6), e01302–e1315. https://doi.org/10.1128/mBio.01302-15

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., & Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: Soft-rotting enterobacterial plant pathogens. *Analytical Methods*, *8*(1), 12–24. https://doi.org/10.1039/C5AY02550H

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends in Genetics:TIG*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., … Chisholm, S. W. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, *424*(6952), 1042–1047. https://doi.org/10.1038/nature01947

Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., & Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*, *239*(2), 226–235. https://doi.org/10.1016/J.JTBI.2005.08.037

Schmutzer, M., & Barraclough, T. G. (2019). The role of recombination, niche-specific gene pools and flexible genomes in the ecological

speciation of bacteria. *Ecology and Evolution*, *9*(8), 4544–4556. https://doi.org/10.1002/ece3.5052

Sela, I., Wolf, Y. I., & Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences*, *113*(41), 11399–11407. https://doi.org/10.1073/pnas.1614083113

Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., Polz, M. F., & Alm, E. J. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science (New York, N.Y.)*, *336*(6077), 48–51. https://doi.org/10.1126/science.1218198

Stolyar, S., & Marx, C. J. (2019). Align to Define: Ecologically meaningful populations from genomes. *Cell*, *178*(4), 767–768. https://doi.org/10.1016/J.CELL.2019.07.026

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955 https://doi.org/10.1073/pnas.0506758102

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., & Pati, A. (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Research*, *43*(14), 6761–6771. https://doi.org/10.1093/nar/gkv657

Vos, M., & Eyre-Walker, A. (2017). Are pangenomes adaptive or not? *Nature Microbiology*, *2*(12), 1576. https://doi.org/10.1038/s41564-017-0067-5

Wiedenbeck, J., & Cohan, F. M. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews*, *35*(5), 957–976. https://doi.org/10.1111/j.1574-6976.2011.00292.x

Wiser, M. J., Ribeck, N., & Lenski, R. E. (2013). Long-term dynamics of adaptation in asexual populations. *Science*, *342*(6164), 1364–1367. https://doi.org/10.1126/science.1243357

Wolf, J. B. W., Künstner, A., Nam, K., Jakobsson, M., & Ellegren, H. (2009). Nonlinear Dynamics of Nonsynonymous (dN) and Synonymous (dS) substitution rates affects inference of selection. *Genome Biology and Evolution*, *1*, 308–319. https://doi.org/10.1093/gbe/evp030

Yan, W., Wei, S., Wang, Q., Xiao, X., Zeng, Q., Jiao, N., & Zhang, R. (2018). Genome rearrangement shapes *Prochlorococcus* ecological adaptation. *Applied and Environmental Microbiology*, *84*(17), e01178–e1218. https://doi.org/10.1128/AEM.01178-18

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yang, Z., & Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, *17*(1), 32–43. https://doi.org/10.1093/oxfordjournals.molbev.a026236

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Gardon H, Biderre-Petit C, Jouan-Dufournel I, Bronner G. A drift-barrier model drives the genomic landscape of a structured bacterial population. *Mol. Ecol.*2020;00:1–14. https://doi.org/10.1111/mec.15628

Article 2 / Co-autrice : Loiseau *et al.*, 2018

Loiseau, C., Hatte, V., Andrieu, C., Barlet, L., Cologne, A., De Oliveira, R., Ferrato-Berberian, L., Gardon, H., Lauber, D., Molinier, M., *et al.* (2018) 'PanGeneHome: a web Interface to analyze microbial pangenomes', *arXiv preprint arXiv:1805.05597*.

Bioinformatics Computational & Systems Biology

http://elynsgroup.com

Research Article | Journal of Bioinformatics, Computational and Systems Biology | Open Access

PanGeneHome : A Web Interface to Analyze Microbial Pangenomes

Camille Loiseau#, Victor Hatte#, Charlotte Andrieu, Loic Barlet, Audric Cologne, Romain De Oliveira, Lionel Ferrato-Berberian, Hélène Gardon, Damien Lauber, Mélanie Molinier, Stéphanie Monnerie, Kissi N'Gou, Benjamin Penaud, Olivier Pereira, Justine Picarle, Amandine Septier, Antoine Mahul, Jean-Christophe Charvy, and François Enault*

Department of Biology, Clermont Auvergne University, Clermont-Ferrand, F-63000, France

Received Date: October 24, 2017, **Accepted Date:** November 23, 2017, **Published Date:** November 30, 2017.

***Corresponding author:** François Enault, UMR CNRS 6023 Microorganisms: Genome and Environment, Build. A, 24 avenue des Landais, 63177 Aubière Cedex. La France, Tel: 33-(0)473- 407-471; Fax: 33-(0)473-407-670: E-mail: francois.enault@uca.fr.*

Abstract

PanGeneHome is a web server dedicated to the analysis of available microbial pangenomes. For any prokaryotic taxon with at least three sequenced genomes, PanGeneHome provides (i) conservation level of genes, (ii) pangenome and core-genome curves, estimated pangenome size and other metrics, (iii) dendrograms based on gene content and average amino acid identity (AAI) for these genomes, and (iv) functional categories and metabolic pathways represented in the core, accessory and unique gene pools of the selected taxon. In addition, the results for these different analyses can be compared for any set of taxa. With the availability of 615 taxa, covering 182 species and 49 orders, PanGeneHome provides an easy way to get a glimpse on the pangenome of a microbial group of interest. The server and its documentation are available at http://pangenehome.lmge.uca.fr.

Keywords: Prokaryote pangenome; Comparative genomics; Bioinformatics; Web site

Introduction

Recent advances in DNA sequencing technology led a rapid accumulation of microbial genomic data. Due to their inherent genomic plasticity, microbial species are now described throughout their pangenome and not just through a unique reference genome ([1–4]; for a review, see [5]). A microbial pangenome is composed of core genes (shared by all strains), dispensable/accessory genes (conserved in two or more strains) and genes unique to single strains. By comparing the number and conservation of genes across multiple genomes, researchers can thus gain insights into the genomic diversity, dynamics and evolution of a microbial taxon. Furthermore, the functional annotations of the core, accessory and unique genes can also be informative.

Several standalone tools (e.g. PGAP [6], PANNOTATOR [7], PanGP [8], Roary [9] and BPGA [10]) and web servers (e.g. Panseq [11], PGAT [12] and PanWeb [13]) dedicated to pangenome analysis have been developed recently and offer the possibility to compute pangenome analysis for genomes provided by a user (for a review of the different existing tools, see [14]). Among these, PGAT [12] is the only web site offering pre-computed pangenome analysis but only nine genus (representing 244 genomes) are available. Thus, as collecting genomes and running these tools implies a significant effort on the user side, we developed PanGeneHome, a Web server that offers precomputed pangenome analysis at a large scale for already sequenced genomes. Browsing PanGeneHome, pangenome analysis results can be directly accessed for any taxonomic level in the bacterial and archaeal trees for which the collection of publicly available genomes is sufficient (> 2 genomes).

Methods

PanGeneHome provides users with a suite of tools for analyzing the pangenome of a selected taxon. To this end, the 2,674 bacterial and 167 archaeal complete genomes available in KEGG [15] were processed (Jan 2015). For any taxon (node or leaf) of the bacterial and archaeal trees that include at least three genomes, the pangenome was determined using KEGG Orthologous Clusters (OCs). These OCs were constructed using the 8,912,641 protein coding genes in all complete genomes (~ 3 % of the genes are encoded in plasmids). The KEGG functional categories to which these OCs are affiliated were included. KEGG orthologous (KO) groups [15], was also used to determine distances based on gene content and on the AAI of the shared genes.

The first step for the user is to select a taxon of interest, at any taxonomic level (phylum, class, order, family, genus or species level). This selection is made through a browsable and searchable tree, encoded using the jQuery plugin jsTree (https://www.jstree.com/). The different PanGeneHome sections described below were constructed as flat-file databases.

Gene Conservation

For each taxon, the gene conservation was determined in two ways: (i) the conservation of all OCs of the taxon and (ii) the conservation of OCs in an average genome of this taxon. For a given taxon, the conservation of an OC is the percentage of genomes of this taxon in which this OC appears. These percentages were subsequently used to compute the distribution of the gene conservation for each genome, and these distributions were averaged for all the genomes of the considered taxon. These two results are displayed through interactive histograms with the Highcharts JavaScript library (http://www.highcharts.com).

Pan- and Core-Genome Curves

The pan- and core-genome curves display respectively the total number of different OCs and the number of conserved OCs when considering an increasing number of genomes. When the number of possible genome combination was too large (e.g. there is more than 17 thousand billion different combination of 10 genomes out of 100), a random subset of 1,000 combinations was used. For each genome number considered, the average pan- and core-genome numbers is plotted with the standard deviation being represented by shaded zones around these two curves. Here again, Highcharts library (www.highcharts.com) was used to display these curves.

In addition, pangenome size, closedness and diversity were estimated using the R package micropan [16]. The Chao method was used to estimate the pan-genome size, a method that gives « *a conservative estimate, i.e. it tends to be on the smaller side of the true size* » [16]. To predict if a pangenome is open or closed, a Heaps law type of model was used and if the alpha value is below one, the pangenome is considered as open (see [17] for details). Finally, the genomic fluidity [18] was computed to quantify the pangenome diversity.

Gene Content Tree

The distance between two genomes was defined as the fraction

of genes unique to one of the two strains [19]. In details, for two genomes A and B, we counted the number of genes of A not present in B divided by the total number of genes of A. The dendrogram for each taxon was then determined by applying the PHYLIP Neighbor Joining method [20] to the corresponding distance matrix. To be able to compare evolutionarily distant species, KEGG KOs and not OCs were here considered. The tree for all bacteria is not available as it contains too many genomes (2674) to be visualized. These trees are displayed as circular and linear trees using the jsPhyloSVG JavaScript plugin [21] and full species name can be obtained by mousing over the corresponding leaf.

AAI

The AAI of the shared genes between all genome pairs was determined as in [22]. Shortly, the identity percentage of all proteins inside a KO were determined using BLASTp [23], and for all genome pairs, the average amino acid identity was computed using all their shared proteins. The AAI between a pair of genomes A and B is thus the average of the amino acid identity using all the pairs of proteins of A and B that are present in the same orthologous group. For each taxon, the resulting AAI matrix was then used to build a dendrogram with the neighbor-joining method (PHYLIP suite [20]). Here again, KEGG KOs were used to enable the comparison of evolutionarily distant species, and dendrograms are displayed using the jsPhyloSVG plugin [21].

Functional Analysis of Core, Accessory and Unique Genes

The functional annotations of the core, accessory and unique genes, defined here by the OC clustering, can also be displayed and compared. To this end, the KEGG pathway database was used [15]. First, the number and percentages of genes involved in the main categories (e.g. « Metabolism », « Genetic information processing », etc...) of this database were calculated for core, accessory and unique genes and displayed as histograms. Second, similar results were computed for a lower level of the pathway database (e.g. « Carbohydrate metabolism », « Replication and repair », etc.) through curves. These interactive histograms and curves were developed using the Highcharts JavaScript library (www.highcharts.com).

Pan- and Core-Genome Comparison

Multiple taxa can be selected through an interactive tree, and the corresponding pan- and core-genome curves (defined previously in section 2) are displayed. The different metrics computed are also provided in a table.

Core Function Comparison

Here again, multiple taxa can selected through an interactive tree, and the functional annotations of the core genes of each taxon are displayed through Highcharts histograms.

Results and Discussion

PanGeneHome is a web server where pangenome analyses are available for all taxon that contain at least three sequenced genomes. The 2,841 genomes of KEGG allowed us to determine the pangenome of 10 phyla, 16 classes, 49 orders, 112 families, 164 genera and 182 species. Among these, 100 and 50 genera have respectively at least 5 and 10 sequenced genomes. This is to our knowledge the first web server where pangenomes are processed for all available genomes and for any taxonomic level.

Details on the Protein Clustering Methods

The identification of orthologs is an important cornerstone for pangenome analysis. Here, two different clustering methods available in KEGG were used. The main difference between these methods is the granularity of the orthologous groups they produce :

- The OCs are constructed by automatically clustering proteins based on their sequence similarities and using a quasi-clique-based method [24]. All protein coding genes are thus included in the clustering and it produces fine-grained clusters. Indeed, the 8,912,641 KEGG proteins are clustered into 358,067 OCs (295 OCs are larger than 1,000 proteins) and 474,400 proteins remained as singletons.
- Complementary to this clustering into OCs, KEGG proteins are also assigned to KOs based on cross-species genome comparison using the KOALA (KEGG Orthology and Links Annotation) system [15]. The KOs produced do not include all proteins and are much larger than OCs: the 4,381,566 proteins assigned to a KO are clustered into 8,252 different KOs. As a comparison, the same proteins are clustered into 160,669 different OCs.

Contrary to KOs, OCs include all genes and were thus used to determine the different categories of a pangenome in a precise manner (core, accessory and unique genes). As KOs group even distantly related homologs (that are separated into several Ocs), KOs were used in gene content and AAI methods in order to compare evolutionarily distant genomes.

Pangenehome through an Example

To illustrate the results that can be obtained with PanGeneHome, microbial species with various characteristics were here chosen. We selected three species described to have a closed pan-genome, namely *Bacillus anthracis* [17,25], *Buchnera aphidicola* [26] and *Campylobacter jejuni* [27], alongside three species that were described to have an open pangenome, namely *Bacillus thuringiensis* [28], *Propionibacterium acnes* [29] and *Prochlorococcus marinus* [17,30], and the species on which the concept of pangenome was initially tested, *Streptococcus agalactiae* [31].

The pangenome curves (Figure 1) and metrics (Table 1) are very different for these seven species considered. Indeed, the curves for *B. thuringiensis* and *P. marinus* keep increasing, even when more than 10 genomes are considered. Moreover, all metrics point out this trend as the estimated size of their pangenome are large (21,427 and 6,492 genes), their fluidity is larger than the one of the other species (> 0.18) and their Heap value lower (< 0.7). All this indicate that these species do have an open pangenome. Conversely, the pangenome curves of *B. anthracis* and *B. aphidicola* seems to

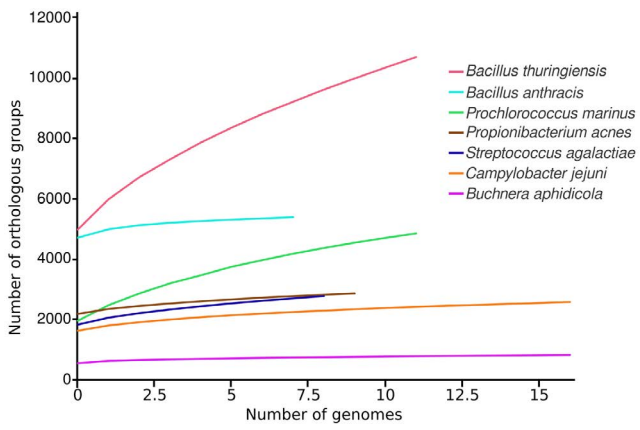


Figure 1: Pan-genome curves of seven different bacterial species. The average pan-genome size (i.e. the average number of total orthologous groups) is shown for the indicated number of genomes.

Taxonomy	#Genomes	Chao	Heaps Alpha	Heaps Intercept	Fluidity Mean ± deviation
Bacillus thuringiensis	12	21427	0.616	1499	0.188 ± 0.051
Prochlorococcus marinus	12	6492	0.644	792	0.272 ± 0.103
Propionibacterium acnes	10	3363	0.705	230	0.08 ± 0.023
Campylobacter jejuni	17	3426	0.812	280	0.11 ± 0.047
Streptococcus agalactiae	9	4036	0.842	414	0.137 ± 0.017
Buchnera aphidicola	17	1001	1.342	176	0.15 ± 0.124
Bacillus anthracis	8	5932	1.703	922	0.059 ± 0.023

Estimated pangenome size (Chao), closedness (Heaps Alpha and Intercept) and diversity (Genomic fluidity), computed using the R package micropan [16]. Taxon with a heaps alpha value below one are considered to have an open pangenome.

Table 1: Estimated pangenome metrics for different bacterial species.

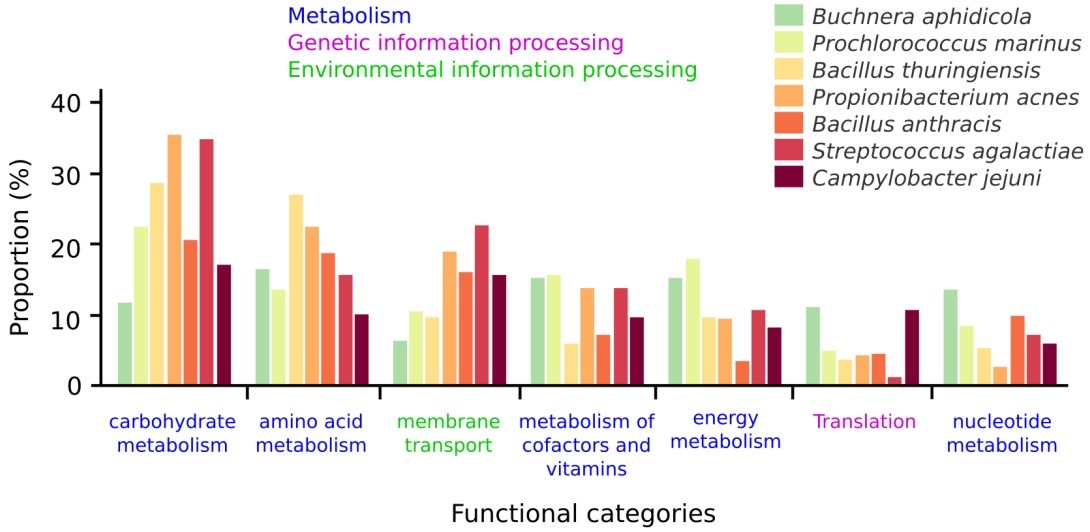


Figure 2: Functional categories of the core genes for seven bacterial species. The proportion of core genes involved in the different functional categories are displayed for each of the seven bacterial species selected. The categories are colored accordingly to the broader functional level to which they belong.

reach a plateau. Metrics also tend to show this trend, as their Heaps value are the only ones greater than one. The estimated pangenome of *B. anthracis* (5,932 genes) is this large because the individual genome of these bacteria are large (4,700 genes per genome), and its genomic fluidity is low (0.06). These two very different species in term of genome size and way of life (intracellular / free living) can be described as having a closed pangenome. Finally, the three last species considered here (*C. jejuni*, *P. acnes* and *S. agalactiae*) have similar trends : their pangenome curve have the same slope, their Heap value are lower than the 1 threshold (between 0.7 and 0.85) and the genomic fluidity is between 0.08 and 0.14. Thus, *C. jejuni*, described to have a closed pangenome [32], seems here quite similar to *S. agalactiae* and *P. acnes* that are described to have open pangenomes [17,29]. Moreover, these last two species do not present a pangenome as open as *B. thuringiensis* and *P. marinus* (Figure 1; Table 1). These results show the importance of using the same annotation and clustering methods to have comparable results, as the granularity of the clustering can have a dramatic impact on the pangenome estimate and metrics. It also highlights that each curve taken individually can lead to different interpretation, and comparing results for several species can provide additional information.

Another possibility offered by PanGeneHome is to compare the functional potential of core genes for the selected taxons, as in [33]. When considering the same 7 species, the core gene functions of *Buchnera aphidicola* are the most different to the ones of other species (Figure 2: average correlation of 0.67 between *B. aphidicola* core functional profiles and other species profiles). Indeed, nearly half of the annotated core genes of *Buchnera aphidicola* are involved in "Genetic information Processing", with 70 of the 157 core genes identified in the 17 genomes of this species being implied in translation. This result is not surprising as *B. aphidicola* is an endosymbiont of aphids that encode less than 600 genes and has lost lots of metabolic potential [34] such as anaerobic respiration, synthesis of phospholipids, complex carbohydrates, etc... The most similar species in terms of functional potential of their core genes are *P. acnes* and *S. agalactiae* (correlation of 0.93 for their functional profiles), two species having comparable pangenome closedness. More surprisingly, *Bacillus thuringiensis* and *Bacillus anthracis* are also similar in terms of functional potential of their core genes (correlation of 0.85) despite having opposite trends in terms of pangenome closedness. These two species belong to the Bacillus cereus group (NCBI taxonomy ID = 86661), and are actually thought to be part of the same species [35,36]. The AAI analysis of this "Bacillus cereus group" show that the sequenced strains of *B. anthracis* are more closely related to each other than the *B. thuringiensis* genomes (Figure 3). This last point might be due to the fact that *B. anthracis* strains are selected for culture and sequencing because of a precise phenotype (high toxicity). Selecting only very closely strains based on this phenoype might narrow the diversity of this group and artificially result in a closed pangenome. The fact that the core genes of these two species have similar functions reinforce the fact that the genomic characteristics of these two species might not be so different and that *B. anthracis* should be considered here as a sub-species.

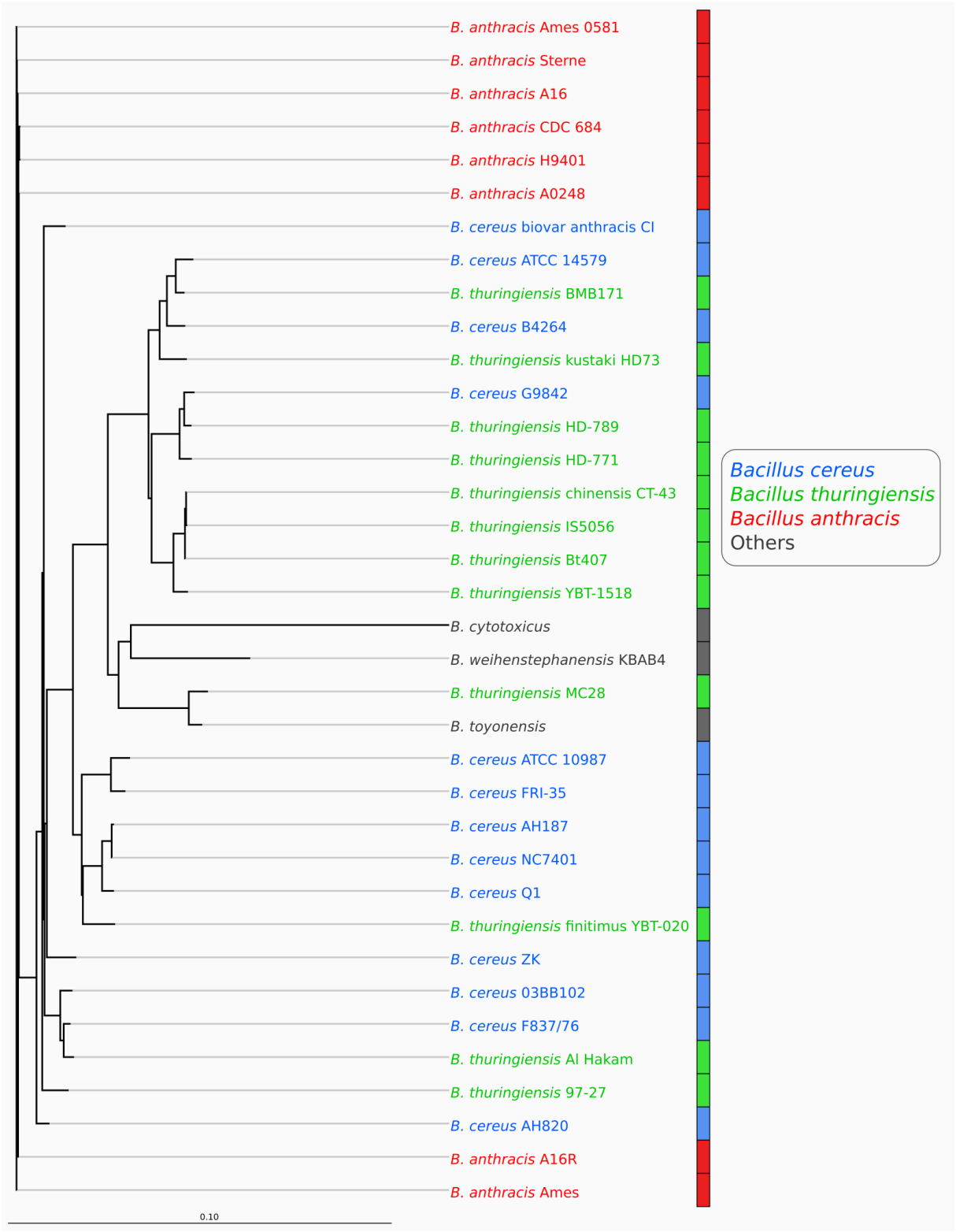


Figure 3: Dendrogram based on the AAI. Dendrogram based on the average amino acid identity of the shared genes (AAI) between all genome pairs of the Bacillus cereus group. The neighbor-joining method (PHYLIP suite [20]) was applied to the AAI matrix to build a dendrogram.

Conclusion

A pangenome describes the full complement of genes in a clade or taxon. Even if pangenomes are typically analyzed at the species level, such analyses can be informative at any taxonomic level. PanGeneHome generates visualizations and metrics for pangenomes for all possible microbial clade, and as many as 615 taxa are available for analysis, including for example 182 different species and 49 different orders.

Pangenome metrics (size, diversity, closedness, etc...) are highly dependent on the genome annotation and protein clustering

methods used. Pangenome studies often focus on one species and the results presented in separate studies are thus hardly comparable. Here, the annotation and clustering methods used were the same for all genomes, and pangenome results can be directly compared. Moreover, as highlighted by the results obtained for two Bacillus species, pangenome results should analyzed in regard of the diversity existing inside each taxon considered. Indeed, considering only evolutionarily close strains for a species will result in a low genomic fluidity and a closed pangenome, and analysis such as AAI should help in deciphering these evolutionary distances. Thus, PanGeneHome provides a comprehensive and uniform framework with a user-friendly interface to explore pangenomes for any microbial taxon, and should help microbiologists to quickly get a glimpse on the genomic plasticity and diversity for a clade of interest.

Considering the fast growing number of microbial genomes, the PanGeneHome tool will need to be updated regularly.

Authors' Contributions

CL, VH, CA, LB, AC, RDO, LFB, HG, DL, MM, SM, KNG, BP, OP, JP and AS developed the Web site and CL and VH finalized its development. AM and JCC took care of the informatic infrastructure. FE conceived the study, coordinated the work and wrote the manuscript. All authors read and approved the final manuscript.

Competing Interests

The authors have declared no competing interests.

Acknowledgements

The authors thank Simon Roux for his careful reading of the manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Udaondo Z, Molina L, Segura A, Duque E, Ramos JL. Analysis of the core genome and pangenome of Pseudomonas putida. Environ Microbiol. 2016;18(10):3268-3283. doi: 10.1111/1462-2920.13015.
2. Bhardwaj T, Somvanshi P. Pan-genome analysis of Clostridium botulinum reveals unique targets for drug development. Gene. 2017;623:48-62. doi: 10.1016/j.gene.2017.04.019.
3. Wee WY, Dutta A, Choo SW. Comparative genome analyses of mycobacteria give better insights into their evolution. PLoS One 2017;12. doi.org/10.1371/journal.pone.0172831.
4. Uchiyama I, Albritton J, Fukuyo M, Kojima KK, Yahara K, Kobayashi I. A Novel Approach to Helicobacter pylori Pan-Genome Analysis for Identification of Genomic Islands. PLoS One. 2016;11. doi.org/10.1371/journal.pone.0159419.
5. Rouli L, Merhej V, Fournier PE, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect 2015;7:72-85. doi: 10.1016/j.nmni.2015.06.005.
6. Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. PGAP: pan-genomes analysis pipeline. Bioinformatics. 2012;28(3):416-8. doi: 10.1093/bioinformatics/btr655.
7. Santos AR, Barbosa E, Fiaux K, Zurita-Turk M, Chaitankar V, Kamapantula B, et al. PANNOTATOR: an automated tool for annotation of pan-genomes. Genet Mol Res. 2013;12(3):2982-9. doi: 10.4238/2013.August.16.2.
8. Zhao Y, Jia X, Yang J, Ling Y, Zhang Z, Yu J, et al. PanGP: a tool for quickly analyzing bacterial pan-genome profile. Bioinformatics. 2014;30(9):1297-9. doi: 10.1093/bioinformatics/btu017.
9. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3. doi: 10.1093/bioinformatics/btv421.
10. Chaudhari NM, Gupta VK, Dutta C. BPGA- an ultra-fast pan-genome analysis pipeline. Sci Rep. 2016;6. doi:10.1038/srep24373.
11. Laing C, Buchanan C, Taboada EN, Zhang YX, Kropinski A, Villegas A, et al. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics. 2010;11:461. doi: 10.1186/1471-2105-11-461.
12. Brittnacher MJ, Fong C, Hayden HS, Jacobs MA, Radey M, Rohmer L. PGAT: a multistrain analysis resource for microbial genomes. Bioinformatics. 2011;27(17):2429-30. doi: 10.1093/bioinformatics/btr418.
13. Pantoja Y, Pinheiro K, Veras A, Araújo F, Lopes de Sousa A, Guimarães LC, et al. PanWeb: A web interface for pan-genomic analysis. PLoS One. 2017;12(5):e0178154. doi: 10.1371/journal.pone.0178154.
14. Xiao J, Zhang Z, Wu J, Yu J. A brief review of software tools for pangenomics. Genomics Proteomics Bioinformatics. 2015;13(1):73-6. doi: 10.1016/j.gpb.2015.01.007.
15. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38(Database issue):D355-60. doi: 10.1093/nar/gkp896.
16. Snipen L, Liland KH. micropan: an R-package for microbial pan-genomics. BMC Bioinformatics. 2015;16:79. doi: 10.1186/s12859-015-0517-0.
17. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opinions Microbiol 2008; 12:472–77. doi. org/10.1016/j.mib.2008.09.006.
18. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics. 2011;12:32. doi: 10.1186/1471-2164-12-32.
19. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. Nat Genet 1999;21:108-10.
20. Felsenstein J. PHYLIP - phylogeny inference package (Version 3.2). Cladistics. 1989;5:164-166.
21. Smits SA, Ouverney CC. jsPhyloSVG: A Javascript Library for Visualizing Interactive and Vector-Based Phylogenetic Trees on the Web. PLoS One 2010;5(8). doi: 10.1371/journal.pone.0012267.
22. Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. J Bacteriol 2005;187(18):6258-64. doi: 10.1128/JB.187.18.6258-6264.2005.
23. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389-3402.
24. Nakaya A, Katayama T, Itoh M, Hiranuka K, Kawashima S, Moriya Y, et al. KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. Nucleic Acids Res. 2013;41(Database issue):D353-7. doi: 10.1093/nar/gks1239.
25. Mbengue M, Lo FT, Diallo AA, Ndiaye YS, Diouf M and Ndiaye M. Pan-genome analysis of Senegalese and Gambian strains of Bacillus anthracis. African Journal of Biotech. 2016;15(45):2538-2546.
26. Snipen L, Almøy T, Ussery DW. Microbial comparative pan-genomics using binomial mixture models. BMC Genomics. 2009;10:385.
27. Lefébure T, Bitar PD, Suzuki H, Stanhope MJ. Evolutionary dynamics of complete Campylobacter pan-genomes and the bacterial species concept. Genome Biol Evol. 2010;2:646-55. doi: 10.1093/gbe/evq048.
28. Fang Y, Li Z, Liu J, Shu C, Wang X, Zhang X, et al. A pangenomic study of Bacillus thuringiensis. J Genet Genomics. 2011;38(12):567-76. doi: 10.1016/j.jgg.2011.11.001.
29. Tomida S, Nguyen L, Chiu BH, Liu J, Sodergren E, Weinstock GM, et al. Pan-genome and comparative genome analyses of propionibacterium acnes reveal its genomic diversity in the healthy and diseased human skin microbiome. MBio. 2013;4(3):e00003-13. doi: 10.1128/mBio.00003-13.
30. Biller SJ, Berube PM, Lindell D, Chisholm SW. Prochlorococcus: the structure and function of collective diversity. Nat Rev Microbiol. 2015;13(1):13-27. doi: 10.1038/nrmicro3378.
31. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et

al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 2005;102(39):13950-5.

32. Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F. The bacterial pan-genome:a new paradigm in microbiology. Int Microbiol. 2010;13(2):45-57.

33. Yang X, Li Y, Zang J, Li Y, Bie P, Lu Y, Wu Q. Analysis of pan-genome to identify the core genes and essential genes of Brucella spp. Mol Genet Genomics. 2016;291(2):905-12. doi: 10.1007/s00438-015-1154-z.

34. Gomez-Valero L, Silva FJ, Christophe Simon J, Latorre A. Genome reduction of the aphid endosymbiont Buchnera aphidicola in a recent evolutionary time scale. Gene. 2007;389(1):87-95.

35. Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, Mock M, et al. Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis--one species on the basis of genetic evidence. Appl Environ Microbiol. 2000;66:2627-30.

36. Okinaka RT, Keim P. The Phylogeny of Bacillus cereus sensu lato. Microbiol Spectr. 2016;4(1). doi: 10.1128/microbiolspec.

Corresponding author: *François Enault, UMR CNRS 6023 Microorganisms: Genome and Environment, Build. A, 24 avenue des Landais, 63177 Aubière Cedex. La France, Tel: 33-(0)473- 407-471; Fax: 33-(0)473-407-670: E-mail: francois.enault@uca.fr.*

Received Date: October 24, 2017, **Accepted Date:** November 23, 2017, **Published Date:** November 30, 2017.

Citation: Loiseau C, Hatte V, Andrieu C, Barlet L, Cologne A, et al. (2017) PanGeneHome : A Web Interface to Analyze Microbial Pangenomes. J Bioinf Com Sys Bio 1(2): 108.

Article 3 / Co-autrice : Biderre-Petit et *et al.*, 2019

Biderre-Petit, C., **Taib, N.**, **Gardon, H.**, **Hochart, C.** and **Debroas, D.** (2019) 'New insights into the pelagic microorganisms involved in the methane cycle in the meromictic Lake Pavin through metagenomics', *FEMS Microbiology Ecology*, 95(3). doi: 10.1093/femsec/fiy183.

FEMS Microbiology Ecology, 95, 2018, fiy183

doi: 10.1093/femsec/fiy183
Advance Access Publication Date: 7 September 2018
Research Article

RESEARCH ARTICLE

New insights into the pelagic microorganisms involved in the methane cycle in the meromictic Lake Pavin through metagenomics

Corinne Biderre-Petit*, Najwa Taib†, Hélène Gardon, Corentin Hochart and Didier Debroas

Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Génome et Environnement, F-63000 Clermont-Ferrand, France

*Corresponding author: Corinne Biderre-Petit. 1, Impasse Amélie Murat -Bat BioA 63178 Aubière (France). Tel: (+33) 473405139; E-mail: corinne.petit@uca.fr

†Present address: Hub Bioinformatique et Biostatistique, Unité de Biologie évolutive de la cellule microbienne, Département de Microbiologie, Institut Pasteur, Paris, France.

One sentence summary: Metagenomic analyses revealed prominent roles for *Methanoregula* and *Methylobacter* in Lake Pavin methane-cycling.

Editor: Gary King

ABSTRACT

Advances in metagenomics have given rise to the possibility of obtaining genome sequences from uncultured microorganisms, even for those poorly represented in the microbial community, thereby providing an important means to study their ecology and evolution. In this study, metagenomic sequencing was carried out at four sampling depths having different oxygen concentrations or environmental conditions in the water column of Lake Pavin. By analyzing the sequenced reads and matching the contigs to the proxy genomes of the closest cultivated relatives, we evaluated the metabolic potential of the dominant planktonic species involved in the methane cycle. We demonstrated that methane-producing communities were dominated by the genus *Methanoregula* while methane-consuming communities were dominated by the genus *Methylobacter*, thus confirming prior observations. Our work allowed the reconstruction of a draft of their core metabolic pathways. Hydrogenotrophs, the genes required for acetate activation in the methanogen genome, were also detected. Regarding methanotrophy, *Methylobacter* was present in the same areas as the non-methanotrophic, methylotrophic *Methylotenera,* which could suggest a relationship between these two groups. Furthermore, the presence of a large gene inventory for nitrogen metabolism (nitrate transport, denitrification, nitrite assimilation and nitrogen fixation, for instance) was detected in the *Methylobacter* genome.

Keywords: meromictic lake; metagenomics; methane cycle; methanogens; methanotrophs; metabolic reconstruction

INTRODUCTION

Methane (CH_4) is the second most important greenhouse gas after carbon dioxide (CO_2) but with a global warming potential that is 21-fold higher. About one third of the total CH_4 emissions in the atmosphere come from natural sources, dominated by wetland emissions (watercourses, swamps, ponds, oceans and lakes). Lakes are estimated to contribute 6–16% of total natural CH_4 emissions (Bastviken *et al.* 2004; Kirschke *et al.* 2013). The release from lakes is therefore considerable (Bastviken *et al.* 2011), exceeding that from oceans despite the comparatively small portion of inland waters on the Earth's surface (<1%)

Received: 13 March 2018; **Accepted:** 6 September 2018

(Downing *et al.* 2006). In lacustrine systems, most of the CH_4 originates from the activity of methanogenic archaea through decomposition of organic matter under anoxic conditions (Conrad, Klose and Noll 2009). After production, as much as 80–90% of the CH_4 is captured by aerobic and anaerobic methanotrophs that use this gas as a sole carbon (C) and energy source for growth before it escapes to the atmosphere (Knittel and Boetius 2009; Ettwig *et al.* 2010; Chistoserdova 2015). Most of the CH_4 oxidation takes place in the top layer of the sediment, in which CH_4 and dioxygen (O_2) form steep counter gradients (Auman *et al.* 2000). Hence, by mitigating CH_4 emissions, microorganisms act as an efficient biological CH_4 filter (Cole *et al.* 1994; Bastviken *et al.* 2004).

Methane has been the focus of intensive research over the last few decades because of its ecological and applied interests. This work has resulted in a detailed understanding of the principal processes linked to its global cycle and the apparition of hundreds of genome sequences for both methanotrophic and methanogenic microorganisms in public databases. Nevertheless, genomic data on methanotrophs are far less numerous in comparison to that of methanogens, mainly due to the very recent discovery of novel biogeochemical processes and microorganisms still having poorly characterized metabolisms (Wu *et al.* 2011; Hernandez *et al.* 2015). For instance, it has recently been demonstrated that aerobic methanotrophs belonging to *Proteobacteria* were capable of coupling CH_4 oxidation to nitrate (NO_3^-) reduction under O_2 limitation (Kits, Klotz and Stein 2015), thereby highlighting their unexpected metabolic flexibility and assigning these bacteria a new role at the metabolic intersection of C and nitrogen (N) cycles. Furthermore, data available for lakes are still scarce compared with other systems, even though these environments are characterized by dynamic cycling of CH_4, serving both as its major sources and major sinks. The principal studies have focused on a few models and their sedimentary compartments (Rahalkar *et al.* 2009; Chistoserdova and Lidstrom 2013), but few studies have addressed pelagic freshwater environments (Pernthaler 2017). However, the importance of this latter compartment is certainly largely underestimated for the overall understanding of CH_4 cycling as evidenced by the active CH_4 production in the oxygenated waters of many lakes (Schulz *et al.* 2001; Grossart *et al.* 2011; Tang *et al.* 2016) as well as the intensive CH_4 oxidation in anoxic waters in stratified lakes (Karr *et al.* 2006). Moreover, permanently stratified lakes (e.g. meromictic) present another interesting characteristic, since only a negligible amount of CH_4 is released into the atmosphere compared with the more classical shallow lakes where most of the CH_4 is consumed at the oxic-anoxic boundary in the water column (Borges *et al.* 2011).

On account of their particular properties, meromictic lakes represent excellent field laboratories to follow biologically mediated redox processes, especially those linked to CH_4 cycling. In Lake Pavin, metabarcoding surveys were conducted in order to highlight which microorganisms play a critical role in the most significant CH_4 pathways. These studies revealed that most of the methanogens were closely related to the genus *Methanoregula* in the *Methanomicrobiales* class and the methanotrophs to the genus *Methylobacter* (Biderre-Petit *et al.* 2011b; Borrel *et al.* 2011). However, the methods used, based on single marker amplification, only gave insight into the specific richness within these relevant uncultivated species, but none into their genomic structure. To overcome these limitations, metagenomics has recently become a powerful tool for collecting information on microbial communities. It provides an all-inclusive picture of their specific functional pathways and allows us to investigate their distribution in response to different environmental conditions in their natural habitats (Kalyuzhnaya *et al.* 2008; Quince *et al.* 2017). In our study, we used a strategy which enabled us to decipher the metabolic potential of the microorganisms involved in methanotrophy and methanogenesis along the Lake Pavin water column through a two-step based method of metagenomic sequencing analysis. The first step consisted of a read-centric approach in order to identify the entire genetic repertoire of planktonic methanotrophs and methanogens, and thus perform a large-scale, strain-level analysis. The second step was an assembly-based approach to resolve the genomic organization for the dominant related populations thriving in this ecosystem.

MATERIALS AND METHODS

Lake Pavin water samples

Lake Pavin is a small (~0.44 km^2), deep (92 m), almost circular (~750 m diameter) lake of volcanic origin formed 6900 years ago and located in the Massif Central (France) (45°29.740 N, 2°53.280 E), at an altitude of 1197 m above sea level. Four sampling depths with contrasting O_2 tensions were selected: the surface (4 m) with high O_2 tension; the oxycline with low O_2 tension (53 m); and two depths in the anoxic zone, one close to the oxic-anoxic interface (65 m; chemocline) and one deeper (80 m), previously shown to contain a high CH_4 concentration (Lehours *et al.* 2005; Biderre-Petit *et al.* 2011b; Lopes *et al.* 2011). Water samples were collected in July 2013 from a platform positioned above the deepest point of the lake (92 m deep). A volume of 20 L was collected using a Van Dorn bottle as described previously (Biderre-Petit *et al.* 2011a). Immediately on collection, samples were transferred into sterile bottles and transported to the laboratory on ice within 0.5 to 3 h. Serial filtrations (25-, 15- and 1.2-μm pore size membrane filters) were performed on each sample to remove large particles and eukaryotic cells. After filtration, tangential ultrafiltration (Amicon pump) was used to concentrate the filtrate containing microorganisms to a final volume of 1 L. As a last step, the microbial biomass was collected on 0.2-μm pore size (pressure <100 mbar) polycarbonate filters (47 mm TSPT Millipore filters, Billerica, MA) and stored at -80°C to await nucleic acid extraction.

Physico-chemical parameters

Temperature and dissolved O_2 profiles were obtained using a multiparameter probe ProOdO™ (Ysi, Germany). For determination of sulfate (SO_4^{2-}), NO_3^- and ammonium (NH_4^+) concentration, three replicates × 50 ml of sampled waters were filtered through a 0.2-μm syringe filter and stored frozen at -20°C to await analysis using Spectroquant reagent standard kits (Merck, Germany) (Fig. S1, supplementary material).

DNA extraction and sequencing

Genomic DNA (gDNA) was extracted using the standard phenol-chloroform method as previously described (Biderre-Petit *et al.* 2011b). DNA concentrations were estimated using a spectrophotometer (NanoDrop ND-1000, Nanodrop Products, Wilmington, DE). The metagenomic library preparation and sequencing were performed by GATC Biotech (Konstanz, Germany) from ~6 μg of gDNA for each sample. Sequencing was performed using the Illumina HiSeq-2000 paired-end technology (2 × 100 bases)

which generated ~95 to 124 × 10^6 reads depending on the sample, comprising a total of ~41 gigabases (Gb) of sequence data (Table 1).

Read-centric analysis

To facilitate analysis, especially in terms of computing time, a sub-sample of each raw dataset containing 2 × 10^6 randomly picked Illumina reads was generated and then processed through the MG-RAST server pipeline (V3). More than 95% of the reads were retained after quality-filtering based on length, ambiguous bases and quality scores. They were then annotated using MG-RAST default parameters (E-value cutoff 10^{-5}, minimum identity cutoff 60% and minimum alignment length cutoff 15 amino acids for proteic features) (Meyer *et al.* 2008). Sequencing statistics are shown in Table 1. The sequences of genes encoding key enzymes of the metabolisms of interest were extracted from sub-sampled metagenomic datasets through keyword research. For this, specific searches for a selection of 268 enzymes (206 covering central C and N metabolisms for methanotrophs (Table S1, Figs S2 and S3, supplementary material) and 62 covering the main methanogenic steps (Table S2, Fig. S4, supplementary material) were performed inside the MG-RAST web browser through keyword filter use. The workbench was used to download the FASTA files with reads corresponding to all selected proteins. Given the non-specificity of most of the selected enzymes for a unique metabolic group, the affiliation of all reads downloaded from MG-RAST was manually curated by BLASTX against the NR protein database via the web interface and executed with default parameters (https://blast.ncbi.nlm.nih.gov/blast, Johnson *et al.* 2008). Only reads with the best hits for methanogens and methanotrophs as well as a drastic decrease of similarity with other taxa were retained for further analysis.

Metagenome assembly

For each full read dataset, raw reads were trimmed using FASTX-Toolkit (v 0.0.13.1; Pearson *et al.* 1997) with a quality cutoff of 30 and 50 bp as a minimal sequence length. Trimmed reads were then assembled using IDBA-UD version 1.0.9, which was run in its default iterative mode (k-min = 20, k-max = 100, k-step = 20) (Peng *et al.* 2012). From 287 042 to 512 204 contigs were produced depending on the sample (Table 1, Fig. S5, supplementary material). In the following step, the reads identified by read-centric analysis were blasted against contigs using BLASTN (version 2.2.26) executed with default parameters. Only the contigs with a read alignment of E-value $<10^{-4}$ were selected. However, given the possibility of cross-reactions and chimeric contigs, the outputs from these automated analyses were manually curated by BLASTN and BLASTX against the NR databases via the NCBI site (https://blast.ncbi.nlm.nih.gov/blast) and executed with default parameters. Of the contigs initially identified by BLASTN, up to 19% of those selected for the methanotrophs (312 contigs in total) and 81% for the methanogens (386 contigs) were finally retained. Gene prediction for the remaining contigs was performed using a MetaGeneAnnotator (Noguchi, Taniguchi and Itoh 2008). All predicted translated genes were compared to Refseq non-redundant (prokaryotes + viruses (O'Leary *et al.* 2016)) and KEGG (Kanehisa *et al.* 2016) databases using BLASTP (Altschul *et al.* 1990) with an E-value threshold of 10^{-5} and a percentage identity of 60%. The rRNA genes in contigs were identified with BLASTN (E-value $<10^{-5}$, identity ≥97%) against the SILVA rRNA database v. 123 (Quast *et al.* 2013). The location of the contigs mapped onto the reference genomes was then visualized using the CGView Server (http://stothard.afns.ualberta.ca/cgview_server/; Grant and Stothard 2008). The BLASTN parameters, used in the CGView Server to compare the reference genome sequences with the contig sets, had an E-value cutoff of 10^{-3}, an alignment length cutoff of 100 bp and a percentage identity cutoff of 60%.

Nucleotide accession number

Sequence data are available in GenBank under the accession numbers KY06104-KY06107 and MF076238-MF076545 for the methanotrophic contigs and KY994152-KY994537 for the methanogenic contigs. The raw sequences were archived in MG-RAST under the accession numbers mgm4580845.3 to mgm4580848.3 for the metagenomic reads and the accession numbers mgm4557752.3.100, mgm4557753.3.100, mgm4581104.3.100 and mgm4581106.3.100 for the assembled contigs.

RESULTS

An overview of the microbial community composition in the Lake Pavin water column

According to the read-centric analysis, the domain Bacteria numerically dominated the composition of the microbial communities at all four of the depths sampled. More than 93% of the taxonomically assigned reads matched with bacteria, with a few representatives of archaea (range 0.1–3.7%), eukaryotes (0.7–3.8%) and viruses (0.6–3.9%) (Table 1). Archaeal reads were more abundant in the anoxic waters (range 3.1–3.7% against 0.1% in oxic waters). The rRNA gene affiliation agreed with the broad taxonomic picture provided by the functional genes with 91–99% of rRNA genes affiliated with Bacteria, whereas Archaea and Eukarya were minor components (up to 6.4% and 4%, respectively) (Table 1). Overall, the most abundantly recovered bacterial rRNA reads were consistent with those usually found in freshwater, with a domination of *Proteobacteria* (59.8%; range 36.6–73.8%), *Actinobacteria* (12.9%; range 3.3–23%), *Bacteroidetes* (8.4%; range 2.6–21.3%), and, to a lesser extent, *Verrucomicrobia* (4.1%; range 2.2–7.6%), *Firmicutes* (1.9%; range 0–4.8%) and *Cyanobacteria* (1.9%; range 0.4–6.4%) (Fig. 1A and B). The bacterial community was also composed of many less abundant phyla (<5% of the community) characterized by a more limited spatial distribution. The largest diversity was observed at the depth 65 m and the lowest at 4 m (Fig. 1B). The former depth is part of the chemocline, a layer known to offer a variety of ecological niches and thereby generally harboring a highly diverse microbial community. Concerning Archaea, corresponding rRNA reads were only detected in the monimolimnion and were all affiliated to the hydrogenotrophic *Methanomicrobiales* (Fig. 1A), mainly to the genus *Methanoregula*.

As *Proteobacteria* account for most of the rRNA reads, we focused on this phylum for a comparative study of its potential involvement in the N, sulfur (S) and CH_4 geochemistry throughout the water column. We used the relative abundance of the related genes and literature knowledge as proxies of the potential relevance of each population *in situ*. For the S cycle, reads affiliated with known sulfur-oxidizing bacteria (SOB) were detected from 53 m to 80 m, with the genera *Sulfuricella* and *Sulfuritalea* (*β-Proteobacteria*) in the oxycline and the genera *Sulfuricurvum, Sulfurimonas* and *Sulfurovum* (*ε-Protetobacteria*) in the anoxic layers. As expected, sulfate-reducing bacteria (SRB), with

Table 1. Sequencing statistics of free-assembly and assembled metagenomes.

Free-assembly metagenomes		4 m	53 m	65 m	80 m
Raw datasets	Number of reads	108 099 890	120 529 997	123 680 966	95 654 671
High-quality reads	Number of reads after QC control	102 305 570	114 673 466	119 466 504	84 488 684
	Total Gbp	10.1	11.3	11.7	8.1
Sub-sampled datasets treated with MG-RAST (V3)	MG-RAST ID	mgm4580845.3	mgm4580848.3	mgm4580846.3	mgm4580847.3
	Number of randomly selected reads	2000 000	2000 000	2000 000	2000 000
	Number of reads after QC control	1967 724	1986 645	1990 038	1915 809
	Total Mbp	194.9	196.7	194.9	187.7
	Mean sequence length (bp)	99 ± 6	99 ± 6	98 ± 8	98 ± 8
	Mean GC ratio (%)	51 ± 12	51 ± 11	57 ± 12	57 ± 11
	Reads with predicted feature (%)	93.3	92.8	92.3	92.3
	Number of predicted proteins	1513 681	1750 233	1767 635	1686 082
	Number of identified proteins	333	407	436	492
	Functionally assigned proteins (%)	79.9	78.8	77.8	79.1
	Taxonomy-classified reads (%)				
	Bacteria	93.8	98.0	94.2	95.4
	Archaea	0.1	0.1	3.7	3.1
	Eukarya	3.8	0.7	1.3	0.7
	Viruses	4.9	1.0	0.6	0.6
	Unclassified	0.1	0.1	0.2	0.1
	Number of rRNA genes (16S/18S and 23S/28S)	249	419	381	298
	Bacteria (%)	94.4	99.1	93.2	91.3
	Archaea (%)	0	0	3.7	6.4
	Eukarya (%)	4	0.9	3.1	1.6
	Not affiliated	1.6	0	0	0.7
Assembled metagenomes		4 m	53 m	65 m	80 m
Contiguous sequences	MG-RAST ID (preprocess passed)	mgm4557753.3.100	mgm4581106.3.100	mgm4581104.3.100	mgm4557752.3.100
	Number of contigs	287 042	474 901	512 204	413 815
	Total Gbp	0.16	0.47	0.54	0.17
	Largest contig	5 055	299 470	290 012	4 146
	N50	1 336	1 694	1 970	1 025

Desulfobacca (*δ-Proteobacteria*) as the main genus, were exclusively detected in the anoxic waters (Fig. 2A). For the N cycle, the potential for nitrification was mostly observed at the oxycline with the *Nitrosomonadales* (*β-Proteobacteria*) as the prevailing ammonium-oxidizing bacteria (AOB) (Fig. 2B) and the phylum *Nitrospirae* as the dominant nitrite-oxidizing bacteria (NOB) (Fig. 1B). As denitrifying bacteria (NRB) are phylogenetically diverse, and distributed over at least 60 genera, they were therefore likely to occur in all samples, with the *β-Proteobacteria* as the primary candidates; principally the genera *Burkholderia*, *Cupriavidus* and *Ralstonia* (Fig. 2B). As for the CH_4 cycle, two proteobacterial types, both recognized as methane-oxidizing bacteria (MOB) (type I MOBs belonging to *γ-Proteobacteria* (*Methylococcaceae*) and type II MOBs belonging to *α-Proteobacteria* (*Methylocystaceae* and *Beijerinckiaceae*)), were found in the rRNA pool. The former was the most abundant with *Methylococcaceae* reads detected both at 53 m and 65 m, whereas only a few sequences were affiliated with the latter at 4 m. These belonged to the genus *Methylocystis* within the family *Methylocystaceae* (Fig. 2C). The co-occurrence of the non-methanotrophic methylotrophic *Methylotenera* (*β-Proteobacteria*) in the same areas as *Methylococcaceae* could suggest a potential relationship between these two groups (Fig. 2C).

Methanogenesis pathway reconstruction

From the metagenomic data gathered in Lake Pavin, we reconstructed a draft of the core metabolic pathways used by planktonic methanogens for CH_4 production. To do so, we first established a list of 62 genes coding for proteins and involved in

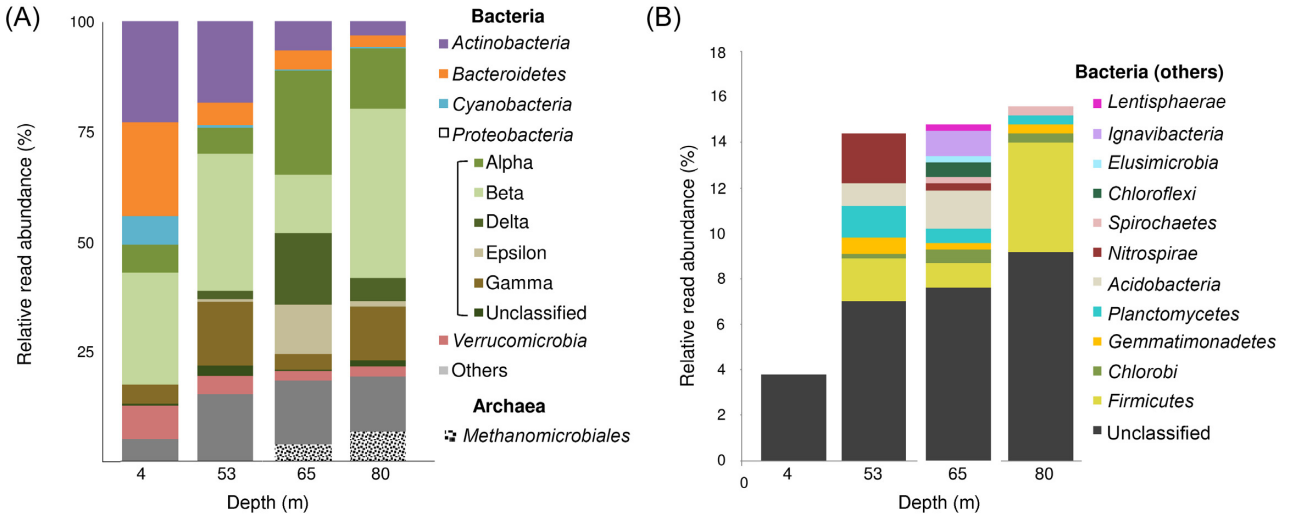


Figure 1. Taxonomic affiliation of rRNA sequences across depths. The sequences were assigned at the phylum-level, except the *Proteobacteria*, which were represented by class. **(A)** All bacterial and archaeal phyla represented by >5% of the total sequences in any metagenome. 'Others' corresponds to the unclassified bacterial sequences and phyla represented by <5% of the total sequences. **(B)** Bacterial phyla represented by <5% of the total sequences in any metagenome with variable abundance as a function of depth.

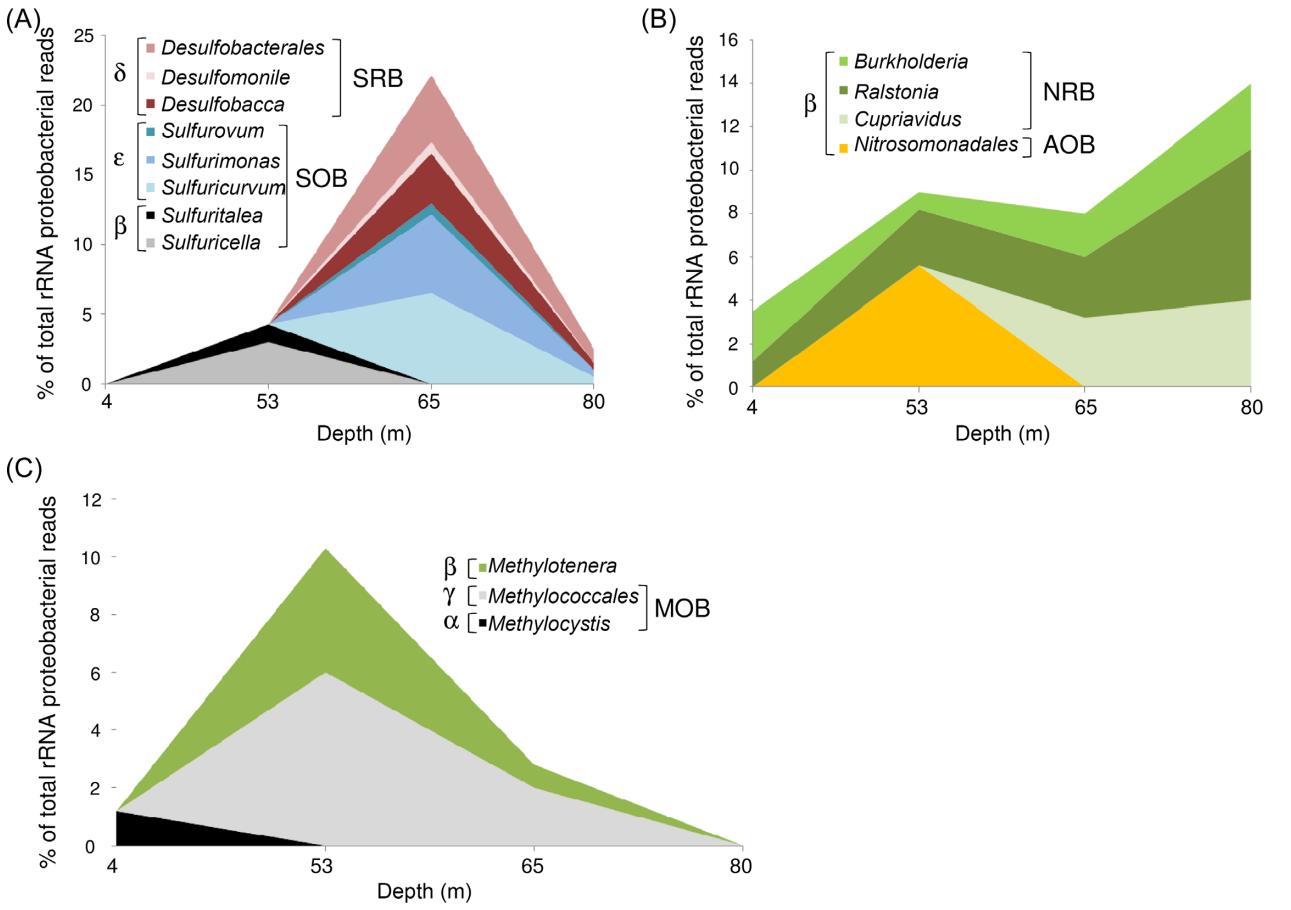


Figure 2. Functional activities of proteobacterial classes in the water column of Lake Pavin linked with the principal biogeochemical cycles based on rRNA sequence (16S and 23S) affiliation. **(A)** Taxonomic groups potentially involved in sulfur metabolisms, **(B)** taxonomic groups potentially involved in nitrogen metabolisms, and **(C)** taxonomic groups potentially involved in methane consumption (*Methylocystis* and *Methylococcaceae*) as well as the abundance of the non-methanotrophic, methylotrophic *Methylotenera* genus. AOB, ammonium-oxidizing bacteria; MOB, methane-oxidizing bacteria; NRB, nitrate-reducing bacteria; SOB, sulfur-oxidizing bacteria; SRB, sulfate-reducing bacteria.

the pathways of methanogenesis (Fig. S4, Table S2, supplementary material) and then used all possible appellations known in scientific literature for each protein as queries against MG-RAST read annotations. The current study showed a strong correlation between our functional profile, methanogenic genes being found only in anoxic metagenomes (at depths of 65 and 80 m), and the CH_4 concentrations measured along the water column of Lake Pavin, as seen in previous studies (Lehours *et al.* 2005; Biderre-Petit *et al.* 2011b; Lopes *et al.* 2011). Up to 0.03% of the reads (1183) of these two metagenomes matched with 52 out of the 62 genes present in the established list. Of note, 55.1% of these were retrieved at a depth of 65 m and the other 44.9% at a depth of 80 m (Fig. 3A). Most of the reads (99.8%) were assigned to methanogenic archaea, mainly the order *Methanomicrobiales* (96.3%), followed by *Methanosarcinales* (1.1%) (Fig. 3B, Fig. S6, supplementary material), which suggests a dominance of hydrogenotrophic methanogenesis over acetoclastic methanogenesis. The remaining 0.2% were affiliated with non-methanogenic archaea belonging to the order *Thermoplasmatales*. At the genus level, most of the methanogenic reads were identified as the hydrogenotrophic *Methanoregula* (66.3%) (Fig. S6, supplementary material).

Using the methanogenic reads as queries, a total of 386 contigs with a size ranging from 229 to 6143 bp were isolated from the assembled metagenomes (Table 2). Of these, 49.7% were retrieved at 65 m and 50.3% at 80 m (Fig. 4A). They also had similar size distribution whatever the depth (Fig. 4B). A total of 526 coding DNA sequences (CDSs) were predicted from the contigs, with an average length of 491 bp (1851 bp for the largest), corresponding to a protein-coding content of 97% (Table S3, supplementary material), as well as six rRNA genes and three tRNA genes (Table S4, supplementary material). Compared with the overall G+C content (54.1%), the intergenic regions had a lower G+C percentage (44%) and the gene sequences had a higher G+C percentage (54.7%) (Table 2). The rRNA operon was organized in a bacteria-like 5'-16S-23S-5S-3' transcriptional unit and displayed the highest identity with *Methanoregula* strains (Table S4, supplementary material). Like the reads, the CDSs had the highest frequency of high-scoring BLAST hits with the genus *Methanoregula* (88.4%), with an average amino acid identity of 82.6% (range 54.7–100%). The frequencies of top BLASTP hits with other groups were, most importantly, *Methanolinea* (4.2%), *Methanosphaerula* (3.4%) and *Methanoculleus* (1.5%), with the former two also being affiliated with the family *Methanoregulaceae* (Table S3, supplementary material). Most of the CDSs (90.5%) were associated to functions related to the CH_4 metabolism initially targeted; only a low proportion (<5%) of the identified KEGG Orthologous (KO) genes was uncategorized (Fig. 5A). All of the CDSs encoded 77 distinct enzymes. The mapping of all contigs onto the *Methanoregula formicicum* SMSPT (CP003167) reference genome using CGView comparison tool software (Grant and Stothard 2008) showed that they covered up to 85 kbp of this genome, scattered over 24 genomic regions on the chromosome (Fig. 6A). Only a few rearrangements were observed in comparison with the reference genome (Fig. 6B).

From the read and contig analysis, all of the genes required for growth on H_2 and CO_2 were uncovered, whether it is those encoding enzymes found in the cytoplasmic cell fraction (steps 1–5 and 7; Fig. S4, supplementary material) or those encoding the membrane-bound methyltransferase (MtrA-H) (step 6; Fig. S4, supplementary material). Although the Mtr complex is used both by the hydrogenotrophic and the acetoclastic methanogens, corresponding sequences invariably showed the best matches with the former. From contig overlaps, all operons encoding the key enzymatic complexes involved in the different steps of the hydrogenotrophic pathway were reconstructed, i.e. that of the coenzyme M reductase (*mcrBDCGA*), the methyltransferase (*mtrEDCBAFGH*) and the formylmethanofuran dehydrogenase (*fmd/fwdFDBAC*), the latter being present in three scattered copies (Fig. 6B). Two separate copies of the component A2 (*atw2*), predicted to encode an ATP-binding protein required in the activation of the Mcr complex, were also detected. The electron-bifurcating complex (a [Ni-Fe] hydrogenase specific to hydrogenotrophic methanogens), which is composed by the cytoplasmic F420 non-reducing hydrogenase (MvhADG) and the heterodisulfide reductase (HdrABC), was also present, constituted by the *mvhD* subunit located directly downstream of the Hdr operon (Fig. 6B). In contrast, no sequence (read or contig) for the second gene class of Hdr complex (*hdrED*), found in acetoclastic methanogens, was detected. The metagenomes also contained genes encoding three other [Ni-Fe] hydrogenases which are required for the reduction of ferredoxin, whether it is the Eha and Ehb hydrogenases found in hydrogenotrophic methanogens (Gao and Gupta 2007) or the homologous Ech enzyme common to all methanogens. Only the read datasets revealed the sequences for Eha and Ehb subunits while the contigs allowed the reconstruction of a partial Ech complex, organized in three adjacent genes (subunits B, C and D). Finally, a complete gene set for a cytoplasmic F420-dependent hydrogenase (FrhABDG), a key enzyme catalyzing the reversible reduction of coenzyme F420 with H_2, and a second separate copy for the *frhB* gene, were reconstructed (Fig. 6B).

Regarding C metabolism, in autotrophic H_2-utilizing methanogens, carbon monoxide (CO) is reported to work as a C source within the anabolic reductive acetyl-CoA (AcCoA) pathway, also known as the Wood-Ljungdahl (WL) pathway. In this study, two copies of the operon encoding the carbon monoxide complex (Cdh) responsible for CO_2 fixation and AcCoA formation (one complete (*cdhABCDE*) and one partial (*cdhCDE*)) were detected in the metanogenomic datasets (Fig. 6B). The enzymes responsible for AcCoA synthesis from acetate, i.e. the acetyl-CoA synthetase (Acs) and phosphotransacetylase (Pta), an alternative to the autotrophic process, were also present in the metagenomes. However, the acetate kinase (Ack) was not detected in any of the read or contig datasets (Fig. S4, supplementary material). We also found the pyruvate ferredoxin oxidoreductase (Por) enzyme (*porGDAB*) responsible for the conversion of AcCoA to pyruvate (Fig. 6B).

Methanotrophy pathway reconstruction

Applying the same method as used for methanogenesis, we reconstructed a draft of the core metabolic pathways of CH_4 oxidation and C assimilation for the planktonic MOBs present in Lake Pavin. We also included, in our survey, the inventory of the genes involved in the N cycle on account of the current growing interest in the impact of this cycle on MOB growth and their CH_4 oxidation capabilities. To do so, we established a list of 206 proteins (107 for C metabolism and 99 for N cycle) (Table S1, Figs S2 and S3, supplementary material) and then processed as previously described. Up to 0.009% metagenomic reads (735 reads), at the four sampling depths, matched with 76 out of the 107 genes involved in central C metabolism. Most of them were recovered from the 53 m metagenome (81.5%), which agreed with the sharp decline in the CH_4 concentration observed close to this depth (Lopes *et al.* 2011), followed by 65 m (13.5%), 80 m (4.1%) and 4 m (0.9%) (Fig. 3A). Concerning the N cycle, methanotrophic genes were detected in 53 m and 65 m datasets (Table

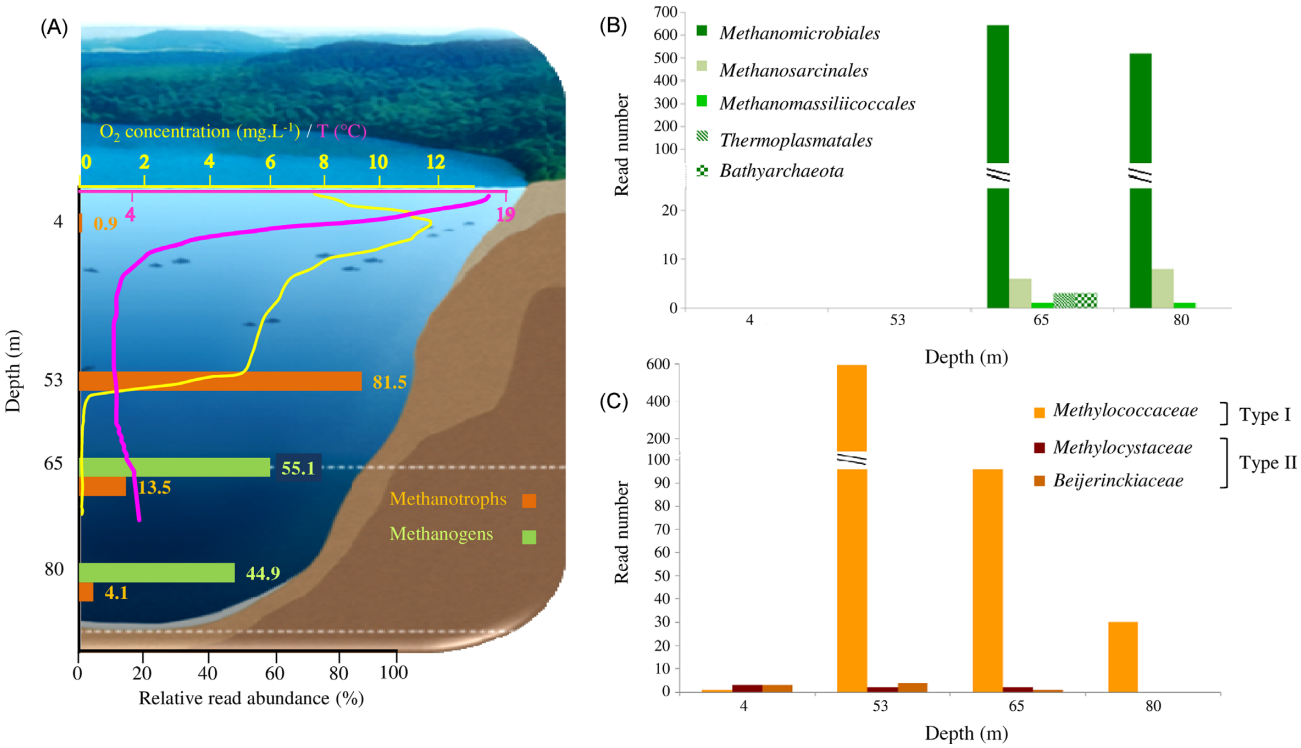


Figure 3. **(A)** Vertical cross-section of the water column of Lake Pavin showing temperature and dissolved oxygen content as well as read distribution associated to methanogens and methanotrophs. **(B)** Read abundance of archaeal orders depending on the depth. **(C)** Read abundance of type I and II MOBs depending on the depth.

Table 2. Contig statistics for methanogens and methanotrophs.

	Methanogens	Methanotrophs
Number of selected contigs	386	312
Assembly size (bp)	271 065	593 276
Mean G+C ratio	54.1%	41.3%
Minimum contig length (bp)	229	236
Maximum contig length (bp)	6143	12 298
Mean contig length (bp)	718	1901
N50 of contigs	777	2730
N90 of contigs	384	907
Number of rRNA genes	6	7
Number of tRNA genes	3	1
Number of CDS	526	622
Mean G+C ratio of CDS	54.7%	41.9%
Mean CDS length (bp)	491	875
Number of unique proteins	77	321
Mean G+C ratio of intergenic regions	44%	33.9%

bp: base pair; CDS: coding DNA sequence.

S5, supplementary material) with up to 0.003% of the reads (152 reads) matching with 14 out of 99 targeted genes; 96.7% of them were from 53 m and 3.3% from 65 m. The taxonomic affiliation showed that type I MOBs dominated, representing more than 98.3% of the methanotrophic reads (Fig. 3B), with *Methylobacter*-like sequences representing the most significant proportion of this group (39.4% of total type I MOB sequences; Fig. S7A and B, supplementary material).

Using MOB reads as queries, a total of 312 contigs ranging from 236 to 12 298 bp and representing a total of 593.27 kb of DNA sequence were isolated from the assembled metagenomes (Table 2); 77.9% of them were retrieved from 53 m, 20.8% from 65 m and 1.3% from 80 m; none was isolated from 4 m (Fig. 4A, Table S6, supplementary material). The longest contigs were identified at 65 m with a mean size of 2800 bp against 1688 bp for 53 m and 646 bp for 80 m (Fig. 4B). A total of 622 CDSs was predicted (Table S7, supplementary material), with an average length of 875 bp (3399 bp for the largest), corresponding to a protein-coding content of 91.7%, as well as seven rRNA genes and one tRNA gene. As seen with the methanogens, the intergenic regions had a lower G+C percentage (33.9%) and the gene sequences a higher G+C percentage (41.9%) compared with the overall G+C content (41.3%) (Table 2). No rRNA operon could be reconstructed from assembled data; 16S rRNA genes had 99 to 100% identity to *Methylobacter* strains and 23S rRNA genes, 92 to 98% identity to *Methylomonas* strains (Table S6, supplementary material).

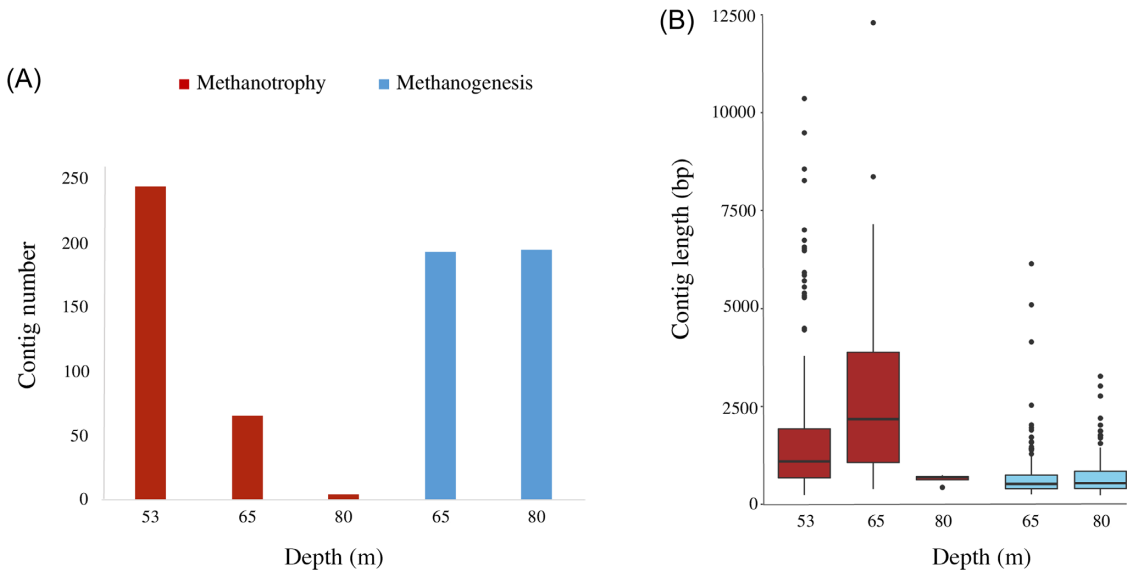


Figure 4. Contig distribution analysis. **(A)** Contig number distribution depending on metabolic groups and depths. **(B)** Boxplots of contig size distribution according to metabolic groups and depths.

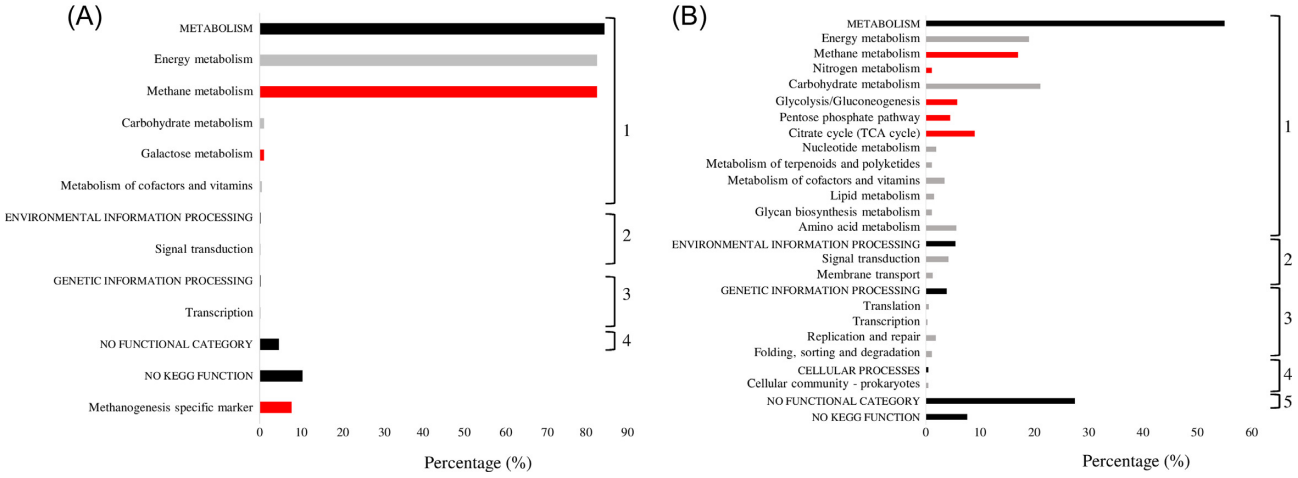


Figure 5. KEGG pathway assignment of CDSs. **(A)** Distribution of methanogenic CDSs to different KEGG subcategories belonging to four main categories (hierarchical categories (HC) noted in capital letters). **(B)** Distribution of methanotrophic CDSs to different KEGG subcategories belonging to five HCs. In each HC, the most represented subcategories appear in grey for the functional categories and in red for the pathways. Each CDS was classified into a single category. No functional category: CDSs poorly characterized in the KEGG database. No KEGG function: CDSs without KEGG affiliation.

Like the reads, CDSs had the highest frequency of high-scoring BLAST hits with the species *Methylobacter tundripaludum* 21/22 (79.1%) with an average amino acid identity of 81.3% (range 46.8–100%). The frequencies of top BLASTP hits with other groups were as follows: *Methylosarcina* (9.2%), *Methylomicrobium* (6.9%), *Methylomonas* (2.9%) and *Methylomarinum* (1.9%) (Table S7, supplementary material). Even though, as expected, the majority of the CDSs (40.1%) were associated with the targeted metabolisms (mainly carbohydrates with glycolysis, pentose phosphate pathway and TCA cycle (19.3%) and energy (specifically to CH_4 (19%) and N (1.1%) metabolisms)), this approach highlighted genes involved in many other important processes (Fig. 5B). Compared with methanogenesis, a large proportion (>27%) of the identified KO genes remained uncategorized (Fig. 5, Table S7, supplementary material). All of the CDSs encoded 361 distinct enzymes. The mapping of all of the contigs onto *M. tundripaludum* 21/22 reference genome showed that they covered up to 381 kbp, scattered over 111 unique regions (Fig. 7A). Gene organization conservation in both order and orientation was observed for most of the genomic regions, except in seven locations where rearrangements were detected, mostly deletions (Fig. 7B).

From the read and CDS analysis, most of the gene-encoding enzymes responsible for CH_4 oxidation to CO_2 were uncovered: (i) both the particulate and soluble forms of the methane monooxygenases (pMMO and sMMO, respectively), the key enzymes responsible for the CH_4 oxidation to methanol; (ii) the Ca^{2+}-dependent (Mxa) and the Ln^{3+}-dependent (XoxF) forms of methanol dehydrogenases (MDHs), another hallmark of the methanotrophy responsible for the oxidation of methanol to formaldehyde; and (iii) enzymes required for the pathways involving the two pterin cofactors, i.e. the tetrahydrofolate (H_4F) and the tetrahydromethanopterin (H_4MPT), both responsible for

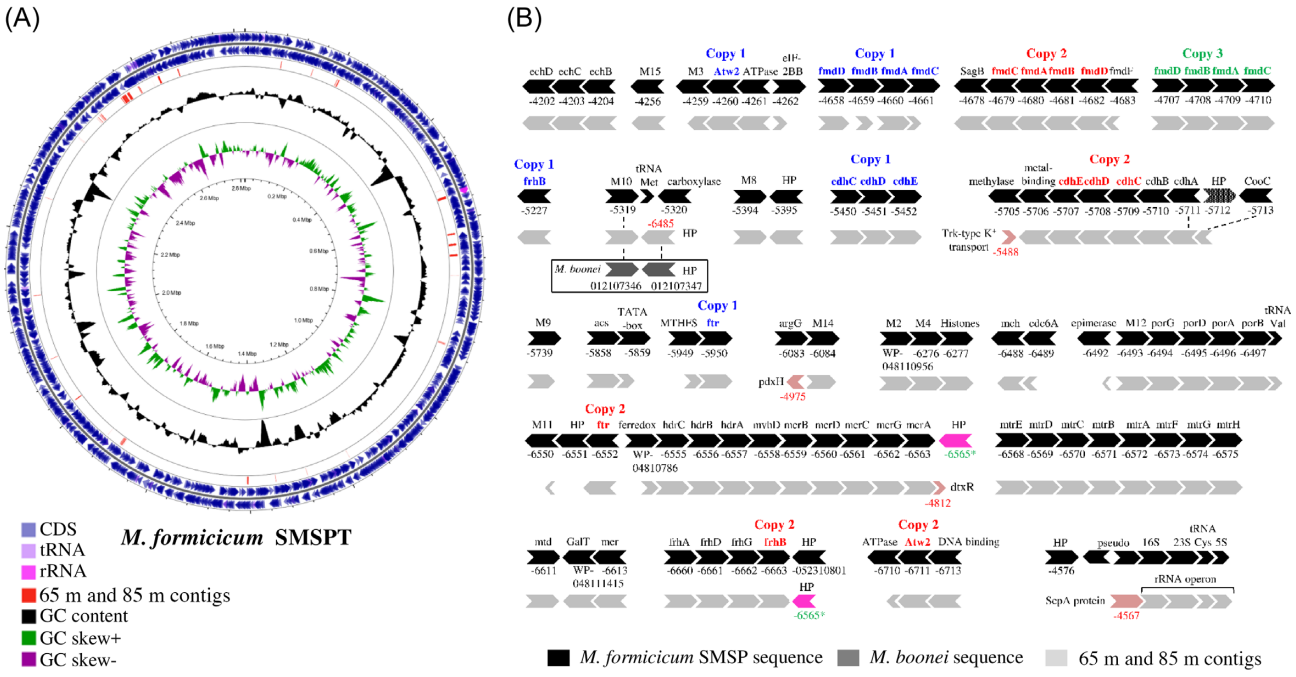


Figure 6. Graphical representation of contig-mapping onto *Methanoregula formicicum* SMSP chromosome used as genome reference (CP003167). **(A)** Circular map of *M. formicicum* SMSP chromosome performed with a CGview comparison tool (Grant and Stothard 2008). From outer to inner circle: genes on forward strand; genes on reverse strand; genes carried by contigs; GC content and GC skew. **(B)** Representation of the reconstruction of genomic regions for methanogens from contig overlaps. *M. formicicum* SMSP used as the genome reference is represented in black, *Methanoregula boonei* (CP000780) in dark grey. The accession number is indicated under each gene with a dash replacing WP_01528. The name of the genes in multiple copies is represented in color. The genes in light red in contigs indicate substitution in comparison with reference while the genes in pink indicate a position change. HP: hypothetical protein.

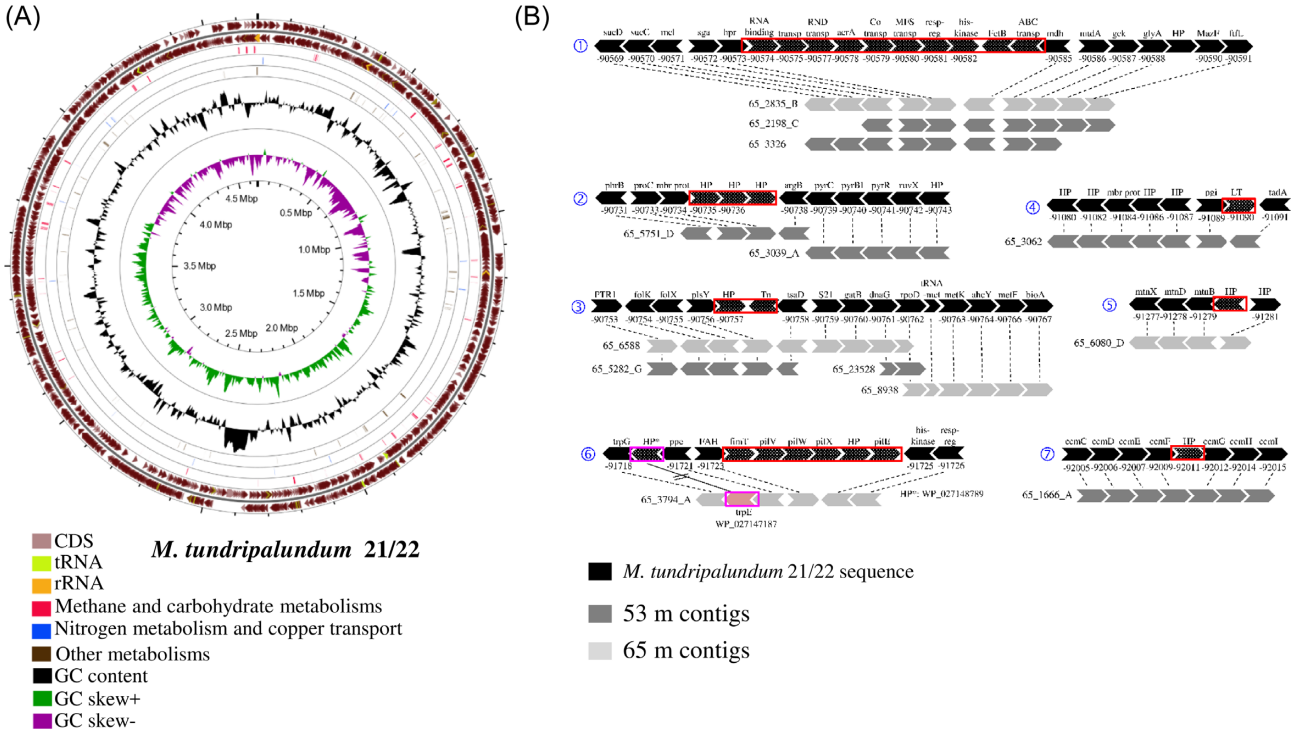


Figure 7. Graphical representation of contig-mapping onto *Methylobacter tundripaludum* 21/22 chromosome used as genome reference (JMLA01000001). **(A)** Circular map of *M. tundripaludum* 21/22 chromosome performed with a CGview comparison tool (Grant and Stothard 2008). From outer to inner circle: genes on forward strand; genes on reverse strand; genes carried by contigs and related to methane and carbohydrate metabolisms; CDSs related to nitrogen metabolism and copper transport; all remaining CDSs; GC content and GC skew. **(B)** Representation of contigs with a different gene organization in comparison with reference. Genes are represented by arrows. Red rectangles indicate genes present on reference genome but not in contigs. Pink rectangles indicate a gene substitution at one position. Contigs isolated from 53 m depths are represented in dark grey and those isolated from 65 m depth in light grey. Genome reference is represented in black. Accession number is indicated under each gene with a dash replacing WP_0068. HP: hypothetical protein.

formaldehyde dissimilation to CO_2 (Fig. S2, Table S6, supplementary material). Furthermore, the expression and the activity of pMMO are known to be tightly regulated by copper (Cu) (for a review, see Semrau, DiSpirito and Yoon 2010). In our study, CopA, a Cu-ATPase involved in Cu efflux and CopZ, a Cu chaperone, both known to be organized within a Cu-regulated operon (Sithisak *et al.* 2007), were found. As for the MDHs, the Mxa cluster showed the same highly conserved gene organization as other organisms, with the *mxaF* gene being linked to genes *mxaJ* and *mxaG* (Table S6, supplementary material). However, the small catalytic subunit *mxaI* was not detected as the assembled contig carrying the complex was not sufficiently long.

In the MOBs, C assimilation takes place either at the level of formaldehyde via the ribulose monophosphate (RuMP) pathway or at the level of methylene-H_4F and CO_2 via the serine cycle. Regarding the RuMP cycle, most of the genes encoding the enzymes involved both in the Embden-Meyerhof-Parnas (EMP) variant and the collateral Enthner-Doudoroff (EDD) variant were present in the metagenomes. The enzymes responsible for the inter-conversions of the C3 intermediate metabolites (pyruvate and phosphoenolpyruvate) resulting from formaldehyde conversion through RuMP were also detected, as well as those for a complete tricarboxylic (TCA) cycle. In contrast, none of the sequences detected for Ack, Pta and Acs enzymes, which are required for acetate conversion to AcCoA, were classified as MOBs. The components of the dissimilatory RuMP cycle for formaldehyde oxidation and RuMP regeneration were also present (Fig. S2, supplementary material). As for the serine cycle, almost all of the related genes were uncovered from the contigs (Fig. S2, supplementary material). They clustered on the same genomic fragment (Table S6, supplementary material). Only the malyl-CoA-lyase (Mcl), the truly specific enzyme of the serine cycle, and the hydroxypyruvate reductase (HPR), were missing from our metagenomic data. Though this cycle is considered as the hallmark of type II MOBs, corresponding sequences invariably showed better matches with type I MOBs. Finally, regarding the inventory of essential genes encoding N metabolisms, the methanotrophic genes involved in NO_3^- transport (*narK*), denitrification (*nirJ* and *norB*), nitrite assimilation (*nirBD*), N fixation (*nifDK*), NH_4^+ transport (*amt*) and assimilation (*glnBDE* and *ntrBC*), and, potentially, nitrification (*hao* homologous sequences), were represented (Fig. S3, Table S6, supplementary material).

DISCUSSION

In this study, we investigated the microbial taxa involved in the CH_4 cycle which are inhabiting the waters of Lake Pavin through the combination of assembly-free and assembly-based analyses of metagenomic data. This two-tiered strategy was adopted because of the high microbial diversity found in the lake waters, whereas methanogens and methanotrophs comprize a minor fraction of the natural community. Indeed, methanogens were estimated to contribute to less than 2% of the total Lake Pavin anoxic water community (Lehours *et al.* 2005) while planktonic MOBs contribute up to 5% of the total cell numbers in most studied lakes (Carini *et al.* 2005; Jones and Lennon 2009; Oswald *et al.* 2015; Samad and Bertilsson 2017). In these conditions, an approach only based on metagenomic assembly analysis would have required many more computational resources and been much more time-consuming without the possibility of assembling whole genomes for these guilds owing to insufficient coverage (Quince *et al.* 2017). Furthermore, despite its democratization, the use of traditional metagenomics to specifically investigate key metabolic functions in highly diversified microbial communities is still scarce. Most of the metagenomic approaches developed for these issues rely on substrate-specific labeling methods such as stable isotope probing (SIP) before sequencing. However, though used to study benthic methylotrophic populations (Kalyuzhnaya *et al.* 2008; Chistoserdova 2011; Beck *et al.* 2013), this method has never been applied on pelagic populations.

Methanoregula as dominant CH_4 producers in anoxic waters

In the present work, the anoxic metagenomic datasets harbored a low proportion of archaea (with a high of 3.7% of taxonomy-classified reads), with 99.7% of archaeal selected reads assigned to methanogenic taxa, mostly *Methanomicrobiales* (Fig. 3). This order, widely reported from planktonic environments (Crevecoeur, Vincent and Lovejoy 2016), was designated as an indicator group for freshwater (Auguet, Barberan and Casamayor 2010) and, consequently, considered among the main players in the pivotal ecological functions within this kind of habitat. In our study, no depth-related change in terms of methanogenic read abundance or diversity was observed. This implies that these CH_4 producers are likely to occur at relatively constant proportions across the anoxic waters and would contribute, at least in part, to the CH_4 measured in the waters of Lake Pavin, as previously demonstrated in other lakes (Iversen, Oremland and Klug 1987; Crowe *et al.* 2011). The genus *Methanoregula* was the dominant taxon, which corroborates our previous findings based on the use of single markers (Biderre-Petit *et al.* 2011b; Denonfoux *et al.* 2013). This genus, for which 16S rRNA sequences were recovered in most freshwater lake clone libraries (Borrel *et al.* 2011), is suspected to have poor salt tolerance (Tong *et al.* 2017).

On the basis of the CDSs information, a draft of methanogenesis pathways was successfully reconstructed for the methanogens in Lake Pavin with the identification of 55 genes, all with encoding proteins directly or indirectly involved in hydrogenotrophic processes. As for the other *Methanoregula* sequenced, the phylotypes living in Lake Pavin are expected to utilize H_2/CO_2 for growth, highlighting the importance of hydrogenotrophic methanogenesis in this ecosystem in comparison with the other methanogenic pathways. Concerning central carbon metabolism, we note the presence of the genes for a Cdh complex, known to allow CO_2 fixation by the Wood-Ljungdahl pathway. Beyond autotrophic metabolism, the methanogens in Lake Pavin also contained an acetyl-CoA synthetase (Acs) and a phosphotransacetylase (Pta), known to be involved in acetate activation, which has been demonstrated previously in other hydrogenotrophic methanogens (Kouzuma *et al.* 2017). This could be an advantage under limited growth conditions. Moreover, a large set of hydrogenases was found (Ech, Mvh, Eha, Ehb and Frh). As with most *Methanomicrobiales*, the complementary [NiFe] hydrogenase (MvhAG) was not identified, but the MvhD subunit was present, tightly bound to HdrABC operon. The best explanation was that MvhAG could be replaced by FrhAG (identified as well) to form a functional complex to catalyze methanogenesis through electron bifurcation (Anderson *et al.* 2009; Thauer *et al.* 2010; Kaster *et al.* 2011; Browne *et al.* 2017; Gilmore *et al.* 2017). However, that association has still not been experimentally verified. Finally, the reconstructed Mcr operon, the key enzyme of methanogenesis, displayed the same genomic organization as that described using the capture approach, with the *dtxR* gene (encoding a metal-dependent

repressor) located directly upstream (Denonfoux *et al.* 2013). Considering that this organization differs from what is generally observed for sequenced *Methanoregula*, it could suggest that it reflects the adaptation of the methanogens to their environment in Lake Pavin.

The assembly-based analyses of the metagenomic data revealed several variants with nucleotide differences of a few per cent for most of the functional genes characterized for methanogens, thus resulting in an important marker redundancy and a short size for most of the reconstructed contigs. Such a functional diversity suggests the co-existence of many closely related populations for the hydrogenotrophic *Methanoregula* group living in Lake Pavin. To sum up, in this study, the functional genes, with the potential to shed light onto strain-level heterogeneity, appeared to be much better-performing markers yielding to a taxonomic resolution not achievable by ribosomal RNA genes. Indeed, all recovered rRNA gene sequences (16S, 23S and 5S rRNA genes) were identical, while rRNA operons only differed by a few nucleotides in spacer regions. Hence, some of these strains could form separate groups with distinct ecological relevance. However, such conclusions must be considered with care because characterizing strains from assembled metagenomes alone remains challenging on account of the many possible biases such as sequencing errors, fragmentary contigs and even the quality of the assembly (Kelley and Salzberg 2010; Wang, Ye and Tang 2012; Shapiro and Polz 2014).

Methylobacter as dominant MOBs under hypoxic and suboxic conditions

Planktonic MOBs, a functional guild found in all studied lakes regardless of status, depth or water chemistry (Ross *et al.* 1997; Eller *et al.* 2005; Sundh, Bastviken and Tranvik 2005; Kojima, Fukuhara and Fukui 2009; Samad and Bertilsson 2017), constitute an important part of freshwater ecosystems and are considered a key link between sediment and pelagic C flow. However, despite the intensive research carried out over the last few years, major gaps still exist in our fundamental knowledge of this key microbial group. In stratified lakes, it consumes most of the CH_4 at the oxic-anoxic interface, but also in anoxic waters, sometimes well below the chemocline (Biderre-Petit *et al.* 2011b; Peura *et al.* 2012; Blees *et al.* 2014). In this work, the analysis of the metagenomes in Lake Pavin supports previous metagenomic findings for stratified lakes (Peura *et al.* 2015), i.e. a high frequency of methanotrophic genes in the hypoxic and upper anoxic zones while practically absent near the surface. *Methylobacter* was the dominant taxon and thus could be the main active methanotroph species, even under anoxic conditions. This postulation is in accordance with the growing number of studies that point to its dominant role in lakes, whether they are stratified (Taipale, Jones and Tiirola 2009; Blees *et al.* 2014) or not (Grossart *et al.* 2011; Tsutsumi *et al.* 2011; Ullrich *et al.* 2016), as well as in sediments (Kalyuzhnaya *et al.* 2008; Dumont, Pommerenke and Casper 2013; Oshkin *et al.* 2015; Martinez-Cruz *et al.* 2017). Low temperatures were proposed to be a significant driver of its predominance (Tsutsumi *et al.* 2011). Furthermore, comparisons of sample-specific datasets uncovered the presence of distinct populations and richness closely related to *Methylobacter tundripaludum* depending on the depth. This suggests a vertical shift in its assemblage along the water column with potentially different ecophysiological characteristics for *Methylobacter* strains, leading to niche differentiation, as hypothesized in other studies (Tsutsumi *et al.* 2011; Oshkin *et al.* 2015).

In this work, 361 gene-encoding proteins involved in the C, N and energy pathways were identified thus allowing a partial reconstruction of the *Methylobacter* metabolism. This included all three subunits of the pMMO, the key enzyme of CH_4 oxidation that was organized in an operon *pmoCAB* as in the other type I MOBs (Trotsenko and Murrell 2008). However, the reads detected for sMMO have better matches with *Methylomonas* and *Crenothrix* than with *Methylobacter*. This also includes the two distinct MDH systems, i.e. the classical Ca^{2+}-dependent MxaF-type and the Ln^{3+}-containing XoxF-type. Moreover, it was recently demonstrated that the presence of non-methanotrophic methylotrophs, particularly *Methylotenera,* could induce a change in the expression of MOB MDHs, with a downregulation of the dominant XoxF-type and an upregulation of the MxaF-type (Krause *et al.* 2017). This switch, resulting in a higher methanol release in medium, could benefit methylotroph growth through a cross-feeding mechanism (He *et al.* 2012; Beck *et al.* 2013). The presence of the pelagic *Methylotenera* in this study positively correlated with that of *Methylobacter*, hence suggesting a potential similar cross-feeding mechanism. Furthermore, their co-location with the peak of NO_3^- (Fig. S1, supplementary material) and low-O_2 concentrations (Fig. 3) could be explained by a niche adaptation conferred by a putative denitrification capability as previously shown in other ecosystems (Kalyuzhnaya *et al.* 2008; Stein and Klotz 2011; Chistoserdova 2015; Oshkin *et al.* 2015; Oswald *et al.* 2016; Martinez-Cruz *et al.* 2017). Indeed, the coupling of NO_3^- respiration to CH_4 oxidation has previously been reported for *Methylobacter* and other MOBs, alone (Kits, Klotz and Stein 2015) or in cooperation (Beck *et al.* 2013; Dumont, Pommerenke and Casper 2013; Zhu *et al.* 2016). However, further investigation is still needed to demonstrate that these species engage in cooperative behavior and the importance of the N cycle in this cooperation. It is interesting to note that the methanotrophs that thrive in Lake Pavin have a rich inventory for nitrogen oxide metabolism.

Both the EDD- and EMP-variants of the RuMP cycle, required by type I MOBs to assimilate formaldehyde into biomass, were predicted in our assembled metagenomes. Which of these variants is primarily employed is still today a debated question, though a recent analysis pointed towards the EMP variant as the major route for single C assimilation, contrary to existing assumptions (Kalyuzhnaya *et al.* 2013). We also showed that the *Methylobacter* in Lake Pavin is endowed with the majority of the serine cycle genes, thus agreeing with recent genomic and proteomic data (de la Torre *et al.* 2015; Ullrich *et al.* 2016; Padilla *et al.* 2017). On the other hand, the interconnection of a partial serine cycle and the EMP variant might represent an alternative strategy to increase the conversion efficiency of AcCoA production from CH_4 (Kalyuzhnaya, Puri and Lidstrom 2015) which, once generated, could enter the TCA cycle for which all essential genes were detected. The presence of putative mixed-acid fermentation and H_2 production genes in the *Methylobacter* genome also opens up a possibility for fermentation under O_2-limiting conditions as recently proposed for a *Methylomicrobium* species (Kalyuzhanaya *et al.* 2013). Indeed, the switching capability from a respiratory mode to a fermentation mode understandingly represents an advantage for organisms living in stratified environments such as Lake Pavin, where availability of O_2 and other electron acceptors varies. Hence, the very broad metabolic flexibility of MOBs in Lake Pavin might be the key to their presence in the hypoxic and suboxic zones, as this allows them to adjust to shifting environmental settings (Oswald *et al.* 2016).

CONCLUSION

Overall, this study expands the current genomic knowledge of pelagic CH_4 producers and utilizers in Lake Pavin. In the absence of cultivated species from the Lake Pavin water column, metagenomics, with the partial reconstruction of the core parts required for methanogenesis and methanotrophy, was revealed to be much more informative than the PCR-dependent approaches based on the use of single marker genes. Indeed, although the latter methods have significantly contributed to the advancement of our understanding of the uncultivated methanogens and MOBs in Lake Pavin (Lehours *et al.* 2007; Biderre-Petit *et al.* 2011b), they didn't provide any insight into their genomic structures. This outcome is of particular importance when considering the microbial communities thriving along the steep redox gradients of O_2-minimum zones as these habitats are predicted to expand in response to climate change. Furthermore, the overlapping patterns between the methanogens and the MOBs in the suboxic area of Lake Pavin are consistent with the simultaneous production and consumption of CH_4. This knowledge is critical for predicting links among greenhouse gas, C and nutrient fluxes in such ecosystems. However, as all of our results are based on metagenomics (DNA level), further approaches based on metatranscriptomics or metaproteomics are required in future studies to explore the principal functions of the living microorganisms involved in the CH_4 production in the water masses of Lake Pavin.

SUPPLEMENTARY DATA

Supplementary data are available at FEMSEC online.

ACKNOWLEDGMENTS

We thank Agnès Vellet for efficient technical assistance and Cécile Lepère for reviewing the English version of the manuscript.

FUNDING

The work of C.H. was supported by the Agence Nationale de la Recherche (ANR) through the project EUREKA (ANR-14-CE02-0004-01). H.G. is supported by a grant from the French Ministry for Higher Education and Research.

Conflicts of interest. None declared.

REFERENCES

Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.

Anderson I, Ulrich LE, Lupa B *et al.* Genomic characterization of methanomicrobiales reveals three classes of methanogens. *PLoS One* 2009;**4**:e5797.

Auguet J-C, Barberan A, Casamayor EO. Global ecological patterns in uncultured Archaea. *ISME J* 2010;**4**:182.

Auman AJ, Stolyar S, Costello AM *et al.* Molecular characterization of methanotrophic isolates from freshwater lake sediment. *Appl Environ Microbiol* 2000;**66**:5259–66.

Bastviken D, Cole J, Pace M *et al.* Methane emissions from lakes: Dependence of lake characteristics, two regional assessments, and a global estimate: LAKE METHANE EMISSIONS. *Glob Biogeochem Cycles* 2004;**18**:n/a – n/a.

Bastviken D, Tranvik LJ, Downing JA *et al.* Freshwater methane emissions offset the continental carbon sink. *Science* 2011;**331**:50–50.

Beck DAC, Kalyuzhnaya MG, Malfatti S *et al.* A metagenomic insight into freshwater methane-utilizing communities and evidence for cooperation between the *Methylococcaceae* and the *Methylophilaceae*. *PeerJ* 2013;**1**:e23.

Biderre-Petit C, Boucher D, Kuever J *et al.* Identification of sulfur-cycle prokaryotes in a low-sulfate lake (Lake Pavin) using *aprA* and 16S rRNA gene markers. *Microb Ecol* 2011a;**61**:313–27.

Biderre-Petit C, Jézéquel D, Dugat-Bony E *et al.* Identification of microbial communities involved in the methane cycle of a freshwater meromictic lake: Methane cycle in a stratified freshwater ecosystem. *FEMS Microbiol Ecol* 2011b;**77**:533–45.

Blees J, Niemann H, Wenk CB *et al.* Micro-aerobic bacterial methane oxidation in the chemocline and anoxic water column of deep south-Alpine Lake Lugano (Switzerland). *Limnol Oceanogr* 2014;**59**:311–24.

Borges AV, Abril G, Delille B *et al.* Diffusive methane emissions to the atmosphere from Lake Kivu (Eastern Africa). *J Geophys Res* 2011;**116**, DOI: 10.1029/2011JG001673.

Borrel G, Jézéquel D, Biderre-Petit C *et al.* Production and consumption of methane in freshwater lake ecosystems. *Res Microbiol* 2011;**162**:832–47.

Browne P, Tamaki H, Kyrpides N *et al.* Genomic composition and dynamics among Methanomicrobiales predict adaptation to contrasting environments. *ISME J* 2017;**11**:87–99.

Carini S, Bano N, LeCleir G *et al.* Aerobic methane oxidation and methanotroph community composition during seasonal stratification in Mono Lake, California (USA). *Environ Microbiol* 2005;**7**:1127–38.

Chistoserdova L, Lidstrom ME. Aerobic Methylotrophic Prokaryotes. In: Rosenberg E, DeLong EF, Lory S *et al.* (eds). *The Prokaryotes*. Springer: Berlin, Heidelberg, 2013, 267–85.

Chistoserdova L. Methylotrophs in natural habitats: current insights through metagenomics. *Appl Microbiol Biotechnol* 2015;**99**:5763–79.

Chistoserdova L. Methylotrophy in a lake: from metagenomics to single-organism physiology. *Appl Environ Microbiol* 2011;**77**:4705–11.

Cole JJ, Caraco NF, Kling GW *et al.* Carbon Dioxide Supersaturation in the Surface Waters of Lakes. *Science* 1994;**265**:1568–70.

Conrad R, Klose M, Noll M. Functional and structural response of the methanogenic microbial community in rice field soil to temperature change. *Environ Microbiol* 2009;**11**:1844–53.

Crevecoeur S, Vincent WF, Lovejoy C. Environmental selection of planktonic methanogens in permafrost thaw ponds. *Sci Rep* 2016;**6**:1–10, DOI: 10.1038/srep31312.

Crowe SA, Katsev S, Leslie K *et al.* The methane cycle in ferruginous Lake Matano: Methane cycle in ferruginous Lake Matano. *Geobiology* 2011;**9**:61–78.

De la Torre A, Metivier A, Chu F *et al.* Genome-scale metabolic reconstructions and theoretical investigation of methane conversion in Methylomicrobium buryatense strain 5G(B1). *Microb Cell Factories* 2015;**14**:188, DOI: 10.1186/s12934-015-0377 -3.

Denonfoux J, Parisot N, Dugat-Bony E *et al.* Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration. *DNA Res* 2013;**20**:185–96.

Downing JA, Prairie YT, Cole JJ *et al.* The global abundance and size distribution of lakes, ponds, and impoundments. *Limnol Oceanogr* 2006;**51**:2388–97.

Dumont MG, Pommerenke B, Casper P. Using stable isotope probing to obtain a targeted metatranscriptome of aerobic methanotrophs in lake sediment: SIP-metatranscriptomics of methanotrophs. *Environ Microbiol Rep* 2013:757–64.

Eller G, Deines P, Grey J *et al.* Methane cycling in lake sediments and its influence on chironomid larval Î´13C. *FEMS Microbiol Ecol* 2005;**54**:339–50.

Ettwig KF, Butler MK, Le Paslier D *et al.* Nitrite-driven anaerobic methane oxidation by oxygenic bacteria. *Nature* 2010;**464**:543–8.

Gao B, Gupta RS. Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. *BMC Genomics* 2007;**8**:86.

Gilmore SP, Henske JK, Sexton JA *et al.* Genomic analysis of methanogenic archaea reveals a shift towards energy conservation. *BMC Genomics* 2017;**18**:639, DOI: 10.1186/s12864-017-4036-4.

Grant JR, Stothard P. The CGView Server: a comparative genomics tool for circular genomes. *Nucleic Acids Res* 2008;**36**:W181–4.

Grossart H-P, Frindte K, Dziallas C *et al.* Microbial methane production in oxygenated water column of an oligotrophic lake. *Proc Natl Acad Sci* 2011;**108**:19657–61.

He R, Wooller MJ, Pohlman JW *et al.* Diversity of active aerobic methanotrophs along depth profiles of arctic and subarctic lake water column and sediments. *ISME J* 2012;**6**:1937–48.

Hernandez ME, Beck DAC, Lidstrom ME *et al.* Oxygen availability is a major factor in determining the composition of microbial communities involved in methane oxidation. *PeerJ* 2015;**3**:e801.

Iversen N, Oremland RS, Klug MJ. Big Soda Lake (Nevada). 3. Pelagic methanogenesis and anaerobic methane oxidation. *Limnol Oceanogr* 1987;**32**:804–14.

Johnson M, Zaretskaya I, Raytselis Y *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res* 2008;**36**:W5–W9.

Jones S, Lennon J. Evidence for limited microbial transfer of methane in a planktonic food web. *Aquat Microb Ecol* 2009;**58**:45–53.

Kalyuzhnaya MG, Lapidus A, Ivanova N *et al.* High-resolution metagenomics targets specific functional types in complex microbial communities. *Nat Biotechnol* 2008;**26**:1029–34.

Kalyuzhnaya MG, Puri AW, Lidstrom ME. Metabolic engineering in methanotrophic bacteria. *Metab Eng* 2015;**29**:142–52.

Kalyuzhnaya MG, Yang S, Rozova ON *et al.* Highly efficient methane biocatalysis revealed in a methanotrophic bacterium. *Nat Commun* 2013;**4**:2785, DOI: 10.1038/ncomms3785.

Kanehisa M, Sato Y, Kawashima M *et al.* KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;**44**:D457–62.

Karr EA, Ng JM, Belchik SM *et al.* Biodiversity of Methanogenic and Other Archaea in the Permanently Frozen Lake Fryxell, Antarctica. *Appl Environ Microbiol* 2006;**72**:1663–6.

Kaster A-K, Moll J, Parey K *et al.* Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea. *Proc Natl Acad Sci* 2011;**108**:2981–6.

Kelley DR, Salzberg SL. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics* 2010;**11**:544.

Kirschke S, Bousquet P, Ciais P *et al.* Three decades of global methane sources and sinks. *Nat Geosci* 2013;**6**:813–23.

Kits KD, Klotz MG, Stein LY. Methane oxidation coupled to nitrate reduction under hypoxia by the Gammaproteobacterium *M ethylomonas denitrificans*, sp. nov. type strain FJG1: Denitrifying metabolism in *M. denitrificans* FJG1. *Environ Microbiol* 2015;**17**:3219–32.

Knittel K, Boetius A. Anaerobic oxidation of methane: Progress with an unknown process. *Annu Rev Microbiol* 2009;**63**:311–34.

Kojima H, Fukuhara H, Fukui M. Community structure of microorganisms associated with reddish-brown iron-rich snow. *Syst Appl Microbiol* 2009;**32**:429–37.

Kouzuma A, Tsutsumi M, Ishii S *et al.* Non-autotrophic methanogens dominate in anaerobic digesters. *Sci Rep* 2017;**7**:1510, DOI: 10.1038/s41598-017-01752 -x.

Krause SMB, Johnson T, Samadhi Karunaratne Y *et al.* Lanthanide-dependent cross-feeding of methane-derived carbon is linked by microbial community interactions. *Proc Natl Acad Sci* 2017;**114**:358–63.

Lehours A-C, Bardot C, Thenot A *et al.* Anaerobic Microbial Communities in Lake Pavin, a Unique Meromictic Lake in France. *Appl Environ Microbiol* 2005;**71**:7389–400.

Lehours A-C, Evans P, Bardot C *et al.* Phylogenetic Diversity of archaea and bacteria in the anoxic zone of a meromictic lake (Lake Pavin, France). *Appl Environ Microbiol* 2007;**73**:2016–9.

Lopes F, Viollier E, Thiam A *et al.* Biogeochemical modelling of anaerobic vs. aerobic methane oxidation in a meromictic crater lake (Lake Pavin, France). *Appl Geochem* 2011;**26**:1919–32.

Martinez-Cruz K, Leewis M-C, Herriott IC *et al.* Anaerobic oxidation of methane by aerobic methanotrophs in sub-Arctic lake sediments. *Sci Total Environ* 2017;**607-608**:23–31.

Meyer F, Paarmann D, D'Souza M *et al.* The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008;**9**:386.

Noguchi H, Taniguchi T, Itoh T. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Res Int J Rapid Publ Rep Genes Genomes* 2008;**15**:387–96.

O'Leary NA, Wright MW, Brister JR *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**:D733–45.

Oshkin IY, Beck DA, Lamb AE *et al.* Methane-fed microbial microcosms show differential community dynamics and pinpoint taxa involved in communal response. *ISME J* 2015;**9**:1119.

Oswald K, Milucka J, Brand A *et al.* Aerobic gammaproteobacterial methanotrophs mitigate methane emissions from oxic and anoxic lake waters: Methane oxidation in Lake Zug. *Limnol Oceanogr* 2016;**61**:S101–18.

Oswald K, Milucka J, Brand A *et al.* Light-dependent aerobic methane oxidation reduces methane emissions from seasonally stratified lakes. *PLoS One* 2015;**10**:e0132574.

Padilla CC, Bertagnolli AD, Bristow LA *et al.* Metagenomic binning recovers a transcriptionally active gammaproteobacterium linking methanotrophy to partial denitrification in an anoxic oxygen minimum zone. *Front Mar Sci* 2017;**4**:23, DOI: 10.3389/fmars.2017.00023.

Pearson WR, Wood T, Zhang Z, Comparison of DNA sequences with protein sequences. *Genomics* 1997;**46**:24–36.

Peng Y, Leung HCM, Yiu SM *et al.* IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 2012;**28**:1420–8.

Pernthaler J. Competition and niche separation of pelagic bacteria in freshwater habitats. *Environ Microbiol* 2017;**19**:2133–50.

Peura S, Eiler A, Bertilsson S *et al.* Distinct and diverse anaerobic bacterial communities in boreal lakes dominated by candidate division OD1. *ISME J* 2012;**6**:1640.

Peura S, Sinclair L, Bertilsson S *et al*. Metagenomic insights into strategies of aerobic and anaerobic carbon and nitrogen transformation in boreal lakes. *Sci Rep* 2015;**5**:12102.

Quast C, Pruesse E, Yilmaz P *et al*. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:D590–6.

Quince C, Walker AW, Simpson JT *et al*. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017;**35**:833–44.

Rahalkar M, Deutzmann J, Schink B *et al*. Abundance and activity of methanotrophic bacteria in littoral and profundal sediments of lake constance (Germany). *Appl Environ Microbiol* 2009;**75**:119–26.

Ross JL, Boon PI, Ford P *et al*. Detection and quantification with 16S rRNA probes of planktonic methylotrophic bacteria in a floodplain lake. *Microb Ecol* 1997;**34**:97–108.

Samad MS, Bertilsson S. Seasonal variation in abundance and diversity of bacterial methanotrophs in five temperate lakes. *Front Microbiol* 2017;**8**:1–12, DOI: 10.3389/fmicb.2017.00142.

Schulz M, Faber E, Hollerbach A *et al*. The methane cycle in the epilimnion of Lake Constance. *Fundam Appl Limnol* 2001;**151**:157–76.

Semrau JD, DiSpirito AA, Yoon S. Methanotrophs and copper. *FEMS Microbiol Rev* 2010;**34**:496–531.

Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol* 2014;**22**:235–47.

Sitthisak S, Knutsson L, Webb JW *et al*. Molecular characterization of the copper transport system in Staphylococcus aureus. *Microbiology* 2007;**153**:4274–83.

Stein LY, Klotz MG. Nitrifying and denitrifying pathways of methanotrophic bacteria. *Biochem Soc Trans* 2011;**39**:1826–31.

Sundh I, Bastviken D, Tranvik LJ. Abundance, activity, and community structure of pelagic methane-oxidizing bacteria in temperate lakes. *Appl Environ Microbiol* 2005;**71**:6746–52.

Taipale S, Jones R, Tiirola M. Vertical diversity of bacteria in an oxygen-stratified humic lake, evaluated using DNA and phospholipid analyses. *Aquat Microb Ecol* 2009;**55**:1–16.

Tang KW, McGinnis DF, Ionescu D *et al*. Methane production in oxic lake waters potentially increases aquatic methane flux to air. *Environ Sci Technol Lett* 2016;**3**:227–33.

Thauer RK, Kaster A-K, Goenrich M *et al*. Hydrogenases from methanogenic archaea, nickel, a novel cofactor, and H_2 storage. *Annu Rev Biochem* 2010;**79**:507–36.

Tong C, Cadillo-Quiroz H, Zeng ZH *et al*. Changes of community structure and abundance of methanogens in soils along a freshwater–brackish water gradient in subtropical estuarine marshes. *Geoderma* 2017;**299**:101–10.

Trotsenko YA, Murrell JC. Metabolic aspects of aerobic obligate methanotrophy⋆. *Advances in Applied Microbiology*. Vol **63**. Elsevier; 2008:183–229.

Tsutsumi M, Iwata T, Kojima H *et al*. Spatiotemporal variations in an assemblage of closely related planktonic aerobic methanotrophs: Spatiotemporal variations in planktonic methanotrophic assemblage. *Freshw Biol* 2011;**56**:342–51.

Ullrich N, Casper P, Otto A *et al*. Proteomic evidence of methanotrophy in methane-enriched hypolimnetic lake water: Proteomics of Lake Methanotrophs. *Limnol Oceanogr* 2016;**61**:S91–100.

Wang M, Ye Y, Tang H. A *de bruijn* graph approach to the quantification of closely-related genomes in a microbial community. *J Comput Biol* 2012;**19**:814–25.

Wu ML, Ettwig KF, Jetten MSM *et al*. A new intra-aerobic metabolism in the nitrite-dependent anaerobic methane-oxidizing bacterium *Candidatus* 'Methylomirabilis oxyfera.' *Biochem Soc Trans* 2011;**39**:243–8.

Zhu J, Wang Q, Yuan M *et al*. Microbiology and potential applications of aerobic methane oxidation coupled to denitrification (AME-D) process: A review. *Water Res* 2016;**90**:203–15.

Article 4 / Co-autrice : Biderre-Petit *et al.*, 2020

Biderre-Petit, C., **Hochart, C.**, **Gardon, H.**, **Dugat-Bony, E.**, **Terrat, S.**, **Jouan-Dufournel, I.** and **Paris, R.** (2020) 'Analysis of bacterial and archaeal communities associated with Fogo volcanic soils of different ages', *FEMS Microbiology Ecology*, 96(7). doi: 10.1093/femsec/fiaa104.

FEMS Microbiology Ecology, 96, 2020, fiaa104

doi: 10.1093/femsec/fiaa104
Advance Access Publication Date: 28 May 2020
Research Article

RESEARCH ARTICLE

Analysis of bacterial and archaeal communities associated with Fogo volcanic soils of different ages

Corinne Biderre-Petit[1,*], Corentin Hochart[2], Hélène Gardon[1], Eric Dugat-Bony[3], Sébastien Terrat[4,†], Isabelle Jouan-Dufournel[1] and Raphaël Paris[5]

[1]CNRS, Laboratoire Microorganismes: Génome et Environnement, Université Clermont Auvergne, F-63000 Clermont-Ferrand, France, [2]CNRS, Laboratoire d'Ecogéochimie des Environnements Benthiques (LECOB), Observatoire Océanologique de Banyuls, Sorbonne Université, F-66650 Banyuls sur Mer, France, [3]INRAE, AgroParisTech, UMR SayFood, Université Paris-Saclay, F-78850, Thiverval-Grignon, France, [4]Agroécologie, AgroSup Dijon, INRAE, Université Bourgogne Franche-Comté, F-21000 Dijon, France and [5]CNRS, IRD, OPGC, Laboratoire Magmas et Volcans, Université Clermont Auvergne, F-63000 Clermont-Ferrand, France

*Corresponding author: 1 impasse Amélie Murat-Bat BioA 63178 Aubière. Tel: 33 4 73 40 51 39; Fax: 33 4 73 40 76 70; E-mail: corinne.petit@uca.fr

One sentence summary: First insights into the archaeal and bacterial diversity colonizing the soils filling microsites on Fogo basaltic lava flows of different ages.

Editor: Petr Baldrian

†Terrat Sébastien, http://orcid.org/0000-0001-5209-6196

ABSTRACT

Basaltic rocks play a significant role in CO_2 sequestration from the atmosphere during their weathering. Moreover, the primary microorganisms that colonize them, by providing mineral elements and nutrients, are shown to promote growth of diverse heterotrophic communities and plants, therefore positively impacting Earth's long-term climate balance. However, the first steps of microbial colonization and subsequent rock weathering remain poorly understood, especially regarding microbial communities over a chronological sequence. Here, we analyzed the microbial communities inhabiting the soil developed in crevices on lava flows derived from different eruptions on Fogo Island. Investigated soils show typically low carbon and nitrogen content and are relatively similar to one another regarding their phylogenetic composition, and similar to what was recorded in large soil surveys with dominance of *Actinobacteria* and *Proteobacteria*. Moreover, our results suggest a stronger effect of the organic carbon than the lava flow age in shaping microbial communities as well as the possibility of exogenous sources of bacteria as important colonizers. Furthermore, archaea reach up to 8.4% of the total microbial community, dominated by the Soil Crenarchaeotic Group, including the ammonium-oxidizer *Candidatus Nitrososphaera* sp. Therefore, this group might be largely responsible for ammonia oxidation under the environmental conditions found on Fogo.

Keywords: Fogo; volcanic soil; basalt; various-aged lava flows; bacterial communities; 16S rRNA gene amplicon sequencing

INTRODUCTION

Volcanic environments are widely distributed on Earth and deciphering the diversity and characteristics of life that they harbor offers us the opportunity to improve our understanding of terrestrial ecosystem formation. Volcanic lava flows are unique ecosystems as they are initially sterile mineral material recovering a preexisting landscape where volcanic

Received: 21 February 2020; **Accepted:** 27 May 2020

eruption acts as a reset factor for life. They are initially highly oligotrophic with low total nitrogen (TN) and total organic carbon (TOC) contents. After eruption, these virgin substrates are available for primary succession (Sato *et al.* 2004) and microorganisms are among the primary colonizers of these newly exposed volcanic substrates (Gomez-Alvarez, King and Nüsslein 2007; Kelly *et al.* 2010). Despite their importance, the origin and identity of these microorganisms are often largely unknown. However, it was demonstrated that some of them may be brought by atmospheric inputs (Womack, Bohannan and Green 2010) through precipitations and dry deposition such as aeolian dust and plant debris (Chuvochina *et al.* 2011; Fujimura *et al.* 2012; DeLeon-Rodriguez *et al.* 2013). It was also suggested that these atmospheric depositions may also provide important sources of C and nutrients to such depleted terrestrial ecosystems (Waldrop *et al.* 2004; Fujimura *et al.* 2012; Rime, Hartmann and Frey 2016). On these young substrates, microbial communities play a central role in rock weathering and early ecosystem development mostly because of the capabilities of the autotrophic species to fix C and N, and to use photo- (i.e. *Cyanobacteria*; Freeman *et al.* 2009) or chemoautotrophic energy-generation systems, which makes them competitive in early stages of succession (Nemergut *et al.* 2007; Sato *et al.* 2009; Fujimura *et al.* 2012; Kerfahi *et al.* 2017). Microbial activities induce the accumulation of TOC in volcanic soil, promoting the growth of diverse heterotrophic communities and changing the physicochemical properties of surrounding microenvironments (Fujimura *et al.* 2016).

Among all weathering processes, those affecting basaltic substrates received more attention in the past years because of their significant role in the global carbonate–silicate cycle (Dessert *et al.* 2003; Daval 2018). Indeed, silicate weathering is responsible for the sequestration of 1.5–3.3 $\times$ 10^8 tons $year^{-1}$ of CO_2 (Hilley and Porder 2008), which could significantly impact the evolution of atmospheric carbon content, knowing that ~60% of the Earth's crust consists of basalts. Because of this characteristic, the transformation of silicates into carbonates from basaltic rocks (i.e. carbonation reaction) has also become a major technological challenge to mitigate environmental consequences of the rising levels of anthropogenic CO_2 emissions. This has resulted in the development of several geological carbon storage plants all over the world, the goal of which being to enable an efficient CO_2 mineralization under the form of stable carbonates by burying this greenhouse gas in deep basaltic rocks. However, Trias *et al.* (2017) demonstrated that these CO_2 injections also resulted in important modifications over short time periods of the deep microbial ecosystems, enhancing the growth of chemolithoautotrophic bacteria, which can promote autotrophic C-fixation. These microbial responses are likely to prevail over much slower abiotic geochemical ones, contributing to CO_2 conversion into much more labile biomass than expected. However, our understanding of the contribution of the microbial biosphere to the C-sequestration, and more generally to silicate weathering both in surface and subsurface, is still in its beginning and further studies are needed.

Despite more intensive research carried out over the last decades, relatively few data are currently available on the diversity of the microbial communities colonizing volcanic rocks, either on fresh deposits or on older and weathered materials. Recent molecular investigations of both types of substrates revealed diverse microbial communities that contained significant proportions of *β-Proteobacteria* for the fresh deposits whereas *α-Proteobacteria*, *Actinobacteria* and *Acidobacteria* are found in the other types of volcanic rocks (Gomez-Alvarez, King and Nüsslein 2007; Kelly *et al.* 2010, 2011, 2014; Byloos *et al.* 2018). However, finding out whether changes in species composition are predictable as soil develops and how shifts in microbial community composition correlate with nutrient dynamics along a chronosequence are still ongoing questions. Most data currently available were obtained from Icelandic and Hawaiian volcanoes, two of the most active volcanic regions on Earth. In contrast, studies regarding the ecology of Fogo Island, another active volcano in the Cape Verde archipelago, remain scarce (Olehowski *et al.* 2008) and mostly focused on vegetation (Brochmann 1997; Romeiras *et al.* 2015; Marques *et al.* 2017; Larrue, Paris and Etienne 2020) and avifauna (Barone and Hering 2010). Some of these studies also highlight an enrichment of Fogo soil in some trace and rare earth elements (Marques *et al.* 2017, 2019). However, no ecological study was performed on volcanic rock-associated microorganisms. In the present study, we sought to redress the lack of characterization of bacterial and archaeal populations inhabiting Fogo soils, by investigating those established in microsites formed on alkaline basaltic rocks, using an in-depth sequencing approach. The examined lava flows were of different ages and exposed on the north-east side of the island, with overall typically low plant coverage. We also assessed the influence of environmental factors on the established microbial communities.

MATERIALS AND METHODS

Geological and ecological setting

Fogo (15.0 N, 24.5 W) is the fourth largest island (476 km^2) of the Cape Verde archipelago, located 400–600 km off West Africa. It is one of the most active volcanoes of the Atlantic Ocean and the only active one of Cape Verde for the last few centuries. Fogo Island corresponds to a shield volcano that exhibits a slightly asymmetric conical shape, being truncated atop by a summit depression opened to the east (Fig. 1). This 8-km-large horseshoe-shaped caldera (so-called Chã das Caldeiras) was formed during an episode of flank collapse ~70 000 years ago (Foeken, Day and Stuart 2009; Paris *et al.* 2011; Ramalho *et al.* 2015). Post-collapse volcanism was dominantly located inside the collapse scar, thus building a new volcanic edifice culminating at the Pico do Fogo (2830 m a.s.l.). Based on the historical record, the mean recurrence interval between eruptions is 19.8 years, with individual intervals ranging from 1 to 94 years. The volcanic rocks of Fogo are silica-undersaturated (38–44% SiO_2) mafic rocks with high alkalinity (2–12% $Na_2O + K_2O$) (Gerlach *et al.* 1988; Doucelance *et al.* 2003; Hildner, Klügel and Hauff 2011).

The semi-arid climate of Fogo Island experiences two seasons: a cool dry season (from November to June) with very low rainfall, and a warmer season with occasional rainfalls (from July to October) (Olehowski *et al.* 2008). As the other islands of the archipelago, Fogo is subjected to dominant north-easterly trade winds. Therefore, the north-eastwards side of the island facing the prevailing winds endures higher rainfall (>600 mm/y) and erosion rates than the leeward southwestern side (~160 mm/y). Above 1300 m a.s.l., a lower rainfall occurs due to the trade wind inversion. A high annual variability in precipitation may occur and an overall decrease in precipitation of 15–30% was observed since the 1970s. Fogo is therefore characterized by very few perennial streams, most rainfall rapidly running to the ocean, evaporating or being used by plants, with the remainder infiltrating through permeable rock to recharge the underlying aquifers.

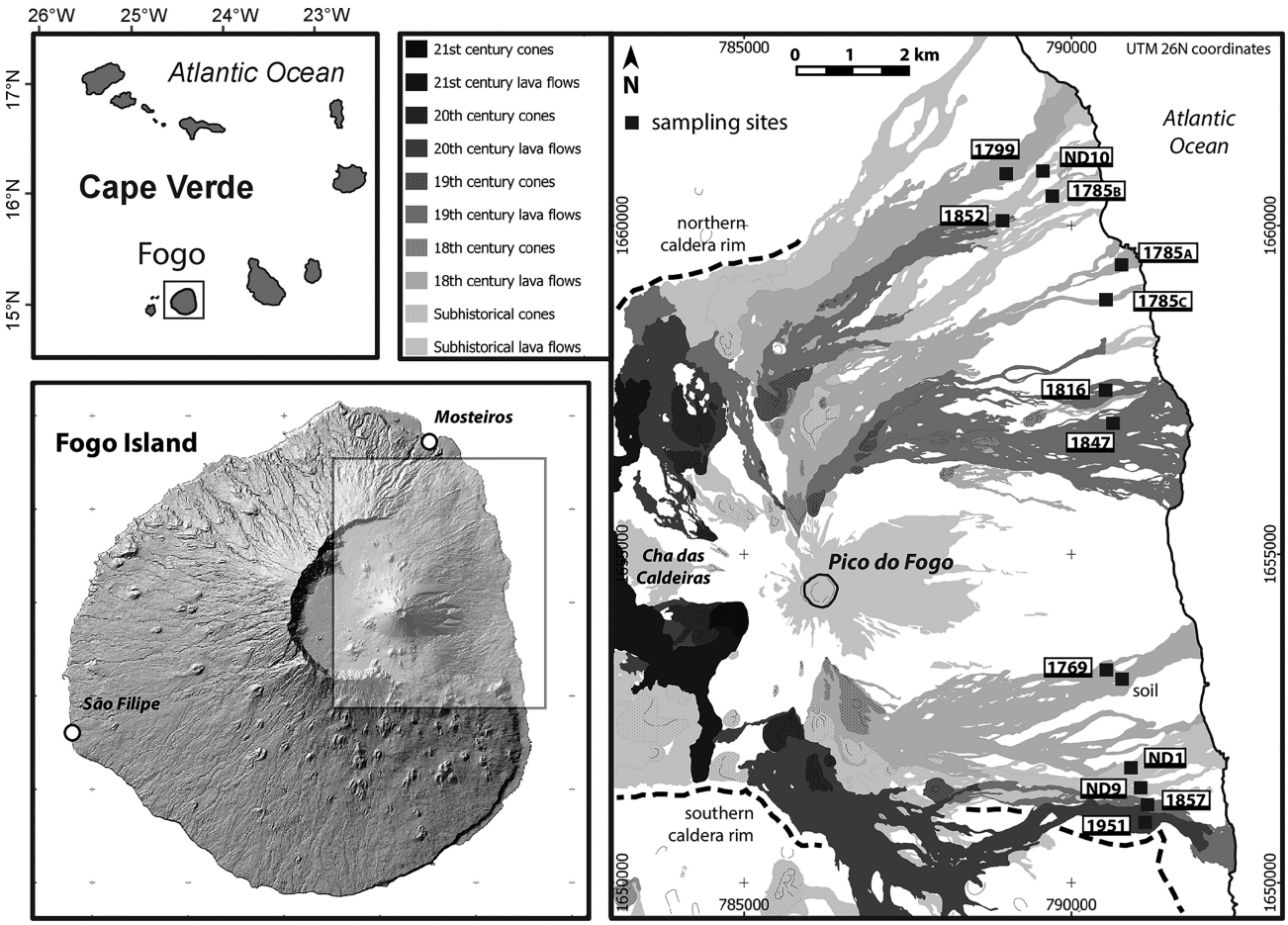


Figure 1. An overview of Fogo Island and location of sampling sites (modified from Larrue, Paris and Etienne 2020). The upper inset shows the geographical position of Fogo Island in the Cape Verde archipelago. The left lower inset indicates the overall sampling zone on the east side of the island. The inset on the right shows the location of the sampling sites on lava flows of different eruptive events. Bold numbers on the investigated lava flows refer to year of eruption. ND indicates undated lava flows. The 'soil' sample is the unique sample not covered by historical or subhistorical lava flow (called 'Soil' in the survey).

Because of water scarcity, the natural vegetation developed on the island is dominated by the savanna type (Olehowski *et al.* 2008). The spatial variability of rainfall leads to island-scale difference in vegetation distribution. The coastal zone is mostly dominated by semi-arid grasslands that rise up to 300 m on the northeastern side and up to 700 m on the southwestern side, followed by a shrub formation.

The soils developed on the island fall into four major soil categories according to the Food and Agriculture Organization's classification (FAO): Regosols, Cambisols, Andosols and Lithosols. The grain size distribution of these soils ranges from dominant coarse material to silty loam material with no cohesion and a higher susceptibility to wind erosion (Marques *et al.* 2014). The most developed soils are found in the highest regions of the steep northeastern side, where precipitations are more important, whereas arid southern areas are dominated by undeveloped shallow soils. Indeed, soil formation was shown as mostly resulting from aeolian processes and steep slopes, chemical and physical weathering being limited by the semi-arid climate and the hardness of the basic volcanic material, respectively (Smolikowski, Puig and Roose 2001).

Soil sampling

This study was conducted on soils developed in crevices on lava flow surface of the eastern flank of Fogo Island, during the wet season in September 2015 and 2016 (Table 1). A total of seven various-aged lava flows (1769, 1785, 1799, 1816, 1847, 1852 and 1951; these dates correspond to the eruption year) and three non-dated lava flows (eruption events prior to 1769 and named ND) were investigated (Table 1). These lava flows shared similar morphological and petrological characteristics. They correspond to basanitic a'a lavas, i.e. lavas with a chaotic surface made of blocks formed during the fragmentation of the cooling crust during lava emplacement. For most eruptions, only one lava channel was sampled (labeled A), whereas three channels were sampled for 1785 and 1816 eruptions (labeled A to C; lava flowing downslope often split into different channels). The distance between each site ranged from 50 to 9980 m (Table S1, Supporting Information). One to three tiny representative crevices (labeled 1 to 3) of homogeneous size (maximum size 0.25 m^2) and of shallow depth were sampled on each channel. Only one soil sample was recovered per crevice except two samples (1785-C1 and C1′) that were taken from the same crevice. The distance

Table 1. Physicochemical variables from volcanic soil samples and brief description.

				Geographical locations		Physicochemical variables								Site description			
Sample	Volcanic eruption date	Sampling date	DNA (mg/g of soil)	Longitude	Latitude	Elevation (meter)	Slope (°)	Exposition (°)	Exposition	TOC (%)	TN (%)	pH (H_2O)	pH (1 M KCl)	Vegetation	Soil type	Moss	Vegetation
1951_A1	1951	13/09/2015	0.64	-24.293246 W	14.918113 N	317	10	90	E	0.29	nd	8	5.7	N	S	N	N
1857_A1	1857	13/09/2015	1.5	-24.293685 W	14.919677 N	320	15	62	NE	0.71	0.06	8.1	5.8	N	S	Y	N
1857_A2	1857	13/09/2015	1.14	-24.293488 W	14.919861 N	312	15	62	NE	0.24	nd	7.8	5.8	P	S	N	P
1857_A3	1857	13/09/2015	0.26	-24.293327 W	14.919781 N	305	11	62	NE	0.59	0.04	7.9	5.7	N	S	Y	N
1852_A1	1852	14/09/2016	28	-24.312128 W	15.000363 N	429	22	55	NE	0.78	0.07	8.7	6.1	N	S	Y	N
1852_A2	1852	14/09/2016	1.68	-24.312237 W	15.000403 N	430	23	55	NE	0.49	0.05	8.7	6.1	P	O	Y	P
1847_A1	1847	13/09/2016	0.12	-24.298097 W	14.971663 N	306	17	74	ENE	0.68	0.06	7.8	6	V	S	N	V
1847_A2	1847	13/09/2016	0.47	-24.297982 W	14.971666 N	300	17	74	ENE	0.27	nd	8.2	6	P	S	Y	P
1847_A3	1847	13/09/2016	0.99	-24.298024 W	14.971666 N	300	17	74	ENE	0.45	0.04	7.9	6.5	N	S	Y	N
1816_A1	1816	14/09/2015	0.51	-24.301053 W	14.976282 N	371	20	78	ENE	0.31	nd	8.2	5.9	N	S	Y	N
1816_A2	1816	14/09/2015	1.74	-24.301044 W	14.976226 N	372	20	78	ENE	0.95	0.1	7.4	6	N	S	Y	N
1816_B1	1816	14/09/2015	4.11	-24.300815 W	14.975785 N	368	22	72	ENE	1.46	0.13	8.2	5.9	V	S	Y	V
1816_B2	1816	14/09/2015	4.2	-24.300774 W	14.975796 N	359	22	72	ENE	2.89	0.29	7.4	6	P	S	Y	P
1816_C1	1816	14/09/2015	0.91	-24.300568 W	14.975176 N	365	24	87	E	0.49	0.04	7.8	5.7	V	O	N	V
1816_C2	1816	14/09/2015	1.33	-24.300604 W	14.975010 N	363	24	80	ENE	0.41	nd	7.6	5.9	P	S	Y	P
1816_C3	1816	14/09/2015	2.6	-24.300608 W	14.974928 N	363	24	80	ENE	0.82	0.07	7.6	5.7	V	S	N	V
1799_A1	1799	14/09/2016	3.28	-24.311672 W	15.006347 N	329	26	81	ENE	2.03	0.13	7.9	5.8	V	S	N	V
1799_A2	1799	14/09/2016	2.15	-24.311753 W	15.006292 N	329	26	81	ENE	1.98	0.16	7.8	5.8	P	S	Y	P
1785_A1	1785	14/09/2015	1.16	-24.296907 W	14.992669 N	137	14	71	ENE	0.28	nd	7.7	5.6	P	O	N	P
1785_A2	1785	16/09/2015	1.64	-24.297655 W	14.992783 N	150	14	68	ENE	0.21	nd	7.8	5.8	P	S	N	P
1785_B1	1785	14/09/2015	7.02	-24.306521 W	15.002944 N	240	15	60	NE	3.81	0.32	6.8	6	V	S	N	V
1785_B2	1785	16/09/2015	6.66	-24.307060 W	15.002640 N	264	19	60	NE	1.39	0.13	8.1	6	P	S	N	P
1785_C1	1785	13/09/2016	2.33	-24.299086 W	14.988658 N	202	16	71	ENE	0.72	0.06	7.6	5.5	V	S	Y	V
1785_C1′	1785	13/09/2016	9.58	-24.299086 W	14.988658 N	202	16	71	ENE	1.07	0.11	7.5	5.7	V	S	Y	V
1769_A1	1769	01/09/2015	0.88	-24.296824 W	14.939195 N	281	19	80	ENE	0.62	0.07	7.2	5.7	N	S	Y	N
1769_A2	1769	13/09/2015	0.63	-24.297123 W	14.939151 N	300	19	80	ENE	0.46	nd	7.6	5.6	V	O	N	V
1769_A3	1769	13/09/2015	0.61	-24.297016 W	14.939091 N	291	19	74	ENE	0.24	nd	8.1	5.5	P	S	N	P
ND1_A1	prior to 1769	16/09/2015	0.86	-24.295857 W	14.922385 N	350	16	107	ESE	0.67	0.05	8	5.5	P	S	N	P
ND1_A2	prior to 1769	16/09/2015	1.41	-24.295970 W	14.922216 N	355	14	107	ESE	2.19	0.19	7.6	5.3	V	O	N	V
ND9_A1	prior to 1769	15/09/2016	3.13	-24.294376 W	14.921758 N	321	14	72	ENE	1.98	0.18	8.6	5.9	V	S	Y	V
ND9_A2	prior to 1769	15/09/2016	2.74	-24.294346 W	14.921799 N	319	14	72	ENE	1.69	0.11	7.9	5.9	V	S	N	V
ND9_A3	prior to 1769	15/09/2016	6.6	-24.294256 W	14.921774 N	321	14	72	ENE	2.05	0.18	8	5.9	V	S	Y	V
ND10_A1	prior to 1769	15/09/2016	5.26	-24.307171 W	15.006887 N	196	16	57	NE	3.06	0.27	7.9	5.8	V	O	N	V
Soil	-	13/09/2015	-	-24.296483 W	14.938871 N	276	18	80	ENE	1.92	0.17	7.6	5.5	V	O	N	V

nd: under threshold
W: west; N: north; E: east; ENE: east-northeast
TOC: total organic carbon; TN: total nitrogen
FAO: Food and Agriculture Organization of the United Nations
S: more or less rocky and shaded; O: open and not rocky
N: no vegetation; P: poorly vegetated; V: vegetated
N: no moss; Y: moss
Sample called 'Soil': sample located outside a lava flow, whether dated or not, hence not associated to a year (–).

between crevices sampled within the same channel ranged from 0.1 to 80 m (Table S2, Supporting Information). A soil from an area outside any historical or subhistorical lava flow was also collected (called Soil) (Table 1). Except the latter, all soil samples were labeled as followed: year of the eruption event or ND / channel (A, B or C) / crevice number (1, 2 or 3).

Soils developed on historical and subhistorical lava flows of the eastern flank (Fig. 1) are mostly Lithosols, hence dominated by weathered rock fragments and lacking in humus. All soil samples were collected in the coastal region at an elevation of between 196 and 430 m, with slope angle comprised between 10 and 26° (Table 1). As previously described, at this height, the area benefits from an annual precipitation of up to 250 mm/year and is characterized by a vegetation of savanna type, except on the youngest lava flows (including that of 1951) that are not weathered enough to allow plant growth on their surface (Olehowski *et al.* 2008). The first centimeters of soils (5–10 cm) were sampled from the bulk soil after elimination of the first surface millimeters, excluding rhizosphere as much as possible where vegetation was present. They were recovered using a spatula sterilized in ethanol and weighing ∼20–40 g, collected directly into sterile tubes, placed immediately on ice and finally frozen at −20°C until chemical and molecular analyses.

Soil physical and chemical analyses

Air-dried soil samples were milled, sieved through a 2-mm mesh to eliminate large particles and roots, homogenized and then prepared according to the protocol for pretreatment of samples for physical and chemical analyses (ISO 11464). Soil pH was measured both in 1 M KCl and deionized water soil suspension (ISO 10390) (1:5 w/v). TOC and TN content were determined by dry combustion method (ISO 10694 and ISO 13878, respectively), with a FlashEA 1112 ThermoQuest elemental analyzer (Thermo Fisher Scientific, Massachusetts, USA). Briefly, all soil samples were filtered through a 200-μm sieve, dried in the oven at 60°C for 2 h and analyzed by the classical Dumas combustion method. This involved combusting the soil matter in the presence of O_2 into simple molecules or gases such as CO_2 and N_2 and then separating these gases using chromatography techniques (Porapak QS50 equipment).

DNA extraction, 16S rRNA gene amplification, Illumina sequencing and bioinformatics

Total genomic DNA (gDNA) was isolated from 500 mg of the 34 soil samples with the FastDNA SPIN kit (MP Biomedical, Santa Anna, CA) according to the manufacturer's protocol. DNA yield was quantified using a Nanodrop spectrophotometer (ND 1000 Spectrophotometer). To perform amplicon next-generation sequencing, the variable region V4–V5 of the 16S rRNA genes was then amplified using the primers 515F (5′-Barcode-GTGYCAGCMGCCGCGGTA-3′) and 909R (5′-Barcode-CCCCGYCAATTCMTTTRAGT-3′) using the following protocol: 94°C for 5 min; 30 cycles of 94°C for 1 min, 58°C for 45 s and 72°C for 45 s; 72°C for 7 min. The amplification reaction was carried out in a final volume of 30 μL containing 10 ng of gDNA, 3 μL of 10× reaction buffer (Eurobio, Les Ulis, France), 1.2 μL of 50 mM $MgCl_2$, 0.75 U of TaqII DNA polymerase (Eurobio, Les Ulis, France), 200 μM of each dNTP, 0.3 μL of 50 mg/mL BSA, 400 nM (each) of barcoded forward and reverse primers. All amplicons were gel-excised, concentrated and cleaned up using a MinElute Gel extraction kit (Qiagen, Germany) according to the manufacturer's instructions. They were then quantified on a Qubit Fluorometer (HS kit, Invitrogen, Carlsbad, CA). PCR products were pooled with equal molarity and sequenced on an Illumina MiSeq instrument (Miseq platform, 250 bp, paired-end). Sequencing was carried out by GATC Biotech GmbH (Konstanz, Germany).

Reads were demultiplexed into sample site datasets and then processed using FROGs pipeline on Galaxy interface (Escudié *et al.* 2018). After a quality control, the merged reads were filtered to minimize the effects of random sequencing errors: (i) only reads with a length between 400 and 520 bp were kept, (ii) those where the two primers were missing were discarded as well as (iii) those with ambiguous nucleotides. Reads were clustered using Swarm (Mahé *et al.* 2014) with an aggregation distance of 3 ($d = 3$), then chimeric sequences in each sample were removed using VSEARCH tool (Rognes *et al.* 2016). Each operational taxonomic unit (OTU) that accounted for <0.005% of the total set of sequences was discarded, as previously recommended (Bokulich *et al.* 2013). The taxonomy assignment was performed with NCBI blastn+ against the pre-processed SILVA123-16S database.

Statistical analyses

The bacterial data set was first rarefied at the sequencing depth of the smallest library (9385 reads) to remove the effect of sample size bias on bacterial community composition. Community structure and composition analyses were performed by processing the OTU table in the R environment (v.3.5.1) with the packages Phyloseq v.1.22.3 (McMurdie and Holmes 2013), vegan v.2.4.5 (Oksanen *et al.* 2017) for community analysis and ggplot2 v.2.2.1 for graphic representation. Shannon and Simpson indexes were calculated for the alpha diversity assessment. The Good's coverage was calculated to corroborate the adequate sampling depth.

For beta diversity, statistical dependence between samples was measured using the Spearman's rank correlation coefficient with the R package NbClust v.3.0 (Charrad *et al.* 2014). Also, principal coordinate analysis (PCoA) (Gower 1966) was used to assess the differences in bacterial communities between sites based on Bray–Curtis distance matrix (Bray and Curtis 1957). The linear discriminant analysis (LDA) effect size (LEfSe, v1.0.8.post1) method, as implemented in the online program (http://huttenhower.sph.harvard.edu/galaxy/root/index), was used with the relative abundance data to ascertain any significant differences in taxonomic abundance between the different subgroups determined using Spearman method. The threshold for the logarithmic LDA score was 3.5. Canonical correspondence analysis (CCA) was performed to determine the most significant environmental variables shaping bacterial community structure and composition. Because TOC and TN variables were redundant based on their collinearity (determined using variance inflation factor or VIF), TN was removed from the analysis.

RESULTS

Physicochemical characteristics of Fogo soils

The physicochemical characteristics of the soils developed in tiny, shallow crevices on lava flow surfaces of different ages are summarized in Table 1. It is assumed that these soils were formed by weathering of the silica-undersaturated (basanitic) rocks composing the lava flows, with additional aeolian inputs. Soil pH varied from 6.8 to 8.7, most of the samples being alkaline (71% of samples with pH ≥ 7.7). Soils collected in crevices on the

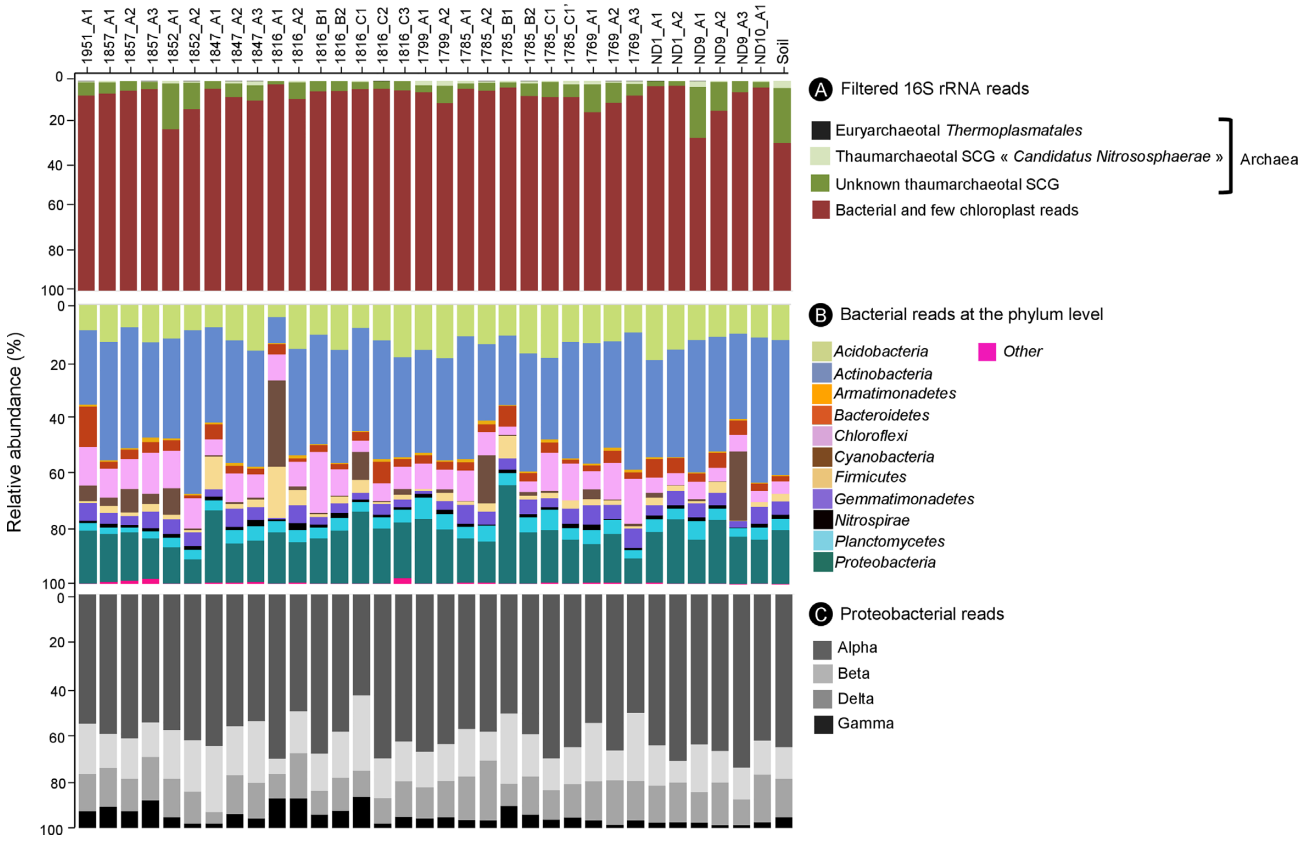


Figure 2. Relative abundances of microbial taxa in Fogo soils inferred from Illumina sequencing of the variable region V4–V5 of 16S rRNA genes. **(A)** For archaea and bacteria: the samples are ordered from the youngest lava flows on the left to the oldest ones on the right (1951 to ND- and soil); **(B)** For bacteria at the phylum level: segments composing each bar are mean number of reads in the indicated taxa normalized to the total number of reads in each library; others included phyla representing <2% of reads in one sample, i.e. *Chlorobi*, *Deinococcus*, *Elusimicrobia*, *Fibrobacteres*, *Hydrogenedentes*, JL-ETNP-Z39, *Latescibacteria*, SHA-109 and *Verrucomicrobia*; **(C)** For *Proteobacteria*: segments composing each bar are mean number of reads in the indicated class normalized to the total number of proteobacterial reads in each library.

historical lava flows were unvegetated or poorly vegetated. The oldest soils were the most colonized by low grass species and mosses ($P < 0.05$, χ^2-test; Figures S1 and S2, Supporting Information). Soil samples were also characterized by an overall low level of TOC (62% of them containing < % of TOC; values ranging from 0.21 to 3.81%) and of TN (~56% with values <0.1% and ~26% below the detection limit). Overall, the highest TOC contents were observed for soils with a higher vegetation (Kruskal–Wallis one-way ANOVA, $P < 0.05$, post hoc Dunn's test vegetated vs poorly vegetated, $P = 0.005$) and for those collected on the oldest lava flows ($P = 0.03$). Moreover, a significant positive correlation was observed between the DNA concentration and TOC (Spearman's rho = 0.779, $P = 8.96e^{-08}$). The DNA concentration was also significantly greater for the vegetated soils (Kruskal–Wallis one-way ANOVA, $P < 0.05$, post hoc Dunn's test vegetated vs poorly vegetated, $P = 0.005$).

Microbial community composition and diversity

Amplicon sequencing targeting the V4–V5 region of the 16S rRNA gene was used to characterize the bacterial and archaeal communities in soil samples. Archaeal sequences represented from 1.4 to 29.3% of total sequences across all sites (Fig. 2A; Table S3, Supporting Information) after filtering out low-quality reads and chimeras (~1.03 10^5 archaeal reads in total from all samples; 453 to 26 132 per sample). There were from 16 to 33 archaeal OTUs per sample (34 unique OTUs in total determined using Swarm parameters) that belonged to *Thaumarchaeota* (32 OTUs; ~99.7% of archaeal sequences) and *Euryarchaeota* (2 OTUs). The highest diversity and abundance were for the oldest soil sample (i.e. Soil). All thaumarchaeotal reads belonged to the Soil Crenarchaeotic Group (SCG), which made up 96.3–100% of all archaeal reads at the different sampling locations. Almost a third of SGC OTUs had 98–100% identities with the ammonia-oxidizing genus *Candidatus Nitrososphaera* (10.4% of all archaeal sequences; Fig. 2A), the remaining being unclassified SCG. The two euryarchaeotal OTUs found in this study were unclassified *Thermoplasmatales*. They were detected in only 18 samples (Fig. 2A; Table S3, Supporting Information) where they accounted from 0.02 to 3.7% of archaeal reads depending on the sample. As the archaeal diversity was low and the community composition rather similar in all samples, no further subsampling was undertaken.

Regarding bacterial communities, 9385 to 75 446 reads were obtained per sample (1.13 10^6 reads in total) after eliminating low quality-reads, chimeras, chloroplast and archaeal reads (Table S4, Supporting Information). As the number of reads in different samples differed significantly, community diversity, similarity and structure were analyzed using the data rarefied to the depth of the smallest bacterial library (9385 reads for the sample 1852_A1). Hence, after subsampling, the total number of bacterial OTUs per sample ranged from 587 to 1430 (2656 unique OTUs in total). Sampling over two years did not show influence on the community composition (Wilcoxon–Mann–Whitney test,

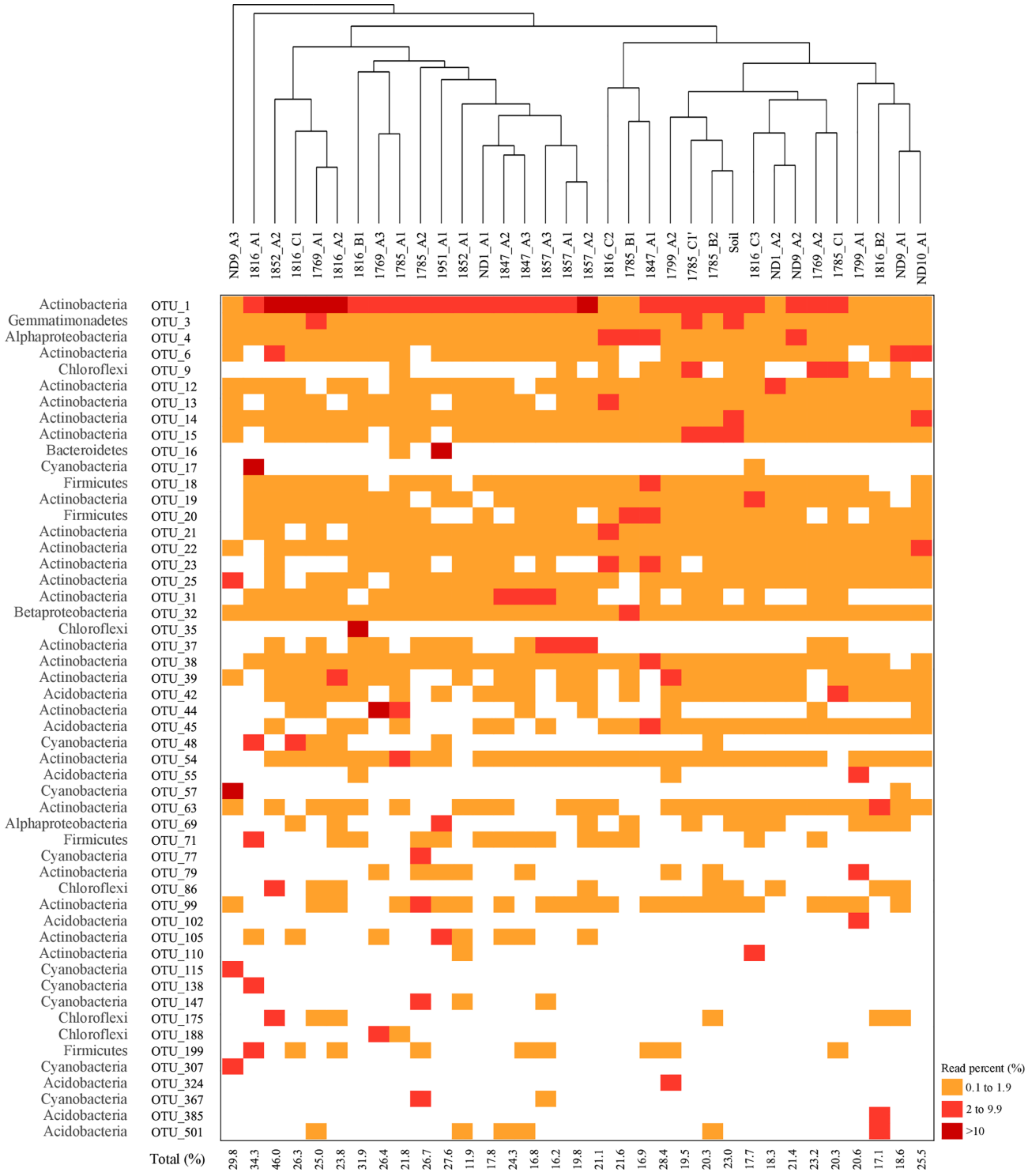


Figure 3. Heatmap of OTUs with relative abundance ≥ 2% in at least one sample collected on the Fogo lava flows. The heatmap represents the relative percentages of each bacterial OTU. The legend for the heatmap is provided in the lower right corner. White: no reads detected for the sample; orange: OTU representing from 0.1 to 1.9% of total reads for the sample; red: OTU representing from 2 to 9.9% of total reads for the sample; and dark red: OTU representing >10% of total reads for the sample.

$P = 0.29$). Moreover, the observed OTUs and the diversity estimators differed depending on the sample, the values for the Shannon's index varying from 4.61 to 6.30 and those for the Simpsons diversity index, from 0.9040 to 0.9958. Moreover, sampling completeness assessed by Good's coverage estimator for each data set gave values ranging from 95.3 to 98.3%, indicating that the vast majority of bacterial diversity was captured.

In total, 21 bacterial phyla were identified (Fig. 2B), of which 11 were present in all soils investigated. For these phyla, almost half of the reads (47.5%) were not classified at the genus level.

Overall, the soils largely contained the same phyla, but with differences in their relative abundance. The dominant phyla (>1% of each community composition) were *Actinobacteria* (9.5–58.7%), *Proteobacteria* (7.5–35.5%), *Acidobacteria* (3.9–19.9%), *Chloroflexi* (3.1–20.9%), *Bacteroidetes* (1.3–15.5%) and *Planctomycetes* (2.3–7.2%), followed by *Gemmatimonadetes* (0.6–6.8%), *Nitrospirae* (0.3–2.3%) and *Armatimonadetes* (0.2–1.4%). *Cyanobacteria* and *Firmicutes* were abundant in a few samples only, reaching 30.7 (0–30.7%) and 18.8% (0.2–18.8%), respectively, of all reads in some soil samples. Regarding the two most abundant phyla, *Actinobacteria* was dominated by the genera *Crossiella* (16.1% of actinobacterial sequences), *Solirubrobacter* (7.8%), *Rubrobacter* (4.1%), *Patulibacter* (3.8%) and *Geodermatophilus* (2.8%) while *Proteobacteria* were essentially represented by the α-subdivision (43.1–71.3% of proteobacterial reads) (Fig. 2C). For the *Cyanobacteria,* OTUs were mainly assigned to the genera *Chroococcidiopsis* (31.4% of the OTUs) and *Coleofasciculus* (formely *Microcoleus*; 6.8% of the OTUs). Finally, *Nitrospirae*, which accounted for 1.2% of all bacterial reads, was the only nitrite-oxidizing bacteria (NOB) group detected in this study.

Among the 2656 unique OTUs, only 52 showed relative abundance $\geq$ 2% from at least one sample. These OTUs comprised from 11.9 to 46% of all reads per sample (Fig. 3). Overall, the major fraction of these 52 OTUs was represented by *Actinobacteria* (40.4%) followed by *Cyanobacteria* (17.3%), *Acidobacteria* (15.4%), *Chloroflexi* (9.6%) and *Firmicutes* (7.7%) (Table 2). Most of them shared the highest similarity with uncultured bacteria often isolated from a lava tube, crater wall, lava cave wall, crust and basaltic rock. The three most abundant OTUs, detected in all soil samples, were affiliated with species of the genera *Crossiella* (*Actinobacteria*, OTU1, 5.5% of all reads) and *Microvirga* (*α-Proteobacteria*, OTU4, 1%), and the phylum *Gemmatimonadetes* (OTU3, 1.1%) (Fig. 3, Table 2).

Assessment of factors influencing bacterial community composition

Overall, the relative abundance of the dominant bacterial phyla was correlated with some environmental parameters when using Pearson's rank correlation coefficients. For instance, the relative abundance of *Proteobacteria* was negatively correlated with pH ($R = -3.63$, $P < 0.001$), whereas that of *Bacteroidetes* was positively correlated with the exposition ($R = 2.13$, $P < 0.05$). Furthermore, TOC, TN and terrain slope were also significantly correlated (positively or negatively) with the presence of several phyla (Table 3).

The patterns of bacterial community composition across all soil samples were visualized using the Spearman's rank correlation coefficient based on the pairwise Euclidean distances as well as a PCoA ordination based on the Bray–Curtis distances. The former allowed us to visualize the correlation strength between samples (Figure S3, Supporting Information) and to group them into four distinct subgroups (Figure S4, Supporting Information). These subgroups were not clearly based on the age of the lava flows, but the extreme samples, i.e. the oldest ones noted ND (prior to 1769) and the youngest one noted 1951, were clustered separately and shared the lowest correlations (Figure S3, Supporting Information). We further calculated Bray–Curtis distances for all pairs of samples and used PCoA for ordination (Fig. 4). The same tendency was observed, with no clustering patterns according to the eruptive event date and the oldest and youngest samples again clustered at opposite poles. But a clear difference in the vegetation cover was observed between these most distant samples. Moreover, the four subgroups resulting from the Spearman's rank correlation coefficient were clearly separated according to the two first axis (Fig. 4). However, because of the limited robustness of the analysis (up to 30% of explained variability), all assumptions drawn from it need to be carefully interpreted. LDA effect size was then performed to obtain further insights into the differentiation of the four identified subgroups and confirmed that each had its own indicator taxa at the class level (Fig. 5). *Anaerolineae* and *Thermomicrobia* within *Chloroflexi* as well as *Cytophagia* within *Bacteroidetes, Rubrobacteria* within *Actinobacteria* and *Cyanobacteria* were specific to the subgroup colored in red (as code used in Fig. 4). *Ktedonobacteria* within *Chloroflexi* as well as *Sphingobacteria* and *Phycisphaerae* within *Bacteroidetes* were all common to the purple subgroup while unclassified *Gammaproteobacteria* was specific to the blue subgroup and the novel class-level lineages *Thermoleophilia* within *Actinobacteria* to the green subgroup. Additionally, CCA was used to relate the soil properties (pH, C, N, altitude, slope, exposition, vegetation and moss) with the soil bacterial community composition data. The CCA explained 28% of the total variance ($P = 0.002$; 10 000 permutations), of which 8.8% was described by CCA1 and CCA2. CCA results showed that differences in microbial community composition had the strongest correlation with C content (and therefore N content) (axis 1) as well as slope (axis 2), which were found to be the only significant explanatory environmental variables ($P = 9.9\ e^{-05}$ and 0.03, respectively) (Fig. 6). Though TOC was mostly correlated with the oldest samples (those noted ND- and the sample 'Soil' located outside the lava flows), it was also correlated with some younger lava flows. The slope was correlated with samples of almost all ages.

DISCUSSION

In the recent years, the study of volcanic rock has attracted substantial interest, notably owing to their involvement in processes that can affect climate across silicate weathering reactions, for instance, allowing CO_2 sequestration in basalt (Dessert *et al.* 2003; Daval 2018). Though the microorganisms were shown to be the key players of these processes, information regarding their diversity in terrestrial volcanic deposits remains scarce and mainly focused on a few areas such as Iceland (Kelly *et al.* 2010, 2014; Olsson-Francis *et al.* 2012; Cockell, Kelly and Marteinsson 2013; Byloos *et al.* 2018) and Hawaii (King 2003; Dunfield and King 2004; Nanba, King and Dunfield 2004; Gomez-Alvarez, King and Nüsslein 2007; Weber and King 2010; Wall *et al.* 2015; Cockell *et al.* 2019); microbial composition of volcanic rock on Fogo Island was never investigated before.

Bacterial diversity and composition

Differences in lava flow composition, as well as the age of the deposits, were shown to contribute to shape the microbial communities in terrestrial volcanic rocks. For instance, several studies have demonstrated an increase in the phylogenetic diversity and richness in oldest rock samples in comparison to the youngest, on short (over a few dozens of years; Nemergut *et al.* 2007) or long (over several dozens of thousands of years; Lysnes *et al.* 2004; Tarlera *et al.* 2008; Kelly *et al.* 2010) timescales. Contrary to these findings, the bacterial communities we detected in the oldest soils studied did not specifically display an increase in their diversity and richness (Table S4, Supporting Information). Hence, the variation observed between our soil samples may be

Table 2. Phylogenetic affiliation and isolation source of dominant bacterial OTUs obtained from volcanic soil samples.

OTU[a]	Sample no. with OTU ≥2%	% of all bacterial reads	Closest sequence from NCBI nucleotide/Accession number	Similarity (%)[b]	Taxonomy	Isolation source
OTU_1	26	5.5	Uncultured bacterium/LT855029	100	Actinobacteria / Crossiella	Spain Canary Island, lava tube
OTU_4	4	1	*Microvirga* sp. strain THG-JDS1,4/KY912078	100	Alphaproteobacteria / Microvirga	Korea, soil
OTU_3	3	1.1	Uncultured bacterium clone SOY85/FJ415702	100	Gemmatimonadetes / Gemmatimonadaceae	China, black loam soil, alkaline pH
OTU_6	3	0.8	Uncultured bacterium clone ZG7041_SP6,ab1 /EU007617	100	Actinobacteria / Solirubrobacterales	China, soil
OTU_15	3	0.7	Uncultured bacterium clone 11–860/KJ013364	100	Actinobacteria / Solirubrobacterales	China, soil
OTU_9	3	0.6	Uncultured soil bacterium clone C022/AF507680	99	Chloroflexi / Ktedonobacteria	Arizona, forest soils
OTU_31	3	0.5	Uncultured bacterium clone Ovdat61e06/JF295637	99	Actinobacteria / Crossiella	Israel, arid soil
OTU_37	3	0.3	Uncultured bacterium clone A3–332/KC554838	99	Actinobacteria / Rubrobacter	China, soil of Yanshan Mountain
OTU_14	2	0.6	Uncultured bacterium clone 15/LC015790	100	Actinobacteria / Solirubrobacter	Japan, sandy soil
OTU_23	2	0.5	*Modestobacter* sp. strain KNN46–8/KY510682	99	Actinobacteria / Modestobacter	Chile, soil
OTU_39	2	0.5	Uncultured actinobacterium clone H1_2/JF681927	100	Actinobacteria / Arthrobacter	Brazil, soil
OTU_44	2	0.5	Uncultured bacterium clone MD20123c10/JN615891	99	Actinobacteria / Euzebya	Portugal, lava cave wall
OTU_20	2	0.4	*Bacillus* sp. strain RT12/MK014243	100	Firmicutes / Bacillus	China, rice endophyte
OTU_48	2	0.3	Uncultured bacterium clone YF311/KF037806	99	Cyanobacteria / Phormidium	China, soil
OTU_175	2	0.1	Uncultured bacterium clone S2_81/JQ738797	89	Chloroflexi / Roseiflexus	India, crater wall surface rocks
OTU_22	1	0.7	Uncultured bacterium clone Site2_295/JN694016	100	Actinobacteria / Solirubrobacterales	USA, soil
OTU_12	1	0.6	Actinobacteria bacterium strain 1011–14,1/MG807549	100	Actinobacteria / Mycobacterium	China, marine sponge
OTU_21	1	0.5	Uncultured bacterium clone S15/HM439478	100	Actinobacteria / Plantactinospora	USA, citrus grove soil
OTU_25	1	0.5	Uncultured bacterium clone 11–706/KC554691	100	Actinobacteria / Kineosporiaceae	China, soil of Yanshan Mountain
OTU_13	1	0.4	Blastococcus litoris strain GP-S2–8/MH128378	100	Actinobacteria / Blastococcus	South Korea, sea-tidal flat sediment
OTU_17	1	0.4	Uncultured *Pleurocapsa* sp. isolate OTU 00036/KM462585	99	Cyanobacteria / Pleurocapsa	Hawaii, air sample
OTU_18	1	0.4	Bacillus aryabhattai strain NBRC AQ3/MG966501	100	Firmicutes / Bacillus	Pakistan, plant roots
OTU_19	1	0.4	Uncultured bacterium clone C-53/JQ978963	99	Actinobacteria / Patulibacter	China, soil
OTU_32	1	0.4	Uncultured soil bacterium clone A027/JX489853	100	Betaproteobacteria / Ramlibacter	China, soil
OTU_35	1	0.4	Uncultured bacterium isolate/GU591336	86	Chloroflexi / Roseiflexus	Japan, activate sludge
OTU_38	1	0.4	Uncultured bacterium clone D3–341/KC554953	100	Actinobacteria / Geodermatophilus	China, soil of Yanshan Mountain
OTU_54	1	0.4	Uncultured bacterium clone S1_121/JQ738711	100	Actinobacteria / Gaiellales	India, crater wall surface rocks

Table 2. Continued

OTU[a]	Sample no. with OTU ≥2%	% of all bacterial reads	Closest sequence from NCBI nucleotide/Accession number	Similarity (%)[b]	Taxonomy	Isolation source
OTU_16	1	0.3	Uncultured bacterium clone SSD44_C08/JQ358355	100	Bacteroidetes / Chitinophaga	USA, grass decomposition
OTU_57	1	0.3	Uncultured cyanobacterium clone BkfYyyy800/KC463672	95	Cyanobacteria / Cyanobacteria	South Africa, soil crust
OTU_63	1	0.3	Uncultured bacterium clone 140a/KX239175	99	Actinobacteria / Solirubrobacterales	USA, soil
OTU_71	1	0.3	Uncultured bacterium clone BG1–161/JX079105	99	Firmicutes / Tumebacillus	India, contaminated agricultural soil
OTU_42	1	0.2	Uncultured Acidobacteria bacterium clone P2c_E7/GQ120661	100	Acidobacteria / Subgroup 4	Mexico, mine tailings at 2558 m altitude
OTU_45	1	0.2	Uncultured bacterium clone YJ-73/JQ769802	99	Acidobacteria / Subgroup 4	China, soil crust of copper mine
OTU_77	1	0.2	*Myxosarcina* sp. SAG 30,84/KM019956	100	Cyanobacteria / Myxosarcina	Germany, AlgaTerra information system
OTU_86	1	0.2	Uncultured bacterium clone 10D-4/DQ906857	88	Chloroflexi / Roseiflexus	Oman, subsurface soil
OTU_99	1	0.2	Uncultured bacterium clone 11–34/KC554626	99	Actinobacteria / Solirubrobacterales	China, soil of Yanshan Mountain
OTU_115	1	0.2	Oscillatoriales cyanobacterium 49 PC/MF581663	99	Cyanobacteria / Oscillatoriales	Brazil, crust
OTU_147	1	0.2	Scytonema crispum U55-MK38/HF911526	100	Cyanobacteria / Scytonema	Brazil, culture collection
OTU_199	1	0.2	Uncultured bacterium clone 26–2E/KC777223	100	Firmicutes / Bacillus	Antarctica, glacier basal ice
OTU_501	1	0.2	Uncultured bacterium clone Ovdat63g06/JF295489	99	Acidobacteria / Subgroup 4	Israel, arid soil
OTU_55	1	0.1	Uncultured bacterium clone YM-36/JQ769978	99	Acidobacteria / Subgroup 4	China, soil crust of copper mine
OTU_69	1	0.1	*Rhizobium* sp. strain UFLA04–628/MH319982	100	Alphaproteobacteria / Rhizobium	Brazil, root nodule
OTU_79	1	0.1	Uncultured bacterium clone S1_89/JQ738719	99	Actinobacteria / Gaiella	India, crater wall surface rocks
OTU_102	1	0.1	Uncultured Acidobacteriales, clone Plot29-D02/EU202813	99	Acidobacteria / Subgroup 6	Mexico, agricultural soil
OTU_105	1	0.1	Uncultured bacterium clone C11-MZ03_DNA_18 785/LT855052	99	Actinobacteria / Acidimicrobiales	Spain Canary Island, lava tube
OTU_110	1	0.1	Uncultured bacterium clone TG-98/JQ769608	99	Acidobacteria / Subgroup 10	China, oil crust of copper mine
OTU_138	1	0.1	*Scytonema* sp. 1F-PS/KT935473	99	Cyanobacteria / Scytonema	India, freshwater
OTU_188	1	0.1	Uncultured bacterium clone B4220H10/KC563036	95	Chloroflexi / Ktedonobacteria	Tibet, soil
OTU_307	1	0.1	Uncultured bacterium clone abscm03.0.651/JX255253	98	Cyanobacteria / Microcoleus	Mongolia, soil crust
OTU_324	1	0.1	Uncultured bacterium clone 120/EF667429	99	Acidobacteria / Subgroup 4	India, soil
OTU_367	1	0.1	Uncultured bacterium clone 1790–2/AY425770	94	Cyanobacteria / Cyanobacteria	Hawaii, volcanic deposit
OTU_385	1	0.1	Uncultured Acidobacteria, clone KBS_T1_R4_149 264_a1/HM062397	97	Acidobacteria / Subgroup 4	USA, topsoil

[a]Representative 16S rRNA gene sequences of the 52 OTUs with ≥2% relative abundance in at least one sample were BLAST-searched against the GenBank database (www.ncbi.nlm.nih.gov).
[b]Sequence similarity to its closest relative in the GenBank database.
OTU: Operational taxonomic unit

Table 3. Pearson's rank correlation coefficients between dominant bacterial taxa and environmental variables.

	Altitude	Slope	Exposition	TOC	TN	pH (H_2O)
Actinobacteria	0.03	0.77	-1.32	0.27	0.54	1.20
Proteobacteria	-0.34	-0.02	1.38	3.22**	2.59*	-3.63***
Acidobacteria	-0.53	0.97	1.60	1.07	1.25	-1.44
Chloroflexi	0.40	0.032	-1.16	-2.67*	-2.54*	1.78
Planctomycetes	-1.51	1.57	0.78	1.36	1.21	-0.62
Bacteroidetes	0.40	-2.31*	2.13*	-0.05	-0.54	-0.52
Gemmatimonadetes	-1.94	-2.44*	-0.58	-0.84	-0.63	-0.14
Nitrospirae	0.68	0.98	-0.16	1.05	1.73	-1.12
Armatimonadetes	-1.38	-0.80	-0.13	-3.62**	-3.90***	0.12
Cyanobacteria	0.56	-0.48	-0.21	-1.20	-1.25	1.23
Firmicutes	0.43	0.08	0.17	-0.07	-0.12	-0.85

TOC, soil total organic carbon; TN, soil total nitrogen; and ***$P < 0.001$, **$P < 0.01$; *$P < 0.05$.

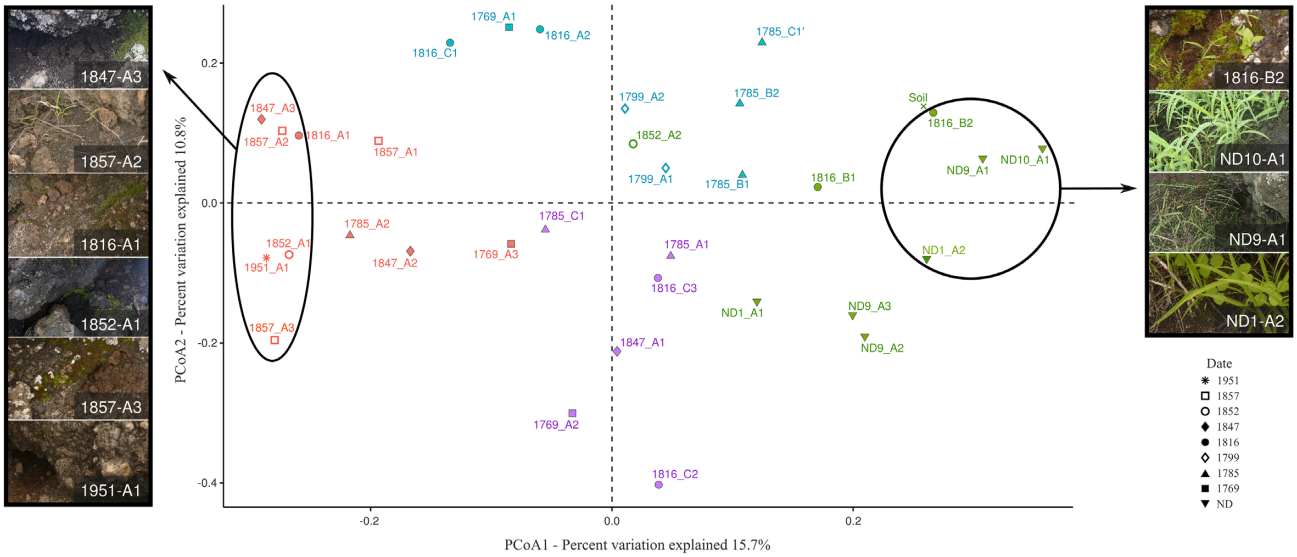


Figure 4. PCoA of the bacterial community structure using Bray–Curtis distances. Colors represent the four different clusters as revealed using Spearman's rank correlation analysis (Figures S3 and S4, Supporting Information). The percentage of variation explained by each ordination axis is indicated. Soil pictures for the most distant samples are shown.

due to differences of microenvironment in which these communities are located. Indeed, many other factors are known to influence microbial diversity such as rock/mineral composition, soil porosity, cavity surface and depth, nutrients, moisture, vegetation and wind (Choe *et al.* 2018), which may thus also have contributed to the differences observed here. Moreover, the short time period between the eruption events (~250 years for dated lavas) might be insufficient and soils on the different flows not developed enough to allow a more extensive bacterial diversity in the oldest soils, as previously reported by Byloos *et al.* (2018) in Icelandic volcanic deposits. By contrast, soil age might have influenced the succession of bacterial populations, the community composition from the youngest (i.e. sample 1951) and the oldest soils (i.e. ND- and Soil samples) being the most distant using statistical and multivariate ordination analyses. This might result from the weathering for a longer period of the oldest age samples, facilitating nutrient release and promoting the growth of different bacterial groups.

Overall, the Fogo volcanic soils were relatively similar to one another in terms of phylogenetic composition, at least at the phylum scale. They all hosted in relative high abundance *Actinobacteria*, *Proteobacteria*, *Acidobacteria*, *Chloroflexi*, *Bacteroidetes* and *Planctomycetes*, similarly to what was recorded in other soil surveys (Janssen 2006; Lauber *et al.* 2009; Nacke *et al.* 2011; Fierer *et al.* 2012; Prober *et al.* 2015). Among them, *Actinobacteria* and *Proteobacteria* were demonstrated to be common representatives of early stages during succession and key inhabitants of basaltic substrates (Cockell, Kelly and Marteinsson 2013; Antony *et al.* 2014; Rughöft *et al.* 2016; Choe *et al.* 2018) because of their important role in rock weathering, decomposition of organic matter and nutrient recycling. Moreover, some members of these phyla have the capacity to not only resist desiccation and temperature fluctuations but also form spores, hence improving their dispersal. They can also develop a filamentous growth, allowing them to invade the rock interstices. All these traits might represent advantages in Fogo soils, as in other volcanic environments (Okoro *et al.* 2009; Cockell *et al.* 2011; Neilson *et al.* 2012; Rughöft *et al.* 2016). Regarding the *Proteobacteria* phylum, the high abundance of its members in early bacterial communities was explained by a wide variety of metabolic traits conferring benefits in environments with limited nutrient resources (i.e. phototrophy, chemolithotrophy and photoheterotrophy). Like various preceding studies on volcanic substrates, *Alphaproteobacteria* was the most abundant class in this phylum and was dominated by the *Rhizobiales*, *Rhodospirillales* and *Sphingomonadales* orders. They are known habitants of soils (Fierer *et al.* 2012)

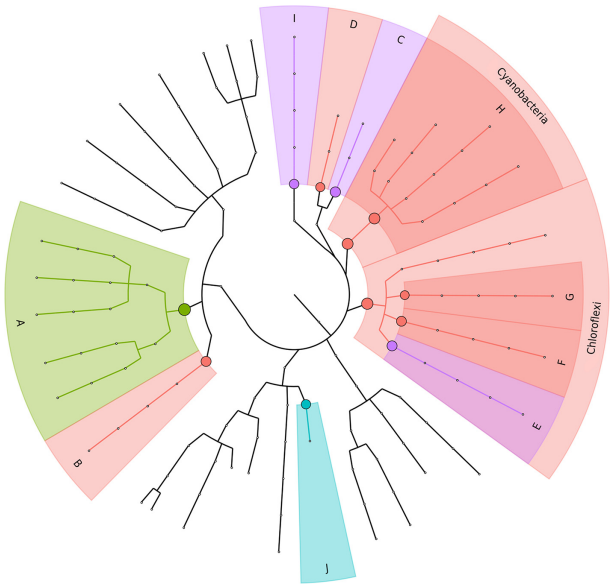


Figure 5. Linear discriminant analysis (LDA) effect size (LEfSe) taxonomic cladogram comparing bacterial communities from the four subgroups determined using Spearman's rank correlation coefficients. Nodes from inside to outside circles represent the bacterial taxon from phylum to genus level, respectively. Significant discriminant taxa to a specific subgroup and its branch areas are highlighted with green, purple, red and blue color (as indicated in Fig. 5). The threshold for the logarithmic LDA score was 3.5. In red: *Cytophagia* within *Bacteroidetes* (D); *Thermomicrobia* (F) and *Anaerolineae* (G) within *Chloroflexi*; *Cyanobacteria* (H); in purple: *Ktedonobacteria* within *Chloroflexi* (E), *Sphingobacteriia* (C) and *Phycisphaerae* (I) within *Bacteroidetes*, (D); in green: *Thermoleophilia within Actinobacteria* (A); in blue: unclassified *Gammaproteobacteria* (J).

but also of rocks at least slightly weathered (Gomez-Alvarez, King and Nüsslein 2007; Cockell *et al.* 2019), such as those found on Fogo surface. Although major soil phyla were detected in Fogo soil samples, surprisingly they lack a significant amount of *Verrucomicrobia*, while they were described as important in many soil bacterial communities (Delgado-Baquerizo *et al.* 2018) and suspected to contribute to weathering in basalt environment (Byloos *et al.* 2018). This underestimation must be taken with care because it might be due to primer bias, many PCR primers being shown to exclude verrucomicrobial 16S rRNA genes (Bergmann *et al.* 2011). Finally, in accordance with previous findings in fumarolic sites (Cockell *et al.* 2019), the prevalence of different phyla in some basaltic sites and not in others could highlight the important role of geographical location or other physicochemical factors in shaping microbial communities of volcanic materials.

Phototrophic bacteria are known to be some of the first colonists of post-volcanic environments (Uroz *et al.* 2009). They include not only oxygenic-dependent species, i.e. *Cyanobacteria*, but also anoxygenic ones (anoxygenic phototrophic bacteria or APBs) that were identified in six bacterial phyla scattered across the domain (Hanada 2016; Ward *et al.* 2018). *Cyanobacteria* were recorded in all samples, representing a significant proportion in few of them (up to 31% of the analyzed reads). This is in accordance with the previous demonstration of their ubiquity in volcanic rocks (Cockell *et al.* 2009) and their important role in weathering and initial soil development (Belnap, Büdel and Lange 2001; Liu *et al.* 2017). In this study, this phylum was dominated by the genera *Phormidium*, *Coleofasciculus*, *Leptolyngbya* and *Chroococcidiopsis* that is consistent with the fact they are widespread rock dwellers and fast-growing *Cyanobacteria* (Olsson-Francis *et al.* 2012). They are also thought to be responsible for the pH increase during the element extraction from rocks. Likewise, our data suggest the presence of diverse APBs in Fogo soils mainly belonging to *Proteobacteria* (in the α- and β-classes), *Chloroflexi* (in the filamentous phototrophic *Chloroflexia* class) and *Gemmatimonadetes*. However, since the two later phyla have only a handful of cultured species, the biochemical potential of most of their members remains uncertain. So, further studies, based on the detection of specific photosynthetic genes or isolation methods, are needed in order to fully reveal the roles of these microorganisms in the investigated soils.

Furthermore, similarly to what was demonstrated for other sites, allochthonous inputs carried by the wind likely participate in soil development and microbial community shaping on Fogo lava flows. This is evidenced at the geological level by the unexpected presence of quartz grains in soils and at the surface of lava flows that are devoid of such minerals, as for all the rocks of the island (Figure S5, Supporting Information). Furthermore, here, we show that the specific taxonomic composition of the microbial communities might be geographically influenced. Indeed, the most abundant OTUs identified in our samples (Table 2) were most similar to sequences from the countries swept by the north-east trade winds. Therefore, this might suggest that atmospheric deposition of microorganisms could be important source of microbial colonizers on Fogo, also highlighting their high dispersal efficiency and viability as shown in previous studies (Smith *et al.* 2013; Hu *et al.* 2018; Tang *et al.* 2018). Likewise, the majority of them were associated with communities from volcanic (lava tube, lava cave, crater wall) and extreme (rock, crust, alkaline soil, arid soil) environments, reflecting the selection of specialized core populations well adapted to such environments.

Nitrogen cycling in Fogo soils

Based on archaeal 16S rRNA gene sequence analysis, Fogo volcanic soils were clearly dominated by *Thaumarchaeota*, mostly the SCG (Group I.1b), which is in accordance with previous studies that identified this group as the main archaeal lineage for the soil environment (Auguet, Barberan and Casamayor 2010; Höfferle *et al.* 2010; Bates *et al.* 2011; Tripathi *et al.* 2015; Liu *et al.* 2019) and crater-wall basalts (Antony *et al.* 2014). The SCG is assumed to contribute significantly to ammonia oxidation (AO) in soils along a broad pH range (Gubry-Rangin *et al.* 2011; Chroňáková *et al.* 2015; Lu, Seuradge and Neufeld 2017), with specialized acidophilic (mostly the genus *Nitrosotalea*) and alkalinophilic (mostly the genus *Nitrososphaera*) lineages (Tourna *et al.* 2011; Gubry-Rangin *et al.* 2015; Oton *et al.* 2016). This agrees relatively well with the affiliation of ~11% of our SCG reads with the alkalinophilic AOA (ammonia-oxidizing archaea) genus Candidatus *Nitrososphaera* and the alkaline properties of the Fogo soils. Moreover, the fact that the majority of the detected SCG reads remained unclassified at the genus level could be explained by the current low number of representative cultured and environmental genomes for this group in public databases. Conversely, *Thermoplasmatales*, belonging to the phylum *Euryarchaeota*, were present in very low abundance in the investigated soils, which is possibly due to their preference toward acidic soils (Angelov and Liebl 2006; Hu *et al.* 2013).

Here, the ammonia-oxidizing communities were clearly dominated by *Archaea* across all samples. Indeed, ammonia-oxidizing bacteria (AOB; *Nitrosomonadaceae* family for β-*Proteobacteria* and *Nitrosococcus* genus for g-*Proteobacteria*) constituted negligible fractions of the reads in all Fogo data sets

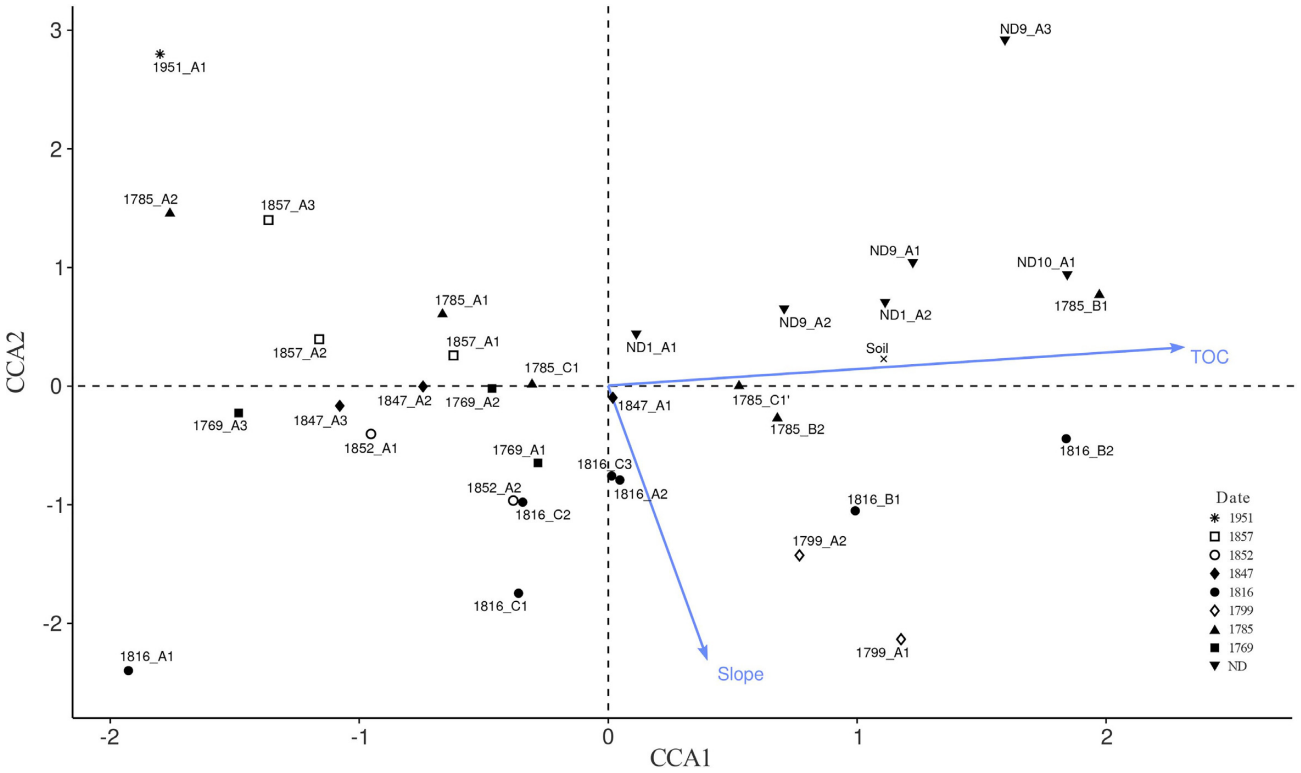


Figure 6. CCA ordination plot for the first two principal dimensions depicting the relationship between the distribution of bacterial clusters, as represented by 16S rRNA gene sequence data, and the environmental parameters used in this study. Correlations between environmental variables and CCA axes are represented by the length and angle of arrows (environmental factor vectors). Abbreviations are the same as in Table 1. Only statistically significant environmental variables are represented.

(0.46 and 0.02% of all bacterial reads, respectively). This is in agreement with the general trend observed in a wide range of soils examined so far (He *et al.* 2007; Di *et al.* 2010; Stopnisek *et al.* 2010; Alves *et al.* 2013), including the volcanic soils (Daebeler *et al.* 2014; Hernandez *et al.* 2015; Rughöft *et al.* 2016) and might be explained by the low substrate content (e.g. TN and TOC) that is thought to benefit AOA over AOB due to a higher substrate affinity (Bates *et al.* 2011; Daebeler *et al.* 2014; Oton *et al.* 2016). This is also consistent with the fact that AOA were shown to be early colonizers of young volcanic soils, where they might play an important role not only in nitrification but also in soil formation and function (Hernandez *et al.* 2015).

Besides the ammonia oxidizers, which catalyze the oxidation of ammonia to hydroxylamine and nitrite, nitrification also comprises NOBs in a last step for transformation of nitrite in nitrate. In our study, only NOBs belonging to the genus *Nitrospira* (phylum *Nitrospirae*) were detected, which agrees with the higher ubiquity and phylogenetic diversity for this genus compared to *Nitrobacter*-NOB type reported in other studies (Xia *et al.* 2011; Pester *et al.* 2014). Moreover, this genus was proposed to be a k-strategist NOB and therefore adapted to environments containing low substrate concentrations (Kim and Kim 2006) as the ones found on Fogo, as opposed to *Nitrobacter* that was proposed to be r-strategist (Wertz, Leigh and Grayston 2012). Our results are also congruent with other studies demonstrating that the dominant form of nitrifying network under low substrate concentrations is AOA / *Nitrospira* (Daebeler *et al.* 2014; Stempfhuber *et al.* 2016).

Moreover, some *Nitrospira* sp. were recently shown to perform complete nitrification (complete ammonia oxidizer or comammox), which has challenged the view of this metabolism as a unique two-step process performed by two distinct functional guilds (Daims *et al.* 2015; van Kessel *et al.* 2015; Shi *et al.* 2018). All comammox belong to the *Nitrospira* lineage II and are indistinguishable from the canonical NOB *Nitrospira*, using 16S rRNA sequences. These comammox can be divided into two monophyletic sister clades based on the sequence of the ammonium monooxygenase subunit A (AmoA), enzyme not found in canonical *Nitrospira* and whose sequence differs from that of canonical ammonia oxidizers. Though most *Nitrospira* identified in Fogo soils are most probably related to canonical NOB, we cannot exclude that some of them could belong to comammox as the latter were detected in a wide range of environmental samples (Pjevac *et al.* 2017; Koch, van Kessel and Lücker 2019) and stably co-existed with AOA (Bartelme, McLellan and Newton 2017). Hence, further studies are needed, notably based on the *amoA* gene marker, to gain insight into the potential contribution of *Nitrospira* to AO in the Fogo Island soils.

CONCLUSION

This survey provides the first data on the archaeal and bacterial diversity colonizing the soils filling microsites on Fogo basaltic lava flows of different ages. Although these communities resembled those commonly found in soils, our work indicates the possibility of exogenous sources of bacteria as important colonizers. Hence, atmospheric deposition (i.e. through aeolian dust, trade winds and rain) may play a crucial role in the development of the recently formed volcanic ecosystems by not only providing source of energy and nutrients but also inoculating these environments with microbial populations well adapted to such harsh conditions (aridity, limited nutrient resources and rock). Moreover, the dominance of AOA compared to AOB in these ecosystems highlights the role of *Archaea* in the volcanic soil N cycle, most likely in combination with nitrite oxidation by *Nitrospira*-related NOB.

SUPPLEMENTARY DATA

Supplementary data are available at FEMSEC online.

FUNDING

GH was supported by the MESR (Ministère de l'Enseignement Supérieur et de la Recherche). Fieldwork on Fogo Island was mostly supported by subsidy from the project PEPS 2015, CNRS, Clermont University Auvergne (UCA), France. This is Laboratory of Excellence ClerVolc contribution number 379.

ACKNOWLEDGEMENTS

We thank Cécile Lepère for her comments and for revising the English of the paper.

DATA AVAILABILITY

Metabarcoding sequence reads generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJEB30019.

Conflicts of Interest. None declared.

REFERENCES

Alves RJE, Wanek W, Zappe A *et al.* Nitrification rates in Arctic soils are associated with functionally distinct populations of ammonia-oxidizing archaea. *ISME J* 2013;**7**:1620–31.

Angelov A, Liebl W. Insights into extreme thermoacidophily based on genome analysis of *Picrophilus torridus* and other thermoacidophilic Archaea. *J Biotechnol* 2006;**126**:3–10.

Antony CP, Shimpi GG, Cockell CS *et al.* Molecular characterization of prokaryotic communities associated with Lonar Crater basalts. *Geomicrobiol J* 2014;**31**:519–28.

Auguet J-C, Barberan A, Casamayor EO. Global ecological patterns in uncultured Archaea. *ISME J* 2010;**4**:182.

Barone R, Hering J. Recent bird records from Fogo, Cape Verde Islands. *Bull Afr Bird Club* 2010;**17**:71–8.

Bartelme RP, McLellan SL, Newton RJ. Freshwater recirculating aquaculture system operations drive biofilter bacterial community shifts around a stable nitrifying consortium of ammonia-oxidizing *Archaea* and comammox *Nitrospira*. *Front Microbiol* 2017;**8**:101.

Bates ST, Berg-Lyons D, Caporaso JG *et al.* Examining the global distribution of dominant archaeal populations in soil. *ISME J* 2011;**5**:908–17.

Belnap J, Büdel B, Lange OL. Biological soil crusts: characteristics and distribution. In: Belnap J, Lange OL (eds). *Biological Soil Crusts: Structure, Function, and Management*. Vol. **150**. Berlin, Heidelberg: Springer, 2001, 3–30.

Bergmann GT, Bates ST, Eilers KG *et al.* The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biol Biochem* 2011;**43**:1450–5.

Bokulich NA, Subramanian S, Faith JJ *et al.* Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* 2013;**10**:57–9.

Bray JR, Curtis JT. An ordination of the upland forest communities of Southern Wisconsin. *Ecol Monogr* 1957;**27**:325–49.

Brochmann C ed. *The Endemic Vascular Plants of the Cape Verde Islands, W. Africa*. Oslo: Botanical Garden and Museum, University of Oslo, 1997.

Byloos B, Monsieurs P, Mysara M *et al.* Characterization of the bacterial communities on recent Icelandic volcanic deposits of different ages. *BMC Microbiol* 2018;**18**:122.

Charrad M, Ghazzali N, Boiteau V *et al.* An R package for determining the relevant number of clusters in a data set. *J Stat Softw* 2014,**61**:1–36.

Choe Y-H, Kim M, Woo J *et al.* Comparing rock-inhabiting microbial communities in different rock types from a High Arctic polar desert. *FEMS Microbiol Ecol* 2018;**94**:fiy070.

Chroňáková A, Schloter-Hai B, Radl V *et al.* Response of archaeal and bacterial soil communities to changes associated with outdoor cattle overwintering. *PLoS One* 2015;**10**:e0135627.

Chuvochina MS, Marie D, Chevaillier S *et al.* Community variability of bacteria in alpine snow (Mont Blanc) containing Saharan dust deposition and their snow colonisation potential. *Microbes Environ* 2011;**26**:237–47.

Cockell CS, Harrison JP, Stevens AH *et al.* A low-diversity microbiota inhabits extreme terrestrial basaltic terrains and their fumaroles: implications for the exploration of Mars. *Astrobiology* 2019;**19**:284–99.

Cockell CS, Kelly LC, Marteinsson V. *Actinobacteria*—an ancient phylum active in volcanic rock weathering. *Geomicrobiol J* 2013;**30**:706–20.

Cockell CS, Kelly LC, Summers S *et al.* Following the kinetics: iron-oxidizing microbial mats in cold icelandic volcanic habitats and their rock-associated carbonaceous signature. *Astrobiology* 2011;**11**:679–94.

Cockell CS, Olsson K, Knowles F *et al.* Bacteria in weathered basaltic glass, Iceland. *Geomicrobiol J* 2009;**26**:491–507.

Daebeler A, Bodelier PL, Yan Z *et al.* Interactions between Thaumarchaea, Nitrospira and methanotrophs modulate autotrophic nitrification in volcanic grassland soil. *ISME J* 2014;**8**:2397–410.

Daims H, Lebedeva EV, Pjevac P *et al.* Complete nitrification by Nitrospira bacteria. *Nature* 2015;**528**:504–9.

Daval D. Carbon dioxide sequestration through silicate degradation and carbon mineralisation: promises and uncertainties. *npj Mater Degrad* 2018;**2**:11.

DeLeon-Rodriguez N, Lathem TL, Rodriguez-R LM *et al.* Microbiome of the upper troposphere: species composition and prevalence, effects of tropical storms, and atmospheric implications. *Proc Natl Acad Sci USA* 2013;**110**:2575–80.

Delgado-Baquerizo M, Reith F, Dennis PG *et al.* Ecological drivers of soil microbial diversity and soil biological networks in the Southern Hemisphere. *Ecology* 2018;**99**:583–96.

Dessert C, Dupré B, Gaillardet J *et al.* Basalt weathering laws and the impact of basalt weathering on the global carbon cycle. *Chem Geol* 2003;**202**:257–73.

Di HJ, Cameron KC, Shen J-P *et al.* Ammonia-oxidizing bacteria and archaea grow under contrasting soil nitrogen conditions: ammonia-oxidizing bacteria and archaea. *FEMS Microbiol Ecol* 2010;**72**:386–94.

Doucelance R, Escrig S, Moreira M *et al.* Pb-Sr-He isotope and trace element geochemistry of the Cape Verde Archipelago. *Geochim Cosmochim Acta* 2003;**67**:3717–33.

Dunfield KE, King GM. Molecular analysis of carbon monoxide-oxidizing bacteria associated with recent Hawaiian volcanic deposits. *Appl Environ Microbiol* 2004;**70**:4242–8.

Escudié F, Auer L, Bernard M *et al.* FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics* 2018;**34**:1287–94.

Fierer N, Leff JW, Adams BJ *et al.* Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci USA* 2012;**109**:21390–5.

Foeken JPT, Day S, Stuart FM. Cosmogenic 3He exposure dating of the Quaternary basalts from Fogo, Cape Verdes: implications for rift zone and magmatic reorganisation. *Quat Geochronol* 2009;**4**:37–49.

Freeman KR, Pescador MY, Reed SC *et al.* Soil CO_2 flux and photoautotrophic community composition in high-elevation, 'barren' soil. *Environ Microbiol* 2009;**11**:674–86.

Fujimura R, Kim S-W, Sato Y *et al.* Unique pioneer microbial communities exposed to volcanic sulfur dioxide. *Sci Rep* 2016;**6**:19687.

Fujimura R, Sato Y, Nishizawa T *et al.* Analysis of early bacterial communities on volcanic deposits on the Island of Miyake (Miyake-jima), Japan: a 6-year study at a fixed site. *Microbes Environ* 2012;**27**:19–29.

Gerlach DC, Cliff RA, Davies GR *et al.* Magma sources of the Cape Verdes archipelago: isotopic and trace element constraints. *Geochim Cosmochim Acta* 1988;**52**:2979–92.

Gomez-Alvarez V, King GM, Nüsslein K. Comparative bacterial diversity in recent Hawaiian volcanic deposits of different ages: microbial diversity of four Hawaiian volcanic deposits. *FEMS Microbiol Ecol* 2007;**60**:60–73.

Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 1966;**53**:325–38.

Gubry-Rangin C, Hai B, Quince C *et al.* Niche specialization of terrestrial archaeal ammonia oxidizers. *Proc Natl Acad Sci USA* 2011;**108**:21206–11.

Gubry-Rangin C, Kratsch C, Williams TA *et al.* Coupling of diversification and pH adaptation during the evolution of terrestrial Thaumarchaeota. *Proc Natl Acad Sci USA* 2015;**112**:9370–5.

Hanada S. Anoxygenic photosynthesis—a photochemical reaction that does not contribute to oxygen reproduction. *Microbes Environ* 2016;**31**:1–3.

He J, Shen J, Zhang L *et al.* Quantitative analyses of the abundance and composition of ammonia-oxidizing bacteria and ammonia-oxidizing archaea of a Chinese upland red soil under long-term fertilization practices. *Environ Microbiol* 2007;**9**:3152.

Hernandez ME, Beck DAC, Lidstrom ME *et al.* Oxygen availability is a major factor in determining the composition of microbial communities involved in methane oxidation. *PeerJ* 2015;**3**:e801.

Hildner E, Klügel A, Hauff F. Magma storage and ascent during the 1995 eruption of Fogo, Cape Verde Archipelago. *Contrib Mineral Petrol* 2011;**162**:751–72.

Hilley GE, Porder S. A framework for predicting global silicate weathering and CO_2 drawdown rates over geologic timescales. *Proc Natl Acad Sci USA* 2008;**105**:16855–9.

Hu H-W, Zhang L-M, Yuan C-L *et al.* Contrasting Euryarchaeota communities between upland and paddy soils exhibited similar pH-impacted biogeographic patterns. *Soil Biol Biochem* 2013;**64**:18–27.

Hu W, Niu H, Murata K *et al.* Bacteria in atmospheric waters: detection, characteristics and implications. *Atmos Environ* 2018;**179**:201–21.

Höfferle Š, Nicol GW, Pal L *et al.* Ammonium supply rate influences archaeal and bacterial ammonia oxidizers in a wetland soil vertical profile: selection of ammonia oxidizers in wetland soil. *FEMS Microbiol Ecol* 2010;**74**:302–15.

Janssen PH. Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl Environ Microbiol* 2006;**72**:1719–28.

Kelly LC, Cockell CS, Herrera-Belaroussi A *et al.* Bacterial diversity of terrestrial crystalline volcanic rocks, Iceland. *Microb Ecol* 2011;**62**:69–79.

Kelly LC, Cockell CS, Piceno YM *et al.* Bacterial diversity of weathered terrestrial icelandic volcanic glasses. *Microb Ecol* 2010;**60**:740–52.

Kelly LC, Cockell CS, Thorsteinsson T *et al.* Pioneer microbial communities of the Fimmvörðuháls Lava Flow, Eyjafjallajökull, Iceland. *Microb Ecol* 2014;**68**:504–18.

Kerfahi D, Tateno R, Takahashi K *et al.* Development of soil bacterial communities in volcanic ash microcosms in a range of climates. *Microb Ecol* 2017;**73**:775–90.

Kim D-J, Kim S-H. Effect of nitrite concentration on the distribution and competition of nitrite-oxidizing bacteria in nitratation reactor systems and their kinetic characteristics. *Water Res* 2006;**40**:887–94.

King GM. Contributions of atmospheric CO and hydrogen uptake to microbial dynamics on recent Hawaiian volcanic deposits. *Appl Environ Microbiol* 2003;**69**:4067–75.

Koch H, van Kessel MAHJ, Lücker S. Complete nitrification: insights into the ecophysiology of comammox Nitrospira. *Appl Microbiol Biotechnol* 2019;**103**:177–89.

Larrue S, Paris R, Etienne S. The use of vascular plant densities to estimate the age of undated lava flows in semi-arid areas of Fogo Island (Cape Verde, Atlantic Ocean). *J Arid Environ* 2020;**173**:104042.

Lauber CL, Hamady M, Knight R *et al.* Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* 2009;**75**:5111–20.

Liu J, Yu Z, Yao Q *et al.* Biogeographic distribution patterns of the archaeal communities across the black soil zone of Northeast China. *Front Microbiol* 2019;**10**:23.

Liu L, Liu Y, Zhang P *et al.* Development of bacterial communities in biological soil crusts along a revegetation chronosequence in the Tengger Desert, northwest China. *Biogeosciences* 2017;**14**:3801–14.

Lu X, Seuradge BJ, Neufeld JD. Biogeography of soil Thaumarchaeota in relation to soil depth and land usage. *FEMS Microbiol Ecol* 2017;**93**:fiw246.

Lysnes K, Thorseth IH, Steinsbu BO *et al.* Microbial community diversity in seafloor basalt from the Arctic spreading ridges. *FEMS Microbiol Ecol* 2004;**50**:213–30.

Mahé F, Rognes T, Quince C *et al.* Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2014;**2**:e593.

Marques R, Prudêncio MI, Abreu MI *et al.* Chemical characterization of vines grown in incipient volcanic soilsof Fogo Island. *Environ Monit Assess* 2019;**3**:128.

Marques R, Prudêncio MI, do Freitas MC *et al.* Chemical element accumulation in tree bark grown in volcanic soils of Cape Verde—a first biomonitoring of Fogo Island. *Environ Sci Pollut Res* 2017;**24**:11978–90.

Marques R, Prudêncio MI, Waerenborgh JC *et al.* Origin of reddening in a paleosol buried by lava flows in Fogo Island (Cape Verde). *J Afr Earth Sci* 2014;**96**:60–70.

McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 2013;**8**:e61217.

Nacke H, Thürmer A, Wollherr A *et al.* Pyrosequencing-based assessment of bacterial community structure along different management types in German forest and grassland soils. *PLoS ONE* 2011;**6**:e17000.

Nanba K, King GM, Dunfield K. Analysis of facultative lithotroph distribution and diversity on volcanic deposits by use of

the large subunit of ribulose 1,5-bisphosphate carboxylase/oxygenase. *Appl Environ Microbiol* 2004;**70**:2245–53.
Neilson JW, Quade J, Ortiz M *et al*. Life at the hyperarid margin: novel bacterial diversity in arid soils of the Atacama Desert, Chile. *Extremophiles* 2012;**16**:553–66.
Nemergut DR, Anderson SP, Cleveland CC *et al*. Microbial community succession in an unvegetated, recently deglaciated soil. *Microb Ecol* 2007;**53**:110–22.
Okoro CK, Brown R, Jones AL *et al*. Diversity of culturable actinomycetes in hyper-arid soils of the Atacama Desert, Chile. *Antonie Van Leeuwenhoek* 2009;**95**:121–33.
Oksanen J, Blanchet FG, Friendly M *et al*. vegan: Community Ecology Package. https://CRAN.R-project.org/package=vegan, 2017, (January 2017, date last accessed).
Olehowski C, Naumann S, Fischer D *et al*. Geo-ecological spatial pattern analysis of the island of Fogo (Cape Verde). *Glob Planet Change* 2008;**64**:188–97.
Olsson-Francis K, Simpson AE, Wolff-Boenisch D *et al*. The effect of rock composition on cyanobacterial weathering of crystalline basalt and rhyolite. *Geobiology* 2012;**10**:434–44.
Oton EV, Quince C, Nicol GW *et al*. Phylogenetic congruence and ecological coherence in terrestrial Thaumarchaeota. *ISME J* 2016;**10**:85–96.
Paris R, Giachetti T, Chevalier J *et al*. Tsunami deposits in Santiago Island (Cape Verde archipelago) as possible evidence of a massive flank failure of Fogos volcano. *Sediment Geol* 2011;**239**:129–45.
Pester M, Maixner F, Berry D *et al*. *NxrB* encoding the beta subunit of nitrite oxidoreductase as functional and phylogenetic marker for nitrite-oxidizing *Nitrospira*: functional and phylogenetic marker for *Nitrospira*. *Environ Microbiol* 2014;**16**:3055–71.
Pjevac P, Schauberger C, Poghosyan L *et al*. AmoA-targeted polymerase chain reaction primers for the specific detection and quantification of comammox Nitrospira in the environment. *Front Microbiol* 2017;**8**:1508.
Prober SM, Leff JW, Bates ST *et al*. Plant diversity predicts beta but not alpha diversity of soil microbes across grasslands worldwide. *Ecol Lett* 2015;**18**:85–95.
Ramalho RS, Winckler G, Madeira J *et al*. Hazard potential of volcanic flank collapses raised by new megatsunami evidence. *Sci Adv* 2015;**1**:e1500456.
Rime T, Hartmann M, Frey B. Potential sources of microbial colonizers in an initial soil ecosystem after retreat of an alpine glacier. *ISME J* 2016;**10**:1625–41.
Rognes T, Flouri T, Nichols B *et al*. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 2016;**4**:e2584.
Romeiras MM, Monteiro F, Duarte MC *et al*. Patterns of genetic diversity in three plant lineages endemic to the Cape Verde Islands. *AoB Plants* 2015;**7**:plv051.
Rughöft S, Herrmann M, Lazar CS *et al*. Community composition and abundance of bacterial, archaeal and nitrifying populations in savanna soils on contrasting bedrock material in Kruger National Park, South Africa. *Front Microbiol* 2016;**7**:1638.
Sato Y, Hosokawa K, Fujimura R *et al*. Nitrogenase activity (acetylene reduction) of an iron-oxidizing Leptospirillum strain cultured as a pioneer microbe from a recent volcanic deposit on Miyake-Jima, Japan. *Microbes Environ* 2009;**24**:291–6.
Sato Y, Nishihara H, Yoshida M *et al*. Occurrence of hydrogen-oxidizing *Ralstonia* species as primary microorganisms in the Mt. Pinatubo volcanic mudflow deposits. *Soil Sci Plant Nutr* 2004;**50**:855–61.
Shi X, Hu H-W, Wang J *et al*. Niche separation of comammox Nitrospira and canonical ammonia oxidizers in an acidic subtropical forest soil under long-term nitrogen deposition. *Soil Biol Biochem* 2018;**126**:114–22.
Smith DJ, Timonen HJ, Jaffe DA *et al*. Intercontinental dispersal of bacteria and archaea by transpacific winds. *Appl Environ Microbiol* 2013;**79**:1134–9.
Smolikowski B, Puig H, Roose E. Influence of soil protection techniques on runoff, erosion and plant production on semi-arid hillsides of Cabo Verde. *Agric Ecosyst Environ* 2001;**87**:67–80.
Stempfhuber B, Richter-Heitmann T, Regan KM *et al*. Spatial interaction of archaeal ammonia-oxidizers and nitrite-oxidizing bacteria in an unfertilized grassland soil. *Front Microbiol* 2016;**6**:1567.
Stopnisek N, Gubry-Rangin C, Hofferle S *et al*. Thaumarchaeal ammonia oxidation in an acidic forest peat soil is not influenced by ammonium amendment. *Appl Environ Microbiol* 2010;**76**:7626–34.
Tang K, Huang Z, Huang J *et al*. Characterization of atmospheric bioaerosols along the transport pathway of Asian dust during the Dust-Bioaerosol 2016 Campaign. *Atmospheric Chem Phys* 2018;**18**:7131–48.
Tarlera S, Jangid K, Ivester AH *et al*. Microbial community succession and bacterial diversity in soils during 77000 years of ecosystem development: microbial community succession in soils. *FEMS Microbiol Ecol* 2008;**64**:129–40.
Tourna M, Stieglmeier M, Spang A *et al*. Nitrososphaera viennensis, an ammonia oxidizing archaeon from soil. *Proc Natl Acad Sci USA* 2011;**108**:8420–5.
Trias R, Ménez B, le Campion P *et al*. High reactivity of deep biota under anthropogenic CO_2 injection into basalt. *Nat Commun* 2017;**8**:1063.
Tripathi BM, Kim M, Tateno R *et al*. Soil pH and biome are both key determinants of soil archaeal community structure. *Soil Biol Biochem* 2015;**88**:1–8.
Uroz S, Calvaruso C, Turpault M-P *et al*. Mineral weathering by bacteria: ecology, actors and mechanisms. *Trends Microbiol* 2009;**17**:378–87.
van Kessel MAHJ, Speth DR, Albertsen M *et al*. Complete nitrification by a single microorganism. *Nature* 2015;**528**:555–9.
Waldrop MP, Zak DR, Sinsabaugh RL *et al*. Nitrogen deposition modifies soil carbon storage through changes in microbial enzymatic activity. *Ecol Appl* 2004;**14**:1172–7.
Wall K, Cornell J, Bizzoco RW *et al*. Biodiversity hot spot on a hot spot: novel extremophile diversity in Hawaiian fumaroles. *MicrobiologyOpen* 2015;**4**:267–81.
Ward LM, Hemp J, Shih PM *et al*. Evolution of phototrophy in the Chloroflexi phylum driven by horizontal gene transfer. *Front Microbiol* 2018;**9**:260.
Weber CF, King GM. Distribution and diversity of carbon monoxide-oxidizing bacteria and bulk bacterial communities across a succession gradient on a Hawaiian volcanic deposit: CO oxidizer diversity across a succession gradient. *Environ Microbiol* 2010;**12**:1855–67.
Wertz S, Leigh AKK, Grayston SJ. Effects of long-term fertilization of forest soils on potential nitrification and on the abundance and community structure of ammonia oxidizers and nitrite oxidizers. *FEMS Microbiol Ecol* 2012;**79**:142–54.
Womack AM, Bohannan BJM, Green JL. Biodiversity and biogeography of the atmosphere. *Philos Trans R Soc B Biol Sci* 2010;**365**:3645–53.
Xia W, Zhang C, Zeng X *et al*. Autotrophic growth of nitrifying community in an agricultural soil. *ISME J* 2011;**5**:1226–36.

Article 5 / Co-autrice : Carles *et al.*, 2019

Carles, L., **Gardon, H.**, **Joseph, L.**, **Sanchís, J.**, **Farré, M.** and **Artigas, J.** (2019) 'Meta-analysis of glyphosate contamination in surface waters and dissipation by biofilms', *Environment International*, 124. doi: 10.1016/j.envint.2018.12.064.

Environment International 124 (2019) 284–293

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Meta-analysis of glyphosate contamination in surface waters and dissipation by biofilms

Louis Carles[a,*], Hélène Gardon[a], Laura Joseph[a], Josep Sanchís[b], Marinella Farré[b], Joan Artigas[a]

[a] Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Génome et Environnement (LMGE), F-63000 Clermont-Ferrand, France
[b] Institute of Environmental Assessment and Water Research (IDAEA-CSIC), C/Jordi Girona, 18-26, 08034 Barcelona, Catalonia, Spain

ARTICLE INFO

Handling Editor: Robert Letcher

Keywords:
Aminomethyl phosphonic acid (AMPA)
Phosphorus
Co-occurrence
Microbial ecotoxicology
Biodegradation
Eutrophication

ABSTRACT

One consequence of the intensive use of glyphosate is the contamination of rivers by the active substance and its metabolites aminomethyl phosphonic acid (AMPA) and sarcosine, inducing river eutrophication. Biofilms are the predominant lifestyle for microorganisms in rivers, providing pivotal roles in ecosystem functioning and pollutant removal. The persistence of glyphosate in these ecosystems is suspected to be mostly influenced by microbial biodegradation processes.

The present study aimed to investigate the tripartite relationship among biofilms, phosphorus and glyphosate in rivers. The first part consists of a co-occurrence analysis among glyphosate, AMPA and phosphorus using an extensive dataset of measurements (n = 56,198) from French surface waters between 2013 and 2017. The second part investigated the capacity of natural river biofilms to dissipate glyphosate, depending on phosphorus availability and the exposure history of the biofilm, in a microcosm study.

A strong co-occurrence among glyphosate, AMPA and phosphorus was found in surface waters. More than two-thirds of samples contained phosphorous with glyphosate, AMPA or both compounds. Seasonal fluctuations in glyphosate, AMPA and phosphorus concentrations were correlated, peaking in spring/summer shortly after pesticide spreading. Laboratory experiments revealed that natural river biofilms can degrade glyphosate. However, phosphorus availability negatively influenced the biodegradation of glyphosate and induced the accumulation of AMPA in water. An increase in alkaline phosphatase activity and phosphorus uptake was observed in glyphosate-degrading biofilms, evidencing the tight link between phosphorus limitation and glyphosate degradation by biofilms.

The results of the present study show that phosphorus not only is a key driver of river eutrophication but also can reduce complete glyphosate degradation by biofilms and favour the accumulation of AMPA in river water. The predominant role of biofilms and the trophic status of rivers must therefore be considered in order to better assess the fate and persistence of glyphosate.

1. Introduction

Glyphosate (*N*-(phosphonomethyl) glycine) is a systemic herbicide exhibiting a broad activity spectrum. This herbicide is mainly used before planting non-genetically modified crops, on glyphosate-resistant crops (Powles and Duke, 2010), in orchards (Maqueda et al., 2017), in minimum-tillage agroecosystems (Buhler, 2014) and in urban areas (Hanke et al., 2010). Since its commercialisation in 1974, glyphosate use has rapidly increased (Duke, 2015). In 2012, approximately 9000 tons of glyphosate was used in France, 127,000 tons in the USA and 700,000 tons worldwide (AGRESTE, 2018; Swanson et al., 2014; US Geological Survey, 2018). This herbicide is authorized at the European Community level (included in Annex I to Directive 91/414/EEC on 2002/07/01 by Commission Directive 2001/99/EC) and at the national level in France, where a total of 192 glyphosate-based formulations are currently approved (E-Phy, 2018). Glyphosate is the most-used active substance in France, and sales remained constant over the period 2011–2015 (AGRESTE, 2018). Broad contamination by glyphosate residues has recently led to awareness of its potential harmful side effects to human health and soil and aquatic ecosystems (Davoren and Schiestl, 2018; Van Bruggen et al., 2018). This global contamination has been a central question in the media and political scenes worldwide concerning the prolongation (or not) of glyphosate's use in the environment.

* Corresponding author at: Laboratoire Microorganismes: Génome et Environnement, 1 Impasse Amélie Murat, TSA 60026, CS 60026, 63178 Aubière Cedex, France.
E-mail address: louis.carles@uca.fr (L. Carles).

https://doi.org/10.1016/j.envint.2018.12.064
Received 10 October 2018; Received in revised form 30 November 2018; Accepted 31 December 2018
Available online 17 January 2019

One consequence of the intensive use of glyphosate is the contamination of various environmental compartments despite the fact that this compound can be degraded by various soil and aquatic microorganisms (Sviridov et al., 2015). The primary degrading product often detected in the environment is AMPA aminomethyl phosphonic acid (AMPA) (Singh and Singh, 2016). Glyphosate can also move through soil and contaminate surface and ground waters (Van Bruggen et al., 2018). In rivers, the dissipation time to eliminate 50% (DT_{50}) of glyphosate varied from 13.8 to 301 days, suggesting a moderate to high persistence of the molecule (EFSA, 2015). In France, AMPA and glyphosate are among the most often quantified compounds in river water (63% and 43% quantification frequency, respectively). Their concentrations can reach up to 558 $\mu g\,L^{-1}$ (AMPA) and 164 $\mu g\,L^{-1}$ (glyphosate) (NAIADES, 2018).

Rivers are dynamic ecosystems that play a key role in nutrient cycling and xenobiotic mitigation (Hanna et al., 2018). Within rivers, biofilms are one of the microbial lifestyle that guarantee essential ecosystem functions and participate in biogeochemical cycles (Battin et al., 2016). The large panel of functions provided by river biofilms is due to their extreme microbial diversity. Biofilms are composed of metazoan organisms, bacteria, archaea, algae, fungi, protozoa and viruses embedded in an extracellular matrix protecting them against environmental stresses such as nutrient starvation and pollutants (Besemer, 2015). River water carries nutrients, organic matter and pollutants, which may constitute important constraints for microorganisms (Battin et al., 2016). In turn, the fact that biofilms are complex assemblages hosting a large variety of microbial species enables biofilms to degrade a large variety of xenobiotics, either by co-metabolism or by mineralization. Biofilms are also dynamic communities that can colonize different types of substrate (e.g. mud, stones, macrophytes) and adapt to xenobiotic exposure (Blanck, 2002).

The glyphosate molecule contains one atom of phosphorus and can therefore be used as a phosphorus source for a variety of microorganisms in biofilms (i.e., bacteria and fungi) (Sviridov et al., 2015). This compound can therefore contribute to river eutrophication (Vera et al., 2010) and reinforce the problem resulting from massive fertilizer utilization (Lasier et al., 2016) can lead to an excess of phosphorus, which interferes with the capacity of biofilms to mitigate glyphosate. The genetic and biochemical bases of glyphosate utilization as phosphorus source for microorganisms have been well described in the literature (Hove-Jensen et al., 2014). Glyphosate is cleaved through the CP-lyase pathway, which involves seven or eight enzyme-catalysed reactions (e.g., phosphonate activation and C-P bound cleavage). The genes encoding the corresponding enzymes are assembled into the *phn* operon, which is widespread among bacterial species.

To date, the effect of phosphorus on glyphosate biodegradation has mostly been studied for isolated microbial strains (Krzysko-Lupicka et al., 2015; McMullan and Quinn, 1994) rather than for natural biofilm communities (Klátyik et al., 2017). The present study investigates the tripartite relationship among biofilms, phosphorus, and glyphosate in rivers. The first part of the study consists of the analysis of an extensive dataset of glyphosate, AMPA and total dissolved phosphorus concentrations recorded in French surface waters between 2013 and 2017 (n = 56,198). The co-occurrence, seasonality and influence of trophic status on glyphosate and AMPA fate in rivers were investigated. The second part of the study was carried out in laboratory microcosms in order to investigate the capacity of natural biofilms to dissipate glyphosate. Biofilms from an upstream site (non-exposed to glyphosate) and a downstream site (chronically exposed to glyphosate) were used to test the influence of i) phosphorus availability in water and in biofilms and ii) glyphosate availability in water on glyphosate dissipation. Our main hypothesis was that glyphosate can be used as a phosphorus source for biofilm microorganisms, though eutrophic conditions might lead to the slower total degradation of glyphosate and accumulation of AMPA in rivers. Besides, biofilms chronically exposed to glyphosate would develop more efficient glyphosate degradation comparing to biofilms non-exposed.

2. Material and methods

2.1. Meta-analysis of glyphosate, AMPA and total phosphorus concentrations in surface waters from France

Data on glyphosate, AMPA and total phosphorus concentrations in surface waters in France were downloaded on 2018/02/16 from the NAIADES public database available at http://www.naiades.eaufrance.fr/acces-donnees#/physicochimie (NAIADES, 2018). This database contains the results of analyses performed by water agencies and environmental consultancies, with a total of 4733 (glyphosate), 4716 (AMPA) and 9326 (total phosphorus) sites located throughout France (excluding French overseas departments and territories). Within this huge database, the research criteria imposed for data selection in our study were as follows: time period: “2013/01/01–2017/12/31”; data qualification: “correct” (i.e. result of the water analysis validated by the freshwater agencies); and fraction: “raw water”. Valid data obtained separately for each parameter (glyphosate samples N = 72,298, AMPA N = 72,277, total phosphorus N = 215,462) were then grouped together (N = 56,198) and filtered again by cases in which quantification of each compound was > average of their corresponding limit of quantification (LQ). This resulted in N = 31,041 for glyphosate (LQ = 0.03 $\mu g\,L^{-1}$), N = 45,552 for AMPA (LQ = 0.02 $\mu g\,L^{-1}$) and N = 199,833 for total phosphorus (LQ = 0.01 $mg\,P\,L^{-1}$).

Co-occurrence analysis was therefore performed in cases where three compounds were detected at the same station AND on the same date. The quantification of glyphosate and AMPA was also classified by trophic status (according to the concentration of total phosphorus in surface water, $\mu g\,P\,L^{-1}$): oligotrophic (< 25, N = 1018), mesotrophic (25–75, N = 6988), eutrophic (75–100, N = 3147) and hypereutrophic (> 100, N = 11,670).

The temporal evolution of glyphosate, AMPA and total phosphorus concentrations was assessed on a monthly basis from January 2013 to August 2017. The average concentration of each parameter was calculated for each month independent of the station (N = 56). The relationship between the AMPA/glyphosate ratio and the concentration of total phosphorus was assessed for a monthly averaged dataset using Spearman's rank correlation test ρ ($P < 0.05$).

2.2. Microcosm study

Colonization of natural biofilms was carried out in the field at the end of spring 2017 in an upstream site (Ups, 45°43′14.4″N 3°01′16.3″E) and a downstream site (Dws, 45°47′44.8″N 3°10′26.8″E) of the Artière River (Puy-de-Dôme region, France) during two weeks (Fig. S1). Both sites were well distinct in terms of (mean for upstream/downstream sites): water velocity 0.21/0.74 $m\,s^{-1}$; water discharge 0.03/1.25 $m^3\,s^{-1}$; dissolved organic carbon 6.50/12.97 $mg\,L^{-1}$; NO_3 3.55/16.70 $mg\,L^{-1}$; soluble reactive phosphorus 0.03/0.32 $mg\,P\,L^{-1}$; dissolved oxygen 10.02/8.15 $mg\,L^{-1}$; conductivity 297.53/654.91 $\mu S\,cm^{-1}$; pH 7.33/7.35; glyphosate 0.00/0.25 $\mu g\,L^{-1}$ and AMPA 0.09/1.08 $\mu g\,L^{-1}$ (Artigas et al., 2017; Rossi et al., 2019). Rocks, boulders, and sand covered the streambed of the upstream site, while rocks and sand were dominant at the downstream site. Frosted glass slides with an area of 21.9 cm^2 were glued onto three different flagstones per site, which were randomly submerged along the stream sections to depths of 10–20 cm.

Twenty-four microcosms consisting of 20 L glass aquariums (rectangular parallelepiped length × width × height 39 × 19 × 23 cm) were used to determine the biodegradation potential of glyphosate in the laboratory. For the acclimation phase, glass slides were detached from flagstones and carefully placed at the bottom of the aquarium. Colonized slides were heavy enough to not to be taken by the water current supplied by the aquarium pump. The side of the slide where the

biofilm was grown was placed in contact with the water column and the light source. The position of biofilms in the microcosm mimicked water depths, current velocity, and light conditions similar to those found in the field during colonization. Half of the aquaria received Ups biofilms, while the other half received Dws biofilms (11 colonized slides per aquarium). Each aquarium containing biofilms was filled with 5 L of filtered (0.5 mm) water collected from the upstream site. This water was characterized by 4.0 mg L^{-1} of NO_3, 100 µg P L^{-1} of soluble reactive phosphorus, 8.2 mg L^{-1} of dissolved organic carbon, 9.2 mg L^{-1} of dissolved oxygen, 244 µS cm^{-1} of conductivity and a pH of 7.3. This water did not contain any traces of glyphosate and AMPA. Water was recirculated in each aquarium by a submerged pump (Newjet 1200, Newa, Italy). Flow rates of aquarium pumps were adjusted in order to obtain water current velocities similar to those measured in the Artière upstream site (200 cm s^{-1}). Temperature and photoperiod were set at 19 ± 1 °C and 13 h light: 11 h dark. The water in the aquaria was completely renewed every 3 days to avoid nutrient limitation in the biofilms during the acclimation phase. The dissolved phosphorus concentration in aquaria containing Ups biofilms was maintained at the initial P concentration at the upstream site (corresponding to the LowP condition, i.e., an average value of 100 µg P L^{-1} during the experiment), whereas the water in Dws aquaria was adjusted to 10 × LowP (corresponding to the HighP condition, 1000 µg P L^{-1}) by adding adequate volumes of 1 M K_2HPO_4 salt solution.

After two weeks of laboratory acclimation for the biofilms, the water was removed, and each biofilm type (Upstream = colonized at a low-P site, Downstream = colonized at a high-P site) were subjected to the four experimental conditions in triplicate according to water phosphorus concentration (LowP and HighP as defined above) and glyphosate concentration (LowG = 10 µg L^{-1} and HighG = 100 µg L^{-1}, nominal concentration) by adding adequate volumes of 10 and 100 g L^{-1} glyphosate aqueous solution prepared with glyphosate PESTANAL® (analytical standard, CAS Number 1071-83-6, Sigma-Aldrich, France) in 0.2 µm filtered water from upstream (15 L per aquarium). These experimental conditions correspond to 1 × and 10 × the mean of actual concentration of glyphosate in surface waters from France. The experimental conditions in our microcosm experiment were therefore chosen to assess realistic environmental contamination conditions. Abiotic controls (0.2 µm filtered water without biofilms) were also carried out for each condition (in triplicate) in order to check the absence of abiotic dissipation of glyphosate. Glyphosate and AMPA were quantified in water at the beginning and at the end of the experiment (day 27).

Water analyses were carried out at 0, 2, 4, 6, 9, 13, 16, 20, 23 and 27 days. One glass slide was removed from each aquarium at 0, 2, 4, 6, 13, 20 and 27 days for biofilm analysis. Scraped biofilms were suspended in 12 mL of sterile 0.8% NaCl solution. After homogenization, aliquots of the biofilm suspension were kept for subsequent analyses.

Nitrate and phosphorus concentrations were maintained throughout the experiment by adding adequate volumes of 1 M KNO_3 (in order to maintain the initial concentration of nitrate in water, i.e. 4.0 mg L^{-1}) and 1 M K_2HPO_4 salt solutions (in order to satisfy the concentration of phosphorus of LowP and HighP described above), respectively.

2.2.1. Water analyses

The concentrations of total phosphorus and soluble reactive phosphorus (SRP) in filtered water (0.45 µm) were determined spectrophotometrically at 890 nm (Murphy and Riley, 1962). The total phosphorus concentration was determined after an additional digestion step: 10 mL of water samples was mixed with 1 mL of digestive reagent (0.185 M potassium persulfate, 0.485 M boric acid in 0.375 M NaOH) and heated at 120 °C for 1 h 30 m.

Glyphosate and AMPA quantification was performed using 100 mL water samples. These samples were immediately sent in refrigerated boxes (4 °C) to the CARSO laboratory (COFRAC agreement number 1-1531, Lyon, France) and kept at 4 °C until glyphosate and AMPA measurements were performed (HPLC/FLD, internal method M_ET143). The method consisted of the analysis of both compounds after a derivatization step (glyphosate and AMPA LQs = 0.1 µg L^{-1}).

2.2.2. Bacterial abundance and phosphorus content in biofilms

Bacterial density was estimated for each experimental condition and sampling time using flow cytometric counts of bacterial cells. Five hundred microliters of biofilm suspension was centrifuged (13,000*g* for 5 min at 4 °C). The pellet was resuspended in 900 µL of sterile PBS pH 7.2 (g L^{-1}: NaCl, 8.5; Na_2HPO_4 1.07; NaH_2PO_4, 0.39) and sonicated twice (40 W, 40 kHz, 30 s) using a sonication bath (model FB 15048, Fisher Scientific, Leicestershire, U.K.). The resulting bacterial suspension was centrifuged again (800*g* for 60 s at 4 °C), fixed with formaldehyde (2% final concentration) and stored at 4 °C until cytometric analysis. Ten microliters of fixed bacterial suspension was diluted 25-fold in TE buffer (10 mM Tris, 1 mM EDTA) and stained with 2.5 µL of SYBR Green I (Molecular Probes) before counting bacterial cells with a BD FACSCalibur flow cytometer (15 mW at 488 nm, Becton Dickinson, USA).

Biofilm phosphorus content was measured spectrophotometrically (Murphy and Riley, 1962) after the digestion of 2 mL of biofilm suspension, using the same protocol described above for total phosphorus quantification in water. Biofilm P content was corrected by the biofilm dry weight. Biofilm dry weight was determined as follows: 2 mL of biofilm suspension were centrifuged (13,000*g* for 5 min at 4 °C), the supernatant was discarded and the pellet was dried over night at 60 °C. The weight of dry pellet was then measured using a precision balance (Sartorius CPA225D, 0.01 mg precision).

2.2.3. Biofilm activity

The potential extracellular alkaline phosphatase activity (APA, EC 3.1.3.1) in biofilms was measured for each experimental condition and sampling time using a methylumbelliferyl-phosphate (MUF-P, Sigma) substrate analogue (Chrost and Krambeck, 1986). Eight hundred microliters of biofilm suspension was mixed with the MUF-P substrate at a known saturation concentration of 0.3 mM. Samples were then incubated for 1 h at 19 °C in the dark with agitation. The enzymatic activity was stopped by adding (1:1, V:V) 0.05 M glycine buffer (pH 10.4), and fluorescence was measured (365 nm excitation, 455 nm emission) with a microplate fluorometer (Fluoroskan™, Thermo Scientific, France).

Phosphorus uptake was calculated from bi-weekly SRP concentration measurements in microcosms containing treated biofilms. Phosphorous uptake (µg P g $(biofilm_{dw})^{-1}$ day^{-1}) was calculated as follows: P_{uptake} = (SRP immediately after P adjustment − SRP after each sampling time interval) / sampling time interval (days).

2.2.4. Glyphosate and AMPA quantification in biofilms

At the end of the experiment (day 27), the biofilms were scraped from the remaining glass slides (total surface = 49.5 cm^2) and lyophilized before glyphosate and AMPA analysis. The dry mass of each biofilm was estimated, and samples were extracted by ultrasound-assisted solid-liquid extraction using a custom method. Briefly, ultrapure water acidified with formic acid (HFor, Sigma-Aldrich, Steinheim, Germany) at pH = 3.0 was added to the extraction tubes at a ratio of 0.10 mL/1.0 mg_{dw} of collected biofilm. The tubes were placed in an ultrasonic bath for 20 min. Afterwards, the extracts were centrifuged (4000 rpm, 10 min). The supernatants were vacuum-filtered through 0.45 µm mesh nylon filters (Whatman, Maidstone, UK) and were derivatized with 9-fluorenylmethoxycarbonyl chloride (FMOC, 97%, Sigma-Aldrich) according to a previously published method (Hanke et al., 2008; Sanchís et al., 2012).

After 2 h of derivatization, the extracts were acidified to pH = 3.0 with $HFor_{(aq)}$ and centrifuged (4000 rpm, 10 min). The supernatants were analysed by liquid chromatography coupled to high resolution mass spectrometry (HPLC-HRMS) with an Acquity UPLC system

(Waters, Milford, MA, USA) and a Q Exactive™ mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA). Separation was achieved with a C18 column (Luna®, 150 × 2.0 mm; particle size, 5 µm, Phenomenex). $HFor_{(aq)}$ (0.1%) and acetonitrile were used as mobile phases at 0.250 mL min^{-1}. Ionisation was carried out with an electrospray ionisation (ESI) source in negative mode. The acquisition was carried out in full-scan mode. The deprotonated molecular ions $[C_{18}H_{17}NO_7P]^-$ (m/z = 390.0748) and $[C_{16}H_{15}NO_5P]^-$ (m/z = 332.0693) were used for FMOC-glyphosate and FMOC-AMPA, respectively.

2.3. Modelling glyphosate dissipation and AMPA formation in microcosms

The dissipation kinetics of glyphosate in water were fitted using OriginPro 2016 software (Origin Lab Corporation, USA) to a simple first-order exponential model (Exp2Mod1), characterized by the following equation:

$$C_t = C_0 e^{-kt} \quad (1)$$

where t is the incubation time, C_t the glyphosate concentration at time t, C_0 the initial concentration of glyphosate and k the rate constant in day^{-1}. DT_{50} is the time required for the concentration to decline to 50% of the initial value.

The kinetics of AMPA formation in water were fitted with the sigmoid function (Boltzmann) of OriginPro 2016, characterized by the following equation:

$$C_t = \frac{C_0 - C_f}{1 + e^{(t-h_{50})k}} + C_f \quad (2)$$

where t is the incubation time, C_t the AMPA concentration at time t, C_0 the initial concentration of AMPA, C_f the final concentration of AMPA, k the rate constant in day^{-1} and h_{50} the time required for the AMPA concentration to reach 50% of ($C_f - C_0$).

2.4. Statistical analyses

Statistical analyses were carried out using OriginPro 2016. The differences in various parameters linked to glyphosate dissipation kinetics (C_0 and DT_{50}), AMPA formation kinetics (C_f), glyphosate and AMPA content in biofilms, phosphorus uptake (P_{uptake}), cell abundance, phosphorus content and phosphatase activity were assessed using three-way ANOVA followed by separate post hoc comparisons (Tukey's test, $P < 0.05$). The three factors tested were site (Ups vs. Dws), phosphorous in water (LowP vs. HighP) and glyphosate initial concentration in water (LowG vs. HighG). The normality and homogeneity of variance were checked prior to ANOVA analysis (Kolmogorov-Smirnov's and Levene's tests, respectively, $P < 0.05$) and data that were not normally distributed were transformed using logarithmic functions.

3. Results

3.1. Concentrations of glyphosate, AMPA and phosphorus in French surface waters between 2013 and 2017

The analysis performed on the public database revealed that phosphorus concentrations are relatively high (mean of 0.1 mg P L^{-1}, maximum 18 mg P L^{-1}), corresponding to sites with high levels of eutrophication. Phosphorus was detected in all the samples analysed (N = 215,462) and was properly quantified (LQ = 0.01 mg P L^{-1}) in 93% of the samples. In parallel, glyphosate and AMPA were also detected in almost all water samples (99.9% each). Our analysis was specifically focused on samples containing at least one of the analytes above its LQ. The Venn diagram showed that co-occurrence (i.e., when compounds were quantified at the same station AND on the same date) was mainly found for glyphosate–AMPA–phosphorus (42.4%), followed by AMPA–phosphorus (22.9%) (Fig. 1). However, very few samples had

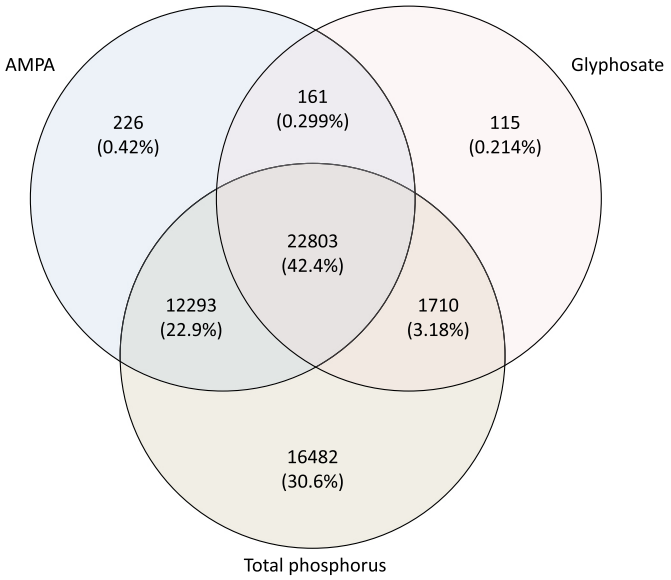


Fig. 1. Co-occurrence of glyphosate, AMPA and phosphorus in surface water in France. Values are the number of water samples in which the corresponding compound was quantified (i.e., result of analysis > LQ). Period of time: January 2013–August 2017 (N = 53,790 measurements).

only the combination glyphosate – phosphorus (3.2%) because glyphosate was most often quantified with its main metabolite AMPA (42.7%) compared to glyphosate without AMPA (3.4%).

The temporal evolution of glyphosate, AMPA and total phosphorus concentrations followed a seasonal pattern that was consistent over the years (Fig. 2). The concentrations of the three compounds increased from spring to summer (the increase in glyphosate occurred first, followed by AMPA and total phosphorus) and decreased from the end of autumn to winter. In terms of phosphorus equivalents, the results showed that the $P_{(AMPA)}$ concentration was always higher than the $P_{(glyphosate)}$ concentration, by a factor of 2 to 10 depending on the season (higher differences were obtained during spring and summer). However, the concentration of phosphorus from both compounds was extremely low compared with the total phosphorus concentration in water ($P_{(AMPA)}$ and $P_{(glyphosate)}$ represent < 0.17% and 0.05% of total P).

Regarding P concentrations, French surface waters containing glyphosate and/or AMPA were classified into 4 different trophic states according to the classification by Dodds et al. (1998). The more eutrophic a site was, the greater were the concentrations of glyphosate and AMPA recorded (Fig. S2). The specific glyphosate and AMPA concentrations in water varied from 0.12 ± 0.02 and 0.13 ± 0.01 µg L^{-1} (oligotrophic) to 0.32 ± 0.05 and 0.90 ± 0.02 µg L^{-1} (hypereutrophic), respectively. The results also indicate that glyphosate and AMPA were quantified in water samples containing < 100 µg P L^{-1}. These compounds were even quantified in oligotrophic water containing very little phosphorous (approximately 10 µg P L^{-1}).

A significant correlation was also observed between the AMPA/glyphosate ratio and total phosphorus concentration in water, as shown by the positive Spearman's rank correlation coefficient $\rho = 0.617$ ($P < 0.001$) (Fig. 3). This correlation indicates that sites with high phosphorus concentrations result in low glyphosate but high AMPA accumulation, since the latter compound still contains an atom of phosphorus. In addition, a positive correlation was observed between glyphosate and AMPA ($\rho = 0.498$, $P < 0.001$).

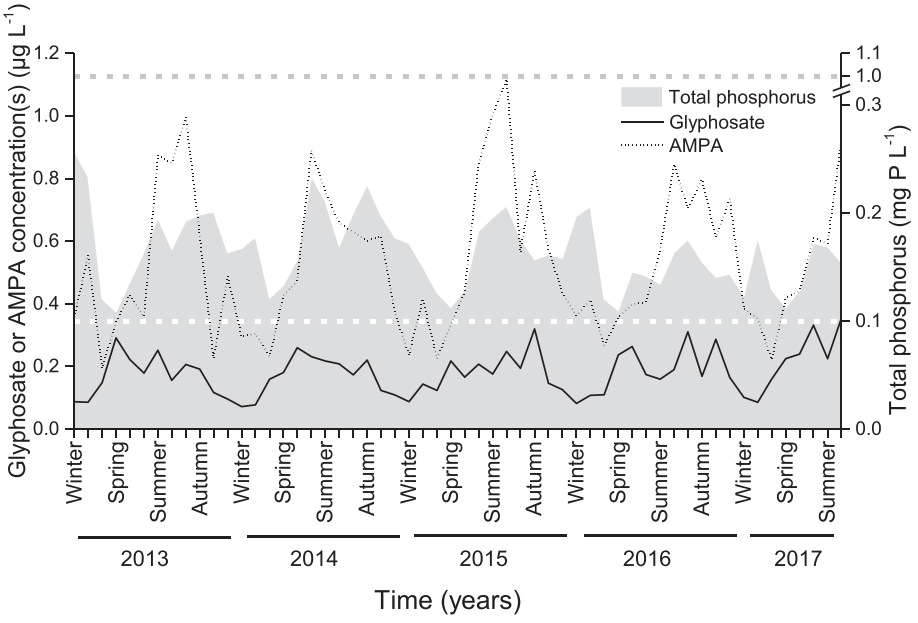


Fig. 2. Seasonal fluctuations in glyphosate, AMPA and phosphorus concentrations in surface water in France between January 2013 and August 2017. A total of N = 22,823 measurements were monthly averaged. The average concentrations of phosphorus used in the microcosm study are indicated by horizontal white (LowP) and grey (HighP) dashed lines.

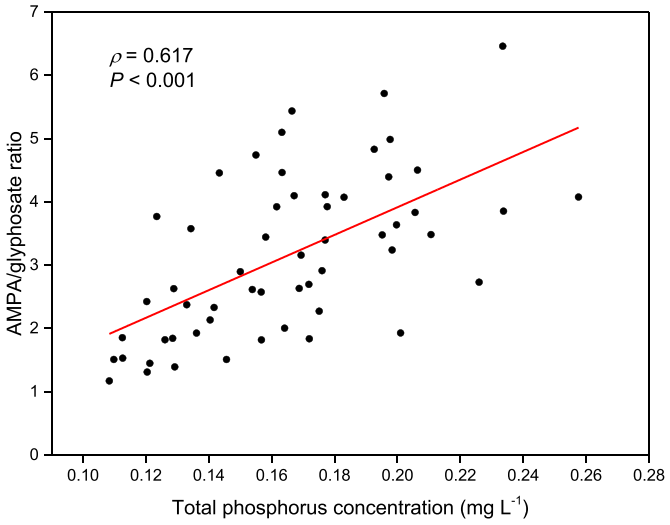


Fig. 3. Correlation between AMPA/glyphosate ratio and total phosphorus concentration in surface water in France. A total of N = 22,823 measurements were monthly averaged from January 2013 to August 2017. The value of Spearman's rank correlation test (ρ) is also indicated in the graph.

3.2. Glyphosate dissipation by river biofilms

Glyphosate and AMPA concentrations in water were monitored for 27 days in microcosms. The initial concentrations of glyphosate measured in aquaria were slightly lower than the nominal concentration. However, as expected, the concentrations of glyphosate in LowG condition (6.5 ± 1.1 µg L^{-1}) were 10 times lower than those in HighG condition (67.0 ± 1.0 µg L^{-1}) at the beginning of the experiment, as shown by C_0 values (Table 1). Throughout the entire experiment, glyphosate dissipation was observed in the presence of biofilms (Fig. 4A and B) but not in the abiotic controls (Table S2), indicating that the dissipation of glyphosate was not explained by abiotic factors such as photolysis and/or adsorption. Glyphosate dissipation curves were well fitted to a simple first-order exponential model in which the rate (k) and DT_{50} were inversely proportional (the higher the k value was, the lower the DT_{50}) (Table 1).

Glyphosate dissipation was faster for the lower concentration of glyphosate (DT_{50} = 4.3 ± 1.3 days) than for the higher concentration (DT_{50} = 23.2 ± 1.4 days) (Table 1, $P < 0.001$), though a strong effect of the concentration of phosphorus in water was also observed on dissipation coefficients (Fig. 4A and B, Table 1). For instance, glyphosate was more rapidly dissipated in LowP water than in HighP water ($P < 0.001$). Consequently, the highest percentages of glyphosate dissipation were obtained for biofilms in LowP and LowG water (Fig. 4A). In the LowP and LowG conditions, the herbicide was completely dissipated after 13 days, with DT_{50} values of 2.27 and 1.95 days for the Ups and Dws biofilms, respectively (Table 1). Concerning the origin of the biofilms, the highest DT_{50} values were obtained in Ups biofilms (Table 1, $P < 0.001$). For instance, the slowest glyphosate dissipation was observed in upstream biofilms subjected to HighP and HighG (3% loss after 27 days). Irrespective of the glyphosate concentration, the influence of water phosphorus on glyphosate dissipation was more marked in upstream biofilms than in downstream biofilms. Indeed, glyphosate DT_{50} in the LowG condition was increased by a factor of 3.7 (Dws) and 5.7 (Ups) from the LowP to the HighP conditions (Table 1).

The formation of AMPA was also modulated by the concentration of phosphorus in water ($P < 0.05$). This effect was more marked in LowG water than in HighG water (Table 1). Thus, in the LowP, LowG water condition, in which glyphosate dissipation was the fastest, AMPA was only a transient intermediate and did not accumulate (not detected after day 9, Fig. 4C). As expected, the final concentration of AMPA was higher in HighG than in LowG conditions (Table 1, $P < 0.001$). Because the transformation of glyphosate into AMPA is equimolar, the proportion of glyphosate transformed into AMPA at the end of the experiment could be determined. Except for the LowP, LowG condition, in which AMPA did not accumulate, calculations revealed that 82 ± 9% of glyphosate was transformed into AMPA in the LowG condition, whereas only 55 ± 5% was transformed in the HighG condition.

Glyphosate and AMPA were quantified in the biofilms at the end of the experiment (day 27). Very low contents of glyphosate (0.005–6.467 µg g_{dw}^{-1}) and AMPA (0.112–0.304 µg g_{dw}^{-1}) were detected in biofilms, the contents represented < 0.1% (glyphosate) and 0.07% (AMPA) of the initial amount of glyphosate added (molar equivalent) (Table S1). This trend was confirmed by the relatively low values of the partitioning coefficient $k(glyphosate)_{biofilm/water}$ obtained for all conditions tested (between $4.9 \cdot 10^{-4}$ and $6.5 \cdot 10^{-2}$ L g_{dw}^{-1}) (Table S1). Moreover, the phosphorus concentration in water did not influence the glyphosate content in biofilms ($P = 0.234$). The AMPA

Table 1
Kinetics of glyphosate dissipation and AMPA appearance in water. The experimental data for glyphosate and AMPA concentrations were fitted with a simple first-order model and the Boltzmann model, respectively, using OriginPro, as described in the Material and methods section. (A) Values of the model parameters dissipation time 50% (DT50), h_{50} and coefficient of determination (R^2) are reported as the mean ± standard error (SE), n = 3. Independently for glyphosate and AMPA, significant differences between conditions for each parameter are indicated by lowercase letters, a < b < c < d (Tukey's test, $P < 0.05$). N/A: not applicable. (B) Results of three-way ANOVA carried out with glyphosate (C_0 and DT_{50}) and AMPA (C_f) parameters. The three factors were site (Ups vs. Dws), phosphorus concentration in water Pwater (LowP vs. HighP) and glyphosate concentration in water Gwater (LowG vs. HighG), significant differences were indicated in bold ($P < 0.05$).

(A)

Glyphosate	Model parameters			
Condition	C_0 (µg L^{-1})	k (day^{-1})	DT_{50} (days)	R^2
Ups_LowP_LowG	5.8 ± 0.7 (a)	0.363 ± 0.102 (d)	2.27 ± 0.68 (a)	0.946 ± 0.031
Ups_LowP_HighG	64.5 ± 3.8 (b)	0.038 ± 0.004 (bc)	18.99 ± 2.40 (bc)	0.696 ± 0.087
Ups_HighP_LowG	7.2 ± 0.2 (a)	0.055 ± 0.004 (bc)	12.78 ± 0.87 (bc)	0.796 ± 0.071
Ups_HighP_HighG	68.6 ± 3.3 (b)	0.004 ± 0.002 (a)	206.36 ± 81.25 (d)	0.070 ± 0.059
Dws_LowP_LowG	6.3 ± 0.9 (a)	0.396 ± 0.097 (d)	1.95 ± 0.40 (a)	0.986 ± 0.004
Dws_LowP_HighG	70.2 ± 1.1 (b)	0.087 ± 0.011 (bc)	8.23 ± 1.05 (bc)	0.888 ± 0.052
Dws_HighP_LowG	6.9 ± 0.6 (a)	0.098 ± 0.007 (c)	7.12 ± 0.54 (b)	0.811 ± 0.042
Dws_HighP_HighG	65.9 ± 2.7 (b)	0.035 ± 0.005 (b)	20.80 ± 3.57 (c)	0.577 ± 0.062

AMPA	Model parameters				
Condition	C_0 (µg L^{-1})	k (day^{-1})	h_{50} (days)	C_f (µg L^{-1})	R^2
Ups_LowP_LowG	N/A	N/A	N/A	N/A	N/A
Ups_LowP_HighG	0.2 ± 0.002 (a)	0.296 ± 0.049 (b)	16.967 ± 0.753 (a)	17.4 ± 3.5 (b)	0.968 ± 0.007
Ups_HighP_LowG	0.2 ± 0.01 (a)	0.239 ± 0.026 (ab)	15.037 ± 1.384 (a)	2.7 ± 0.2 (a)	0.955 ± 0.009
Ups_HighP_HighG	0.2 ± 0.01 (a)	0.160 ± 0.005 (a)	15.256 ± 0.409 (a)	19.3 ± 2.1 (b)	0.879 ± 0.006
Dws_LowP_LowG	N/A	N/A	N/A	N/A	N/A
Dws_LowP_HighG	0.3 ± 0.01 (a)	0.201 ± 0.030 (ab)	16.328 ± 0.662 (a)	21.8 ± 1.6 (b)	0.918 ± 0.034
Dws_HighP_LowG	0.2 ± 0.01 (a)	0.185 ± 0.024 (ab)	14.268 ± 1.317 (a)	3.5 ± 0.3 (a)	0.928 ± 0.009
Dws_HighP_HighG	0.2 ± 0.03 (a)	0.166 ± 0.007 (ab)	14.949 ± 0.683 (a)	24.9 ± 2.7 (b)	0.906 ± 0.004

(B)

3-way ANOVA	Glyphosate				AMPA	
Factors/interactions	C_0		DT_{50}		C_f	
	F value	P value	F value	P value	F value	P value
Site	0.1	0.763	46.9	< **0.001**	4.6	< **0.05**
Pwater	1.9	0.185	134.1	< **0.001**	4.8	< **0.05**
Gwater	1552.1	< **0.001**	181.3	< **0.001**	226.01	< **0.001**
Site * Pwater	1.1	0.318	11.5	< **0.01**	0.2	0.693
Site * Gwater	0.01	0.922	18.5	< **0.001**	3.3	0.090
Pwater * Gwater	1.9	0.188	0.02	0.887	0.05	0.828
Site * Pwater * Gwater	0.002	0.967	2.9	0.112	0.005	0.945

content in biofilms was not influenced by any of the factors tested in this experiment (site, $P = 0.364$; Pwater, $P = 0.405$; glyphosate, $P = 0.780$).

3.3. Bacterial abundance and phosphorus content in biofilm

Bacterial growth in biofilms during glyphosate experiments was assessed by flow cytometry. The initial biofilm cell abundance was higher in biofilms from the downstream site ($2.8 \cdot 10^7 \pm 2.6 \cdot 10^6$ cells cm^{-2}) than in those from the upstream site ($2.4 \cdot 10^6 \pm 2.5 \cdot 10^5$ cells cm^{-2}); among the former samples, most of the surface of the glass slide was completely colonized. However, biofilm growth during the experiment was greater for upstream (increase of 1581.2 ± 197.8%) than for downstream (increase of 66.5 ± 25.4%) biofilms (Fig. S3, $P < 0.001$). More specifically, the bacterial growth in Ups biofilms was higher in LowP water (where phosphorus uptake and APA were higher, see below) than in HighP water (Fig. S3, $P < 0.05$).

The phosphorus content in biofilms was significantly higher in Dws biofilms than in the Ups ones (Fig. 5, $P < 0.001$). During the first 6 days of the experiment, the biofilms exhibiting the lowest P content (approximately 25 mg P g^{-1} of biofilm dry weight) were those from the Ups site subjected to LowP (Fig. S4). The phosphorus content of biofilms reflected the phosphorus concentration in water since biofilms placed in LowP water had a lower P content than those placed in HighP water throughout the entire experiment (Fig. 5, $P < 0.001$). Although glyphosate did not significantly affect the microbial growth and P content of biofilms, independent of the site and the phosphorus conditions in water, the high P demand observed in Ups biofilms coincided with the fastest glyphosate dissipation without AMPA accumulation, suggesting the use of glyphosate and/or AMPA as a phosphorus source for microbial growth.

3.4. Phosphorus utilization by biofilms

Phosphorus uptake was calculated from bi-weekly SRP concentration measurements in microcosms containing treated biofilms. A significantly higher P uptake was observed for biofilms subjected to LowP water than for those subjected to HighP water (Fig. 6, $P < 0.001$). Moreover, the phosphorus consumption was higher for Ups biofilms

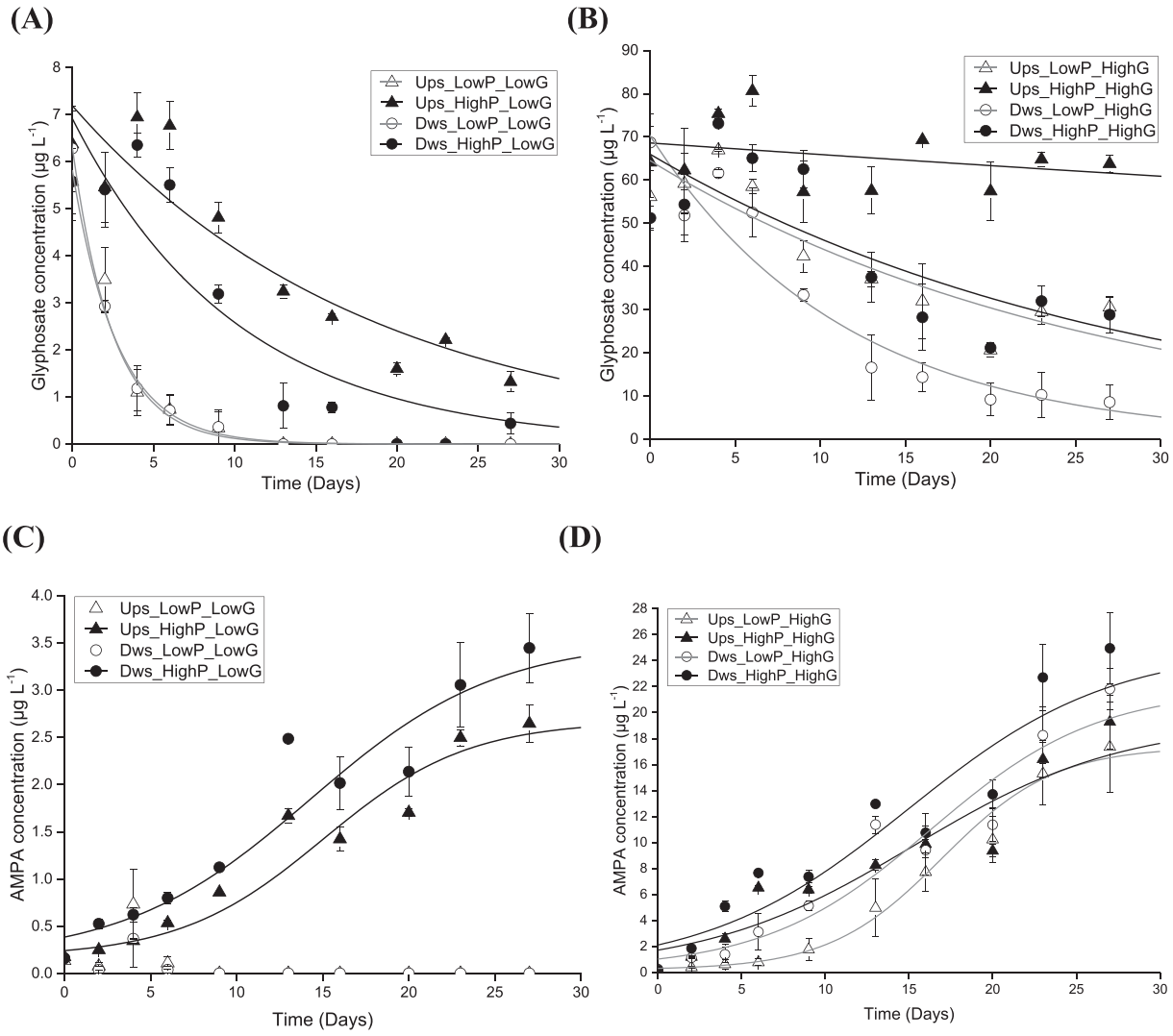


Fig. 4. Kinetics of glyphosate dissipation (A, B) and AMPA formation (C, D) in water. Biofilms from the upstream (Ups) and downstream (Dws) sites were exposed to low (A, C) and high (B, D) glyphosate concentrations and low (LowP) and high (HighP) phosphorus concentrations in water. The experimental data for glyphosate and AMPA concentrations were fitted with simple first-order and Boltzmann models, respectively, using OriginPro as described in the Material and methods section. The lines correspond to the model with mean parameters values for each condition. The values are means ± standard errors (n = 3) of the experimental data.

(colonized in a P-poor area) than for Dws biofilms (colonized in a P-rich area) (Fig. 6, $P < 0.001$). Phosphorus uptake by biofilms did not differ between glyphosate conditions ($P = 0.401$).

Potential APA (EC 3.1.3.1) was measured in biofilms at each sampling time, which permitted the integration of 27 days of activity for each experimental condition. The integrated APA did not differ between the Ups and Dws sites ($P = 0.358$) or between LowG and HighG conditions ($P = 0.116$, Fig. 7). However, similarly to phosphorus uptake, the phosphorus concentration in water was observed to have a significant effect on APA; the activity was the lowest under the HighP water condition ($P < 0.005$). This effect was particularly marked for Ups biofilms in the first 4 days of the experiment (average of 15.4 ± 2 and 3.2 ± 0.2 µmol MUF h^{-1} $g_{dw}{}^{-1}$ for LowP and HighP, respectively) compared with Dws biofilms (average of 4.6 ± 1.3 and 2.3 ± 0.7 µmol MUF h^{-1} $g_{dw}{}^{-1}$ for LowP and HighP, respectively) (Fig. S5).

4. Discussion

After herbicidal application, glyphosate can migrate through soil and reach aquatic compartments. This movement to surface water is limited by biodegradation by soil microorganisms (Sviridov et al., 2015) and sorption to soil particles, the latter being negatively correlated with pH and phosphorus content (Okada et al., 2016). The similarity between the chemical structures of glyphosate and phosphate molecules establishes competition between both compounds towards soil sorption sites (Kanissery et al., 2015). The observed seasonal variations in glyphosate and AMPA in French surface waters coincide with observations made in Canadian streams (Struger et al., 2015): concentrations increase from early spring to summer and decrease from the end of autumn to winter, matching perfectly with the pesticide application calendar for crops. This result suggests that glyphosate transfer between the terrestrial and aquatic ecosystems is relatively fast and that the saturation of soil sorption sites by phosphate probably accelerates this transfer. Indeed, the high level of co-occurrence among phosphorus, glyphosate, and AMPA is reflected by the strong positive correlation found between the AMPA/glyphosate ratio and phosphorus concentration in water (Fig. 3), which suggests that P eutrophication is probably an important factor influencing the balance between glyphosate and AMPA in surface water.

AMPA derived from glyphosate transformation constitutes the main source of AMPA contamination in surface water ecosystems, the other source being AMPA derived from amino-polyphosphonate (commonly used in industrial and household applications as detergents, flame

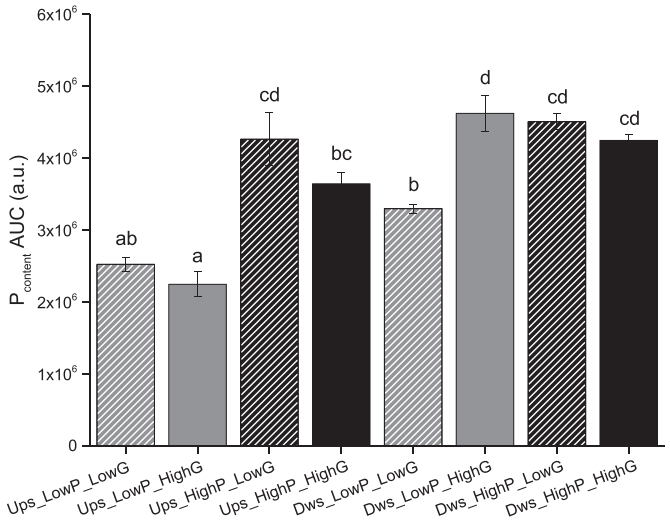


Fig. 5. Biofilm phosphorus content. The area under the curve (AUC), expressed in arbitrary units (a.u.) integrates the curves of phosphorus content over time. The results are reported as the mean ± standard error (SE), n = 3. Significant differences are indicated by lowercase letters, a < b < c < d (Tukey's test, $P < 0.05$).

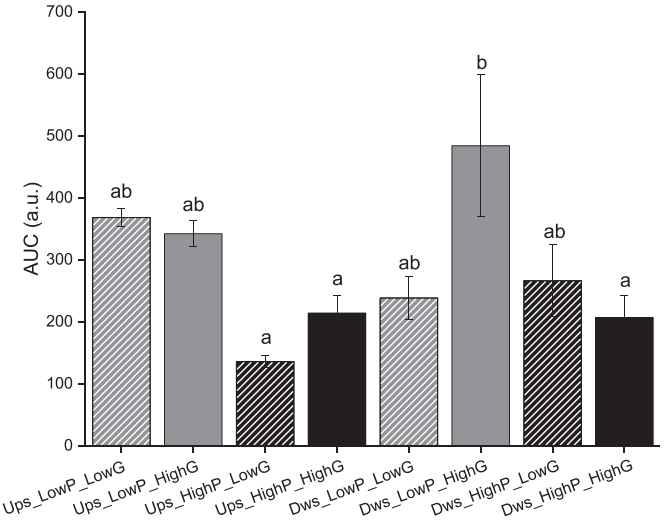


Fig. 7. Phosphatase activity of biofilms. Phosphatase activity curves were integrated with OriginPro 2016 to obtain the area under the curve (AUC), expressed in arbitrary units (a.u.). The results are reported as the mean ± standard error (SE), n = 3. Significant differences are indicated by lowercase letters, a < b (Tukey's test, $P < 0.05$).

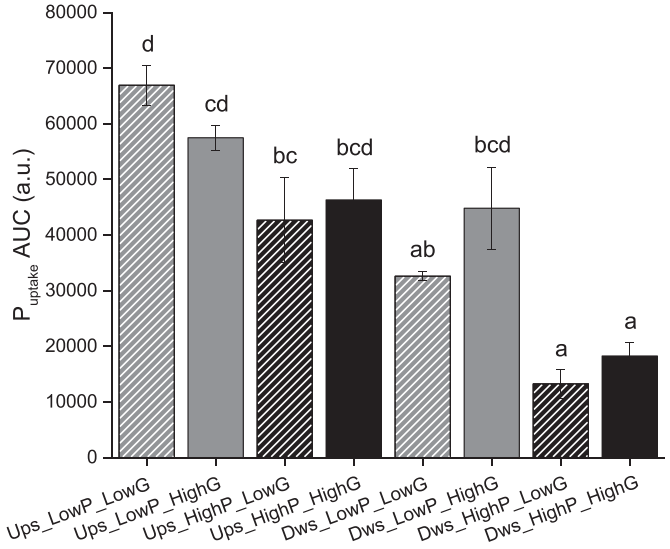


Fig. 6. Phosphorus uptake. Phosphorus uptake curves were integrated with OriginPro 2016 to obtain the area under the curve (AUC), expressed in arbitrary units (a.u.). The results are reported as the mean ± standard error (SE), n = 3. Significant differences are indicated by lowercase letters, a < b < c < d (Tukey's test, $P < 0.05$).

retardants, anticorrosives, etc.) degradation (Grandcoin et al., 2017). The present meta-analysis shows that glyphosate and AMPA co-occurred in water: glyphosate was quantified without AMPA in only 3.4% of 53,790 water samples analysed from 2013 to 2017 in France (Fig. 1), and a positive correlation was observed between glyphosate and AMPA ($\rho = 0.498$, $P < 0.001$). These results are in accordance with a spatially broad occurrence study performed in the USA, in which glyphosate was detected without AMPA in only 2.3% of 3732 water and sediment samples (Battaglin et al., 2014), and with another study that showed a co-occurrence ($\rho = 0.76$) between glyphosate and AMPA in Canadian rivers (Struger et al., 2015). The co-occurrence of glyphosate and AMPA in surface waters seems therefore inevitable and suggests the incomplete mineralization of the molecule, especially at eutrophic sites.

The levels of surface water contamination by pesticides are influenced by a large number of drivers that may be directly influenced by anthropic activities (e.g., agricultural practices, industrial activities) or not (e.g., water flow, temperature, biodegradation by microorganisms). The present study revealed that natural river biofilms possess a strong capability for glyphosate degradation under environmentally realistic conditions (Fig. 4A). Biofilms exposed to low concentrations of glyphosate ($< 10\,\mu g\,L^{-1}$) can dissipate the herbicide completely after 13 days. This capability was also found in a recent study, but only in the presence of formulated glyphosate (Klátyik et al., 2017). In realistic conditions (LowG in our study), the pre-exposure history of biofilms to glyphosate and AMPA does not influence the capacity of biofilm microorganisms to degrade the two molecules (a site effect was indicated only when the HighG condition was added to analyses). These results contradict our hypothesis, and those of many other studies on pesticides (Carles et al., 2017; Mamy et al., 2005; Tuxen et al., 2002) suggesting a fast adaptation of biofilms for glyphosate degradation irrespectively of their history of exposure to the molecule.

The dissolved phosphorus concentration in water has been shown to modulate the dissipation of glyphosate by biofilms, and this activity is well explained by the fact that this herbicide can represent a phosphorus source for microbes (Hove-Jensen et al., 2014). For the first time, this effect has been demonstrated for natural biofilm communities, indicating that similar biochemical effects can be observed at both the population and community scales. The utilization of glyphosate as phosphorus source has not yet been described for natural river biofilms; the microbial strains that are capable of using this herbicide as a sole phosphorus source have been isolated from other environments, mainly soil and activated sludge (Sviridov et al., 2015). Pure strain studies have also shown that glyphosate catabolism can be repressed by phosphorus in *Arthrobacter* sp. (Pipke et al., 1987) and *Pseudomonas* sp. PG2982 (Fitzgibbon and Braymer, 1988). Essentially, the strong effect of phosphorus availability on glyphosate dissipation, AMPA production (Fig. 4), biofilm phosphorus content (Fig. 5), P uptake and APA can be explained by phosphorus limitation within biofilms. Indeed, the fastest dissipation of glyphosate without AMPA production was observed in the LowP condition (Fig. 4 and Table 1). Inversely, glyphosate is mostly transformed into AMPA in eutrophic water in comparison with more P-poor waters. Surprisingly, glyphosate and AMPA were still detected at oligotrophic sites, whereas the two compounds disappeared completely in the LowG/LowP condition for upstream communities in our experiment. These different responses could result from differences in i) the composition of microbial communities, ii) environmental conditions

(e.g., temperature, pH), and/or iii) the chronic input of glyphosate and/or AMPA in surface waters, which was not the case in our microcosm study.

Although variations in environmental conditions and watershed types make it difficult to predict the environmental persistence of glyphosate, some general trends can be highlighted. The literature indicates that glyphosate has low to high persistence in soils under aerobic conditions (DT_{50} ranging from 2.8 to 500.3 days) and high persistence in anaerobic soils (DT_{50} of 135–1000 days) (EFSA, 2015). Some conditions could thus favour the persistence of glyphosate in soil and delay its biodegradation and/or movement to aquatic ecosystems. In water sediments, the persistence of glyphosate is moderate to high (DT_{50} range: 13.82–301 days) (EFSA, 2015). However, the concentration of phosphorus from both glyphosate and AMPA is extremely low compared with the total phosphorus concentration in surface water ($P_{(AMPA)}$ and $P_{(glyphosate)}$ represent < 0.17% and 0.05% of total P). One can therefore identify two main factors influencing the levels of glyphosate contamination in surface water. The first factor is the herbicide input, which is influenced by soil biological activity (biodegradation) and retention capacity (sorption). The second factor is phosphorus availability in water, which decreases glyphosate degradation and promotes AMPA accumulation. However, it cannot be ruled out that P production derived from glyphosate degradation could display a negative feed-back on glyphosate degradation in extremely low P systems. Overall, eutrophication by phosphorus favours the incomplete degradation of glyphosate in aquatic systems. Eutrophication levels should thus be taken into account in the environmental risk assessment of glyphosate and AMPA in surface waters.

Acknowledgements

The authors thank Muriel Joly, Florence Donnadieu and Florent Rossi for their valuable help in water and biofilm sampling.

Funding

This work was supported by the Agence Nationale de la Recherche (grant number ANR-16-CE32-0001-01 BIGLY).

Declaration of interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2018.12.064.

References

AGRESTE, 2018. Données de vente des produits phytosanitaires 2011 à 2015. URL. http://agreste.agriculture.gouv.fr/thematiques-872/productions-vegetales-874/grandes-cultures-fourrages-875, Accessed date: 22 May 2018 (WWW Document).

Artigas, J., Rossi, F., Gerphagnon, M., Mallet, C., 2017. Sensitivity of laccase activity to the fungicide tebuconazole in decomposing litter. Sci. Total Environ. 584–585, 1084–1092. https://doi.org/10.1016/j.scitotenv.2017.01.167.

Battaglin, W.A., Meyer, M.T., Kuivila, K.M., Dietze, J.E., 2014. Glyphosate and its degradation product AMPA occur frequently and widely in US soils, surface water, groundwater, and precipitation. J. Am. Water Resour. Assoc. 50, 275–290.

Battin, T.J., Besemer, K., Bengtsson, M.M., Romani, A.M., Packmann, A.I., 2016. The ecology and biogeochemistry of stream biofilms. Nat. Rev. Microbiol. 14, 251–263. https://doi.org/10.1038/nrmicro.2016.15.

Besemer, K., 2015. Biodiversity, community structure and function of biofilms in stream ecosystems. Res. Microbiol. 166, 774–781. https://doi.org/10.1016/j.resmic.2015.05.006.

Blanck, H., 2002. A critical review of procedures and approaches used for assessing pollution-induced community tolerance (PICT) in biotic communities. Hum. Ecol. Risk. Assess. 8, 1003–1034.

Buhler, D.D., 2014. Weed management. In: Reference Module in Earth Systems and Environmental Sciences. Elsevier. https://doi.org/10.1016/B978-0-12-409548-9.09118-1.

Carles, L., Rossi, F., Joly, M., Besse-Hoggan, P., Batisson, I., Artigas, J., 2017. Biotransformation of herbicides by aquatic microbial communities associated to submerged leaves. Environ. Sci. Pollut. Res. 24, 3664–3674. https://doi.org/10.1007/s11356-016-8035-9.

Chrost, J., Krambeck, H.J., 1986. Fluorescence correction for measurements of enzyme activity in natural waters using methylumbelliferyl-substrates. Arch. Hydrobiol. Stuttg. 106, 79–90.

Davoren, M.J., Schiestl, R.H., 2018. Glyphosate based herbicides and cancer risk: a post IARC decision review of potential mechanisms, policy, and avenues of research. Carcinogenesis. https://doi.org/10.1093/carcin/bgy105.

Dodds, W.K., Jones, J.R., Welch, E.B., 1998. Suggested classification of stream trophic state: distributions of temperate stream types by chlorophyll, total nitrogen, and phosphorus. Water Res. 32, 1455–1462.

Duke, S.O., 2015. Perspectives on transgenic, herbicide-resistant crops in the United States almost 20 years after introduction: perspectives on transgenic, herbicide-resistant crops. Pest Manag. Sci. 71, 652–657. https://doi.org/10.1002/ps.3863.

EFSA, 2015. Conclusion on the peer review of the pesticide risk assessment of the active substance glyphosate: peer review of the pesticide risk assessment of the active substance glyphosate. EFSA J. 13, 4302. https://doi.org/10.2903/j.efsa.2015.4302.

E-Phy, 2018. Le catalogue des produits phytopharmaceutiques et de leurs usages, des matières fertilisantes et des supports de culture autorisés en France. URL. https://ephy.anses.fr/substance/glyphosate, Accessed date: 25 July 2018 (WWW Document).

Fitzgibbon, J., Braymer, H.D., 1988. Phosphate starvation induces uptake of glyphosate by *Pseudomonas* sp. strain PG2982. Appl. Environ. Microbiol. 54, 1886–1888.

Grandcoin, A., Piel, S., Baures, E., 2017. AminoMethylPhosphonic acid (AMPA) in natural waters: its sources, behavior and environmental fate. Water Res. 117, 187–197. https://doi.org/10.1016/j.watres.2017.03.055.

Hanke, I., Singer, H., Hollender, J., 2008. Ultratrace-level determination of glyphosate, aminomethylphosphonic acid and glufosinate in natural waters by solid-phase extraction followed by liquid chromatography–tandem mass spectrometry: performance tuning of derivatization, enrichment and detection. Anal. Bioanal. Chem. 391, 2265–2276.

Hanke, I., Wittmer, I., Bischofberger, S., Stamm, C., Singer, H., 2010. Relevance of urban glyphosate use for surface water quality. Chemosphere 81, 422–429.

Hanna, D.E.L., Tomscha, S.A., Ouellet Dallaire, C., Bennett, E.M., 2018. A review of riverine ecosystem service quantification: research gaps and recommendations. J. Appl. Ecol. 55, 1299–1311. https://doi.org/10.1111/1365-2664.13045.

Hove-Jensen, B., Zechel, D.L., Jochimsen, B., 2014. Utilization of glyphosate as phosphate source: biochemistry and genetics of bacterial carbon-phosphorus lyase. Microbiol. Mol. Biol. Rev. 78, 176–197. https://doi.org/10.1128/MMBR.00040-13.

Kanissery, R.G., Welsh, A., Sims, G.K., 2015. Effect of soil aeration and phosphate addition on the microbial bioavailability of carbon-14-glyphosate. J. Environ. Qual. 44, 137. https://doi.org/10.2134/jeq2014.08.0331.

Klátyik, S., Takács, E., Mörtl, M., Földi, A., Trábert, Z., Ács, É., Darvas, B., Székács, A., 2017. Dissipation of the herbicide active ingredient glyphosate in natural water samples in the presence of biofilms. Int. J. Environ. Anal. Chem. 97, 901–921. https://doi.org/10.1080/03067319.2017.1373770.

Krzysko-Lupicka, T., Krecidlo, L., Koszalkowska, M., 2015. The ability of selected bacteria to grow in the presence of glyphosate. Ecol. Chem. Eng. Chem. Inzynieria Ekol. A 22, 185–193. https://doi.org/10.2428/ecea.2015.22(2)15.

Lasier, P.J., Urich, M.L., Hassan, S.M., Jacobs, W.N., Bringolf, R.B., Owens, K.M., 2016. Changing agricultural practices: potential consequences to aquatic organisms. Environ. Monit. Assess. 188, 672. https://doi.org/10.1007/s10661-016-5691-7.

Mamy, L., Barriuso, E., Gabrielle, B., 2005. Environmental fate of herbicides trifluralin, metazachlor, metamitron and sulcotrione compared with that of glyphosate, a substitute broad spectrum herbicide for different glyphosate-resistant crops. Pest Manag. Sci. 61, 905–916.

Maqueda, C., Undabeytia, T., Villaverde, J., Morillo, E., 2017. Behaviour of glyphosate in a reservoir and the surrounding agricultural soils. Sci. Total Environ. 593, 787–795.

McMullan, G., Quinn, J.P., 1994. In vitro characterization of a phosphate starvation-independent carbon-phosphorus bond cleavage activity in *Pseudomonas fluorescens* 23F. J. Bacteriol. 176, 320–324.

Murphy, J., Riley, J.P., 1962. A modified single solution method for the determination of phosphate in natural waters. Anal. Chim. Acta 27, 31–36.

NAIADES, 2018. Données sur la qualité des eaux de surface. URL. http://www.naiades.eaufrance.fr/acces-donnees#/physicochimie, Accessed date: 23 May 2018 (WWW Document).

Okada, E., Costa, J.L., Bedmar, F., 2016. Adsorption and mobility of glyphosate in different soils under no-till and conventional tillage. Geoderma 263, 78–85.

Pipke, R., Schulz, A., Amrhein, N., 1987. Uptake of glyphosate by an *Arthrobacter* sp. Appl. Environ. Microbiol. 53, 974–978.

Powles, S.B., Duke, S.O., 2010. Glyphosate-resistant Crops and Weeds: Now and in the Future.

Rossi, F., Mallet, C., Portelli, C., Donnadieu, F., Bonnemoy, F., Artigas, J., 2019. Stimulation or inhibition: leaf microbial decomposition in streams subjected to complex chemical contamination. Sci. Total Environ. 648, 1371–1383. https://doi.org/10.1016/j.scitotenv.2018.08.197.

Sanchís, J., Kantiani, L., Llorca, M., Rubio, F., Ginebreda, A., Fraile, J., Garrido, T., Farré, M., 2012. Determination of glyphosate in groundwater samples using an ultrasensitive immunoassay and confirmation by on-line solid-phase extraction followed by liquid chromatography coupled to tandem mass spectrometry. Anal. Bioanal. Chem. 402, 2335–2345.

Singh, B., Singh, K., 2016. Microbial degradation of herbicides. Crit. Rev. Microbiol. 42, 245–261. https://doi.org/10.3109/1040841X.2014.929564.

Struger, J., Van Stempvoort, D.R., Brown, S.J., 2015. Sources of aminomethylphosphonic

acid (AMPA) in urban and rural catchments in Ontario, Canada: glyphosate or phosphonates in wastewater? Environ. Pollut. 204, 289–297. https://doi.org/10.1016/j.envpol.2015.03.038.

Sviridov, A.V., Shushkova, T.V., Ermakova, I.T., Ivanova, E.V., Epiktetov, D.O., Leontievsky, A.A., 2015. Microbial degradation of glyphosate herbicides. Appl. Biochem. Microbiol. 51, 188–195 (Review).

Swanson, N.L., Leu, A., Abrahamson, J., Wallet, B., 2014. Genetically engineered crops, glyphosate and the deterioration of health in the United States of America. J. Org. Syst. 9, 6–37.

Tuxen, N., de Lipthay, J.R., Albrechtsen, H.-J., Aamand, J., Bjerg, P.L., 2002. Effect of exposure history on microbial herbicide degradation in an aerobic aquifer affected by a point source. Environ. Sci. Technol. 36, 2205–2212. https://doi.org/10.1021/es0113549.

US Geological Survey, 2018. Pesticide use maps. URL. https://water.usgs.gov/nawqa/pnsp/usage/maps/show_map.php?year=2012&map=GLYPHOSATE&hilo=L&disp=Glyphosate, Accessed date: 22 May 2018 (WWW Document).

Van Bruggen, A.H.C., He, M.M., Shin, K., Mai, V., Jeong, K.C., Finckh, M.R., Morris, J.G., 2018. Environmental and health effects of the herbicide glyphosate. Sci. Total Environ. 616, 255–268. https://doi.org/10.1016/j.scitotenv.2017.10.309.

Vera, M.S., Lagomarsino, L., Sylvester, M., Perez, G.L., Rodriguez, P., Mugni, H., Sinistro, R., Ferraro, M., Bonetto, C., Zagarese, H., Pizarro, H., 2010. New evidences of Roundup(A (R)) (glyphosate formulation) impact on the periphyton community and the water quality of freshwater ecosystems. Ecotoxicology 19, 710–721. https://doi.org/10.1007/s10646-009-0446-7.

CV – Décembre 2021

Hélène Gardon
Docteure en Bioinformatique

UMR CNRS 6023, LMGE
helene.gardon.pro@gmail.com
Née le 5 Oct. 1991

Expériences de recherche

depuis Oct. 2016 — **Thèse de Recherche**, *UMR CNRS 6023, LMGE,* Clermont-Ferrand, France.
Encadrantes : C. Petit et G. Bronner
Sujet : Étude du pangénome d'une population bactérienne structurée : vers une nouvelle compréhension de l'origine des variations intra-génomiques.
Description : analyse pangénomique d'une population bactérienne environnementale et structurée, évaluation de la variabilité des processus de sélection et transferts de matériel génétique le long de compartiments génomiques.

Janvier – Juillet 2016 — **Stage de Master 2 (6 mois),** *UMR CNRS 6023, LMGE,* Clermont-Ferrand, France.
Encadrante : G. Bronner
Sujet : Analyse de données *single-cell* pour l'exploration de la théorie de réduction des génomes chez les *Archaea* de l'environnement.
Description : évaluation des caractéristiques génomiques liées à une réduction des génomes, estimation de pressions de sélection et biais mutationnels, mesure du contenu en dinucléotides GC et usage des codons.

Mai – Juillet 2015 — **Stage de Master 1 (3 mois)**, SciLifeLab, Department of Cell & Molecular Biology, Suède
Encadrante : S. Andersson
Sujet : A bioinformatic study of *Planctomycetale* bacteria *Gemmata obscuriglobus*: Comparative genomics reveal the presence of secondary metabolic biosynthetic genes.
Description : identification et analyse comparative de clusters de gènes impliqués dans le métabolisme secondaire.

Formation

2014 – 2016 **Master Génétique et Physiologie, Bioinformatique – Analyse et Modélisation des Données**, Université Blaise Pascal, Clermont-Ferrand, France.

2011 – 2014 **Licence Biologie cellulaire et Physiologie**, Université Blaise Pascal, Clermont-Ferrand, France

Communications scientifiques

Articles

1. **H. Gardon**, C. Biderre-Petit, I. Jouan-Dufournel, G. Bronner (2020). A drift-barrier model drives the genomic landscape of a structured bacterial population. Molecular Ecology
2. C. Biderre-Petit, C. Hochart, **H. Gardon**, E. Dugat-Bony, S. Terrat, I. Jouan-Dufournel, R. Paris (2020). Analysis of bacterial and archaeal communities associated with Fogo volcanic soils of different ages. FEMS microbiology ecology, 96 (7), fiaa104
3. L. Carles, **H. Gardon**, L. Joseph, J. Sanchis, M. Farré, J. Artigas (2019). Meta-analysis of glyphosate contamination in surface waters and dissipation by biofilms. Environment International, 124, 284-293.
4. C. Biderre-Petit, N. Taib, **H. Gardon**, C. Hochart, D. Debroas D (2019). New insights into the pelagic microorganisms involved in the methane cycle in the meromictic Lake Pavin through metagenomics. FEMS microbiology ecology, 95 (3), fiy183
5. C. Loiseau, V. Hatte, C. Andrieu, L. Barlet, A. Cologne, R. De Oliveira, L. Ferrato-Berberian, **H. Gardon**, D. Lauber, M. Molinier, S. Monnerie, K. N'gou, B. Penaud, O. Perieira, J. Picarle, A. Septier, A. Mahul, J-C. Charvy, and F. Enault, (2018). Pangenehome / A Web Interface To Analyze Microbial Pangenomes. Journal of Bioinformatics, Computational and Systems Biology, 1(2) : 108.

Communications orales

1. **H. Gardon**, C. Biderre-Petit, I. Jouan-Dufournel, G. Bronner (2019). Dynamique évolutive des compartiments génomiques chez *Prochlorococcus*. *22èmes Journées de l'École Doctorale Sciences de la Vie, Santé, Agronomie et Environnement*, Clermont-Ferrand, France.
2. **H. Gardon**, C. Biderre-Petit, G. Bronner (2018). Caractérisation des mécanismes de la dynamique des génomes chez *Prochlorococcus*. *Rencontre des Microbiologistes du Pôle Clermontois*, Clermont-Ferrand, France.

3. **H. Gardon**, C. Biderre-Petit, G. Bronner (2017). New insights into the plasticity of *Prochlorococcus* genomes. *Colloque de Génomique Environnementale*, Marseille, France.

Posters

1. **H. Gardon**, C. Biderre-Petit, I. Jouan-Dufournel, G. Bronner (2019). Differential patterns of evolutionary dynamics of genomic compartments in co-occurring *Prochlorococcus* populations. *Annual Conference of Society for Molecular Biology and Evolution*, Manchester, UK.
2. **H. Gardon**, C. Biderre-Petit, G. Bronner (2017). New insights into the plasticity of *Prochlorococcus* genomes. *Colloque de l'Association Francophone d'Écologie Microbienne*, Camaret-sur-Mer, France. 3ème prix du meilleur poster.

Autres

Co-fondatrice et membre du conseil d'administration de l'**Organisation des Jeunes Chercheurs en Écologie** (OJCE) – Association créée en Octobre 2021 afin d'aider à l'intégration des jeunes chercheurs et chercheuses dans le monde de la recherche

Encadrement de stages

Juin – Juillet 2019 **Encadrement Master 1 (2 mois)**
Héléna Vassilieff, M1 Bioinformatique – Analyse et Modélisation des Données, Université Clermont Auvergne, France
Sujet : Impact des processus de recombinaison et transferts de gènes sur la différenciation des populations bactériennes.
Description : quantification de la balance entre recombinaison et sélection, répartition spatiale des événements de recombinaison, analyse de l'échange de matériel génétique entre populations.

Juin – Juillet 2018 **Encadrement Master 1 (2 mois)**
Camille Dupont, M1 Bioinformatique – Analyse et Modélisation des Données, Université Clermont Auvergne, France
Sujet : Développement d'une méthode pour mesurer et quantifier le chimérisme dans l'assemblage de métagénomes.
Description : utilisation de l'entropie de l'information et de la divergence de Kullback-Liebler pour la détection du chimérisme, comparaison d'assembleurs et paramètres d'assemblage dans l'introduction de chimérisme

Juin – Juillet 2017 **Encadrement Master 1 (2 mois)**
Clovis Norroy, M1 Bioinformatique – Analyse et Modélisation des Données, Université Clermont Auvergne, France

Sujet : Conception et implémentation d'une base de données pour la mesure et la caractérisation du chimérisme dans l'assemblage de métagénomes.
Description : simulation, puis assemblage, de métagénomes à partir de données *single-cell*, conception d'une base de données pour l'analyse du chimérisme.

Compétences techniques

Informatique	Langages de programmation : R, Bash, AWK, Perl ; Système d'exploitation : Linux
Statistiques	statistique descriptive, statistique inférentielle, analyses multivariées, traitement de données -*omiques*
Langues	Français – langue maternelle Anglais – professionnel

Enseignement

2017 – 2019 **Chargée de missions d'enseignement en bioinformatique et biologie évolutive**, Université Clermont Auvergne

- **L3 Bioinformatique**, TD/TP, 60h eq. TD , responsable : G. Bronner
  Annotations structurale et fonctionnelle de séquences géniques, introduction aux traitement de données de métagénomique, alignement de séquences, phylogénies

- **L3 Introduction à la biologie de l'évolution**, TD/TP, 22h eq. TD, responsable : G. Bronner
  Introduction aux méthodes de reconstructions phylogénétiques, interprétation de phylogénies

- **M1 Génomique comparative**, CM/TD, 30h eq. TD, responsable : G. Bronner
  Méthodes d'analyses comparées des génomes (homologies, phylogénomique, synténie), Génomique comparative des microorganismes (métagénomique, single-cell, analyse de pangénome)

- **M1 Bioanalyse de microbiomes**, TP, 10h eq. TD, responsable : D. Debroas
  Traitement de données de métatranscriptomique, analyse d'expression différentielle de gènes, analyses multivariées.

Résumé

Les modèles d'évolution associés aux concepts de l'espèce mettent en avant des processus de balayage sélectif à l'échelle des gènes ou à l'échelle des génomes. Au cours de cette thèse, une reconsidération des processus à l'origine de la différenciation de populations bactériennes libres environnementales a été réalisée en prenant comme modèle des sous-populations cooccurrentes de l'écotype HLII de *Prochlorococcus*. L'objectif était d'appréhender les forces évolutives à l'origine de la formation et du maintien du pangénome pour des populations bactériennes libres de l'environnement.

Le pangénome de *Prochlorococcus* apparaît ouvert à l'échelle populationnelle. Les gènes *core* et flexibles qui le composent dessinent un paysage génomique caractérisé par des régions conservées et variables. Cette organisation génomique s'accompagne d'une répartition non aléatoire des fonctions portées par les gènes flexibles et d'une dynamique évolutive différentielle, illustrée par une variation des contraintes sélectives et l'identification de points chaud de recombinaison, le long du génome.

Les résultats obtenus au cours de ces travaux mettent en évidence une distinction des trajectoires évolutives d'ensembles de gènes, spécifiques de compartiments génomiques particuliers, dans une population structurée. Ceci est conforme à une évolution de type barrière à la dérive. En outre, la structuration de l'information génétique le long du génome pourrait dépendre de la dynamique des flux de gènes entre les sous-populations, en particulier pour les gènes flexibles. Plutôt qu'une acquisition non aléatoire des gènes en fonction de leur localisation génomique, une probabilité différentielle de rétention des gènes transférés comme conséquence de la fluctuation de la taille efficace de la population le long du génome peut être envisagée.

Mots-clés : *Prochlorococcus*, pangénome, compartiments génomiques, dynamique évolutive, pressions de sélection, recombinaison homologue, transferts horizontaux de gènes, *Ne*, barrière à la dérive

Abstract

Evolutionary models associated with species concepts underlie selective sweep processes at the gene or genome scale. In this thesis, a reconsideration of the processes underlying the differentiation of free-living bacterial populations from the environment was carried out using co-occurring subpopulations of the *Prochlorococcus* HLII ecotype as models. The objective was to understand the evolutionary forces driving the formation and maintenance of the pangenome for environmental free-living bacterial populations.

The *Prochlorococcus* pangenome appears to be open at the population level. The core and flexible genes that make up the pangenome form a genomic landscape characterised by conserved and variable regions. This genomic organisation is accompanied by a non-random distribution of the functions carried by the flexible genes and by differential evolutionary dynamics, illustrated by a variation in selective constraints and the identification of recombination hotspots along the genome.

The results obtained in this work reveal distinct evolutionary trajectories of sets of genes, specific to particular genomic compartments, in a structured population. This is consistent with a drift-barrier model of evolution. Furthermore, the structuring of genetic information along the genome may depend on the dynamics of gene flow between subpopulations, especially for flexible genes. Rather than non-random acquisition of genes according to their genomic location, a differential probability of retention of transferred genes as a consequence of fluctuating effective population size along the genome might be considered.

Keywords: *Prochlorococcus*, pangenome, genomic compartments, evolutionary dynamics, selective pressures, homologous recombination, horizontal gene transfer, *Ne*, drift-barrier