Analyses de variations génomiques liées à la biogéographie des picoalgues Mamiellales

Jade Leconte

université PARIS-SACLAY

Analyses de variations génomiques liées à la biogéographie des picoalgues Mamiellales.

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°577,
Structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat : Sciences de la Vie et de la Santé
Unité de recherche : Université Paris-Saclay, Univ Evry, CNRS, CEA, Génomique métabolique, 91057, Evry-Courcouronnes, France.
Référent : Université d'Évry Val d'Essonne

Thèse présentée et soutenue à Évry, le 28/08/2020, par

Jade LECONTE

Composition du Jury

Christophe AMBROISE Professeur des universités, Université d'Evry Val d'Essonne (UMR 8071)	Examinateur & Président
François-Yves BOUGET Directeur de recherche CNRS, Sorbonne Université (UMR 7232)	Rapporteur & Examinateur
Frédéric PARTENSKY Directeur de recherche CNRS, Sorbonne Université (UMR 7144)	Rapporteur & Examinateur
Ian PROBERT Ingénieur de recherche, Sorbonne Université	Examinateur
Olivier JAILLON Directeur de recherche, CEA (UMR 8030)	Directeur de thèse

Thèse de doctorat

NNT : 2020UPASE013

Remerciements

Je souhaite remercier en premier lieu mon directeur de thèse, Olivier Jaillon, pour son encadrement, ses conseils et sa compréhension depuis mon arrivée au Genoscope. J'ai appris beaucoup à ses côtés, progressé dans de nombreux domaines, et apprécié partager l'unique bureau sans moquette du troisième étage avec lui. Merci également à Patrick Wincker pour m'avoir donné l'opportunité de travailler au sein du LAGE toutes ces années. J'ai grâce à vous deux eu l'occasion de plonger un peu plus loin dans le monde de l'écologie marine à travers le projet *Tara* Oceans, et j'en suis plus que reconnaissante.

Je tiens également à remercier les membres de mon jury de thèse, à la fois mes rapporteurs François-Yves Bouget et Frédéric Partensky qui ont bien voulu évaluer ce manuscrit, ainsi que Christophe Ambroise et Ian Probert qui ont également volontiers accepté de participer à ma soutenance. Une pensée spéciale pour Hervé Moreau, qui avait rejoint mon jury avec enthousiasme, j'aurais aimé qu'il puisse être avec nous également.

Un grand merci à tous mes collègues du Genoscope, ceux qui sont partis avant moi comme ceux qui resteront un peu plus longtemps. J'en oublierai forcément dans cette liste, mais ces cinq dernières années ont été pavées de nombreuses bonnes rencontres. Merci à Thomas, Yoann et Sarah, mes prédécesseurs, pour leurs nombreux conseils et leur amitié, merci à Kevin, Romuald, Julie et Paul, camarades thésards qui étaient là pour la majorité de ma thèse et avec qui j'ai partagé de nombreux bons moments, merci à Éric, Betina, et Quentin, le « bureau d'à côté » où j'ai toujours pu aller chercher opinions et bonne ambiance, merci à Benjamin de nous avoir rejoint si souvent dans nos expéditions culinaires, merci à Tom pour tous tes conseils. Un merci particulier à Samuel et Jana, avec qui j'ai tour à tour partagé mon bureau, venir travailler à vos côtés a toujours été un plaisir.

Merci aux membres de mon comité de thèse, Gwenaël Piganeau et Lionel Guidi, pour m'avoir guidée au cours de cette thèse et donné votre avis sur mes travaux. Merci aux nombreuses autres personnes avec qui j'ai eu l'occasion de collaborer, de *Tara* et d'ailleurs, sur différents projets passionnants. Merci à Catherine Sarlande, Nancy Delpech et Catherine Contrepois pour votre aide dans de nombreuses tâches administratives.

Enfin, merci à tous mes proches. Merci à ma mère, fan numéro un du projet *Tara*, et à mon père, qui m'ont toujours soutenue dans ce que j'entreprenais. Merci à ma sœur, qui commence tout juste sa propre thèse et va brillamment la réussir. Et merci à Tabatha qui partage ma vie pour le meilleur et pour le pire depuis de nombreuses années déjà, pour tout ce qu'elle fait pour moi. Merci à tous.

Table des matières

Introduction

I. Le plancton, acteur majeur de notre écosystème

1. Diversité des organismes planctoniques

Le "plancton" correspond aux organismes qui vivent dans de grandes étendues d'eau et sont incapables de nager contre le courant[1] . Ce terme regroupe donc des espèces marines de tailles très variables. Bien qu'un grand nombre soient microscopiques, les dimensions des virus et organismes planctoniques peuvent aller du picomètre pour les premiers à plusieurs mètres de long pour certaines méduses. Le plancton inclut à la fois des virus, des bactéries, des archées, des protistes et des métazoaires, et est donc défini par une niche écologique plus que par une classification phylogénétique ou taxonomique.

Parmi les groupes de micro-organismes qui constituent le plancton, on retrouve donc d'abord des procaryotes, les bactéries et les archées, qui ne possèdent ni noyau ni membrane interne, sauf pour l'embranchement des cyanobactéries qui ont une invagination de la membrane externe. On trouve ensuite des eucaryotes, des cellules plus complexes contenant un noyau et d'autres organites tels que les chloroplastes et les mitochondries, possédant leur propre ADN, délimités par des membranes. Les eucaryotes incluent les protistes, dont font partie les Mamiellales, et les métazoaires, plancton animal caractérisé par sa multicellularité et une alimentation hétérotrophe.

Le terme protiste désigne les eucaryotes unicellulaires, qui représentent plus de 85% de la diversité des eucaryotes[2]. On retrouve parmi eux les cinq supergroupes d'eucaryotes (basés sur la classification phylogénomique de Burki et al, 2014)[3] : les Amibozoaires, des protistes hétérotrophes se déplaçant majoritairement par contraction cytoplasmique, les Excavés, partageant une structure appelée cytosome permettant l'ingestion de fines particules alimentaires, les Opisthocontes, résultant du rapprochement récent de plusieurs taxons dont les champignons et les métazoaires, les SAR, contenant les Straménopiles, Alvéolés et Rhizaires, qui sont donc extrêmement diversifiés, et enfin les Archaeplastida ou lignée verte, possédant un chloroplaste et capables de photosynthèse, dont font partie les Mamiellales (Figure 1).

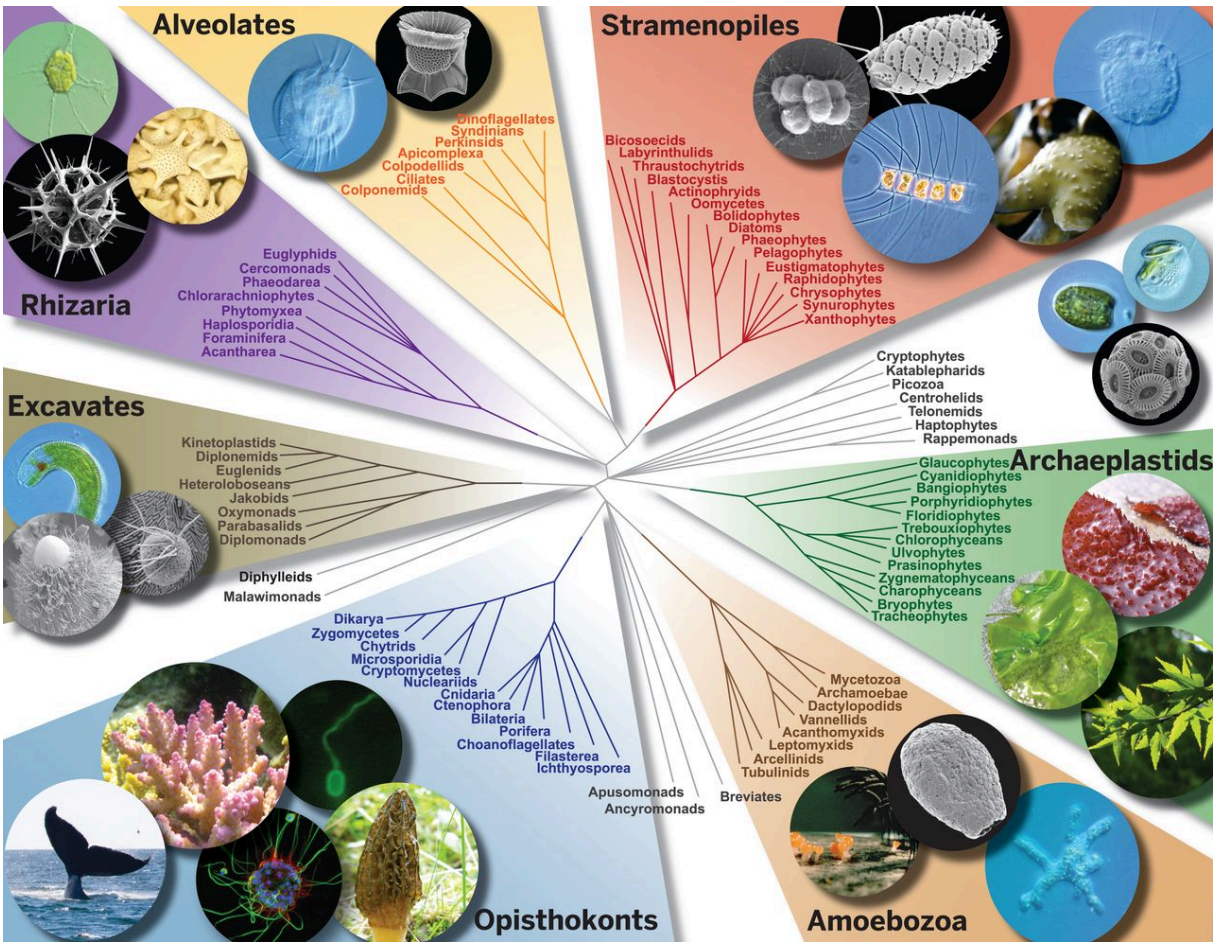


Figure 1 : Schéma de l'arbre phylogénétique des eucaryotes. Sept "supergroupes" sont indiqués par des couleurs différentes, les trois supérieurs parfois rassemblés en un groupe "SAR". Chacun contient de nombreuses lignées de protistes marins. Les images périphériques soulignent la diversité à la fois microbienne et multicellulaire. (Figure extraite de Worden et al, 2015[4]).

On peut aussi différencier le plancton permanent, dont tous les stades de vie sont passés sous forme planctonique, du plancton temporaire qui ne sera planctonique qu'à l'état d'embryon ou de larve comme les oursins, ou la plupart des poissons qui une fois adultes vivent sur les fonds marins ou nagent librement. Bien que définis par leur incapacité à lutter contre les mouvements d'eau, les organismes planctoniques sont généralement mobiles, soit par contractions de leurs corps, soit par la présence d'organes locomoteurs tels que les cils ou les flagelles ce qui leur permet de rester en suspension mais également de se déplacer pour attraper des

proies. Certains planctons sont capables de migrations verticales allant jusqu'à plusieurs centaines de mètres d'amplitude[5].

2. Interactions entre le plancton et l'environnement

On retrouve au niveau trophique deux principaux groupes de plancton : le phytoplancton et le zooplancton. Les premiers sont des autotrophes, la fraction végétale du plancton capable de réaliser la photosynthèse et donc de se nourrir de carbone inorganique en utilisant la lumière comme source d'énergie. Ils sont donc des producteurs primaires à la base de la chaine alimentaire marine, qui seront ensuite consommés par du zooplancton herbivore, alimentant à son tour le zooplancton carnivore, correspondant donc tous deux à la fraction animale du plancton. Ils serviront ensuite de ressource aux plus grands prédateurs tels que les poissons ou les mammifères marins[6].

En plus de ces organismes autotrophes ou hétérotrophes, on retrouve également parmi le plancton un grand nombre de mixotrophes, capables d'alterner entre la photosynthèse et l'intégration de proies selon les ressources nutritives disponibles dans leur environnement[7]. Certains réalisent la photosynthèse de manière dominante, ou d'autres à l'inverse sont majoritairement phagotrophes, mais sont capables d'utiliser le second mécanisme en complément pour survivre[8]. Certains organismes sont également capables de vivre en symbiose, par exemple certains phytoplanctons avec de plus larges protistes hétérotrophes tels que des foraminifères ou des radiolaires[9].

Le phytoplancton comporte à la fois des micro-algues unicellulaires et des bactéries photosynthétiques vivant dans les eaux de surface, qui représentent environ 1% de la biomasse mondiale mais sont responsables de presque 50% de la production d'oxygène, au même titre que l'ensemble des plantes terrestres[10]. De plus, ce plancton végétal permet l'absorption d'une grande partie du dioxyde de carbone atmosphérique: approximativement 30% du CO_2 produit par l'ensemble des activités anthropogéniques est absorbé par les océans[11]. Ces organismes sont donc directement liés à la limitation de l'acidification des océans et du réchauffement climatique.

Si le plancton meurt avant d'être mangé, les débris cellulaires vont couler des couches supérieures d'eau vers les fonds de l'océan faisant ainsi partie de la neige marine, de la même manière notamment que les déchets fécaux, pour être stockés

dans les sédiments. Ce processus, appelé pompe biologique, participe à faire des océans les plus importants puits de carbone sur Terre, des réservoirs naturels qui absorbent plus de carbone qu'ils n'en relâchent. Il s'agit du processus majeur permettant une distribution verticale du carbone, complété par ce qu'on appelle la pompe physique ou pompe de solubilité. Celle-ci participe à la distribution du carbone en profondeur, le refroidissement des eaux de surface sur les hautes latitudes favorisant la capacité du dioxyde de carbone atmosphérique à se dissoudre et augmentant sa densité, ce qui va permettre à ces particules lourdes de couler vers les fonds marins, limitant leur contact avec l'atmosphère.

Il est cependant à noter qu'une grande partie du carbone qui va passer dans la pompe biologique est décomposée en chemin par les bactéries hétérotrophes, qui vont reminéraliser une partie de ce carbone organique en CO_2 par la respiration et réincorporer une partie de la matière organique dissoute à la chaîne trophique : c'est la boucle microbienne. Cette boucle microbienne permet un recyclage efficace de la matière organique dissoute et augmente l'activité photosynthétique dans les systèmes limités par la concentration en nutriments[12,13]. Le zooplancton produit également du CO_2 par respiration (Figure 2).

Le plancton impacte donc l'environnement mais les variations environnementales peuvent également avoir un effet en retour. Par exemple, il a été montré que l'acidification des océans pourrait avoir un effet de rétroaction négative, entrainant une baisse d'efficacité dans le stockage du carbone chez le phytoplancton austral avant la fin du siècle[14]. La composition des communautés planctoniques et les propriétés biogéochimiques des masses d'eau sont ainsi étroitement liées dans un ensemble qu'on peut qualifier de "seascape" ou paysage marin[15].

Généralement, la croissance du plancton est limitée par la disponibilité des nutriments dans les gyres tropicaux et subtropicaux, tandis que la luminosité est souvent le facteur limitant dans les gyres subarctiques[16]. Ces paramètres impactent le phytoplancton, qui à son tour impactera les plus hauts niveaux de la chaine alimentaire. L'impact du réchauffement de l'eau sur les populations de phytoplancton est également un paramètre à prendre en compte, les variations de température tendant à diminuer la richesse des communautés[17] et à affecter le métabolisme de ces organismes[18]. Malgré la présence suffisante de macronutriments tels que le nitrate ou le phosphate, certaines régions ne contiennent que peu de phytoplancton. Elles corres-

pondent typiquement à des régions oligotrophes, pauvres en micronutriments et notamment en fer, un facteur limitant pour la croissance de ces organismes[19].

Il existe donc un lien direct entre d'une part la croissance, la biodiversité et le métabolisme du phytoplancton et d'autre part les conditions environnementales dans lequel il se trouve. En retour ce phytoplancton impacte son environnement en stockant le carbone dissout et en se plaçant à la base de la chaine alimentaire marine. Dans le cadre de cette thèse, nous nous intéresserons en particulier à une famille de protistes appartenant au phytoplancton : les Mamiellales.

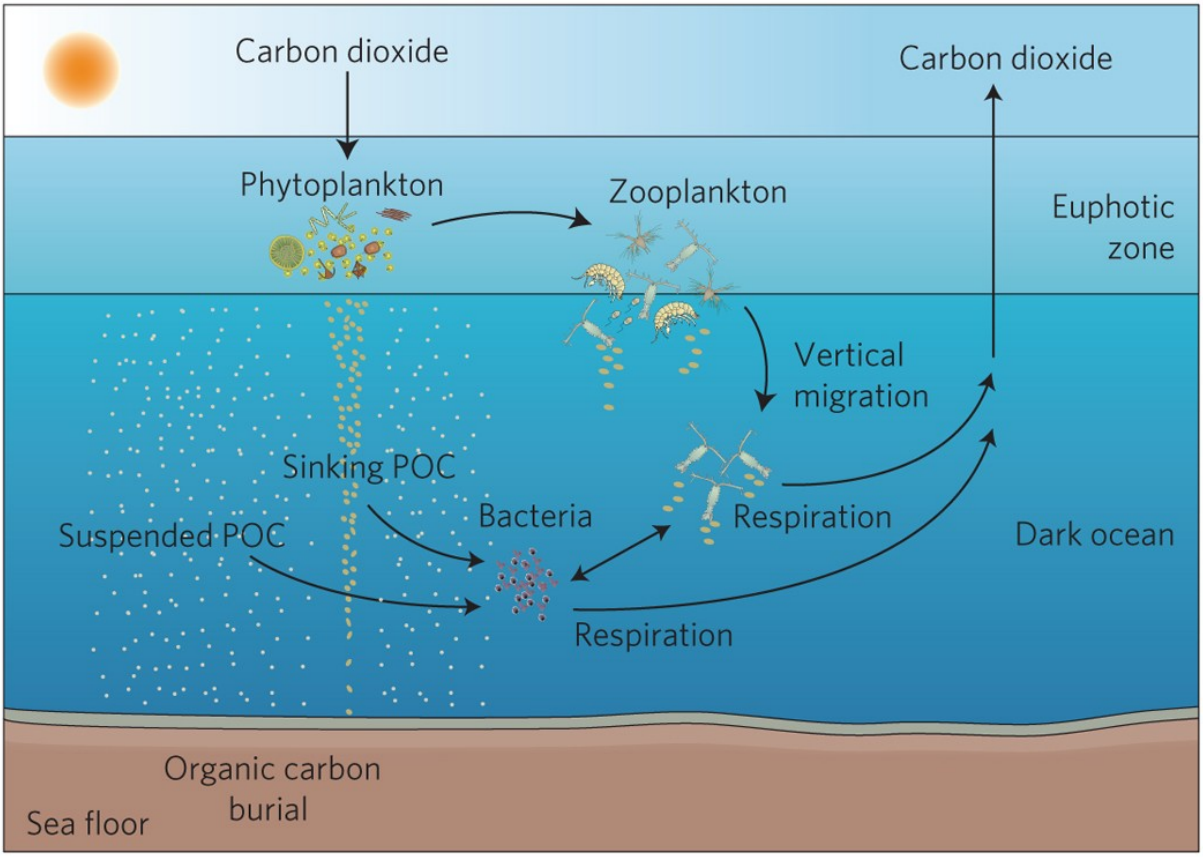


Figure 2 : Pompe biologique. Le phytoplancton dans la zone euphotique fixe le CO_2 en utilisant l'énergie solaire. Le carbone organique particulaire (POC) produit est consommé par le zooplancton herbivore, ou par des bactéries hétérotrophes se nourrissant des débris de phytoplancton. Le carbone organique est ensuite reconverti en CO_2 par respiration. Environ 1% seulement de la production de surface atteint les fonds marins. (Figure extraite de Herndl & Reinthaler, 2013)[20].

3. Cycle de vie du phytoplancton

Les espèces appartenant au phytoplancton peuvent alterner dans leur cycle de vie entre quatre phases distinctes : une phase de croissance, de reproduction sexuelle, de quiescence et de mort cellulaire. Ces transitions sont un élément clé pour comprendre l'histoire des espèces et impactent leur distribution écologique ainsi que leurs fonctions biogéochimiques. Pourtant, ce cycle de vie représente l'un des aspects les moins compris de la biologie des algues à l'heure actuelle, notamment car certaines de ces phases peuvent être épisodiques et difficiles à observer, comme par exemple les cycles de reproduction sexuelle[21].

Il a longtemps été assumé que les protistes n'avaient pas de phases sexuelles[22,23], les transitions vers ces phases étant souvent complexes à induire et contrôler en conditions de laboratoire[21], ce qui pose problème lors de l'utilisation de méthodes d'observation classiques, comme la microscopie optique. Le manque de partenaires ou la faible fréquence d'interactions sexuelles en laboratoire pourraient également être des facteurs entravant l'observation des stades sexuels[24]. De plus, en particulier pour les protistes picoplanctoniques, la difficulté est encore plus grande à cause de leur petite taille et du manque de caractéristiques morphologiques suffisantes pour permettre d'observer des différences entre les différentes phases.

Les outils apportés par la génomique moderne ont cependant permis de fortes présomptions sur la capacité de divers groupes de protistes à entrer en phase sexuelle (syngamie et méiose), tel que les Amibozoaires, les Dinoflagellés, les Diatomées, les Chrysophytes et les Mamiellales. Pour cela, des assemblages génomiques et des transcriptomes ont été utilisés pour dépister des gènes spécifiques ou liés au sexe impliqués dans des processus tels que la fusion de la membrane, la caryogamie et la méiose[25–32]. De plus, des données provenant de populations ont également montré des preuves de recombinaison[33,34].

Actuellement le sexe est donc considéré comme une caractéristique ancienne et omniprésente de la vie eucaryote[35]. D'après plusieurs études, la sexualité serait en effet l'état ancestral chez les eucaryotes, et l'asexualité serait apparue plus tard, et à plusieurs reprises de manière indépendante[36,37]. L'ensemble de ces analyses génomiques constituent des preuves indirectes de reproduction sexuelle, indiquant une "sexualité cryptique". Il manque donc encore des observations directes d'interactions

et de cycles sexuels pour de nombreuses espèces[24,35], et notamment pour les Mamiellales[34,38].

II. Les Mamiellales, membres clés du phytoplancton

1. Phylogénie des Mamiellales

Les Mamiellophyceae sont des algues vertes unicellulaires et représentent l'une des classes eucaryotes les plus importantes écologiquement[38]. Ils se découpent en trois ordres: les Monomastigales qui sont des organismes trouvés en eaux douces, les Dolichomastigales qui sont très diversifiés dans les eaux marines[39] mais représentent une partie mineure des Mamiellophyceae, et les Mamiellales, la lignée la plus importante[40,41], présents typiquement dans les eaux côtières[42] mais également trouvés dans la zone épipélagique des eaux océaniques[43] où ils peuvent atteindre des densités allant jusqu'à 10^3–10^5 cellules par mL[44]. Ils dominent également le phytoplancton eucaryote dans les eaux Arctiques[45].

Les Mamiellales se distinguent en deux familles, les Mamiellaceae et les Bathycoccaceae, qui comportent certaines des Chlorophytes les plus communes. La première regroupe le genre très répandu *Micromonas* ainsi que les moins connus *Mantoniella* et *Mamiella*. La seconde inclut le genre *Bathycoccus* également commun, et *Ostreoccocus*, l'eucaryote ayant la plus petite taille connue (Figure 3).

Au niveau morphologique, les Mamiellophyceae sont des organismes de moins de 3µm, dont une des caractéristiques bien que non-conservée dans certains genres, est la présence d'écailles organiques non-minéralisées (essentiellement polysaccharide) en forme de toile d'araignée. Ces écailles recouvrent des portions de la surface cellulaire ainsi que du flagelle s'il est présent chez *Bathycoccus*, *Mamiella*, *Mantoniella* et *Dolichomastix*. *Micromonas* et *Ostreococcus* ne possèdent pas d'écailles, *Bathycoccus* et *Ostreococcus* sont dépourvus de flagelles.

Les genres les plus étudiés parmi les Mamiellophyceae sont *Bathycoccus*, *Micromonas* et *Ostreococcus*. Ces organismes sont abondants en milieu océanique et peuvent être relativement aisément cultivés en laboratoire, facilitant leur adoption en tant que modèles biologiques et écologiques.

La classification des espèces du genre *Micromonas* a récemment été réévaluée[46], séparant ainsi la très répandue et diversifiée *Micromonas pusilla* en quatre espèces différentes: *Micromonas pusilla, Micromonas commoda*, *Micromonas bravo*, et *Micromonas polaris*, ainsi que deux espèces candidates. Il existe également quatre espèces d'*Ostreococcus*, *Ostreococcus lucimarinus*, *Ostreococcus tauri*[47], *Ostreococcus* spp. RCC809, et *Ostreococcus mediterraneus*[48], ainsi que deux espèces de *Bathycoccus*, *Bathycoccus prasinos*[40] et TOSAG39-1[49], aussi appelés clades BI et BII, qui possèdent contrairement aux autres Mamiellales toutes deux la même séquence d'ARN ribosomique 18S, les rendant impossibles à différencier par ce marqueur génomique classiquement utilisé.

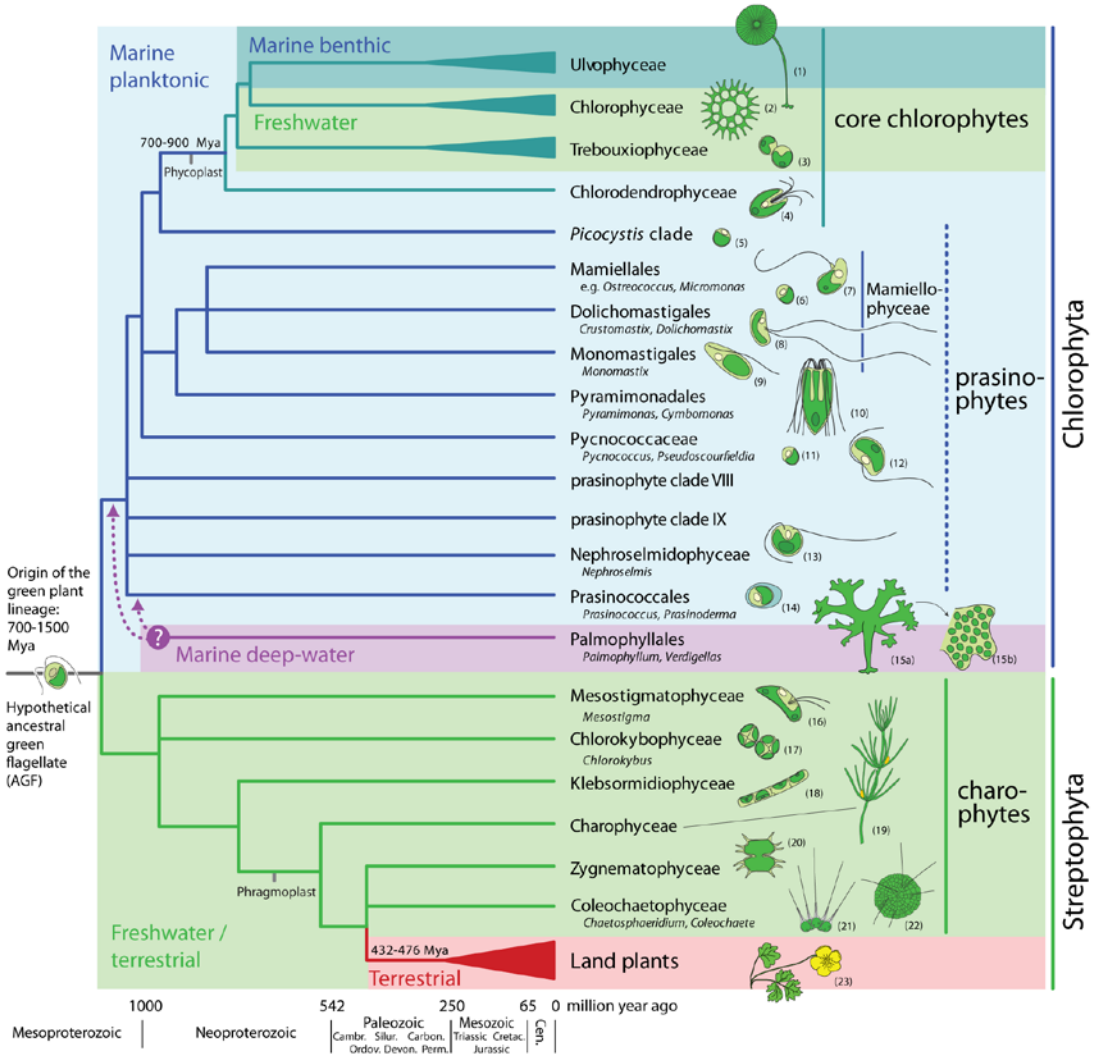


Figure 3 : Arbre phylogénétique des principales lignées vertes, en particulier aquatiques, basé sur des évidences moléculaires (Figure extraite de Leliaert et al, 2011[50]).

2. Biogéographie des Mamiellales

Les Mamiellales, même au sein d'un même genre, semblent se distinguer au niveau des niches environnementales qu'ils occupent. Par exemple, pour *Micromonas*, *M. pusilla* est typiquement océanique[51], *M. commoda* et *M. bravo* sont des espèces plutôt côtières[52], tandis que *M. polaris* est une espèce trouvée en milieu arctique dans lequel elle peut être totalement dominante[45].

Des analyses visant à étudier la réponse à la luminosité ont permis de séparer deux types d'*Ostreococcus* : adaptés aux hautes luminosités (*O. tauri*, *O. mediterraneus* et *O. lucimarinus*) et adaptés aux basses luminosités (*O.* RCC809)[53]. De plus, des études de la séquence ribosomique 18S suggèrent une différence de niche entre *O. tauri*, détecté plutôt dans les zones côtières ou les lagons, et *O. lucimarinus* et *O.* RCC809 qui sont plus largement répartis en haute mer[54,55]. Quant à *O. mediterraneus*, identifié plus récemment, il se trouve être l'espèce dominante parmi les Mamiellales en mer Méditerranée et dans des eaux chaudes sur des sites côtiers des deux rives de l'Atlantique[55].

Enfin concernant *Bathycoccus*, la souche RCC1105 correspondant à *Bathycoccus prasinos* a été définie comme adaptée à des milieux froids, abondant en milieu tempéré et arctique[45,56] tandis que TOSAG39-1 peut au contraire se trouver dans des environnements plus chauds[49,57]. Ces différences de préférences environnementales seront étudiées plus en profondeur dans le cadre de cette thèse.

Ostreococcus, contrairement à *Micromonas* et *Bathycoccus*, n'a jamais été reporté dans les eaux arctiques bien qu'il soit présent dans des eaux adjacentes recouvertes de glace saisonnièrement, telles que la mer Baltique[58] ou la mer Blanche[59].

De manière générale, le phytoplancton est défini comme abondant et très répandu[60,61] (Figure 4) et son succès écologique est en particulier démontré par un taux de croissance rapide et une contribution estimée à la production primaire[62]. Ces espèces de petite taille semblent également augmenter en abondance dans les eaux arctiques, tandis qu'on observe une diminution de plus gros phytoplanctons ces dernières années dans un contexte de changements climatiques[63]. Cependant, des analyses moléculaires indiquent que des modifications relativement mineures dans la disponibilité des nutriments ou de la luminosité peuvent entrainer des variations importantes d'abondance de Mamiellales[64]. Vu leur influence sur les écosystèmes planc-

toniques et leur apparente sensibilité, il est important d'étudier leur diversité ainsi que leur biogéographie à large échelle.

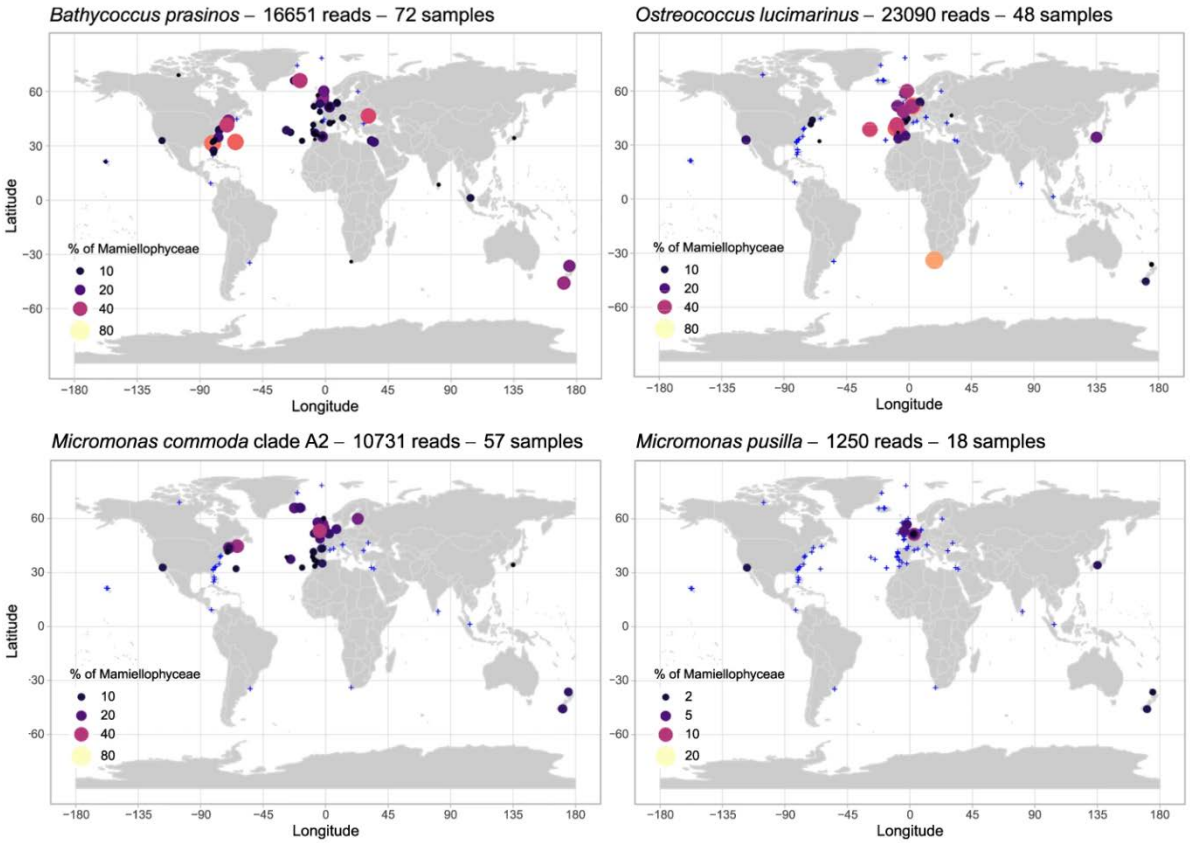


Figure 4 : Distribution des métabarcodes de quelques Mamiellales à partir des données du projet OSD (*Ocean Sampling Day*). Chaque point représente un échantillon, la taille et la couleur représentent la proportion de l'espèce parmi les Mamiellophyceae (Figure adaptée de Tragin et Vaulot, 2019[55]).

3. Génomes des Mamiellales

Les génomes des Mamiellales sont généralement de relative petite taille et compacts : en moyenne, moins de 200 paires de bases non-codantes séparent deux gènes. La taille de leur génome peut aller de 12.56Mb pour *O. tauri* [25],15Mb pour *B. prasinos*[65], et jusqu'à 22Mb pour *M. pusilla*[66]. Une caractéristique conservée entre tous les génomes de Mamiellales séquencés à ce jour est la présence de deux chromosomes "outliers", un premier de grande taille (BOC, *Big Outlier Chromosome*) et un second de plus petite taille (SOC, *Small Outlier Chromosome)*, présentant notamment un taux de GC plus bas que le reste du génome et un contenu atypique en gènes, dont la fonction est encore peu connue pour la plupart.

Cependant, il y a une surreprésentation consistante dans le SOC de gènes impliqués dans le transport, la synthèse et les modifications des carbohydrates[67] et il est probable que leurs contenus en GC, leurs fortes proportions en gènes spécifiques et leurs rythmes d'évolution plus rapide soient maintenus par les mêmes pressions évolutives[65,66,68].

Alors que le SOC peut être entièrement considéré comme outlier, le BOC ne fonctionne pas de la même manière. En effet, ce n'est pas l'ensemble du chromosome qui est considéré comme différent, mais une région marquée par un faible contenu en GC, entourée de deux régions à haut contenu en GC. La fonction des régions outlier du BOC reste obscure, mais ces caractéristiques seraient en accord avec l'hypothèse qu'elles puissent constituer un chromosome sexuel[65,69]. A ce jour, aucun autre chromosome outlier n'a été détecté chez un autre groupe d'algue verte, rendant leurs propriétés uniques et leurs rôles potentiels particulièrement intéressants.

Environ les trois-quarts des gènes codant des protéines présentent des similarités avec des gènes de fonctions connues, et la moitié des gènes sont homologues à des gènes de plantes. D'un point de vue phylogénétique, les Mamiellales divergent très tôt de la lignée verte (Figure 3) et généralement le nombre de gènes dupliqués est bas comparé aux plantes terrestres. Des analyses suggèrent que les *Micromonas* seraient les membres les plus anciens du groupe, *Bathycoccus* aurait émergé ensuite, suivi par Ostreococcus apparaissant dans la famille Bathycoccaceae.[38]

Le scénario le plus probable est que ces algues aient évolué sous une pression de sélection qui aurait conduit à une réduction de la taille de leur génome, entrainant de faibles nombres de familles de gènes, certaines disparaissant complètement si elles n'étaient pas absolument nécessaires à leur survie, comme par exemple les gènes de synthèse du flagelle chez *Ostreococcus*[70]. On pourrait supposer que cette tendance allant vers la réduction du génome permet une augmentation du ratio entre surface et volume, ce qui maximiserait l'efficacité d'absorption des nutriments[71], ou qui permettrait d'échapper à certains prédateurs préférant des proies plus grandes, telles que les huitres[72], *Ostreococcus* étant typiquement abondant dans les lagons ou elles sont cultivées.

Il est probable qu'en conséquence de cette sélection, les Mamiellales perdent certains traits les distinguant, et on peut donc s'attendre à une convergence morphologique chez des espèces différentes aux contraintes écologiques semblables et bio-

logiquement adaptées à un environnement similaire tel que les conditions de température ou la disponibilité des nutriments[48,53,73]. Cependant, on ne peut pas exclure l'hypothèse alternative, une évolution divergente de ces espèces par adaptation à différentes niches à partir d'ancêtres identiques.

III. De la génomique à la métagénomique, de nouvelles méthodes pour étudier les organismes dans l'environnement naturel

1. Métabarcoding

Le séquençage de "codes-barres" génétiques est une méthode très utilisée en écologie afin d'identifier les espèces présentes dans un échantillon environnemental sur la base de la séquence de certains gènes marqueurs. Dans l'idéal, la séquence du gène marqueur doit présenter une très faible variabilité entre individus d'une même espèce, mais une forte variabilité inter espèces. La distance entre les séquences doit également refléter la distance phylogénétique entre les espèces[74].

Typiquement, ce sont les gènes codant les sous-unités du ribosome qui servent de marqueurs, car leur conservation permet l'utilisation d'amorces universelles mais ils diffèrent suffisamment entre les espèces pour une identification taxonomique relativement précise. C'est souvent le gène codant la petite sous-unité du ribosome qui est utilisé, 16S chez les procaryotes et 18S chez les eucaryotes, ou même uniquement une séquence encore plus courte pour les seconds, les régions hypervariables V4 ou V9 (Figure 5).

Malgré la variabilité de ces gènes, la nature malgré tout conservée de ces régions fait que l'on sous-estime la diversité d'espèces dans une communauté, comme cela est le cas pour le zooplancton[75]. En effet ils ne sont malheureusement parfois pas suffisamment résolutifs pour différencier plusieurs espèces proches[76]. C'est par exemple le cas pour *Bathycoccus prasinos*, dont deux espèces présentent exactement la même séquence 18S.

D'autres marqueurs peuvent dans ce cas être utilisés pour essayer de les séparer, comme les séquences ITS (Internal Transcribed Spacers)[77] situées entre les gènes de la petite et de la grande sous-unité du ribosome.

L'étude de ces séquences permet donc d'étudier des échantillons complexes pour comparer leur composition taxonomique. C'est une méthode relativement simple et peu coûteuse, car l'amplification est ciblée grâce aux amorces et seul le gène marqueur est séquencé.

L'identification de la taxonomie d'une séquence repose essentiellement sur des bases de données existantes, telle que PR[2],"Protist Ribosomal Reference", qui recense de nombreuses séquences de référence liées à une assignation taxonomique, très utilisée pour l'identification de protistes[78]. Cette méthode permet aussi d'assigner des espèces inconnues à un taxon d'ordre supérieur à condition bien sûr que le gène marqueur soit suffisamment conservé pour être détecté. Il est aussi possible de créer de nouveaux groupes uniquement sur la base de leur séquence marqueur, on parle alors d'OTU pour *Operational Taxonomic Unit*. Il devient difficile d'identifier le phylum d'un organisme lorsque les séquences de référence ont moins de 80% d'identité avec celle analysée.

Le nombre de séquences assignées à une espèce permet une approximation de l'abondance relative de l'organisme dans l'échantillon, mais malheureusement ces estimations ne sont pas toujours exactes. Par exemple, certains organismes peuvent présenter plusieurs copies d'ADNr, faussant l'analyse quantitative.

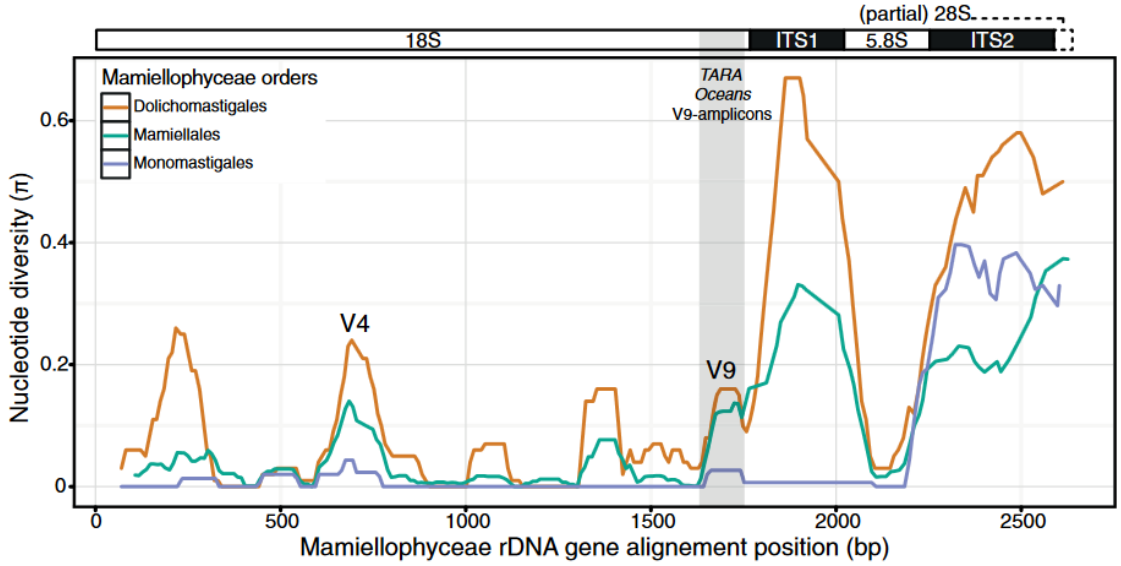


Figure 5 : Diversité nucléotidique de l'ARN ribosomique des Mamiellophyceae. Cette diversité est représentée par le nombre moyen de différences nucléotidiques par site entre deux séquences et a été calculée pour chaque lignée (Dolichomastigales, Mamiellales et Monomastigales) individuellement (Figure extraite de Monier et al, 2016[39]).

2. Métagénomique et métatranscriptomique

La grande majorité des micro-organismes marins ne peut pas être cultivée en laboratoire. De plus, étudier un organisme en culture ne permet pas d'accéder aux variations génomiques liées à son environnement d'origine. Il est donc important, en plus de chercher des moyens de garder ces espèces en laboratoire[79], de pouvoir les étudier à partir d'échantillons naturels, ce qui permet également d'étudier non pas un seul organisme, mais l'ensemble d'une communauté.

La métagénomique est une méthode consistant à séquencer l'ensemble des séquences d'ADN appartenant aux organismes présents dans un échantillon environnemental, tandis que la métatranscriptomique est l'équivalent en séquençage d'ARN, permettant ainsi de connaître dans le premier cas l'ensemble des gènes présents dans un échantillon, et dans le second de savoir lesquels sont exprimés au moment du prélèvement.

Différentes analyses peuvent être réalisées à partir du séquençage d'une communauté d'individus. Il est par exemple possible d'estimer l'abondance de certains gènes ou d'espèces présents dans un échantillon[80], typiquement en utilisant une séquence ou un génome de référence afin de compter les séquences qui en sont suffisamment proches, de manière similaire au métabarcoding vu précédemment, mais sur la base de n'importe quel gène voir d'un génome entier. Il est aussi possible de comparer des gènes présents dans plusieurs échantillons différents afin d'en étudier les variations quantitatives et les différentes mutations. L'équivalent est possible en transcriptomique, permettant donc de voir si un gène est plus ou moins exprimé entre plusieurs échantillons.

Cependant, l'abondance d'un organisme étudié de cette manière sera relative, car dépendante de l'abondance des organismes et de la taille des génomes présents dans la communauté. En effet, notamment chez les eucaryotes, les tailles de génomes peuvent être très variables, allant des 12.56Mb d'O. tauri à environ 200 Gb estimés pour le dinoflagellé *Prorocentrum mican* [81]. Un autre inconvénient de cette méthode est la nécessité d'une référence, on ne peut donc s'intéresser qu'aux abondances de taxons déjà connus.

3. Single-cell

Le Single-cell, ou génomique en cellule unique, consiste à séquencer le génome d'un organisme à partir d'une seule cellule prélevée dans l'environnement, sans étape de culture[82]. On peut ainsi étudier le génome d'un organisme non-cultivable. Cette méthode est également utilisée dans le domaine médical pour le génotypage de lignées cellulaires, comme des cellules cancéreuses[83]. Cette technique passe par l'isolation d'une cellule puis l'amplification de son génome avant de le séquencer puis de l'assembler. L'amplification est nécessaire car une seule cellule ne peut pas contenir suffisamment de matériel génétique pour un séquençage direct. Le Single-cell s'est donc développé avec l'apparition par déplacements multiples ou MDA (*Multiple Displacement Amplification*)[84] qui permet l'obtention de plusieurs microgrammes à partir d'une molécule d'ADN, avec un taux d'erreurs inférieur et des fragments plus grands que la DOP-PCR (*Degenerate Oligonucleotide Primed-Polymerase Chain Reaction*)[85], une autre méthode d'amplification, variante de la PCR utilisant des amorces dégénérées. La MDA a aussi l'avantage d'être isotherme, pouvant donc être réalisée à 30°C, ce qui facilite sa mise en œuvre.

Malheureusement ce type d'amplification présente deux biais majeurs: la production de séquences chimériques[86], formées à partir de deux régions distantes du génome amplifié, et une différence d'amplification entre les fragments, certains l'étant beaucoup plus que d'autres, causant des biais de représentation de ces fragments.

Afin de tenter de palier à ces problèmes, d'autres méthodes sont apparues, comme l'amplification MALBAC (*Multiple Annealing and Looping Based Amplification Cycles*)[87] qui permet d'obtenir une couverture plus uniforme du génome et d'obtenir une haute reproductibilité. En revanche, environ un tiers des variations nucléiques ne sont pas détectées par cette méthode.

Enfin, une autre possibilité pour améliorer le séquençage de cellules est l'utilisation de plusieurs cellules appartenant à la même espèce, augmentant ainsi la quantité de matériel génétique disponible et permettant d'améliorer la détection de variants. Chaque cellule est séquencée indépendamment avant d'être co-assemblées, améliorant ainsi la taille du génome reconstitué par l'apport d'un plus grand nombre de lectures.

IV. Génomique à l'échelle des populations

1. Présentation de l'étude des populations

Une population correspond à un groupe d'organismes appartenant à la même espèce et vivant dans une zone géographique particulière. L'étude de la génomique des populations s'intéresse donc aux différentes caractéristiques génomiques entre plusieurs populations par la comparaison de séquences d'ADN. Originellement proposées en 1998 dans le cadre de l'analyse des maladies humaines[88], ces études sont maintenant devenues incontournables dans l'exploration de l'évolution des génomes. Ce domaine s'est notamment développé suite à l'évolution des technologies de séquençage à haut-débit, permettant la détection de variations à un nombre croissant de positions, mais aussi grâce à de nouvelles approches statistiques développées spécifiquement pour ce type d'études[89–91]. On analyse notamment des phénomènes tels que l'adaptation, qui permet à un organisme de survivre au mieux à son environnement, la spéciation, processus évolutif par lequel des populations deviennent des espèces distinctes, ou encore la structure des populations, par la présence systématique de fréquences alléliques différentes entre deux populations.

Une approche très connue de l'étude des variants est l'étude d'association pangénomique, ou GWAS (Genome-Wide Association Study). Celle-ci consiste à génotyper un grand nombre de marqueurs génétiques chez un grand nombre de sujets, afin d'essayer de lier des mutations particulières à un phénotype. Typiquement, ce type d'approche est appliqué aux maladies génétiques humaines afin de découvrir des variants responsables de ces maladies[92], ou encore à l'agronomie afin d'améliorer certaines cultures[93]. De nouvelles approches de prédiction de mutations, notamment grâce à l'apprentissage automatique ou "machine learning", commencent aujourd'hui à compléter ce type d'études[94]. Cependant il s'agit ici de méthodes visant à lier un variant à un phénotype plutôt qu'à étudier les variants communs à une population afin d'analyser des phénomènes d'adaptation.

La génomique des populations a en particulier été définie comme l'étude simultanée de nombreux loci et régions génomiques afin de mieux comprendre les rôles des processus évolutifs (tels que les mutations, la dérive génétique, le flux de gènes et la sélection naturelle) qui influencent la variation parmi les génomes et les

populations[95,96]. Cette définition met l'emphase sur la compréhension d'effets locus-spécifiques tels que la sélection par rapport à des analyses à l'échelle du génome afin d'améliorer notre compréhension de l'évolution adaptative et de la spéciation.

Lors de l'étude de la structure des différentes populations d'une espèce, la variabilité du génome va donc être un point clé pour la compréhension de leur histoire. Au cours du temps, plusieurs mécanismes évolutifs sont responsables de l'évolution des génomes d'une espèce. Ces mécanismes, tels que la génération de variants structuraux (insertions, délétions, inversions, translocations) ou l'accumulation de mutations ponctuelles (Single Nucleotide Variants, SNV), impactent les génomes à différentes échelles. Alors que les premiers modifient de manière importante leur structure, le second amène à une divergence progressive de la séquence des génomes et ainsi des séquences protéiques entre les populations.

Concernant plus particulièrement le phytoplancton, les théories prévoient qu'à cause de leur grande taille de populations et leur taux de réplication rapide, ces organismes sont capables d'une réponse évolutive dans des délais relativement courts[97]. De plus, plusieurs auteurs ont suggéré qu'une distribution apparaissant cosmopolite et le maintien d'une forte abondance étaient liée à des changements de composition du génome au sein des espèces chez de nombreux phytoplanctons[98–100]. Cette hypothèse a notamment été testée chez les cyanobactéries pour lesquelles l'existence d'écotypes ayant une distribution non-aléatoire liée à leur habitat est bien établie[101,102]. De plus, des phénomènes d'adaptation physiologique et génomique associés à certains écotypes de cyanobactéries ont été documentés, suggérant un rôle de la sélection naturelle et de l'évolution adaptative dans le maintien ou la création de cette distribution non-aléatoire[103–105].

2. Méthodes d'analyse de la structure des populations

Les technologies avancées de séquençage à haut débit tel que RAD-Seq (*Restriction-site Associated DNA sequencing*) permettent désormais d'identifier des centaines, voire des milliers de loci polymorphes sans génomes de référence. Il s'agit d'une méthode alternative au séquençage NGS du génome entier pour la détection de variations d'ADN en ciblant des régions adjacentes à des sites de restriction communs répartis sur tout le génome, réduisant ainsi notablement les coûts de séquen-

çage. Par exemple appliquée à des organismes zooplanctoniques, cette stratégie a permis l'identification de loci sous pression de sélection et de structures de populations significatives en milieu océanique[106], mais n'a malheureusement pas permis de lier ces éléments à des fonctions biologiques par manque de génome de référence.

Les approches de séquençage de génomes entiers, plus coûteuses, ont été majoritairement appliquées aux humains, plantes, animaux ou micro-organismes d'intérêt agronomique ou de santé publique. Elles fournissent une vision globale des régions génomiques ciblées par les processus de sélection et sont moins biaisées que les technologies telles que le RAD-Seq[107,108] basées sur la capture de certaines régions.

Une autre méthode possible pour analyser la diversité d'une espèce à travers plusieurs échantillons environnementaux est la métagénomique. En effet, avec un génome de référence afin de recruter des séquences appartenant à une espèce d'intérêt, il est possible d'étudier sa diversité dans un échantillon environnemental. Ce type d'analyses, moins courant mais émergent ces dernières années[109–111], nécessite la mise en place d'un protocole adapté afin de tenir compte notamment des autres espèces présentes dans les échantillons. Récemment, au niveau planctonique, la structure des populations de copépodes a été étudiée en Méditerranée en utilisant les échantillons de *Tara* Oceans[112]. C'est ce type de méthodologie qui sera également appliquée dans le cadre de cette thèse sur le phytoplancton *Bathycoccus prasinos*.

Au niveau des méthodes statistiques d'analyse de données, la structure des populations ainsi que les loci sous pression de sélection sont majoritairement étudiés à l'aide de statistiques telles que les indices de fixations (Fst, Gst…), les plus connus et utilisés. Ces métriques originellement développées par Wright[113] puis ajustées au cours du temps pour convenir à différents types d'études[114,115] se basent sur l'analyse de la variance des fréquences alléliques entre différentes populations. Mais elles ont initialement été pensées pour des organismes diploïdes, utilisant notamment les variations d'hétérozygotie ce qui n'est pas applicable à nos génomes haploïdes. Des méthodes prenant ces derniers en compte ont émergé, mais elles reposent surtout sur l'étude des positions polymorphiques bialléliques et ne prennent pas forcément en compte les variations fixées dans les populations[116]. De plus, la validité de l'ensemble de ces indices a été très discutée dans le domaine de l'écologie molécu-

laire en tant que mesure de différentiation génétique, car ils se basent en particulier sur des hypothèses fortement dépendantes de modèles démographiques. Par exemple, les populations sont dans la plupart des cas supposées de tailles égales et ayant une absence de structure, ou même étant indépendantes les unes des autres. Ces approximations sont pourtant irréalistes pour des populations naturelles[117,118], c'est pourquoi dans le cadre de l'utilisation d'échantillons environnementaux pour étudier la structure des populations d'un génome tel que *Bathycoccus*, nous éviterons d'utiliser ce type de statistiques. Nous préfèrerons privilégier des approches telles que le nombre de variations d'acides aminés comparé aux variations nucléotidiques, l'étude des scores BLOSUM associés ou des distances génomiques simplement basées sur le contenu allélique afin d'éviter tout apriori sur les populations et l'évolution de l'organisme d'intérêt autre que l'utilisation du code génétique et les propriétés physico-chimiques des acides aminés.

3. Historique des applications aux Mamiellales

Il existe une grande diversité au sein de la classe des Mamiellophyceae[76]. Malgré la proximité phylogénétique des différentes espèces, il est à noter que chez *O. tauri* et *O. mediterraneus*, il existe par exemple de nombreuses souches génétiquement distinctes.

Le taux de mutations spontanées de quelques Mamiellales a été la cible d'une étude récente[119], afin d'essayer pour la première fois d'en obtenir une estimation pour aider à la compréhension des changements évolutifs qu'ont subi ces algues vertes qui possèdent par exemple de nombreux gènes espèce-spécifiques. Ces taux de mutation ont donc été établis pour *B. prasinos*, *M. pusilla*, *O. tauri* et *O. mediterraneus*, et il a été estimé comme attendu qu'ils étaient plus élevés dans les régions intergéniques que dans les régions codantes. Ces taux semblent également augmenter dans les régions ou le contenu en GC du génome dévie de sa valeur moyenne.

Des expériences d'accumulation de mutations ont également été étudiées sur ces mêmes espèces[120], les plaçant en conditions normales ou de stress afin d'étudier leur capacité potentielle d'adaptation. Seul *O. tauri* a montré une diminution de la croissance de sa population prise comme mesure de "fitness" pour cette expérience en conditions classiques, mais l'accumulation de mutations augmente lorsque les lignées sont exposées à des conditions de stress. Les coefficients de sélection, estimés à partir du nombre de divisions cellulaires quotidiennes, varient signi-

ficativement entre les différentes conditions environnementales, montrant des effets bénéfiques et néfastes des mutations spontanées.

Enfin, une étude de génomique des populations a été conduite récemment sur une population naturelle d'*O. tauri*[33], étudiant le polymorphisme entre 13 génomes complets de cellules isolées en mer Méditerranée. Celle-ci a montré une grande diversité génétique entre les organismes analysés, et appuyé notamment l'hypothèse de l'existence d'une position "mating-type" dont il existerait deux versions sur le grand chromosome outlier. Cependant, cette étude reste localisée à une zone géographique restreinte, et ne permet donc pas l'étude des populations dans des conditions environnementales très variées.

V. Le projet *Tara* Oceans

1. Historique des principaux projets océanographiques

C'est à la fin du XIXème siècle, entre 1872 et 1876 qu'est lancée la première expédition d'océanographie moderne : l'expédition du H.M.S. Challenger, un navire britannique. Elle avait pour but d'améliorer notre connaissance des fonds marins et d'étudier la répartition de la vie animale, ainsi que les paramètres physico-chimiques dans différents environnements. Pour cela, les scientifiques participant avaient mis en place un protocole d'échantillonnage incluant des analyses physiques, chimiques et biologiques. De nombreuses images d'organismes planctoniques ont été dessinées durant ces quatre années par le naturaliste allemand Ernst Haeckel[121] (Figure 6), et environ 4700 espèces jusqu'alors inconnues ont pu être décrites lors de cette expédition, dont certaines à plus de 5500 mètres de profondeur. Mais à cette époque, les différents courants marins et régions océaniques étaient très mal connus, et les méthodes de quantification des organismes trop complexes à mettre en place.

De nos jours, le séquençage à haut débit permet une toute autre approche de ce type d'expéditions scientifiques. En 2004, c'est le *Global Ocean Sampling*[122] qui est lancé afin d'étudier la diversité microbienne grâce à la génomique et de mieux comprendre les rôles des microorganismes dans les écosystèmes naturels. Durant deux ans, des échantillons ont été prélevés dans l'océan Pacifique et dans l'Atlantique Nord à bord du Sorcerer II, chacun séparé de plus de 200 miles afin de permettre l'étude d'environnements différents. L'analyse des données récoltées a notamment

permis de démontrer le manque de connaissances et d'informations taxonomiques sur les micro-organismes marins[123], mais aussi de détecter de nouvelles familles de protéines. Il a également été noté que des différences génomiques existaient entre les espèces abondantes dans différents environnements, suggérant un mécanisme d'adaptation des micro-organismes. Des échantillons ayant un contenu similaire peuvent également se trouver à de très grandes distances géographiques. Ce projet a donc fourni une grande quantité de données à la communauté scientifique, permettant de réaliser l'ampleur de la diversité des communautés planctoniques mais suggérant aussi l'étude du lien entre celles-ci et les paramètres environnementaux qui composent leur milieu naturel[124,125].

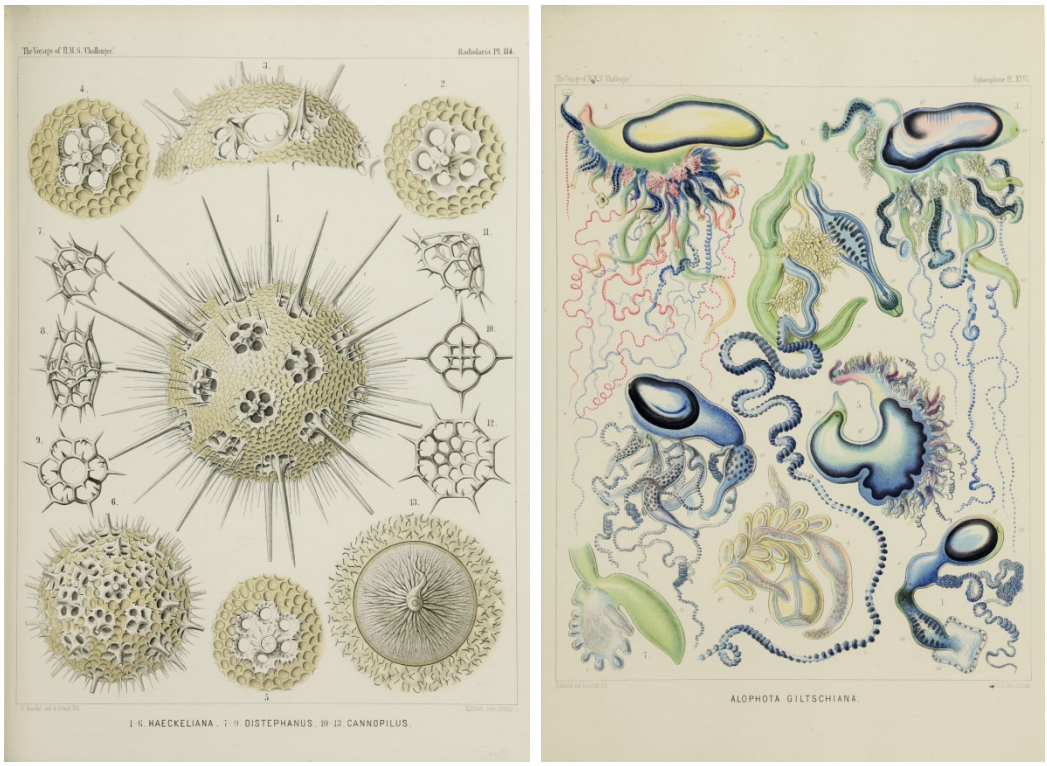


Figure 6 : Dessins d'organismes planctoniques réalisés par Ernst Haeckel durant l'expédition du H.M.S. Challenger. Images extraites de *Art Forms from the Abyss*[121].

Le projet *Tara* Expéditions est à l'origine de plusieurs expéditions scientifiques qui ont eu lieu depuis sa création en 2003, au moment de l'achat de la Goélette *Tara* (Figure 7) par Étienne Bourgois. Il lance ce projet afin d'encourager une prise de conscience de la fragilité des environnements marins et de développer une connaissance de haut niveau sur l'océan. La toute première expédition, baptisée *Tara* Arctic, commence en septembre 2006 pour une dérive d'un an et demi sur l'océan Arctique afin d'étudier les phénomènes de changements climatiques des hautes latitudes. Sa

coque en forme de "noyau d'olive" et en aluminium est prévue pour résister aux pressions extrêmes de la banquise ainsi qu'aux faibles températures, et peut donc se laisser prendre par la glace avant de dériver.

C'est ensuite que l'expédition *Tara* Oceans fut lancée en Septembre 2009 pour se terminer en Mai 2012, avec pour objectif d'étudier les écosystèmes planctoniques des océans ouverts à l'échelle globale, c'est-à-dire pour tous les organismes présents (virus, bactéries, archées, eucaryotes unicellulaires, zooplancton), dans tous les océans. Puis dans la continuité de cette étude de la diversité océanique, le bateau est reparti échantillonner le cercle polaire, cette seconde partie de l'expédition a débuté en Mai 2013, le bateau réalisant une circumnavigation de l'océan arctique de 25 000 km en 6 mois.

Depuis, plusieurs autres expéditions ont eu lieu comme *Tara* Méditerranée en 2014, analysant l'impact des micro-plastiques sur l'écosystème de la mer Méditerranée[126]. Plus récemment, *Tara* Pacific s'est concentrée sur l'avenir du corail marin dans un contexte de réchauffement climatique, avec une expédition ayant eu lieu de 2016 à 2018 à travers l'océan Pacifique[127].

Dans le cadre de cette thèse, l'ensemble des échantillons analysés provient de l'expédition *Tara* Oceans, incluant ceux de la seconde partie récoltée dans l'océan arctique, ayant été rendus disponibles au cours de la thèse.

Figure 7 : Goélette Tara, ayant servi à la collecte des échantillons lors de plusieurs expéditions dont notamment *Tara* Oceans.

2. Le parcours de *Tara* Oceans

La goélette *Tara* a donc parcouru les océans afin de prélever des échantillons planctoniques provenant de nombreux sites sur l'ensemble du globe. Le choix du parcours de l'expédition (Figure 8) a été motivé par des considérations scientifiques et météorologiques précises. En effet, il fallait choisir un parcours qui permette de passer par des régions océaniques particulièrement intéressantes et variées, pour tenter d'établir par exemple une corrélation entre la structure des écosystèmes planctoniques et l'environnement. *Tara* étant un voilier, il fallait de plus sur une période de deux ans et demi, arriver au bon endroit au bon moment afin de profiter de conditions météorologiques permettant d'effectuer l'échantillonnage. Par exemple, il a fallu utiliser au maximum les vents et courants portants afin d'éviter la saison des cyclones ou typhons dans les régions tropicales.

Le navire est donc parti de Lorient, prélevant les premiers échantillons le long des côtes du Portugal avant de se diriger vers le canal de Gibraltar et d'entrer en Méditerranée. Le trajet se poursuit par la mer Rouge, qui subit une forte évaporation et a donc des eaux particulièrement chaudes et salées. Au nord, elles sont relativement oligotrophes (pauvres en plancton photosynthétique) alors qu'elles s'enrichissent en nutriments venant de l'océan Indien et donc en vie planctonique au sud. Le nord de l'océan Indien est une zone acide qui constitue aussi le réservoir d'eaux profondes pauvres en oxygène le plus important du monde.

Tara est ensuite passée par le canal du Mozambique avant de progresser vers le Cap de Bonne Espérance en suivant le courant des Aiguilles, ou Agulhas. Ce courant marque la délimitation entre l'Indien et l'Atlantique sud. Au contact des eaux de l'Atlantique, plus froides, le courant des Aiguilles génère au niveau du cap des tourbillons de plusieurs centaines de kilomètres de diamètres, appelés anneaux d'Agulhas. Des prélèvements ont été effectués au niveau d'un second environnement d'intérêt, l'upwelling du Benguela (zone de remontées d'eaux froides riches en nutriments).

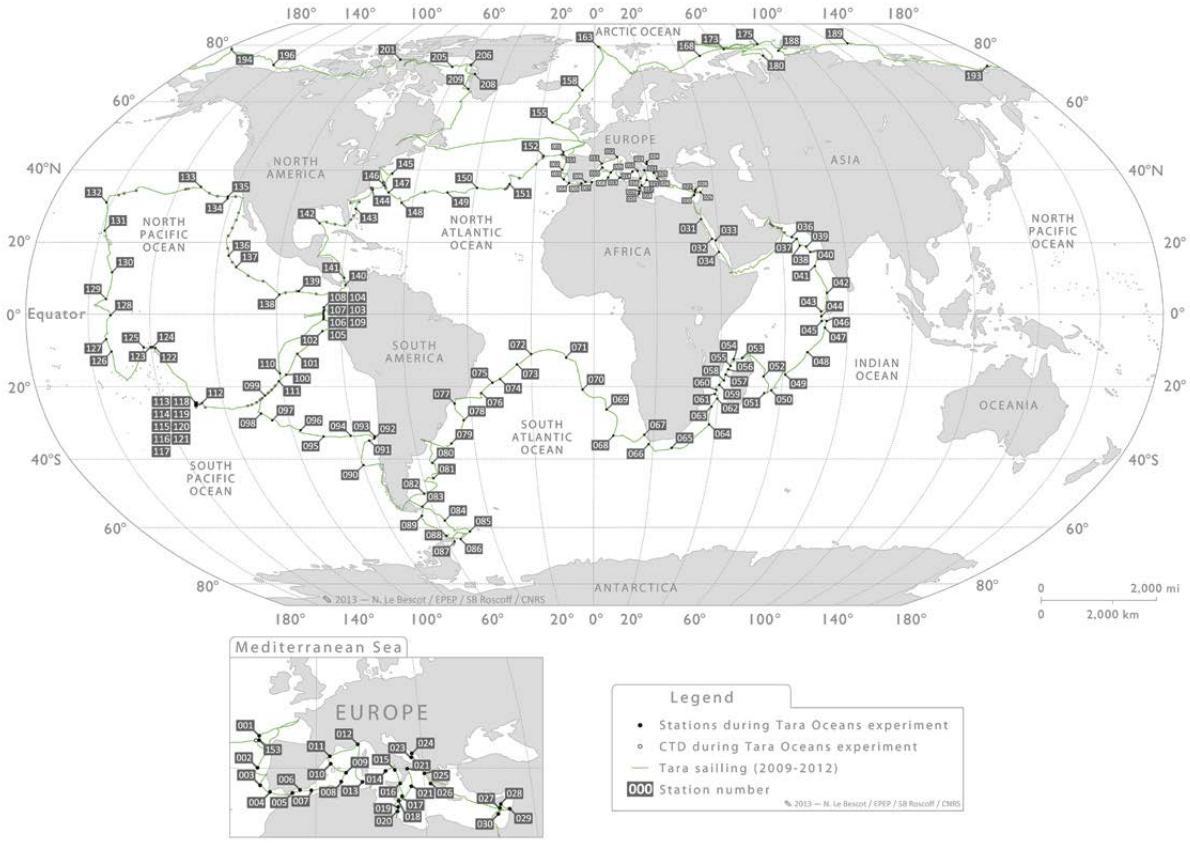


Figure 8 : Parcours de l'expédition *Tara* Oceans (2009-2013). En vert est indiqué le trajet du bateau, annoté d'étiquettes pour chaque station de prélèvement.

Le bateau a alors suivi le grand gyre sud-atlantique jusqu'à Buenos Aires, qui comprend le courant sud-équatorial, le courant du Brésil et les régions oligotrophiques, avant de se diriger vers l'océan austral pour échantillonner plusieurs points du canal de Drake et de la mer de Weddell. Le parcours s'est poursuivi vers le nord, passant le long de la côte Ouest d'Amérique du Sud qui comprend un nouvel upwelling, celui du Chili, avant d'entamer la traversée du système désertique du gyre du Pacifique sud-est connu pour être pauvre en nutriments, puis le gradient d'acidité ente l'île de Pâques et les Galápagos. La transition s'est ensuite effectuée vers l'Ouest et les îles Gambier, avant de remonter vers le nord et la seconde zone majeure du Pacifique, le gyre subtropical, plus riche en nutriments. C'est également une région de forte concentration en plastiques dérivants, le célèbre "continent du plastique", notamment entre Hawaii et San Diego. L'expédition a alors suivi les côtes d'Amérique du nord vers le canal de Panama, prélevant au passage l'upwelling de Californie. Cette région est connue pour ses eaux anoxiques proches de la surface, ce qui n'est pas le cas de l'autre côté du canal. Les eaux de surface se refroidissant progressive-

ment le long du courant qui circule des Caraïbes vers le Gulf Stream, tout en devenant plus salées, avec une faible concentration en chlorophylle. Des prélèvements ont également été effectués dans le golfe du Mexique avant de rejoindre l'est de la Floride pour longer les Etats-Unis et enfin partir le long du Gulf Stream. Cette zone permet de voir comment les organismes sont dispersés par le courant, avant de passer sur la caractérisation du gyre de l'Atlantique nord pour enfin regagner le port de Lorient.

Lors de cette dernière expédition *Tara* Oceans, l'océan Arctique avait manqué dans l'effort de collecte de plancton réalisé sur tous les océans de la planète. Il y avait donc un intérêt important à compléter ces études de biodiversité par la province Arctique, c'est pourquoi *Tara* après quelques mois est parti compléter en six mois le parcours précédant d'une circumnavigation par les passages du nord-est et du nord-ouest (Figure 9). Le trajet s'est essentiellement effectué en lisière de banquise, là où l'activité planctonique est la plus importante. Le bateau est donc à nouveau parti de Lorient, suivant les courants nord-atlantiques jusqu'au nord de la Norvège, la température baissant rapidement, avant de suivre les côtes de la Russie en passant par l'archipel de François-Joseph dans la mer de Barents.

L'expédition gagne alors une région plus difficile couverte par la glace entre la mer de Kara et celle de Laptev. La seconde est une région comportant par endroits des eaux très profondes, cette station d'échantillonnage est donc particulièrement intéressante. *Tara* traverse alors une partie de la mer de Sibérie Orientale se trouvant sous l'influence de nombreuses masses d'eau douce provenant des grandes rivières de Sibérie, avant de passer du côté de l'Alaska et de la mer de Beaufort. L'étape suivante est le suivi des côtes canadiennes jusqu'à la baie de Baffin, à proximité du Groenland, couverte de glace une grande partie de l'année.

En redescendant en direction du passage nord-ouest vers l'atlantique, des échantillons sont notamment prélevés dans des « brines », eaux de surface très froides issues de l'hiver précédent qui coulent jusqu'à atteindre une eau de même densité. Finalement, le bateau traverse le détroit de Davis et rejoint la mer du Labrador, traversée par trois courants principaux : un courant froid qui remonte le long de la côte du Groenland, un autre qui descend le long du Labrador et enfin un troisième d'origine atlantique plus chaud. Ce seront les derniers prélèvements pour *Tara* qui regagne ensuite, une fois encore, le port de Lorient en passant par l'Atlantique nord.

Figure 9 : Détails de la seconde partie de l'expédition *Tara* Oceans se concentrant sur le cercle polaire (2013). Le trajet du bateau est indiqué en orange.

3. Méthode d'échantillonnage

Un total de 210 stations d'échantillonnage, dont 154 pour la première partie et 56 pour le cercle polaire, ont été prélevées durant l'expédition *Tara* Oceans. Le choix de la position exacte de chacune de ces stations a été réalisé en temps réel, en utilisant les données des instruments de bord ainsi que des données satellitaires afin de sélectionner au mieux les régions d'intérêt[128]. L'échantillonnage s'est concentré sur trois profondeurs différentes. La sub-surface d'abord, entre 3 et 9m de profondeur, et la DCM (*Deep Chlorophyll Maximum*) se situant à la profondeur avec la concentration en chlorophylle la plus importante au moment du prélèvement, généralement entre 20 et 100m. Enfin des échantillons ont également été récupérés en zone mésopélagique, entre 300 et 400m, où la lumière est presque absente.

Pour chaque station, un protocole a été mis en place afin de récolter divers types d'organismes, les séparant en différents filtres de tailles correspondants. Par exemple les plus petites fractions correspondent aux communautés virales (0-0.2µm),

puis bactériennes (0.22-3µm). Viennent ensuite les protistes (0.8-5µm) et le zooplancton réparti sur plusieurs filtres (5-20µm, 20-180 µm, 180-2000µm). Cette liste représente les principaux filtres (Figure 10) mais n'est pas exhaustive, quelques échantillons correspondant à des tailles différentes. De plus pour l'expédition du cercle polaire les fractions ont, pour certaines, été modifiées : on retrouve notamment des filtres 0.8µm-2000µm, retirant simplement les plus petits organismes, le filtre 0.8-5µm disparait, et le filtre 5-20µm devient systématiquement 3-20µm.

Il existe une corrélation négative entre la taille des organismes et leur abondance en milieu marin, c'est pourquoi plus l'organisme est de grande taille ou peu abondant, plus le volume d'eau filtré doit être important pour récolter suffisamment de matériel génétique (Figure 10). La méthode utilisée pour ce filtrage est également différente selon les organismes : pour les plus petites tailles, des bouteilles Niskin sont plongées dans l'eau, tandis que des filets présentant différentes tailles de mailles permettent de récolter de plus grands organismes.

La mesure des paramètres environnementaux des différentes colonnes d'eau échantillonnées a été réalisée avec la rosette CTD plongeant dans l'eau qui contient les bouteilles Niskin, permettant ainsi l'acquisition de valeurs telles que la température ou les nutriments.

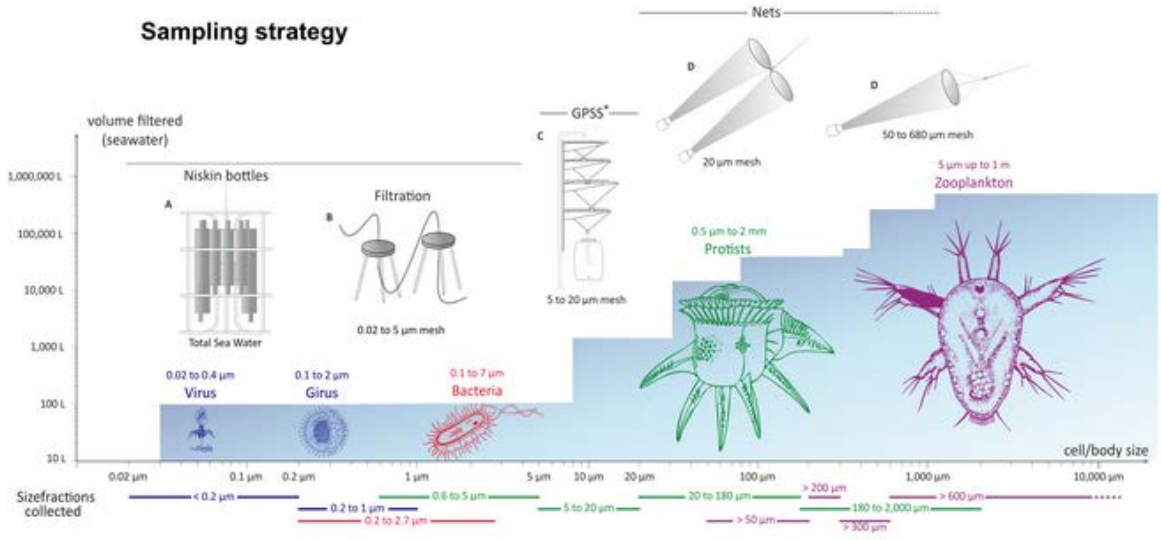


Figure 10 : Organisation du protocole de prélèvement durant l'expédition *Tara* Oceans en fonction des organismes cibles. Plus la taille des organismes est importante (axe des abscisses), plus le volume d'eau filtré (bleu) est important. La méthode de filtration (pompe péristaltique ou filet) varie également selon la taille des organismes. (Figure extraite de Karsenti et al, 2011[129])

Afin de réaliser le séquençage métagénomique pour chacun des échantillons, le matériel biologique a été récupéré sur une membrane filtrante correspondant à une fraction de taille, puis l'ADN a été extrait à partir d'un protocole étudié pour effectuer une lyse des cellules et des noyaux de protistes et de métazoaires[130]. Les fragments d'ADN ont ensuite été fractionnés en séquences de 300 bp en moyenne et préparés pour du séquençage pairé (paired-end), majoritairement sur des plateformes Illumina HiSeq2000 et HiSeq2500 HiSeq2500, produisant en moyenne 160 millions de lectures de 2×101 bp par échantillon.

L'ADN purifié a également été utilisé pour obtenir des gènes d'ARNr 16S et 18S, permettant le séquençage de ces métabarcodes afin d'avoir une meilleure vision globale des communautés présentes dans les échantillons, et enfin des cellules ont été isolées par cytométrie en flux afin de pouvoir les analyser par génomique en cellule unique.

Objectif de la thèse

Cette thèse s'est effectuée dans le cadre de multiples études visant à une meilleure compréhension du lien entre l'environnement et les organismes planctoniques. L'objectif est donc d'utiliser quelques organismes modèles majeurs du plancton afin d'analyser leur lien à l'environnement dans des échantillons environnementaux. Pour cela, nous allons donc nous intéresser en particulier aux espèces appartenant à l'ordre des Mamiellales en nous servant de génomes de référence croisés avec les échantillons métagénomiques et de métabarcodes provenant du projet *Tara* Oceans. Comme nous l'avons abordé dans la partie précédente, ces organismes sont d'un grand intérêt écologique ainsi que de très bons modèles d'étude, notamment en raison de leur abondance qui nous permet de travailler de manière optimale avec des échantillons de ce type. Le projet couvrant un grand nombre de positions géographiques et disposant d'informations sur les paramètres physico-chimiques correspondant à chaque échantillon génomique, il nous sera ainsi possible d'étudier à la fois leur répartition géographique, la structure de leurs populations, et la corrélation de ces informations avec leur environnement. Afin de comparer ces résultats avec d'autres types d'organismes, nous élargirons également nos analyses à d'autres espèces.

Dans le premier chapitre, nous nous concentrerons donc sur les Mamiellales, en commençant par *Bathycoccus* dont le génome d'une seconde espèce a pu être séquencé grâce à la méthode de génomique en cellule unique, en utilisant les échantillons de métabarcodes et métagénomiques de la première partie de *Tara* Oceans, sans le cercle polaire. Nous élargirons ensuite l'étude de biogéographie appliquée ici à d'autres Mamiellales pour lesquels nous possédons des génomes de référence. Nous évoquerons enfin les variations observées sur le grand chromosome outlier entre les échantillons, discutant ainsi la question de la région "mating-type" de ces espèces.

Par la suite, dans un second chapitre, grâce à l'accès aux résultats génomiques de la seconde partie de l'expédition se concentrant sur le cercle polaire, nous étudierons avec le même type de données la présence des Mamiellales dans les eaux froides.

Dans le troisième chapitre, nous pourrons nous concentrer sur les variations génomiques de l'espèce *Bathycoccus prasinos*, allant ainsi à une échelle plus fine pour étudier la structure des populations dans des environnements très différents, allant de températures négatives à environ vingt-cinq degrés.

Enfin, le quatrième et dernier chapitre nous donnera l'occasion de discuter l'impact de l'environnement sur d'autres espèces en prenant le modèle des Straménopiles, des organismes hétérotrophes. Nous passerons également à une échelle beaucoup plus large, celle des communautés, avant de brièvement évoquer la question de l'impact du réchauffement climatique prévu dans les années à venir sur l'ensemble de ces organismes planctoniques.

Chapitre 1 : Biogéographie des Mamiellales à partir de données métagénomiques

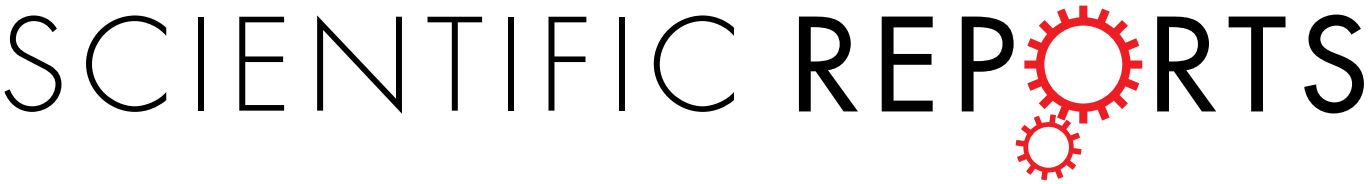
I. Article 1 : Survey of the green picoalga *Bathycoccus* genomes in the global ocean

Bathycoccus prasinos est une algue verte unicellulaire cosmopolite de l'ordre des Mamiellales, et donc un contributeur important à la production primaire. Il s'agit d'un bon modèle d'étude dans l'environnement en raison de son abondance et c'est pourquoi nous avons voulu observer sa répartition géographique à partir des données *Tara* Oceans. Jusqu'alors, bien que l'existence de deux écotypes trouvés à des températures différentes de *Bathycoccus* aient été suggérés dans plusieurs études, leur séquence 18S identique a toujours rendu difficile leur identification, et un seul génome correspondant à la souche RCC1105, provenant de Méditerranée, avait été séquencé.

Cependant, nous avons pu obtenir un génome du second écotype à partir d'un SAG (*Single Amplified Genome*) récolté dans l'Océan Indien durant l'expédition, puis séquencé et annoté au Genoscope. Sa complétion est estimée à 64% et cette seconde espèce a été nommée TOSAG39-1. Dans le cadre de la publication *Survey of the green picoalga Bathycoccus genomes in the global ocean* dans le journal *Scientific Report*, les deux génomes ont été comparés, confirmant qu'ils étaient génétiquement distants et appartenaient à deux espèces distinctes avec notamment une identité protéique à 78% et une synthénie incomplète.

Ma contribution à cette publication a principalement porté sur l'analyse de la distribution des espèces et leur lien à l'environnement. En utilisant 122 échantillons métagénomiques du projet *Tara*, nous avons pu recruter les lectures correspondant à nos génomes de référence afin d'obtenir leurs abondances relatives dans chacun d'entre eux. Cela nous a permis de voir que comme suggéré par les hypothèses précédentes sur les préférences de ces organismes, ils se trouvaient bien dans des niches écologiques distinctes, pouvant parfois se recouper. Le facteur les séparant majoritairement se trouve être la température, RCC1105 étant trouvé dans des eaux froides et tempérées de 5 à 20 degrés environ, tandis que TOSAG39-1 ne se trouve qu'à partir

de 15 degrés et existe dans des échantillons allant jusqu'à 28 degrés. De plus, la première espèce est trouvée plus en surface que la seconde et donc à une plus forte luminosité, et les deux semblent absentes des échantillons du Pacifique, océan connu pour sa faible concentration en fer. Les données supplémentaires correspondant à cet article sont disponibles en Annexe 1.



OPEN

Survey of the green picoalga *Bathycoccus* genomes in the global ocean

Received: 28 April 2016
Accepted: 03 November 2016
Published: 30 November 2016

Thomas Vannier[1,2,3], Jade Leconte[1,2,3], Yoann Seeleuthner[1,2,3], Samuel Mondy[1,2,3], Eric Pelletier[1,2,3], Jean-Marc Aury[1], Colomban de Vargas[4], Michael Sieracki[5], Daniele Iudicone[6], Daniel Vaulot[4], Patrick Wincker[1,2,3] & Olivier Jaillon[1,2,3]

Bathycoccus **is a cosmopolitan green micro-alga belonging to the Mamiellophyceae, a class of picophytoplankton that contains important contributors to oceanic primary production. A single species of** ***Bathycoccus*** **has been described while the existence of two ecotypes has been proposed based on metagenomic data. A genome is available for one strain corresponding to the described phenotype. We report a second genome assembly obtained by a single cell genomics approach corresponding to the second ecotype. The two** ***Bathycoccus*** **genomes are divergent enough to be unambiguously distinguishable in whole DNA metagenomic data although they possess identical sequence of the 18S rRNA gene including in the V9 region. Analysis of 122 global ocean whole DNA metagenome samples from the** ***Tara*****-Oceans expedition reveals that populations of** ***Bathycoccus*** **that were previously identified by 18S rRNA V9 metabarcodes are only composed of these two genomes.** ***Bathycoccus*** **is relatively abundant and widely distributed in nutrient rich waters. The two genomes rarely co-occur and occupy distinct oceanic niches in particular with respect to depth. Metatranscriptomic data provide evidence for gain or loss of highly expressed genes in some samples, suggesting that the gene repertoire is modulated by environmental conditions.**

Phytoplankton, comprising prokaryotes and eukaryotes, contribute to nearly half of the annual global primary production[1]. Picocyanobacteria of the genera *Prochlorococcus* and *Synechococcus* dominate the prokaryotic component[2]. However, small eukaryotes (picoeukaryotes; $<2\,\mu m$) can be major contributors to primary production[3,4]. In contrast to cyanobacteria, the phylogenetic diversity of eukaryotic phytoplankton is wide, with species belonging to virtually all photosynthetic protist groups[5]. Among them, three genera of green algae belonging to the order Mamiellales (class Mamiellophyceae[6]), *Micromonas, Ostreococcus* and *Bathycoccus* are particularly important ecologically because they are found in a wide variety of oceanic ecosystems, from the poles to the tropics[7–12]. The cosmopolitan distribution of these genera raises the questions of their diversity and their adaptation to local environmental conditions. These genera exhibit genetic diversity: for example, there are at least three genetically different clades of *Micromonas* with different habitat preferences[12,13]. One ecotype of *Micromonas* seems to be restricted to polar waters[8,14]. *Ostreococcus* which is the smallest free-living eukaryotic cell known to date with a cell size of $0.8\,\mu m$[15] can be differentiated into at least four clades. Two *Ostreococcus* species have been formerly described: *O. tauri* and *O. mediterreanus*[15,16]. Among these *Ostreococcus* clades, different strains seem to be adapted to different light ranges[17]. However, the ecological preferences of *Ostreococcus* strains are probably more complex, implying other environmental parameters such as nutrients and temperature[9].

The genus *Bathycoccus* was initially isolated at 100 m from the deep chlorophyll maximum (DCM) in the Mediterranean Sea[18] and cells with the same morphology (body scales) had been reported previously from the Atlantic Ocean[19]. *Bathycoccus* has been since found to be widespread in the oceanic environment, in particular in coastal waters[20,21], and one genome sequence from a coastal strain is available[22]. Metagenomic data have suggested the existence of two *Bathycoccus* ecotypes[10,11,23], recently named B1 and B2[11]. These two ecotypes have

[1]CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [2]CNRS, UMR 8030, CP5706 Evry, France. [3]Université d'Evry, UMR 8030, CP5706 Evry, France. [4]Sorbonne Universités, UPMC Université Paris 06, CNRS, UMR7144, Station Biologique de Roscoff, 29680 Roscoff, France. [5]National Science Foundation, 4201 Wilson Blvd., Arlington, VA 22230, USA. [6]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. Correspondence and requests for materials should be addressed to P.W. (email: pwincker@genoscope.cns.fr). or O.J. (email: ojaillon@genoscope.cns.fr)

SAG Assembly	Total Size (Mb)	N50 (kb)	NG50[1] (kb)	Genome Completion (%)
A	3.5	14.8	NA	30.8
B	4.7	14.5	NA	27.7
C	3.7	24.1	NA	21.5
D	4.1	18.1	NA	26.0
(A) + (B) + (C) + (D)[2]	8.0	16.6	0.9	44.6
Combined ABCD[3]	10.1	14.1	6.0	64.0

Table 1. Assembly summaries of TOSAG39-1. [1]The longest assembly contigs covering together half of the genome size (15 Mbp) are each longer than the NG50. This evaluation was not possible for the four individual cell assemblies for which the total assembly sizes are shorter than half of the genome size. [2]A + B + C + D corresponds to a non-redundant merging of contigs from individual assemblies. [3]Combined ABCD corresponds to the co-assembly process.

identical 18S rRNA sequences and therefore cannot be discriminated when using metabarcodes such as the V4 or V9 regions of the 18S rRNA genes[10]. However information on the ocean-wide distribution and the ecological preferences of these two ecotypes are lacking.

Mapping of metagenomic reads onto whole genomes (fragment recruitment) has been shown to be an efficient way to assess the distribution of oceanic bacterial populations[24,25]. The paucity of eukaryotic genomes and metagenomes has prevented this approach to be applied on a large scale to eukaryotes. Therefore the determination of the geographical distribution and ecological preferences of marine eukaryotic species has relied on the use of marker genes such as 18S rRNA or ITS (internal transcribed spacer)[26] and more recently on metabarcodes[27]. One major problem is the absence of reference genomes for many marine eukaryotes as a consequence of the difficulty to cultivate them. To overcome this limitation, Single Cells Genomics is a very promising approach[28,29]. However, this approach has been largely used for bacteria[30] and numerous technical challenges have limited the recovery of eukaryotic genomes with this approach[28,31–33]. The most complete assembly obtained so far is for an uncultured stramenopile belonging to the MAST-4 clade and contains about one third of the core eukaryotic gene set[33]. Recently, the *Tara* Oceans expedition collected water samples from the photic zone of hundreds of marine sites from all oceans and obtained physicochemical parameters, such as silicate, nitrate, phosphate, temperature and chlorophyll[34–36]. This expedition also led to the massive sequencing of the V9 region from 18S ribosomal gene providing a description of the eukaryotic plankton community over wide oceanic regions[27]. During this expedition a large number of metagenomic data and single-cell amplified genomes (SAGs[37]) have also been acquired. Here, we introduce a novel genome assembly for *Bathycoccus* based on the sequence assembly of four SAGs obtained from a *Tara* Oceans sample collected in the Arabian Sea. Comparison of this assembly with the reference sequence of *Bathycoccus* strain RCC1105[22] unravels substantial genomic divergence. We investigated the geographical distributions of these two genomes by mapping onto them the short reads of a large set of metagenomes obtained in multiple marine basins from the *Tara* Oceans survey[35,38]. We also determined the genomic properties and habitat preferences of these two *Bathycoccus*.

Results

Genome structure of *Bathycoccus* TOSAG39-1. We obtained a new *Bathycoccus* SAG assembly (TOSAG39-1) by the single cell genomics approach from four single cells collected from a single sample during the *Tara* Oceans expedition. We presumed these cells were from the same population and combined their genomic sequences to improve the assembly. The length of the final combined-SAGs assembly is 10.3 Mb comprising 2 345 scaffolds. Half of the assembled genome lies in 179 scaffolds longer than 13.6 kb (N50 size). This assembly covers an estimated 64% of the whole genome when considering the proportion of identified eukaryotic conserved genes[39]. We verified that this combined SAG assembly has longer cumulative size, and a larger representation of the genome than each assembly obtained from sequences of a single-SAG. We also merged the four assemblies from single-SAGs and, after removing redundancies, we obtained a substantially lower genomic representation than for the combined-SAGs strategy (Table 1). We mapped the reads of each SAG-sequencing onto the final assembly to examine whether genomic variability among the sampled population might have affected the quality of the assembly. We did not detect any major genomic variability; contigs can be formed by reads from different cells (Supplementary Figure S1). In total, half of the assembly (52.2%) was generated by reads from a single cell and one third (30.5%) by two cells.

The approximate estimated genome size is 16 Mb and GC content is 47.2%, similar to what has been reported for RCC1105 (15 Mb and 48%, respectively). We predicted 6 157 genes (Supplementary Table 1), representing a higher gene density compared to RCC1105 (622 *vs.* 520 genes per Mb), probably because of the higher fragmentation of the SAG assembly (the coding base density is conversely higher in TOSAG39-1, 742 *vs.* 821 kb/Mb for the two assemblies, respectively, Supplementary Table 1). The photosynthetic capacity of TOSAG39-1, presumed from the chlorophyll autofluorescence in the cell sorting step, was verified by the presence of plastid contigs (removed during quality control filtering) and by the presence of nuclear photosynthetic gene families (encoding RuBisCo synthase, starch synthase, alternative oxidase and chlorophyll a/b binding proteins) in the final assembly.

Previous comparisons of Mamiellales genomes demonstrated global conservation of chromosomal locations of genes between *Bathycoccus*, *Ostreococcus* and *Micromonas*[22]. These genera all possess outlier chromosomes (one part of chromosome 14 and the entire chromosome 19 for *Bathycoccus*) that display an atypical GC% and numerous small, unknown, non-conserved genes. We detected almost perfect co-linearity between non-outlier

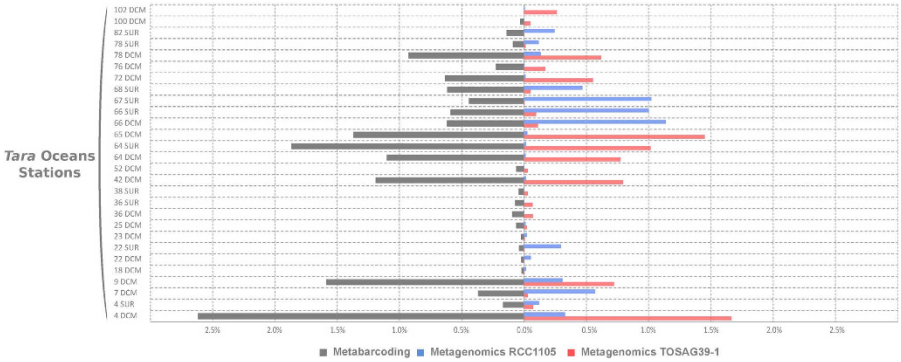


Figure 1. Comparisons of relative abundances of *Bathycoccus* in the 0.8–5 μm size fraction samples from *Tara* Oceans stations. Left: relative 18S rRNA V9 amplicons abundance (percent of reads). Right: relative metagenomic abundances (percent of metagenomic reads) from direct mapping of metagenomic reads onto two genome sequence assemblies (strain RCC1105 and TOSAG39-1, single cell assembly from an Indian Ocean sample). Stations and depth (Surface or DCM) are indicated on the Y axis.

chromosomes of RCC1105 and orthologous regions of TOSAG39-1 scaffolds (Supplementary Figure S2). However, there is a significant evolutionary divergence between the genomes: the orthologous proteins are only 78% identical on average (Supplementary Figure S3). Only 26 genes are highly conserved (>99% identity), they are distributed on 14 chromosomes (including outlier chromosome 14) and did not display any clustering. As expected, chromosome 19 did not fit this pattern: we could not align most of its genes by direct BLAST comparison. Some traces of homology were observed for nine genes (62% protein identity). One of the twenty longest scaffolds of TOSAG39-1 had characteristics similar to chromosome 19. This scaffold could not be aligned to RCC1105 and has the lowest GC content (0.44 *vs.* 0.48% for the other scaffolds on average).

Manual curation of alignments to analyze synteny along the twenty longest TOSAG39-1 scaffolds showed that 90% of genes are collinear between the two genomes, 5% are shared outside syntenic blocks, and 5% are specific to TOSAG39-1. The three rRNA genes (18S or small subunit (SSU), 5S, 23S or large subunit (LSU)), used as phylogenetic markers in many studies, are identical between the two genomes. The SSU and LSU genes of TOSAG39-1 have introns. The SSU intron (440 bp) is at the same position as in RCC1105, but is only 91% similar. The LSU intron (435 bp) is only present in TOSAG39-1. The internal transcribed spacers (ITS) are different between the two TOSAG39-1 and the RCC1105 assemblies (82% and 86% for ITS1 and ITS2, respectively) but closer to those of two *Bathycoccus* oceanic strains from the Indian Ocean (RCC715 and RCC716) (Supplementary Figure S4) and of a metagenome from the Atlantic Ocean DCM[40]. We also looked at the plastid 16 S marker gene[41] and to the PRP8 intein gene that has been proposed as markers for *Bathycoccus*[10]. The plastid 16 S sequences of the two *Bathycoccus* genomes share 92% identical nucleotides, and PRP8 is lacking from the TOSAG39-1 assembly.

We were able to determine the affiliation of three metagenomes[23,40] containing *Bathycoccus* and two *Bathycoccus* transcriptomes of the MMETSP database[42] (Supplementary Figures S5). Metagenomes T142 and T149 from the South East Pacific[23] and transcriptome MMETSP1399 (strain CCMP1898, which is the type strain for *Bathycoccus prasinos*) correspond, or are closely related to RCC1105. The tropical Atlantic Ocean metagenome[40] and transcriptome MMETSP1460 (strain RCC716 from the Indian Ocean) correspond, or are closely related to TOSAG39-1. Direct amino acid BLAST[43] comparison of TOSAG39-1 and RCC1105 versus metagenomes T142 and T149 demonstrates the presence of additional genomes in these samples that were obtained by flow cytometry sorting of natural picoplankton populations (Supplementary Figure S5).

Oceanic distribution of *Bathycoccus* genomes. We analyzed the worldwide distribution of the two *Bathycoccus* genomes using metagenomic samples from the *Tara* Oceans expedition. Metagenomic short reads obtained from 122 samples taken at 76 sites and covering 24 oceanic provinces were mapped onto the two *Bathycoccus* genomes RCC1105 and TOSAG39-1. Among the four eukaryotic size fractions sampled in this expedition (0.8–5 μm, 5–20 μm, 20–180 μm, 180–2000 μm) statistically significant mapping was only obtained for the 0.8–5-μm fraction, which matches the cellular size of *Bathycoccus* (1.5–2.5 μm[18]). The percentage of filtered mapped metagenomic reads for every gene and station was used to estimate the relative genomic abundance of *Bathycoccus*. We compared final counts of genome abundances with counts based on amplicon sequences of the V9 region of the 18S rRNA gene[27] which does not distinguish RCC1105 from TOSAG39–1 because their 18S rRNA gene sequences are identical. The V9 data demonstrated the wide distribution of *Bathycoccus* in marine waters, with maximum relative abundance reaching 2.6% of all reads. The *Bathycoccus* metabarcode was represented by more than 1% of reads in 13% of the samples. *Bathycoccus* sequences were detected in whole metagenome reads from the same samples where *Bathycoccus* was detected with 18S rRNA metabarcodes (Fig. 1). For each sample displaying a V9 signal, we detected the presence of the genomes of either RCC1105, TOSAG39-1, or both. In addition, the relative abundances estimated from V9 metabarcodes were correlated with the sum of the relative genomic abundances of TOSAG39-1 and RCC1105 (Supplementary Figure S6). Therefore, the *Bathycoccus* populations detected by the V9 metabarcode are likely to correspond to these two genomes only, and not to a third yet unknown genome.

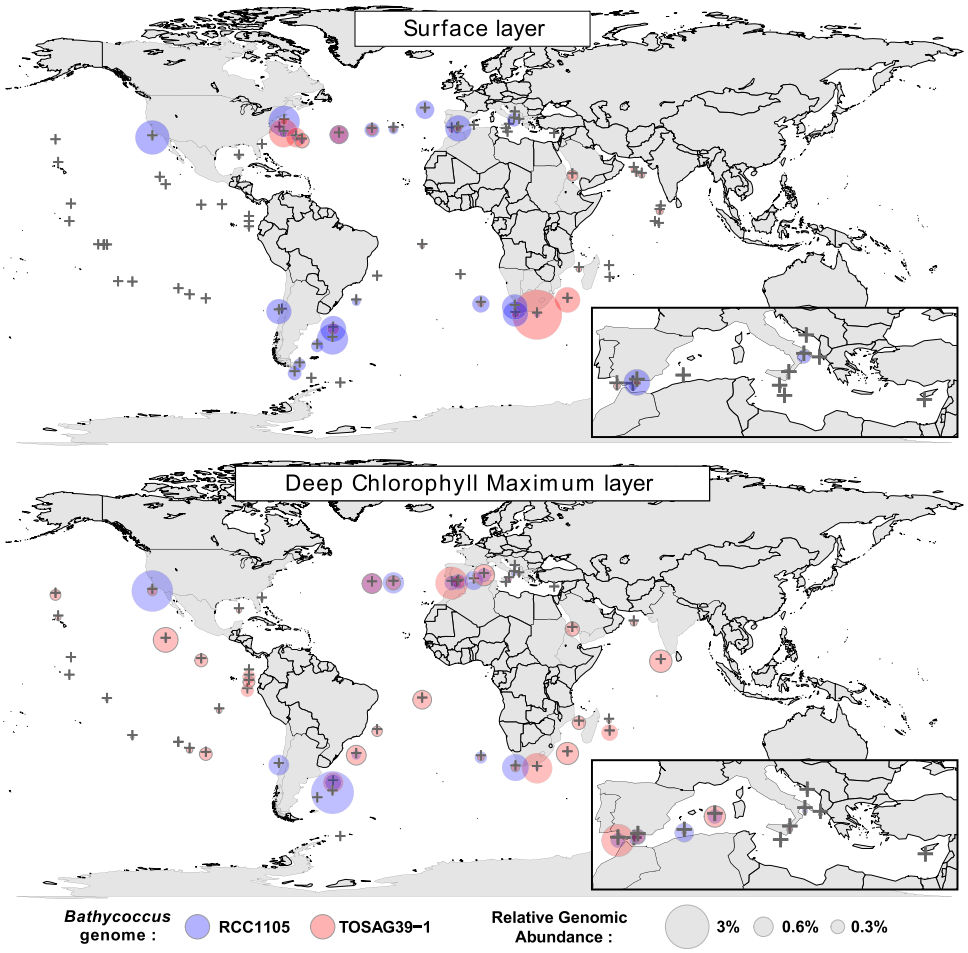


Figure 2. Geographical distribution of two *Bathycoccus* genomes, RCC1105 and TOSAG39-1, along *Tara* Oceans expedition stations from recruitments of metagenomic reads. Top and bottom maps correspond to the surface and deep chlorophyll maximum (DCM) samples respectively. Gray crosses indicate *Tara* Oceans sampling stations and the sizes of the red or blue circles indicate the relative genomic abundances of the two *Bathycoccus* types. We generated this map using R-package maps_2.1-6, mapproj_1.1-8.3, gplots_2.8.0 and mapplots_1.4 (version R-2.13, https://cran.r-project.org/web/packages/maps/index.html).

Among the 58 samples where *Bathycoccus* metagenomics abundances represented more than 0.01% of the total numbers of reads, in 91% of the cases a single genome was dominant, i.e. accounting for more than 70% of the reads. The two *Bathycoccus* showed similar proportions (i.e., between 40% and 60% of the reads) in only two samples (stations TARA_006 and TARA_150 at DCM, Supplementary Figure S7).

The global distribution of the two *Bathycoccus* genomes revealed complex patterns. The RCC1105 genome was found mainly in temperate waters, both at the surface and at the DCM, whereas TOSAG39-1 appeared more prevalent in tropical zones and at the DCM (Fig. 2). TOSAG39-1 was found in surface water in only five winter samples from the Agulhas and Gulf Stream regions at stations undergoing strong vertical mixing (Supplementary Table 2, Supplementary Figure S8). RCC1105 was detected more widely in surface water and was restricted to two narrow latitudinal bands around 40°S and 40°N. Conversely, TOSAG39-1 was found throughout a latitudinal range from 40°S to 39°N (Fig. 2). In particular, TOSAG39-1 was found in the tropical and subtropical regions in the Pacific, Atlantic and Indian Oceans.

In the equatorial and tropical Pacific Ocean, a region characterized by high nutrient and low chlorophyll where phytoplankton is limited by iron[44], *Bathycoccus* was not detected (or only at very low abundance), except close to the Galapagos Islands. We detected opposite trends in the presence of the two *Bathycoccus* along the Gulf Stream: RCC1105 increased from west to east while TOSAG39-1 showed the reverse trend. The two *Bathycoccus* also showed opposite trends at some stations that were relatively close but located on both sides of important oceanographic boundaries. The first case was off South Africa, between stations TARA_065 and TARA_066 (Supplementary Figure S8) located, respectively, in coastal, temperate Atlantic and in Indian subtropical water from the Agulhas current[45].

The second case occurred in winter in the North Atlantic, downstream of Cape Hatteras (US East coast), where station TARA_145 was in cold, nutrient-rich waters north of the northern boundary of the Gulf Stream (also called the Northern Wall for its sharp temperature gradient) and TARA_146 was south of the southern boundary, in the subtropical gyre (Fig. 2 and Supplementary Figure S8).

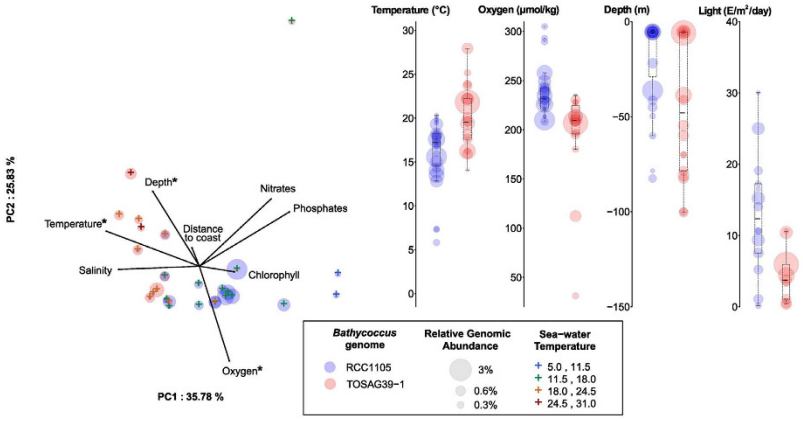


Figure 3. Relationships between environmental parameters and *Bathycoccus* genome abundance. Left: Principal component analysis. We only considered stations where we detected 98% of the genes for one *Bathycoccus* genome, and for which all environmental parameters were available (Oxygen, Nitrates, Phosphates, Chlorophyll, Sampling Depth, Water Temperature and Salinity). Crosses indicate stations, with a color scale corresponding to the water temperature. The distance to coast parameter corresponds to the shortest geographical distance to the coast. The two *Bathycoccus* are distributed along temperature and oxygen axes. Stars indicate parameters that statistically discriminate the two *Bathycoccus*. Right: Range of values of temperature, oxygen and sampling depth for parameters where a significant difference was detected between RCC1105 and TOSAG39-1.

Principal component analysis was used to assess the relationship between the genomic data and environmental parameters determined *in situ*[36] complemented by satellite and climatology data (Supplementary Information). Temperature, oxygen, sampling depth and PAR (photosynthetic active radiation), though with less significant p-values for the latter, were related to the segregation of the two genomes (Fig. 3 and Supplementary Figure S9). The two *Bathycoccus* were found in temperature ranges from 0 to 32 °C and from 7 to 28 °C for RCC1105 and TOSAG39-1, respectively. On average, the TOSAG39-1 genome was found in waters 3 °C warmer than was RCC1105 (21.5 *vs.* 18.4 °C, p-value $< 10^{-3}$, Fig. 3 and Supplementary Figure S10). Abundances were very low below 13 °C for both genomes, and above 22 °C for RCC1105. A similar discrimination was observed for oxygen: TOSAG39-1 was found in samples with lower oxygen content. For example, the TOSAG39-1 genome was abundant in the DCM of station 138 where O_2 was low (31.2 µM, Fig. 3, Supplementary Figures S9 and S10), though no samples originated from anoxic waters[46].

The two *Bathycoccus* were recovered from significantly different ranges of PAR, estimated from weekly averages of surface irradiance measurements extrapolated to depth using an attenuation coefficient derived from local surface chlorophyll concentrations[47] (Fig. 3, Supplementary Figures S9 and S10, Supplementary Information). Both *Bathycoccus* could thrive in winter when the overall light availability is low (Supplementary Figure S8). Nutrient concentrations did not seem to explain the separation between the two *Bathycoccus*. We found RCC1105 in nutrient-rich surface waters and TOSAG39-1 mostly at the DCM in oligotrophic waters, close to the nutricline characterized by a significant upward flux of nutrients[48,49]. While RCC1105 was never abundant below 80 m, TOSAG39-1 extended down to almost 150 m (Fig. 3 and Supplementary Figure S10).

Genomic plasticity. For each genome, we searched for evidence of gene gain or loss by analyzing gene content variations at the different stations. Lost or gained genes could be considered as dispensable genes or as present only in some genomic variants, therefore, characterizing a "pan-genome" analogous to what is observed in bacterial populations[50]. We analyzed the coverage of metagenomic reads that were specifically mapped at high stringency onto one genome and looked for traces of gene loss. To avoid false positives caused by conserved genes, we restricted this analysis to samples where 98% of the genes from one of the two *Bathycoccus* genome sequences were detected, and focused on genes that were detected in the metagenomes of at least four samples, and not detected in at least five samples. Metatranscriptomic data was used to select genes having an expression signal in at least six samples. Using these stringent criteria, we detected about one hundred dispensable genes for each genome (Supplementary Tables 1,4 and 5). Half of the RCC1105 dispensable genes (50/108) are located on chromosome 19, representing 70% of the genes on this chromosome. These genes have shorter coding and intronic regions than other genes (Supplementary Table 1), which is a property of the genes predicted on outlier chromosome 19[22]. Dispensable genes on regular chromosomes also tend to be shorter. Additionally, the distribution of dispensable genes on the genome is not random. Among the 72 genes of chromosome 19, 47 out the 50 dispensable genes are grouped into two long blocks at the chromosome end, leaving the first part of chromosome 19 almost free of dispensable genes (Supplementary Figure S11). Dispensable genes also appear clustered on regular chromosomes. Twenty-one out of 58 dispensable genes are in small cassettes, two to four gene-long, especially on chromosomes 2, 5 and 17 (Fig. 4 and Supplementary Figure S11). We verified the contiguity of the genomic regions around the dispensable genes by alignment with assemblies of metagenomics reads (Supplementary Information). We analyzed the pattern of loss of these dispensable cassettes in samples where

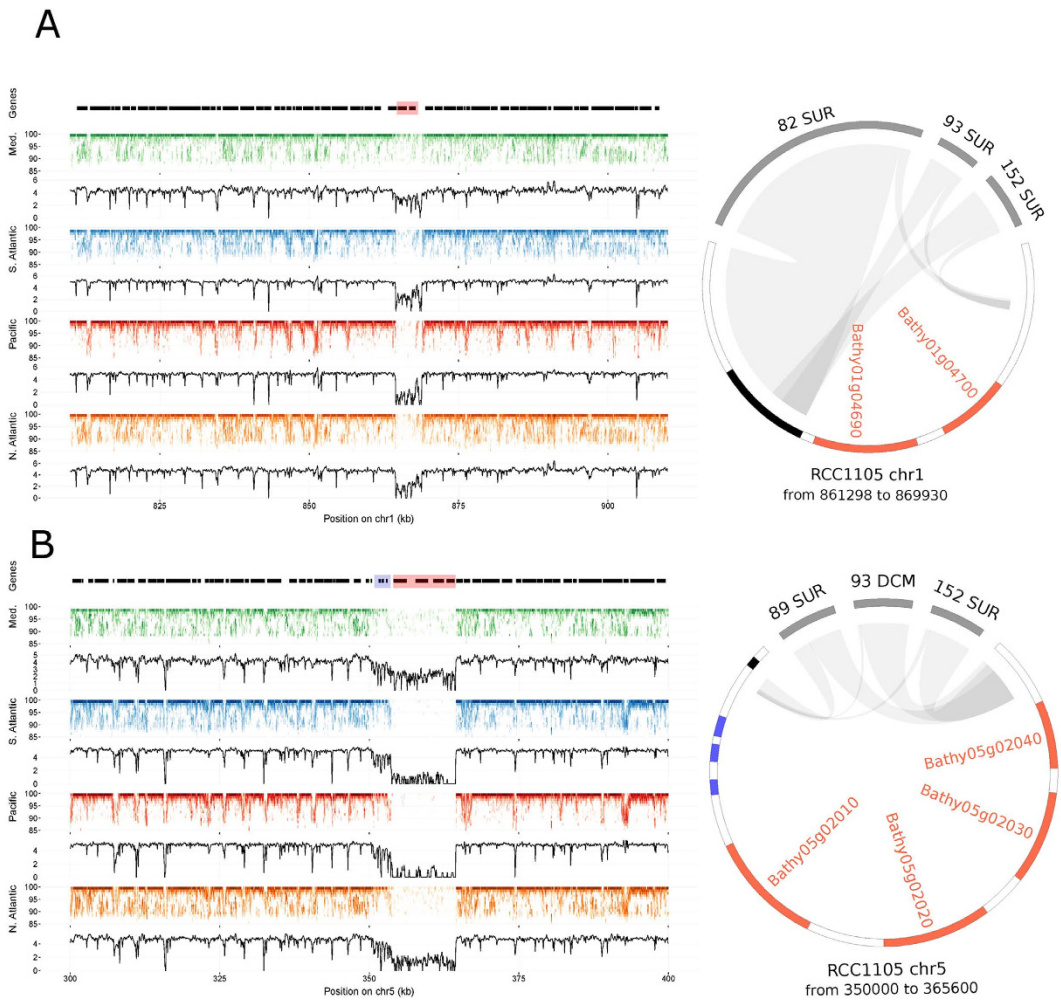


Figure 4. Evidence for cassettes of dispensable genes in *B. prasinos* RCC1105. Left and right sides of the figures represent fragment recruitment and genomic alignments of dispensable gene cassettes, respectively. Fragment recruitments plots are displayed by marine zones (left legend). Each dot corresponds to a given number of mapped reads at a given identity percent (indicated on the Y-axis). The density of mapped read is displayed as the black line plotted below each fragment recruitment plot. Gene positions are represented by black boxes on the top of the first fragment recruitment plot and dispensable genes are highlighted in red. Genomic alignments are represented as circos graphs[79] on which dispensable genes are colored in red, and other genes are represented by black boxes. Left side and right side of the genomic region are connected to metagenomics contigs (gray segments), leaving in-between the locus of the dispensable gene cassette that remains unconnected to any metagenomic contig. Connections correspond to blast alignments positions. (**A**) 100- and 8.6-kb regions of chromosome 1 are represented on a fragment recruitment plot and on the circos graph, respectively. A two gene long cassette is represented. A massive decrease of read coverage appears on the fragment recruitment plot in all oceanic zones except in the Mediterranean Sea, which indicates that the two genes are present only in a sub-population in this basin. A similar pattern is observed in panel (**B**) for four consecutive genes for which fragment recruitment plots representing 100 kb of chromosome 5 suggest a presence in a Mediterranean sub-population and absence in other marine areas. The circos graph represents alignments along the 15.6-kb cassette locus with metagenomics contigs, which resulted in a gap that included three small genes (in blue) in addition to the four automatically detected dispensable genes. Fragment recruitment confirmed a significant, but not total, decrease of read coverage for these three genes in every oceanic zone, indicating that their presence or absence in the two sub-populations was widely distributed.

they were not detected and obtained alignments that included gaps in place of dispensable genes (Fig. 4). Notably, cassette borders were at the same positions in the various samples, showing a low diversity at these loci. This suggests that a common or single breakpoint event occurred in the past. Fragment recruitments plots showed a homogenous decrease of read coverage along the contiguous dispensable genes, confirming that genomic losses or gains occurred at the scale of entire cassettes (Fig. 4 and Supplementary Figure S11). We examined the synteny between RCC1105 and TOSAG39-1 for the regions corresponding to the two cassettes illustrated in Fig. 4. We retrieved the orthologous genes situated around the cassettes in two TOSAG39-1 scaffolds in a clear syntenic relationship, but the cassettes genes were missing.

We observed an incomplete, but marked, depletion of read coverage for three contiguous genes on chromosome 5. These genes immediately precede the longest dispensable gene cassette. This incomplete read coverage depletion indicates that this genomic region only occurs in a sub-population, suggesting a sympatry or at least co-occurence of these two genomic forms. This pattern was observed in every oceanic basin (Fig. 4B) with the longest dispensable gene cassette spanning seven genes.

The function of these dispensable genes is unclear. Only 15 dispensable genes located on RCC1105 non-outlier chromosomes possess a protein Pfam domain (Supplementary Information, Supplementary Table 3). However, several of these genes might be involved in genomic rearrangements because they contain reverse transcriptase and HNH endonuclease domains and this could be linked to their dispensability. Intriguingly, the average relative transcriptomic activity is higher in dispensable genes than in non-dispensable genes (0.73 *vs.* 0.56, Mann-Whitney-Wilcoxon test p-value = 1.52E-4, Supplementary Table 1).

Beside these patterns suggesting gene gains or losses, we examined at a global level the genomic variation within populations of each *Bathycoccus*. This was done by fragment recruitment of the metagenomic reads of *Tara* Oceans samples onto the two reference assemblies. The distributions of nucleotide identities show a weak divergence between the reference assemblies and geographically distant samples, though higher for TOSAG39-1 than for RCC1105 (Supplementary Information, Supplementary Figure S12).

Discussion

We provide a novel *Bathycoccus* genome assembly using a single-cell genomics approach. This assembly is estimated to be 64% complete, which is, to our knowledge, the most complete eukaryotic genome obtained to date by this approach. This relatively high level of completion was reached through the combination of several independent cells originating from the same population. It has been described that the enzymatic amplification of DNA which is inherent to single-cell genomics induces strong biases in sequencing depth along the genome, leading to partial and fragmented assemblies[51]. Here, this caveat appears reduced as the combined-SAGs assembly is significantly more complete than the assembly obtained from each of the individuals SAGs.

This *Bathycoccus* SAG assembly is significantly different from the previously described genome assembly, originating from the coastal Mediterranean strain RCC1105. The former corresponds to the B1 clade and the latter to the B2 clade as, defined recently[11]. Orthologous proteins of these two genomes share only 78% identity, which is similar to the 74% of amino-acid identity shared by the two sequenced *Ostreococcus* isolates which belong to different clades[52].

A previous study[11] estimated a lower genetic distance (82% of identical nucleotides) between the two *Bathycoccus* using metagenomic data. This difference is probably as expected because of the reduced dataset of highly conserved and single copy genes (1 104 genes) considered in the latter analysis. The evolutionary distance that separates the protein coding genes of these two *Bathycoccus* is slightly smaller than the one between two vertebrate lineages separated by more than 400 million years (mammal and fish share 72% of identity[53]) and larger than the one reported between many model organisms (for example, human and mouse share 85% of identity[54,55]). This high divergence in protein coding genes and the frequent genes rearrangement in chromosomes is hardly compatible with chromatid pairing required for intercrossing[56] between the two *Bathycoccus*. Very few genes are highly conserved (>99% identity) between the two *Bathycoccus* and conserved genes are not clustered, which makes active genetic exchange by homologous recombination unlikely. Therefore, although the two *Bathycoccus* share 100% similar rRNA gene sequences, these genomic differences reflect two different, probably cryptic, species. Identical rRNA sequences have been previously reported in the yeast *Saccharomyces cerevisiae sensu stricto* clade[57], or the haptophyte species *Emiliania huxleyi* and *Gephyrocapsa oceanica,* which also have identical 18S rRNA gene sequences, but quite different morphologies[58].

The combination of genomics and environmental data from a large set of oceanic samples revealed the distinct ecological preferences of the two *Bathycoccus* with respect to depth, temperature, light and oxygen. TOSAG39-1 is usually found in warmer but deeper and darker water than RCC1105. TOSAG39-1 seems to be well adapted to the DCM conditions, which would explain its presence in oligotrophic marine zones where nutrients are found deeper.

Numerous marine bacteria show geographical variation of their gene repertoire[59–63] which affects genomic regions that generally represent only a few percent of the total genome[61] and has been proposed, in some cases, to result from horizontal transfer. In *Prochlorococcus*, genomic islands are thought to be related to niche adaptation[63] because they host ecologically important genes[60]. A comparison of two *Prochlorococcus* ecotypes revealed that differences in gene content were related to high-light *vs.* low-light adaptation[64]. Such adaptations have been hypothesized in species closely related to *Bathycoccus*, like *Ostreococcus*[17], but are still a matter of debate[9]. Our data show that the depth and light ranges of the two *Bathycoccus* are different but overlapping, with TOSAG39-1 extending deeper. Interestingly, the surface samples where TOSAG39-1 was detected correspond to sites that undergo vertical mixing (Aghulas and Gulf Stream). Temperature also seemed to influence the distribution of the two *Bathycoccus*, as for example along the Gulf Stream where one type is more prevalent on the West side and is replaced by the other type eastward as water cools down. Among eukaryotes, several examples of correspondence between temperature and geographical distribution have been reported, such as for the heterotrophic MAST-4[26,65] and the Arctic ecotype of *Micromonas*[8]. TOSAG39-1 was also observed at low O_2 concentrations at Costa Rica Dome station 138, an area of high biological production in the East equatorial Pacific[66] where picoplankton can be very abundant[67]. This could reflect the fact that since TOSAG39-1 is better adapted to low light conditions it could be found deeper in the water column where suboxic conditions are developing, rather than having a specific capacity to withstand low O_2.

The wide geographical distribution and relatively high abundance of *Bathycoccus* observed here implies a capability to thrive across a range of ecological niches. Dispensable genes could correspond to the genomic traces of this adaptation. Intriguingly, dispensable *Bathycoccus* genes have genomic features similar to those of

chromosome 19 genes, such as a lower GC content. This suggests that these genes may have been located on chromosome 19 ancestrally and have undergone subsequently inter-chromosomal translocations. A recent experimental evolution experiment of *Ostreococcus tauri* inoculated with a large quantity of virus, Otv5, provided evidence that genes on outlier chromosome 19 are up-regulated in viral-resistant cell lines and that the size of this chromosome varies in resistant lines[68]. Our results on gene content plasticity in Chromosome 19 is consistent with the immunity chromosome hypothesis: frequent events of gene birth and gene loss may thus be the genomic traces of a microalgal – virus evolutionary arm race.

Dispensable genes possess features of so-called *de novo* genes, genes emerging from previously noncoding regions. These genes are an important class of unknown genes and challenge evolutionary sciences[69,70]. It has been hypothesized that cosmopolitan bacteria would hold specific genes or gene variants due to their ecological properties[71]. Cosmopolitan marine lineages are exposed to a range of contrasted environmental constraints, raising the question of their genomic plasticity. The high turnover of a certain class of genes restricted to some environmental conditions might be an evolutionary advantage for rapid acclimation related to being cosmopolitan.

The amplification biases inherent to the Single Cell Genomics approach do not in general allow recovering full genomes from environmental protists. However even incomplete SAG assemblies are sufficient to allow mapping of environmental metagenomes and to determine the distribution of genotypes that are not resolved by traditional marker genes or metabarcodes. In the case of *Bathycoccus* we provide the distribution of two clades, corresponding to the genomes of RCC1105 (clade B1) and to the genome of TOSAG39-1 (clade B2) and identify environmental parameters underlying these distributions. Our observations unfortunately do not cover all oceanic ecosystems, particularly the polar zones. Future analysis of additional genomes and transcriptomes of wild and cultured *Bathycoccus* will improve the accuracy of the environmental niches of the two types of *Bathycoccus*.

Material and Methods

During the *Tara* Oceans expedition[34,35], we collected and cryo-preserved samples at station TARA_039 situated in the Arabian Sea (Supplementary Figure S13, oceanographic conditions are available in reference[36]). In the laboratory, single cells were sorted by flow cytometry based on their size and chlorophyll autofluorescence. Four *Bathycoccus* cells were identified following DNA amplification and 18 S rDNA sequencing[37]. The four amplified genomes (A, B, C, D - Table 1) were individually sequenced using Illumina HiSeq technology, and a suite of tools was used to obtain single-cell final assembly (Supplementary Information). Firstly, individual assemblies were generated using a colored de Bruijn graph-based method[72] and then a final assembly, named here as TOSAG39-1, was generated comprising gap-reduced scaffolded contigs, using SPAdes, SSPACE and GapCloser[73–75] (Supplementary Figure S14). The four cells had identical 18 S sequences and came from the same 4 mL sample, so it is reasonable to presume they were of the same population.

Quality control filters detected and removed contigs or scaffolds that did not correspond to *Bathycoccus* nuclear DNA (Supplementary Figure S14, Supplementary Information). Direct comparisons of sequence assemblies detected putative DNA contamination from other SAGs that were sequenced in the same laboratory and scaffolds corresponding to organelles.

We predicted exon-intron gene structures by integrating various coding regions data. We aligned the reference protein set of the published *Bathycoccus* RCC1105 genome[22] to our assembly. We extracted and sequenced polyA mRNA from *Tara* Oceans samples. We aligned this eukaryote metatranscriptome on TOSAG39-1 assembly. We also used a public protein databank[76] and the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) collection of marine protist transcriptomes[42]. In addition, we performed direct *ab initio* prediction by calibrating and running the Markov model implemented in snap[77]. Integrating and combining all this evidence provided a final set of genes, using a process based on Gmorse software rationale[78]. We evaluated the relative genomic abundance of each genome for two sampled depths (surface and DCM) at the 76 *Tara* Oceans stations (122 samples in total, Supplementary Figure S13) by recruiting metagenomic reads[24]. We mapped metagenomic reads directly from 0.8–5 μm organism-size fraction samples onto genome assemblies, and estimated the relative contribution of each *Bathycoccus* genome in the metagenomes. To obtain a proper genome abundance estimate, we developed methods to select genome-specific signals only (Supplementary Information). We discarded highly conserved genes that were detected by direct sequence comparisons.

A more detailed description of methods is available in the online supplementary information.

References

1. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281,** 237–240 (1998).
2. Scanlan, D. J. *et al.* Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.* **73,** 249–299 (2009).
3. Worden, A. Z., Nolan, J. K. & Palenik, B. Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. *Limnol. Oceanogr.* **49,** 168–179 (2004).
4. Wilkins, D. *et al.* Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics. *Environ. Microbiol.* **15,** 1318–1333 (2013).
5. Vaulot, D., Eikrem, W., Viprey, M. & Moreau, H. The diversity of small eukaryotic phytoplankton (≤3 μm) in marine ecosystems. *FEMS Microbiol. Rev.* **32,** 795–820 (2008).
6. Marin, B. & Melkonian, M. Molecular phylogeny and classification of the Mamiellophyceae class. nov. (Chlorophyta) based on sequence comparisons of the nuclear- and plastid-encoded rRNA operons. *Protist* **161,** 304–336 (2010).
7. Šlapeta, J., López-García, P. & Moreira, D. Global dispersal and ancient cryptic species in the smallest marine eukaryotes. *Mol. Biol. Evol.* **23,** 23–29 (2006).
8. Lovejoy, C. *et al.* Distribution, phylogeny, and growth of cold-adapted picoprasinophytes in arctic seas. *J. Phycol.* **43,** 78–89 (2007).
9. Demir-Hilton, E. *et al.* Global distribution patterns of distinct clades of the photosynthetic picoeukaryote *Ostreococcus*. *ISME J.* **5,** 1095–1107 (2011).
10. Monier, A., Sudek, S., Fast, N. M. & Worden, A. Z. Gene invasion in distant eukaryotic lineages: discovery of mutually exclusive genetic elements reveals marine biodiversity. *ISME J.* **7,** 1764–1774 (2013).

11. Simmons, M. P. *et al.* Abundance and biogeography of picoprasinophyte ecotypes and other phytoplankton in the eastern north pacific ocean. *Appl. Environ. Microbiol.* **82,** 1693–1705 (2016).
12. Foulon, E. *et al.* Ecological niche partitioning in the picoplanktonic green alga *Micromonas pusilla*: evidence from environmental surveys using phylogenetic probes. *Environ. Microbiol.* **10,** 2433–2443 (2008).
13. Worden, A. Z. *et al.* Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324,** 268–272 (2009).
14. Simmons, M. P. *et al.* Intron invasions trace algal speciation and reveal nearly identical arctic and antarctic *Micromonas* populations. *Mol. Biol. Evol.* **32,** 2219–2235 (2015).
15. Chrétiennot-Dinet, M.-J. *et al.* A new marine picoeucaryote: *Ostreococcus tauri* gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia* **34,** 285–292 (1995).
16. Subirana, L. *et al.* Morphology, genome plasticity, and phylogeny in the genus *Ostreococcus* reveal a cryptic species, *O. mediterraneus* sp. nov. (Mamiellales, Mamiellophyceae). *Protist* **164,** 643–659 (2013).
17. Rodríguez, F. *et al.* Ecotype diversity in the marine picoeukaryote *Ostreococcus* (Chlorophyta, Prasinophyceae). *Environ. Microbiol.* **7,** 853–859 (2005).
18. Eikrem, W. & Throndsen, J. The ultrastructure of *Bathycoccus* gen. nov. and *B. prasinos* sp. nov., a non-motile picoplanktonic alga (Chlorophyta, Prasinophyceae) from the Mediterranean and Atlantic. *Phycologia* **29,** 344–350 (1990).
19. Johnson, P. W. & Sieburth, J. M. *In-Situ* morphology and occurrence of eucaryotic phototrophs of bacterial size in the picoplankton of estuarine and oceanic waters. *J. Phycol.* **18,** 318–327 (1982).
20. Collado-Fabbri, S., Vaulot, D. & Ulloa, O. Structure and seasonal dynamics of the eukaryotic picophytoplankton community in a wind-driven coastal upwelling ecosystem. *Limnol. Oceanogr.* **56,** 2334–2346 (2011).
21. Not, F. *et al.* A single species, *Micromonas pusilla* (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. *Appl. Environ. Microbiol.* **70,** 4064–4072 (2004).
22. Moreau, H. *et al.* Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biol.* **13,** R74 (2012).
23. Vaulot, D. *et al.* Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS ONE* **7,** e39648 (2012).
24. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5,** e77 (2007).
25. Hellweger, F. L., van Sebille, E. & Fredrick, N. D. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science* **345,** 1346–1349 (2014).
26. Rodríguez-Martínez, R., Rocap, G., Salazar, G. & Massana, R. Biogeography of the uncultured marine picoeukaryote MAST-4: temperature-driven distribution patterns. *ISME J.* **7,** 1531–1543 (2013).
27. de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348,** 1261605 (2015).
28. Heywood, J. L., Sieracki, M. E., Bellows, W., Poulton, N. J. & Stepanauskas, R. Capturing diversity of marine heterotrophic protists: one cell at a time. *ISME J.* **5,** 674–684 (2011).
29. Lasken, R. S. Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10,** 631–640 (2012).
30. Gasc, C. *et al.* Capturing prokaryotic dark matter genomes. *Res. Microbiol.* **166,** 814–830 (2015).
31. Yoon, H. S. *et al.* Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332,** 714–717 (2011).
32. Martinez-Garcia, M. *et al.* Unveiling *in situ* interactions between marine protists and bacteria through single cell sequencing. *ISME J.* **6,** 703–707 (2012).
33. Roy, R. S. *et al.* Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci. Rep*. **4,** 4780 (2014).
34. Karsenti, E. A journey from reductionist to systemic cell biology aboard the schooner Tara. *Mol. Biol. Cell* **23,** 2403–2406 (2012).
35. Karsenti, E. *et al.* A holistic approach to marine eco-systems biology. *PLoS Biol* **9,** e1001177 (2011).
36. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2,** 150023 (2015).
37. Stepanauskas, R. & Sieracki, M. E. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc. Natl. Acad. Sci. USA*. **104,** 9052–9057 (2007).
38. Bork, P. *et al.* Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. *Science* **348,** 873 (2015).
39. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067 (2007).
40. Monier, A. *et al.* Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. *Environ. Microbiol.* **14,** 162–176 (2012).
41. Decelle, J. *et al.* PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Mol. Ecol. Resour.* **15,** 1435–1445 (2015).
42. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLOS Biol* **12,** e1001889 (2014).
43. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
44. Martin, J. H. *et al.* Testing the iron hypothesis in ecosystems of the equatorial Pacific Ocean. *Nature* **371,** 123–129 (1994).
45. Villar, E. *et al.* Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science* **348,** 1261447–1261447 (2015).
46. Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M. & Stewart, F. J. Microbial oceanography of anoxic oxygen minimum zones. *Proc. Natl. Acad. Sci*. **109,** 15996–16003 (2012).
47. Morel, A. *et al.* Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. *Remote Sens. Environ.* **111,** 69–88 (2007).
48. Cullen, J. J. Subsurface chlorophyll maximum Layers: enduring enigma or mystery solved? *Annu. Rev. Mar. Sci.* **7,** 207–239 (2015).
49. Fernández-Castro, B. *et al.* Importance of salt fingering for new nitrogen supply in the oligotrophic ocean. *Nat. Commun.* **6,** 8002 (2015).
50. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev*. **15,** 589–594 (2005).
51. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17,** 175–188 (2016).
52. Palenik, B. *et al.* The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. USA* **104,** 7705–7710 (2007).
53. Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431,** 946–957 (2004).
54. Makałowski, W., Zhang, J. & Boguski, M. S. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* **6,** 846–857 (1996).
55. Mouse Genome, Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
56. Coleman, A. W. Is there a molecular key to the level of 'biological species' in eukaryotes? A DNA guide. *Mol. Phylogenet. Evol.* **50,** 197–203 (2009).
57. James, S. A., Cai, J., Roberts, I. N. & Collins, M. D. A phylogenetic analysis of the genus *Saccharomyces* based on 18S rRNA gene sequences: description of *Saccharomyces kunashirensis* sp. nov. and *Saccharomyces martiniae* sp. nov. *Int. J. Syst. Bacteriol.* **47,** 453–460 (1997).

58. Bendif, E. M. *et al.* Genetic delineation between and within the widespread coccolithophore morpho-species *Emiliania huxleyi* and *Gephyrocapsa oceanica* (Haptophyta). *J. Phycol.* **50,** 140–148 (2014).
59. Acuña, L. G. *et al.* Architecture and gene repertoire of the flexible genome of the extreme acidophile *Acidithiobacillus caldus*. *PLoS ONE* **8,** (2013).
60. Coleman, M. L. *et al.* Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311,** 1768–1770 (2006).
61. Fernández-Gómez, B. *et al.* Patterns and architecture of genomic islands in marine bacteria. *BMC Genomics* **13,** 347 (2012).
62. Gonzaga, A. *et al.* Polyclonality of concurrent natural populations of *Alteromonas macleodii*. *Genome Biol. Evol.* **4,** 1360–1374 (2012).
63. Kashtan, N. *et al.* Single-Cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344,** 416–420 (2014).
64. Rocap, G. *et al.* Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424,** 1042–1047 (2003).
65. Lin, Y.-C. *et al.* Distribution patterns and phylogeny of marine Stramenopiles in the North Pacific Ocean. *Appl. Environ. Microbiol.* **78,** 3387–3399 (2012).
66. Fiedler, P. C. The annual cycle and biological effects of the Costa Rica Dome. *Deep Sea North Pacific Ocean Res. Part Oceanogr. Res. Pap*. **49,** 321–338 (2002).
67. Ahlgren, N. A. *et al.* The unique trace metal and mixed layer conditions of the Costa Rica upwelling dome support a distinct and dense community of *Synechococcus*. *Limnol. Oceanogr.* **59,** 2166–2184 (2014).
68. Yau, S. *et al.* A viral immunity chromosome in the marine picoeukaryote, *Ostreococcus tauri*. *PLoS Pathog.* Part I **12,** e1005965 (2016).
69. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* **487,** 370–374 (2012).
70. Schlötterer, C. Genes from scratch – the evolutionary fate of de novo genes. *Trends Genet.* **31,** 215–219 (2015).
71. Ramette, A. & Tiedje, J. M. Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microb. Ecol.* **53,** 197–207 (2007).
72. Chitsaz, H. *et al.* Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat. Biotechnol.* **29,** 915–921 (2011).
73. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **19,** 455–477 (2012).
74. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinforma. Oxf. Engl.* **27,** 578–579 (2011).
75. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1,** 18 (2012).
76. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinforma. Oxf. Engl.* **23,** 1282–1288 (2007).
77. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5,** 59 (2004).
78. Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9,** R175 (2008).
79. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).

Acknowledgements

We thank the commitment of the following people and sponsors who made this expedition possible: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, the French Governement 'Investissement d'Avenir' programs Oceanomics (ANR-11-BTBR-0008) and FRANCE GENOMIQUE (ANR-10-INBS-09), Fund for Scientific Research – Flanders, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, ANR (projects POSEIDON/ANR-09-BLAN-0348, PROMETHEUS/ANR-09-PCS-GENM-217, TARA-GIRUS/ANR-09-PCS-GENM-218), EU FP7 MicroB3/No.287589, US NSF grant DEB-1031049 to MES, FWO, BIO5, Biosphere 2, Agnès b., the Veolia Environment Foundation, Region Bretagne, World Courier, Illumina, Cap L'Orient, the EDF Foundation EDF Diversiterre, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, and not least, the *Tara* schooner and its captain and crew. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We acknowledge Samuel Chaffron, Lionel Guidi and Lars Stemmann for help with the environmental parameters, Claude Scarpelli for support with the high-performance computing. We warmly thank Gwenael Piganeau for reading and suggestions on this manuscript. We thank members of the *Tara* Oceans consortium, coordinated by Eric Karsenti, for the creative environment and constructive criticism.

Author Contributions

C.d.V., M.S., P.W. and O.J. designed the study. O.J. wrote the paper, with significant inputs from D.V., T.V. and P.W. M.S. managed the single cell isolation; Y.S. and J.M.A. managed the SAG assembly and gene predictions. T.V. and O.J. analyzed the genomic data, with significant input from J.L., Y.S., S.M., E.P., J.M.A., D.V. and P.W. T.V., J.L., D.V., D.I. and O.J. analyzed the oceanographic data. All authors discussed the results and commented on the manuscript.

Additional Information

Accession codes: This article is contribution number 48 of Tara Oceans. Physicochemical parameters from all Tara Oceans samples are available at Pangea (http://doi.pangaea.de/10.1594/PANGAEA.840721); metagenomics reads can be downloaded at SRA under identification study number PRJEB402 (https://www.ebi.ac.uk/ena/data/view/PRJEB402). The sequences of TOSAG39-1 were deposited and are available at EMBL/DDBL/GenBank under accession number ERA768231.

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Vannier, T. *et al.* Survey of the green picoalga *Bathycoccus* genomes in the global ocean. *Sci. Rep.* **6**, 37900; doi: 10.1038/srep37900 (2016).

II. Article 2 : Genome Resolved Biogeography of Mamiellales

Suite à cette première publication démontrant la différente répartition de deux espèces de *Bathycoccus*, nous avons souhaité étendre ces analyses aux autres organismes de l'ordre des Mamiellales, *Micromonas* et *Ostreoccocus*. Nous avons donc réuni les génomes de référence disponibles pour *O. tauri*, *O. mediterraneus*, *O.* RCC809, *O. lucimarinus*, *M. pusilla* et *M. commoda* en plus de *B. prasinos* et TOSAG39-1. Nous les avons donc étudiés de manière similaire en utilisant les 133 échantillons métagénomiques disponibles provenant toujours de la même expédition.

Malheureusement, *O. tauri* et *O. mediterraneus* n'ont pas du tout été détectés dans nos échantillons. Les deux *Ostreococcus*, deux *Micromonas* et deux *Bathycoccus* restants ont donc été étudiés dans l'article *Genome Resolved Biogeography of Mamiellales* publié dans le journal *Genes*. A ce manuscrit j'ai contribué à l'ensemble des analyses présentées ainsi qu'à la majorité de la rédaction.

L'axe principal de cette étude est l'analyse comparative des localisations et préférences environnementales de ces espèces basée sur la méthode de recrutement des lectures à partir d'une référence, déjà utilisée pour la publication précédente. Les résultats ont confirmé la distribution cosmopolite des Mamiellales et suggèrent une co-occurence des espèces avec notamment deux triplets d'organismes associés à différentes températures. En effet, *Ostreococcus lucimarinus*, *Bathycoccus prasinos* et *Micromonas pusilla* ont été trouvés ensemble dans des eaux plus froides que *Ostreococcus* RCC809, *Bathycoccus* TOSAG39-1 et *Micromonas commoda*, partageant également un certain nombre d'échantillons communs de leur côté.

Un axe secondaire qui a pu être exploré grâce à l'étude des génomes entiers et à la couverture génomique suffisante obtenue dans certaines stations est l'étude des mating-type. En effet, des séquences de gènes des deux types sexuels, appelés *MT+* et *MT-* , ont récemment été décrites[131] pour *O. mediterraneus* et *O. lucimarinus*. Le second étant abondant dans les échantillons *Tara*, nous avons recruté des lectures à partir de ces séquences de gènes liés au sexe et comparé les résultats avec les couvertures du BOC provenant du génome de référence sur lequel se trouvent également ces gènes. Nous avons ainsi pu montrer que la couverture moyenne de la zone

non-outlier correspondait au total des deux mating-types réunis, tandis que la couverture de zone outlier du chromosome ne correspondait qu'à un seul mating-type, la différence entre les deux correspondant donc, par soustraction, à la couverture du second mating-type. En procédant ainsi, nous avons pu analyser l'abondance relative de *MT+* et *MT-* dans les 11 échantillons où *O. lucimarinus* est abondant, montrant une écrasante majorité de *MT+* dans l'ensemble des échantillons mais avec une présence systématique des deux versions. Les échantillons provenant de l'océan austral ainsi que de l'upwelling du Chili contiennent tout de même jusqu'à un tiers de *MT-* ce qui est bien plus que les autres échantillons qui ne présentent qu'environ 5% de *MT-*. Les données supplémentaires correspondant à cet article sont disponibles en Annexe 2.

genes

MDPI

Article

Genome Resolved Biogeography of Mamiellales

Jade Leconte [1,2], L. Felipe Benites [3], Thomas Vannier [1,2], Patrick Wincker [1,2], Gwenael Piganeau [3,*] and Olivier Jaillon [1,2,*]

1 Génomique Métabolique, Genoscope, Institut de Biologie François Jacob, Commissariat à l'Énergie Atomique (CEA), CNRS, Université Évry, Université Paris-Saclay, 91057 Évry, France; jleconte@genoscope.cns.fr (J.L.); thomas.VANNIER@univ-amu.fr (T.V.); pwincker@genoscope.cns.fr (P.W.)

2 Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, 3 rue Michel-Ange, 75016 Paris, France

3 Observatoire Océanologique, UMR 7232 Biologie Intégrative des Organismes Marins BIOM, CNRS, Sorbonne Université, F-66650 Banyuls-sur-Mer, France; benites@obs-banyuls.fr

* Correspondence: gwenael.piganeau@obs-banyuls.fr (G.P.); ojaillon@genoscope.cns.fr (O.J.)

Received: 29 November 2019; Accepted: 3 January 2020; Published: 7 January 2020

Abstract: Among marine phytoplankton, Mamiellales encompass several species from the genera *Micromonas*, *Ostreococcus* and *Bathycoccus*, which are important contributors to primary production. Previous studies based on single gene markers described their wide geographical distribution but led to discussion because of the uneven taxonomic resolution of the method. Here, we leverage genome sequences for six Mamiellales species, two from each genus *Micromonas*, *Ostreococcus* and *Bathycoccus,* to investigate their distribution across 133 stations sampled during the *Tara* Oceans expedition. Our study confirms the cosmopolitan distribution of Mamiellales and further suggests non-random distribution of species, with two triplets of co-occurring genomes associated with different temperatures: *Ostreococcus lucimarinus*, *Bathycoccus prasinos* and *Micromonas pusilla* were found in colder waters, whereas *Ostreococcus* spp. RCC809, *Bathycoccus* spp. TOSAG39-1 and *Micromonas commoda* were more abundant in warmer conditions. We also report the distribution of the two candidate mating-types of *Ostreococcus* for which the frequency of sexual reproduction was previously assumed to be very low. Indeed, both mating types were systematically detected together in agreement with either frequent sexual reproduction or the high prevalence of a diploid stage. Altogether, these analyses provide novel insights into Mamiellales' biogeography and raise novel testable hypotheses about their life cycle and ecology.

Keywords: Mamiellales; biogeography; *Tara* Oceans; sexual reproduction; mating-type; ecogenomics

1. Introduction

Mamiellales is an order of green algae (Chlorophyta) that contains some of the most ecologically important groups of photosynthetic picoeukaryotes in the marine environment [1–3]. They are prevalent and abundant in coastal surface waters and throughout the oceanic euphotic zone [4–7] where populations can reach high densities up to 10^3–10^5 cells per ml [8].

This order is composed of two families. The Bathycoccaceae are represented by the genera *Ostreococcus* and *Bathycoccus*, which are distributed across a range of marine environments [2,3,9]. Within Bathycoccaceae, previous metabarcoding analyses based on regions of the highly conserved 18S ribotype suggested niche differentiation: sequences of *Ostreococcus tauri* are detected in coastal and lagoonal environments while *Ostreococcus lucimarinus* and *Ostreococcus* spp. RCC809 are more broadly distributed in oceanic open regions [6,9]. Recently, a novel *Ostreococcus* clade E group, identified by a different 18S sequence, was found to be the most prevalent Mamiellales in the Mediterranean Sea and in coastal warm temperate sites on both sides of the Atlantic [9]. The genus *Bathycoccus* is composed

of two cryptic species [10,11] with identical 18s rRNA sequences but marked differences in their ITS (Internal Transcribed Spacer) region and highly divergent genome sequences. Indeed, orthologous proteins share 78% amino-acid identity and only 26 highly conserved genes (>99%) [7,12] are found. This is similar to genome divergence reported for different species [13].

The second family of the Mamiellales order is Mamiellaceae, comprising the genera *Micromonas*, *Mantoniella* and *Mamiella* that are widespread from tropical to polar regions. Within Mamiellaceae, the genus *Micromonas* comprises the most described species with a global distribution in coastal and open ocean areas and with species adapted to polar environments [12,14,15]. *Micromonas pusilla* [16] was recently split into four species, namely, *Micromonas bravo*, *Micromonas commoda*, *Micromonas polaris*, *Micromonas pusilla*, and two clades described as candidate species 1 (clade B_4) and candidate species 2 (clade B warm) [9,15].

There are many complete genomes available for Mamiellales [17]; 17 genomes have been sequenced from four *Ostreococcus* species [18–20], while there are two sequenced genomes for *Micromonas*, *M. pusilla* CCMP1545 isolated in 1950 in the North Atlantic sea near Plymouth (England) and *M. commoda* RCC299 isolated in the South Pacific in 1998 [5]. *Bathycoccus* has two reference genomes, one complete genome from the strain RCC1105 [10], isolated in the Banyuls bay, and a 64% complete single amplified genome co-assembly TOSAG39–1 from the Arabian sea [11]. One common feature shared by all Mamiellales genomes is the presence of two unusual "outlier" chromosomes which present striking structural and compositional differences from standard chromosomes [21]. In *O. tauri*, these chromosomes are chromosome 2 (big outlier chromosome or BOC) and chromosome 19 (small outlier chromosome or SOC) [22,23]. In *M. pusilla* CCMP1545 BOC and SOC are represented respectively by chromosome 2 and chromosome 19 while in *M. commoda* RCC299 BOC is chromosome 1 [24] and SOC is chromosome 17. In *Bathycoccus*, the outlier chromosomes are chromosome 14 (BOC) and 19 (SOC) [10]. In all of these genomes, the BOC contains one contiguous low GC content region, flanked by two high GC content regions. The low-GC region in *O. tauri* encodes two highly divergent haplotypes, and thorough phylogenetic analysis of these two haplotypes suggests that this low-GC region designates the mating-type in Mamiellales [10,25]. Gene sequences from the two mating-types have been recently described in two additional *Ostreococcus* species, *O. mediterraneus* and *O. lucimarinus*, confirming an ancient origin of this mating type in this genus [26].

Biogeography of marine organisms faces numerous technical issues owing to the complexity of the ecosystem, especially for microbes. Consequently, knowledge about species distributions has long remained very scarce and limited to visible organisms [27]. Nonetheless, more recent studies based on microscopy or DNA sequencing addressed planktonic organisms, including viruses [28,29], bacteria, unicellular eukaryotes [30,31] and small animals [32]. Most environmental genomics surveys of organisms are based on sequencing of taxonomic marker genes, known as the barcode approach. This approach relies on the specificity of the marker and on the availability of a large taxonomic reference database such as PR2 [33]. It has been demonstrated that commonly used taxonomic marker genes have uneven levels of resolution [34] and only genome sequencing can solve this issue [11]. In that context, metagenomes can be used as resources to discover and quantify the presence of organisms for which other genomic information is available. Single cell genomes or metagenome-assembled genomes are very valuable approaches in that perspective especially for uncultured organisms [35–37]. However, these techniques still in their infancy are biased, and while eukaryotes for which compact and abundant genomes are most likely to be successfully detected, special attention must be paid to the possibility of assembling chimeric genomes [35].

Previous analysis of *Tara* Oceans metagenomes from 122 stations using the two *Bathycoccus* genomes to recruit reads revealed that the two species rarely co-occur and occupy distinct oceanic niches with respect to depth [11]. Here, we leverage available whole genome data in Mamiellales and metagenomic sequences in 133 samples collected from 80 sites from the *Tara* Oceans expedition to investigate the biogeography of Mamiellales and the level of diversity of the putative mating-type chromosome on a global scale.

2. Materials and Methods

2.1. Genomic Resources

Six Mamiellales reference genomes including *Bathycoccus prasinos* RCC1105 and *Bathyococcus* TOSAG39–1, *M. commoda* RCC299 and *M. pusilla* CCMP1545, and *O.* RCC809 and *O. lucimarinus* strain CCE9901 (Table 1), were used to search a large number of open ocean metagenomic samples from the *Tara* Oceans expedition (Table S1) [38]. Genomes from *O. tauri* RCC4221 and *O. mediterraneus* RCC2590 were also searched but were not found in the *Tara* metagenomic dataset. We also used genes positions data from the Pico-PLAZA platform for each genome [17]. This study focused on samples from the 0.8–5 µm organism-size fraction, corresponding to the cell size of Mamiellales, adding up to a total of 79 surface samples as well as 54 samples from the deep-chlorophyll maximum (DCM). Those 133 samples were taken from 80 different sites located in the Mediterranean Sea, Indian, Pacific and Atlantic Oceans.

Table 1. Genome data used.

Species	Source	Genome Size	Sampling Site	Sampling Year
B. prasinos (RCC1105)	pico-PLAZA [17]	15.1 Mb	Mediterranean Sea, France, Banyuls Bay	2006
B. spp. TOSAG39-1	*Tara* Oceans single-cell [11]	10.3 Mb (64% complete)	Indian Ocean, Arabian Sea, Station TARA_039	2010
M. pusilla (CCMP1545)	pico-PLAZA [17]	21.9 Mb	Atlantic Ocean, United Kingdom, near Plymouth	1950
M. commode (RCC299)	pico-PLAZA [17]	20.9 Mb	Pacific Ocean, Equatorial Pacific, New Caledonia	1998
O. lucimarinus (CCE9901)	pico-PLAZA [17]	13.2 Mb	Pacific Ocean, USA, California	2001
O. spp. RCC809	pico-PLAZA [17]	13.3 Mb	Atlantic Ocean, Tropical Atlantic, international waters	1991

Genomic abundance of each genome in those *Tara* Oceans stations was estimated by mapping metagenomic reads onto the reference genome sequences using Bowtie2 2.1.0 aligner with default parameters [39]. We filtered out alignments corresponding to low complexity regions using the dust algorithm [40] and alignments with less than 95% mean identity or with less than 30% of high complexity bases were also discarded. These identity thresholds were compatible with intraspecific levels of diversity around 1% estimated from a population genomics study in *O. tauri* [19] so that all reads recruited to a reference genome could be assumed to belong to the same species, despite intraspecific variation. We then computed relative abundances as the number of reads mapped onto genes normalized by the total number of reads sequenced for each sample. In order to avoid non-specific mapping signals, we defined a set of outlier genes for each genome. Genes with atypical mapping behavior based on the distribution of deviant numbers of recruited metagenomic reads, and organellar genes, were discarded from the analysis similarly to Seeleuthner et al. [35]. Biogeographical maps were plotted using R-packages ggplot2_2.2.1, scales_0.4.1, maps_3.1.1 and ggtree_1.6.11. We computed principal component analysis (PCA) using the vegan_2.4-1 R-package and we used the Spearman correlation coefficient to estimate correlations between relative abundances of species.

2.2. Environmental Analysis

To study the link between geographical distribution of species and abiotic factors, we used the physicochemical parameter values related to the *Tara* Oceans expedition sampling sites available in the PANGAEA database [41], (http://doi.org/10.1594/PANGAEA.875575, https://doi.org/10.1594/PANGAEA.875576).

We extracted the median values for a set of parameters available for each sampling location, including depth, temperature, oxygen, salinity, photosynthetically active radiation (PAR) on the sampling week using a diffuse attenuation coefficient, concentration of nitrates, nitrates + nitrites, phosphate, silicate and chlorophyll *a*. We supplemented this dataset with simulated values for iron and ammonium (using the MITgcm Darwin model [42]).

We tested whether the occurrence of the six Mamiellales in *Tara* Oceans samples was correlated with local physicochemical conditions. For each parameter, we performed a Kruskal–Wallis test with the R base stats package followed by a post-hoc Tukey's test using nparcomp_2.6 for significant parameters (p-value < 0.05).

2.3. Mating Types

We screened the *Tara* Oceans metagenome datasets for the presence of sequences of the recently identified 23 core gene families (GFs) of the two mating types (MTs) of *O. lucimarinus* [26,43] to estimate their ratios among samples.

The two different mating types (*MT*+ and *MT*−) were previously sequenced in the strains BCC118000 (MMETSP0939) and CCE9901 respectively. The latter also corresponded to the reference genome of *O. lucimarinus* used in our biogeography analysis. We mapped metagenomic reads from eleven metagenomes where the relative abundance of *O. lucimarinus* was above 0.1% on the 23 GFs (in total 16 *MT*+ and 41 *MT*− genes sequences) with the same tools and thresholds as described for read recruitment on the complete genomes.

For each sample, we normalized the estimated relative abundance of *MT*+ and *MT*− sequences by dividing the number of recruited metagenomic reads by the total number of reads of the corresponding sample. Secondly, to get a single value of relative abundance for *MT*+ and *MT*−, we averaged the relative abundances of their corresponding genes.

Then, ratios were computed by using these mean relative gene abundances. This approach directly using mating type sequences will be referred from here as the MT genes method.

A second method to estimate the *MT*+ and *MT*− ratios was based on the proportion of metagenomic reads that were mapped on the mating type locus of chromosome 2 (BOC) following the protocol described above.

The vertical coverage of metagenomic reads (number of metagenomic reads cumulatively mapped at a given position) mapped on BOC chromosome was heterogeneous, the genomic region containing the candidate mating type locus, known to have low GC composition, recruited fewer metagenomic reads than the rest of the chromosome (Figure S1). Considering that the reference genome was *MT*−, this pattern suggested that lack of coverage in this region would correspond to the presence of the opposite mating-type in the sample.

We thus estimated the relative read coverages in the low-GC genomic region of chromosome 2, which encodes the putative mating types in Mamielles [19], compared to the standard coverage of this chromosome which corresponds to both mating-types cumulated. Those values were used as a proxy for the relative proportion of the *MT*− strains in the corresponding samples. This will be referred to as the whole chromosome method.

3. Results

3.1. Genome Resolved Distribution of Mamiellales

In order to gain a global view of their geographical distribution, we estimated the proportions of metagenomic reads from each of 133 *Tara* Oceans metagenomes that were recruited to six Mamiellales genomes (Figure 1). Mamiellales were found in 68 out of the 133 samples and were distributed across all ocean regions. Altogether, the six genomes recruited 1.38% of total metagenomic reads of the merged 68 metagenomes, with a maximum of 4.8% in one sample. *Bathycoccus* TOSAG39-1 was found in 43 out of 68 different samples, making it the most cosmopolitan species, followed

by *O.* RCC809, *B. prasinos*, and *M. commoda* (Table 2). The presence of *O. lucimarinus* and *M. pusilla* was detected in 11 and five samples, respectively. However, despite those relatively lower numbers, the sites where they were detected were located in different oceans, consistent with a wide geographical distribution. Taking the proportion of recruited reads to each genome as a proxy for relative species abundance [11,35], the frequencies of the six Mamiellales appeared to be similar, ranging from 0.11% to 0.65% of the total amount of metagenomic reads, with a local maximum represented by *O.* RCC809 that recruited up to 4% of metagenomic reads in a DCM sample of a station situated in the Indian Ocean. The six Mamiellales together recruited 1.4% of all metagenomic reads from 68 stations, but with a maximum of 4.8% of metagenomic reads in any given station.

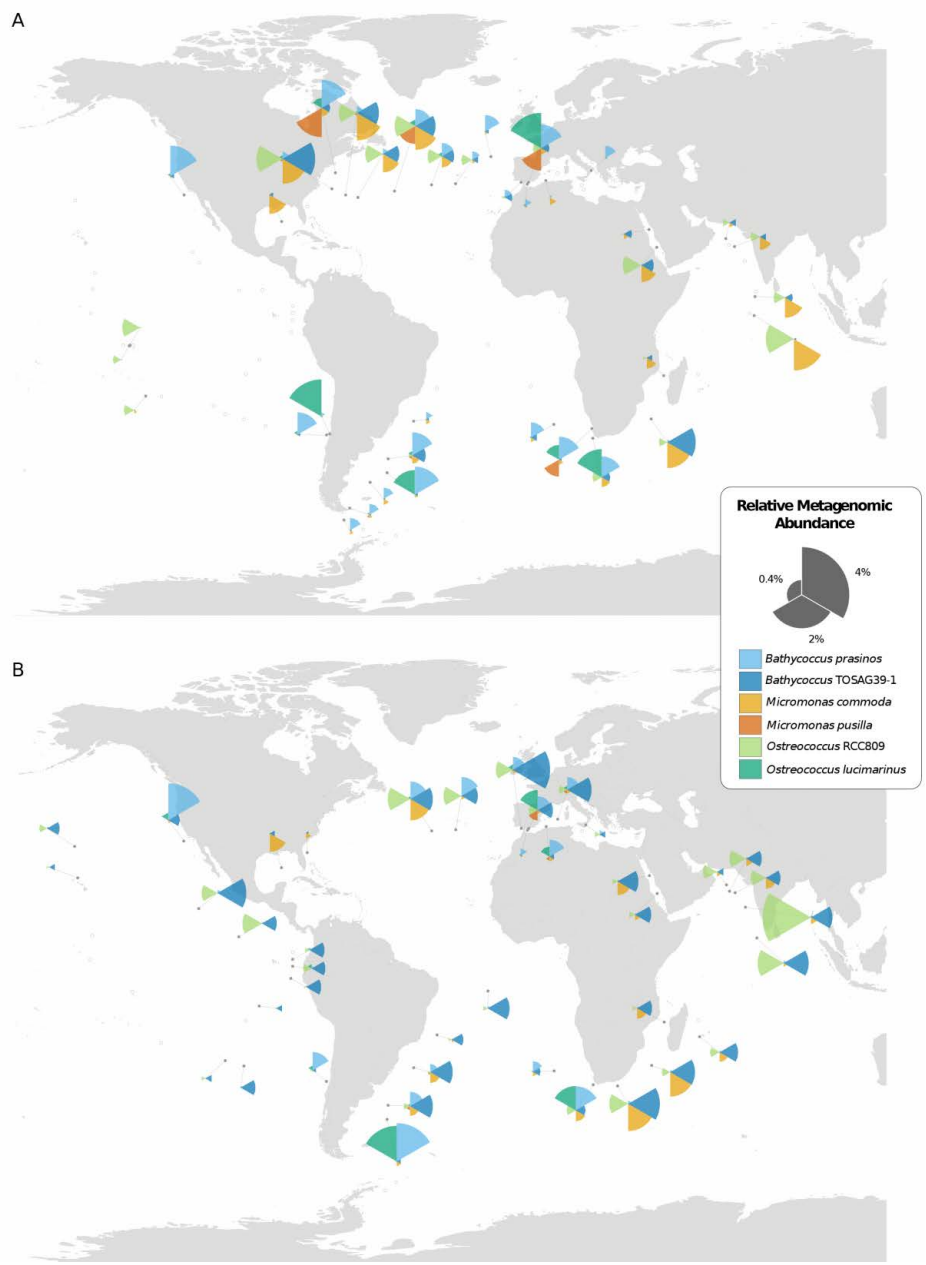


Figure 1. Geographical distribution of the six Mamiellales genomes in Tara Oceans stations from (A) surface and (B) deep-chlorophyll maximum (DCM) waters, as inferred from the relative abundance of recruited metagenomic reads. Samples with less than 0.1% relative abundance of a species are displayed as an empty circle. The sizes of the segments of coxcomb charts indicate the relative genomic abundances of the corresponding Mamiellales.

Table 2. Mamiellales repartition and abundances among *Tara* Oceans samples.

Species	Number of Samples (Abundance > 0.1%)	Percentage of Reads (Merged Samples with Reference)	Maximum Abundance in Sample
B. prasinos	33	0.65%	2.54%
B. TOSAG39-1	43	0.54%	2.41%
M. commoda	27	0.43%	1.70%
M. pusilla	5	0.11%	1.47%
O. RCC809	34	0.54%	4.07%
O. lucimarinus	11	0.35%	2.36%
Total	68	1.38%	4.80%

All species were present in both surface and DCM samples, but three species displayed preference for one water depth. *B. prasinos* and *M. pusilla* were mostly found in surface waters while *B.* TOSAG39-1 was more frequent in DCM waters. The three other species were equally distributed between the two depths. Comparing the basins for which *Tara* Oceans samples were available, surface water from the Eastern Pacific Ocean appeared to be the only area where Mamiellales were not detected at a significant level (no species reaching 0.1% relative abundance).

A visual inspection of the geographical distribution of the six species suggested a pattern of co-occurrence between some of them. For example, in the Indian Ocean, *B.* TOSAG39-1 and *M. commoda* always appeared together with the former being dominant. To address the question of co-occurrences, we computed pairwise correlation tests between all pairs of the six Mamiellales based on theirrespective metagenomic abundances (Table 3). We obtained statistically significant positive correlations between *B.* TOSAG39-1, *M. commoda* and *O.* RCC809 abundances. Also, abundances of O. lucimarinus and B. prasinos were strongly correlated. *M. pusilla* being the least abundant was found in few sites but was detected at the same locations as O. lucimarinus. These two groups of triplets appeared geographically segregated. Finally, we compared the distributions of the six Mamiellales through a principal component analysis (PCA) based on their metagenomic abundances (Figure 2).

Table 3. Correlations between occurrences of the six Mamiellales genomes.

Species	*B. prasinos*	*O. lucimarinus*	*M. pusilla*	*B.* TOSAG39-1	*O.* RCC809
M. commoda	−0.20	−0.14	−0.01	0.37 ***	0.28 *
O. RCC809	−0.19	−0.13	−0.05	0.33 ***	
B. TOSAG39-1	−0.22	−0.21	−0.09		
M. pusilla	0.31	0.24 *			
O. lucimarinus	0.47 ***				

Level of confidence: * p-value < 0.05, *** p-value < 0.001.

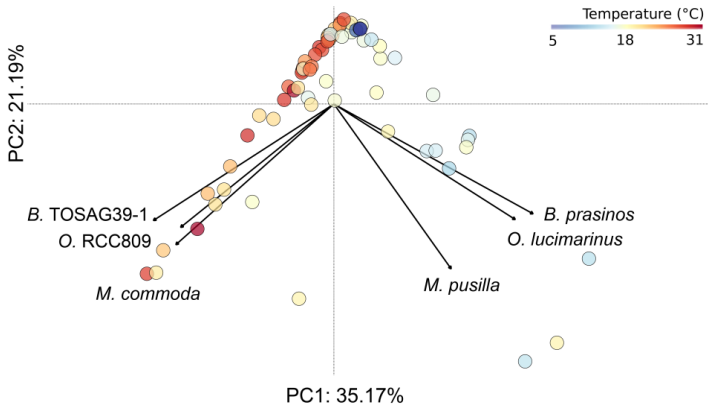


Figure 2. Principal Component Analysis computed on relatived metagenomic abundance of the six Mamiellales. Each circle corresponds to a sample and is colored according to water temperature.

The first two axes of the PCA explain more than half the variance, meaning a good representation of our sample distribution on those components. This analysis showed a very clear pattern of segregation between these two triplets of genomes that have co-occurring abundances. Each triplet was composed of one of the *Bathycoccus*, *Micromonas* and *Ostreococcus*, and the two groups, when they presented high abundances, were found in distinct samples.

3.2. Environmental Variables Linked to Strain Presence

Adding a color corresponding to water temperature for the samples in the PCA figure clearly revealed that the two triplets were found at different ranges of temperatures, with one group found in warmer water than the other. Therefore, we analyzed the ecological preferences of the six species and statistically tested whether available physicochemical parameters were able to discriminate them (Figure 3).

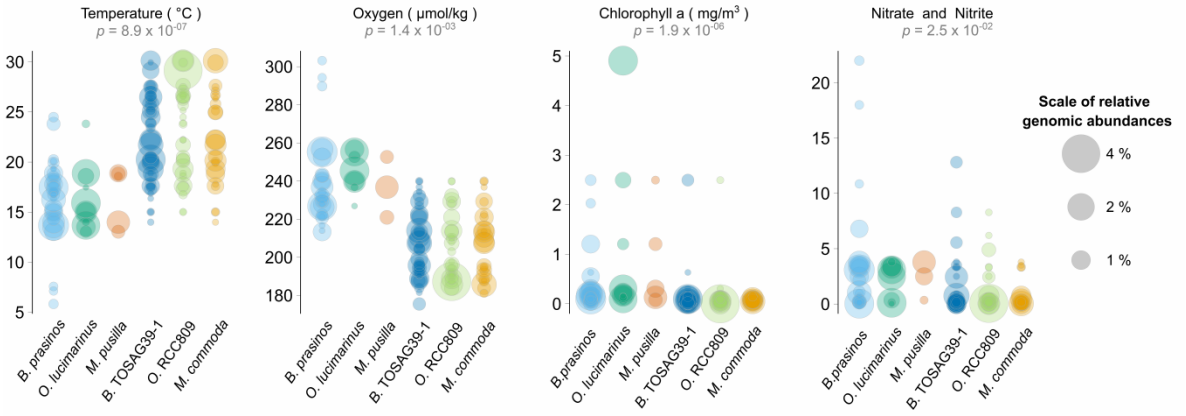


Figure 3. Ranges of values of environmental parameters where a significant difference was detected among Mamiellales species. Each circle corresponds to the relative metagenomic abundance of corresponding species in a given sample. A sample where several species are present is thus represented for these corresponding species (at the same value on the Y axis) but possibly with different circle sizes. p-values correspond to a Kruskal–Wallis test using the six Mamiellales (non-significant environmental parameters are not shown).

From 12 physicochemical environmental parameters, a Kruskal–Wallis test was significant for four of them and temperature was the most significant. As expected, oxygen which is well-known to be directly anti-correlated to water temperature was significant. The third discriminating parameter was chlorophyll a, with a more significant p-value than nitrate and nitrite nutrients (phosphate was not significant). For each environmental parameter, we computed pairwise species statistical comparisons (post-hoc Tukey test) in order to determine whether or not genomes were found in similar environments. This test confirmed different ecological preferences in all paired species from distinct triplets (Table S2).

3.3. Mating Types

We investigated the biogeography of the two candidate mating types of O. lucimarinus by two methods. The first method, the MT whole chromosome method, was based on differential numbers of metagenomic reads that were recruited to mating type (MT) versus non-MT loci. The second method, the MT genes method, was based on the differential numbers of metagenomic reads that could be mapped on MT+ and MT− gene sequences (Methods). The two approaches provided highly correlated estimations of mating type ratios (linear regression r2 = 0.99, Figure S2). However, the estimations based on MT genes suggested higher frequencies of the MT+ as compared to the estimations inferred from the whole chromosome method. This may be due to the number of marker genes in the MT+ gene set which may not be representative of the whole mating type region. Strikingly, we detected the presence of both mating types in each sample where the genome had been found. Moreover, the mating

type ratio was estimated to be variable between sites but systematically biased toward the *MT+* type from 66% up to 97% (Figure 4) using the *MT* genes method. This bias was estimated to be lower, within the 52.3% to 95.2% range, when calculated from the whole chromosome. Using the latter approach, the mating type ratio was close to 50% in both surface and DCM samples of the *Tara* Oceans station 81 (52.3% and 55.0% respectively) which is the most austral site where *O. lucimarinus* was detected. These analyses provide strong evidence of the presence of both mating-types in the samples we analyzed. In addition, the mating type genes were mapped at 99.3% and 99.6% of identity on average on the *MT+* and *MT−* genes respectively, versus 98.6% on the genes located on standard regions of the genome, consistent with their very high conservation level and the lack of putative technical bias due to sequence divergence. Finally, mating type genes were only detected in samples presenting a minimum of 0.1% relative abundance of *O. lucimarinus*. The sum of *MT+* and *MT−* abundances fits the distribution of the total abundance estimated from whole genome read recruitment (Figure S3). From a technical point of view, this comparison also provides evidence that the mating type ratio can be estimated based on differential metagenomic read coverage for any species for which the mating-type region has been identified, even if the sequence of the alternative mating-type or mating-types is not yet available. In Mamiellales, the species with only one sequenced mating-type were *O.* RCC809 [43]

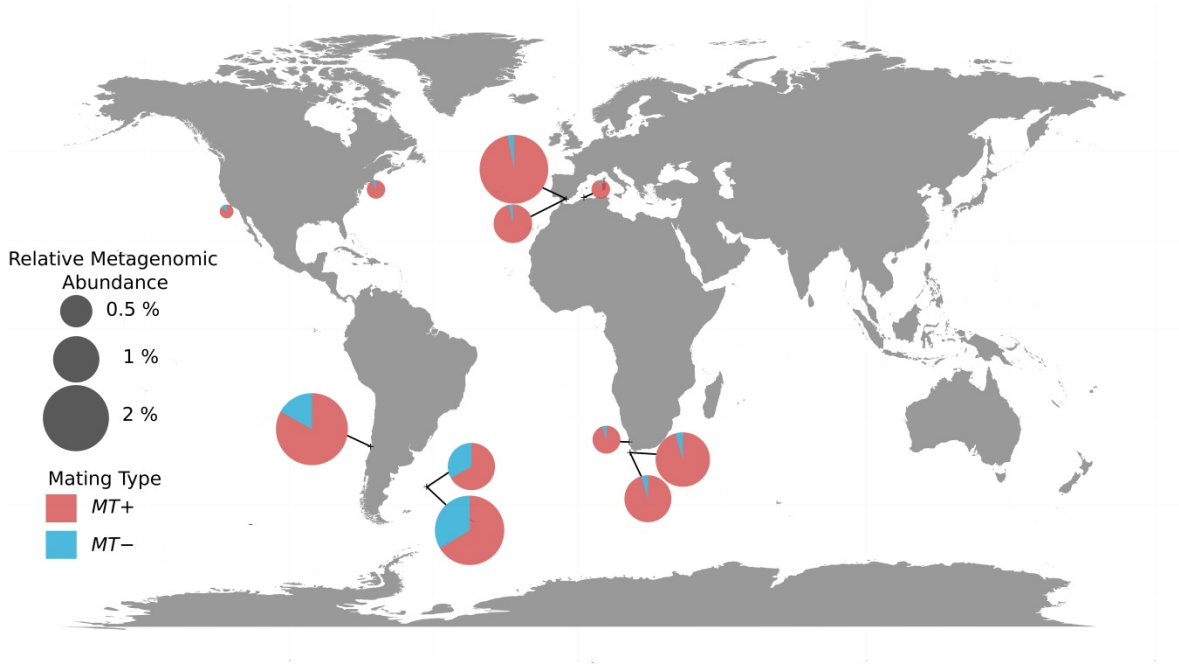


Figure 4. Geographical distribution of the two mating types of *Ostreococcus lucimarinus* in *Tara* Oceans stations. Each pie-chart represents a sample with a size relative to the relative metagenomic abundance of whole genome, and shows the average proportions of the two mating types genes. Ratios were determined using the MT genes method (Methods).

4. Discussion

Following a previous whole-genome-resolved biogeographic approach based on Tara Oceans data and focused on B. prasinos and B. TOSAG39-1 [11], we applied the same approach to four Mamiellales reference genomes from the *Micromonas* and *Ostreococcus* genus and added 11 metagenomes to the previous dataset.

This whole genome resolved biogeographic approach enabled a better taxonomic resolution than the ribotype metabarcoding approach [14,44,45], whose resolution power is limited by the very high conservation of the 18S rDNA sequence [34]. Indeed, the V9 region of the 18S sequences is identical in all species within Ostreococcus and Bathycoccus [45].

Furthermore, screening the *Tara* Oceans metabarcoding dataset [46,47] with a similar detection threshold (minimum 0.1% relative abundance) we did not detect Mamiellales in more samples as compared with the whole genome approach. Indeed, while the whole genome approach detected Mamiellales in 68 out of our 133 samples (51%), the V9 approach detected them in 70 out of 164 available samples (48%) or in 65 out of the 131 samples (50%) that are common between the two datasets. This suggests that the metabarcoding approach did not provide a better geographical resolution than the metagenomic one, probably because we were focusing here on abundant species with relatively small genomes.

The whole genome approach revealed intriguing co-occurrence patterns between triplets of species from the three different genera: *B. prasinos*, *O. lucimarinus* and *M. pusilla* or *B.* TOSAG39-1, *O.* spp. RCC809 and *M. commoda*. It remains to be shown whether these co-occurrences are the consequence of random sampling from the same water masses, as for example within the Indian Ocean or within the Gulf Stream in which we can notice similar patterns along currents, or whether these species are adapted to the same environmental conditions. A test of the neutral hypothesis could be performed by a competition experiment between these six strains at 15 and 22 °C, the mean temperature in which each triplet was detected, and the monitoring of species frequencies over time.

The paucity of these Mamiellales genomes in samples from tropical and sub-tropical Pacific Ocean contrasts with their very broad distribution in other sampled basins. Only *B.* TOSAG39-1 and *O.* RCC809 were detected in tropical and sub-tropical basins and at lower abundance than in most other basins. The oligotrophic conditions of these Pacific areas might explain this pattern, especially the very low concentration in dissolved iron, previously described as the probable factor leading to peculiar phototrophic communities and/or local adaptations [48]. Ferredoxin is an important electron transfer enzyme for photosynthesis but particularly sensitive to iron limitation. This enzyme has been described as lost in the evolution of many lineages where its function is replaced by a stress-resistant isofunctional carrier, flavin-containing flavodoxin [49]. This hypothesis must be confirmed by the sequencing of local strains, but accordingly, we noticed the presence of flavodoxin and did not find ferredoxin in the reference genomes of *O.* RCC809 and *B.* TOSAG39-1. It would also be possible to perform tests in culture to estimate the minimal concentrations of iron and other oligo-nutrients among species and strains. Quantifying the relative abundance of a genome in a natural environment by leveraging metagenomes requires specificity to avoid over- or underestimates. Overestimates could be due to the presence in the same environment of related evolutionary species containing genomic regions conserved between them, leading to an over-recruitment of metagenomic reads. Thus, a preliminary detection of specific genomic regions is necessary. As described in Seeleuthner et al. [35], selecting regions with homogeneous coverage of recruited reads helps to exclude genomic regions that are conserved between genomes present in the same environment. Underestimates could be due to multiple situations such as less abundant species close to the detection threshold and/or a very large genome or a genome for which the reference is in the form of a low-quality assembly. Low quality reference genomes could also lead to an erroneous signal in the case of a chimeric assembly. For these reasons, genomic biogeography is well adapted to the genomes of bacteria and compact eukaryotes such as the Mamiellales. In the near future, this approach will be a key and routine methodology when advances in sequencing, in particular for metagenomes, will allow the detection or even reconstitution of complex and scarce genomes at very low cost.

The recent identification of the sequences of both highly divergent mating types in *Ostreococcus* [26] enabled an additional interpretation of the genome-resolved approach in *O. lucimarinus*. Comparison of the coverage of the mating-type sequences allowed us to infer not only the presence or absence, but also the frequency of the two mating types in the natural environment. Indirect estimation of the minimum frequency of sexual reproduction in a natural population of *O. tauri* has been estimated to be very low (one meiosis every 100,000 mitoses) [19]. This low prevalence of sexual reproduction might be the consequence of the low probability of contact between non-motile cells from the two different mating types in the environment. The indirect estimation of sexual reproduction in natural populations of *O.*

lucimarinus is yet unknown, but the systematic co-occurrence of strains from both mating types suggests that mating does happen. Under clonal reproduction, haploid *Ostreococcus* strains display up to 40% differences in growth rates [19] in different temperature conditions. As a consequence, sustained clonal reproduction for a significant number of generations is expected to distort mating type frequencies in any given environment. Over many different environments and taking population bottlenecks into account we may thus expect a discrete pattern of presence or absence of each mating-type. Information about mating-type frequency in the natural environment is scarce, and the sampling and sequencing of 12 *O. tauri* strains revealed that the 12 strains all belonged to the *MT*+ mating type [19], and more recently the sequencing of four *O.* spp. RCC809 single amplified genomes from the Indian Ocean also all encoded the *MT*+ type [43]. In our study, the mating type ratio seems to be slightly unbalanced with a bias towards a higher frequency of the *MT*+ mating type of *O. lucimarinus* in all 11 sites, while the ratio was close to 50/50 in one station. This suggests a tendency towards a higher prevalence of *MT*+ types in the natural environment. Heterogametic diploids have been observed in other green algae in which sexual reproduction is well studied [50]. Diploid *Ostreococcus* zygotes are therefore expected to be heterogametic, that is containing one *MT*+ and one *MT*− chromosome. However, frequent phases of clonal reproduction of haploid strains with different growth rates [19] are poised to distort mating type ratios. To our knowledge, no diploid Mamiellales has yet been observed and maintained in culture in the lab. This is not surprising since it is yet unknown how to induce meiosis in this group, and because the lab culture conditions have been optimized to sustain haploid microalgal growth. The identification of mating-types in additional species opens many novel research opportunities into the life cycle of these microalgae. Practically, knowledge about the frequency of the two mating types in the natural environment indicates how to scale up isolation protocols to increase the probability of finding both mating types. The hypothesis of a different fitness of the two mating types, suggestive of a cryptic anisogamy, could be tested experimentally for species where both *MT* have been maintained in culture.

In conclusion, the sequencing of additional metagenomes and reference genomes will improve this type of genome-based approach to a wide range of environments in many yet understudied lineages of ecologically relevant phytoplanktonic microalgae, and beyond. This is poised to reveal unprecedented insights into their ecology, i.e., interactions with other species, as well as insights into the genomic basis of their evolution and haploid–diploid life cycle alternation.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/11/1/66/s1, Table S1: Metagenomic samples identifying information, Table S2: Post-hoc Tukey test pairwise *p*-values, Table S3: Relative metagenomic abundances for the six Mamiellales genomes in all samples, Figure S1: Metagenomic read coverage along chromosome 2 of *O. lucimarinus*, Figure S2: Scatterplot of *O. lucimarinus* mating type ratios comparing the two different methods, Figure S3: Barplot comparisons of relative metagenomic abundances.

Author Contributions: P.W., G.P. and O.J. designed the study. J.L., L.F.B., G.P. and O.J. wrote the paper. J.L. and T.V. performed genomic biogeography analysis. J.L. performed environmental analyses. J.L. and L.F.B. performed mating type analysis. All authors have read and agreed to the published version of the manuscript.

Funding: L.F.B. was funded by the EU Horizon 2020 research and innovation program, under the Marie Skłodowska-Curie grant agreement no. H2020-MSCA-ITN-2015-675752. This work was also founded by the ANR project ALGALVIRUS ANR- 17-CE02- 0012.

Acknowledgments: We would like to thank all Genophy group and LAGE members for stimulating discussions on this project. We also thank Noan Le Bescot (TernogDesign) for artwork on Figure 1. *Tara* Oceans would not exist without the *Tara* Ocean Foundation and the continuous support of 23 institutes (https://oceans.taraexpeditions.org/). This article is contribution number 99 of Tara Oceans.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Not, F.; Latasa, M.; Marie, D.; Cariou, T.; Vaulot, D.; Simon, N. A single species, Micromonas pusilla (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. *Appl. Environ. Microbiol.* **2004**, *70*, 4064–4072. [CrossRef]

2. Vaulot, D.; Eikrem, W.; Viprey, M.; Moreau, H. The diversity of small eukaryotic phytoplankton (< or =3 microm) in marine ecosystems. *FEMS Microbiol. Rev.* **2008**, *32*, 795–820. [PubMed]
3. Marin, B.; Melkonian, M. Molecular Phylogeny and Classification of the Mamiellophyceae class. nov. (Chlorophyta) based on Sequence Comparisons of the Nuclear- and Plastid-encoded rRNA Operons. *Protist* **2010**, *161*, 304–336. [CrossRef]
4. Worden, A.; Nolan, J.; Palenik, B. Assessing the Dynamics and Ecology of Marine Picophytoplankton: The Importance of the Eukaryotic Component. *Limnol. Oceanogr.* **2004**, *49*, 168–179. [CrossRef]
5. Worden, A.Z.; Lee, J.-H.; Mock, T.; Rouzé, P.; Simmons, M.P.; Aerts, A.L.; Allen, A.E.; Cuvelier, M.L.; Derelle, E.; Everett, M.V.; et al. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science* **2009**, *324*, 268–272. [CrossRef] [PubMed]
6. Demir-Hilton, E.; Sudek, S.; Cuvelier, M.L.; Gentemann, C.L.; Zehr, J.P.; Worden, A.Z. Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *ISME J.* **2011**, *5*, 1095–1107. [CrossRef] [PubMed]
7. Vaulot, D.; Lepère, C.; Toulza, E.; la Iglesia, R.D.; Poulain, J.; Gaboyer, F.; Moreau, H.; Vandepoele, K.; Ulloa, O.; Gavory, F.; et al. Metagenomes of the Picoalga Bathycoccus from the Chile Coastal Upwelling. *PLoS ONE* **2012**, *7*, e39648. [CrossRef]
8. Countway, P.D.; Caron, D.A. Abundance and distribution of Ostreococcus sp. in the San Pedro Channel, California, as revealed by quantitative PCR. *Appl. Environ. Microbiol.* **2006**, *72*, 2496–2506. [CrossRef]
9. Tragin, M.; Vaulot, D. Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Sci. Rep.* **2019**, *9*, 5190. [CrossRef]
10. Moreau, H.; Verhelst, B.; Couloux, A.; Derelle, E.; Rombauts, S.; Grimsley, N.; Van Bel, M.; Poulain, J.; Katinka, M.; Hohmann-Marriott, M.F.; et al. Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol.* **2012**, *13*, R74. [CrossRef]
11. Vannier, T.; Leconte, J.; Seeleuthner, Y.; Mondy, S.; Pelletier, E.; Aury, J.-M.; de Vargas, C.; Sieracki, M.; Iudicone, D.; Vaulot, D.; et al. Survey of the green picoalga Bathycoccus genomes in the global ocean. *Sci. Rep.* **2016**, *6*, 37900. [CrossRef] [PubMed]
12. Lovejoy, C.; Vincent, W.; Bonilla, S.; Roy, S.; Martineau, M.-J.; Terrado, R.; Potvin, M.; Massana, R.; Pedrós-Alió, C. Distribution, phylogeny, and growth of cold−adapted picoprasinophytes in Arctic SeaS1. *J. Phycol.* **2007**, *43*, 78–89. [CrossRef]
13. Jancek, S.; Gourbière, S.; Moreau, H.; Piganeau, G. Clues about the genetic basis of adaptation emerge from comparing the proteomes of two Ostreococcus ecotypes (Chlorophyta, Prasinophyceae). *Mol. Biol. Evol.* **2008**, *25*, 2293–2300. [CrossRef] [PubMed]
14. Balzano, S.; Marie, D.; Gourvil, P.; Vaulot, D. Composition of the summer photosynthetic pico and nanoplankton communities in the Beaufort Sea assessed by T-RFLP and sequences of the 18S rRNA gene from flow cytometry sorted samples. *ISME J.* **2012**, *6*, 1480–1498. [CrossRef] [PubMed]
15. Simon, N.; Foulon, E.; Grulois, D.; Six, C.; Desdevises, Y.; Latimier, M.; Le Gall, F.; Tragin, M.; Houdan, A.; Derelle, E.; et al. Revision of the Genus Micromonas Manton et Parke (Chlorophyta, Mamiellophyceae), of the Type Species *M. pusilla* (Butcher) Manton & Parke and of the Species *M. commoda* van Baren, Bachy and Worden and Description of Two New Species Based on the Genetic and Phenotypic Characterization of Cultured Isolates. *Protist* **2017**, *168*, 612–635.
16. Butcher, R.W. Contributions to our knowledge of the smaller marine algae. *J. Mar. Biol. Assoc. U. K.* **1952**, *31*, 175–191. [CrossRef]
17. Vandepoele, K.; Van Bel, M.; Richard, G.; Van Landeghem, S.; Verhelst, B.; Moreau, H.; Van de Peer, Y.; Grimsley, N.; Piganeau, G. Pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environ. Microbiol.* **2013**, *15*, 2147–2153. [CrossRef]
18. Blanc-Mathieu, R.; Verhelst, B.; Derelle, E.; Rombauts, S.; Bouget, F.-Y.; Carré, I.; Château, A.; Eyre-Walker, A.; Grimsley, N.; Moreau, H.; et al. An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. *BMC Genom.* **2014**, *15*, 1103. [CrossRef]
19. Blanc-Mathieu, R.; Krasovec, M.; Hebrard, M.; Yau, S.; Desgranges, E.; Martin, J.; Schackwitz, W.; Kuo, A.; Salin, G.; Donnadieu, C.; et al. Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Sci. Adv.* **2017**, *3*, e1700239. [CrossRef]

20. Palenik, B.; Grimwood, J.; Aerts, A.; Rouzé, P.; Salamov, A.; Putnam, N.; Dupont, C.; Jorgensen, R.; Derelle, E.; Rombauts, S.; et al. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7705–7710. [CrossRef]
21. Grimsley, N.; Yau, S.; Piganeau, G.; Moreau, H. Typical Features of Genomes in the Mamiellophyceae. In *Marine Protists: Diversity and Dynamics*; Ohtsuka, S., Suzaki, T., Horiguchi, T., Suzuki, N., Not, F., Eds.; Springer: Tokyo, Japan, 2015; pp. 107–127. ISBN 978-4-431-55130-0.
22. Subirana, L.; Péquin, B.; Michely, S.; Escande, M.-L.; Meilland, J.; Derelle, E.; Marin, B.; Piganeau, G.; Desdevises, Y.; Moreau, H.; et al. Morphology, genome plasticity, and phylogeny in the genus ostreococcus reveal a cryptic species, *O. mediterraneus* sp. nov. (Mamiellales, Mamiellophyceae). *Protist* **2013**, *164*, 643–659. [CrossRef]
23. Yau, S.; Hemon, C.; Derelle, E.; Moreau, H.; Piganeau, G.; Grimsley, N. A Viral Immunity Chromosome in the Marine Picoeukaryote, *Ostreococcus tauri*. *PLoS Pathog.* **2016**, *12*, e1005965. [CrossRef] [PubMed]
24. Verhelst, B.; Van de Peer, Y.; Rouzé, P. The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. *Genome Biol. Evol.* **2013**, *5*, 2393–2401. [CrossRef] [PubMed]
25. Lee, S.C.; Ni, M.; Li, W.; Shertz, C.; Heitman, J. The Evolution of Sex: A Perspective from the Fungal Kingdom. *Microbiol. Mol. Biol. Rev.* **2010**, *74*, 298–340. [CrossRef] [PubMed]
26. Benites, L.F.; Bucchini, F.; Sanchez-Brosseau, F.; Grimsley, N.; Piganeau, G. Evolutionary dynamics of sex-related chromosomes at the base of the green lineage. *Mol. Biol. Evol.* **2019**, in press.
27. Cox, C.; Moore, P.; Ladle, R. *Biogeography: An Ecological and Evolutionary Approach*, 9th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2016; ISBN 978-1-118-96858-1.
28. Brum, J.R.; Ignacio-Espinoza, J.C.; Roux, S.; Doulcier, G.; Acinas, S.G.; Alberti, A.; Chaffron, S.; Cruaud, C.; de Vargas, C.; Gasol, J.M.; et al. Patterns and ecological drivers of ocean viral communities. *Science* **2015**, *348*, 1261498. [CrossRef]
29. Roux, S.; Brum, J.R.; Dutilh, B.E.; Sunagawa, S.; Duhaime, M.B.; Loy, A.; Poulos, B.T.; Solonenko, N.; Lara, E.; Poulain, J.; et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **2016**, *537*, 689–693. [CrossRef]
30. Martiny, J.; Bohannan, B.; Brown, J.; Colwell, R.; Fuhrman, J.; Green, J.; Horner-Devine, C.; Kane, M.; Krumins, J.; Kuske, C.; et al. Microbial biogeography: Putting microorganisms on the map. *Nat. Rev. Microbiol.* **2006**, *4*, 102–112. [CrossRef]
31. Hanson, C.A.; Fuhrman, J.A.; Horner-Devine, M.C.; Martiny, J.B.H. Beyond biogeographic patterns: Processes shaping the microbial landscape. *Nat. Rev. Microbiol.* **2012**, *10*, 497–506. [CrossRef]
32. Madoui, M.-A.; Poulain, J.; Sugier, K.; Wessner, M.; Noel, B.; Berline, L.; Labadie, K.; Cornils, A.; Blanco-Bercial, L.; Stemmann, L.; et al. New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. *Mol. Ecol.* **2017**, *26*, 4467–4482. [CrossRef]
33. Guillou, L.; Bachar, D.; Audic, S.; Bass, D.; Berney, C.; Bittner, L.; Boutte, C.; Burgaud, G.; de Vargas, C.; Decelle, J.; et al. The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **2013**, *41*, D597–D604. [CrossRef] [PubMed]
34. Piganeau, G.; Eyre-Walker, A.; Grimsley, N.; Moreau, H. How and Why DNA Barcodes Underestimate the Diversity of Microbial Eukaryotes. *PLoS ONE* **2011**, *6*, e16342. [CrossRef] [PubMed]
35. Seeleuthner, Y.; Mondy, S.; Lombard, V.; Carradec, Q.; Pelletier, E.; Wessner, M.; Leconte, J.; Mangot, J.-F.; Poulain, J.; Labadie, K.; et al. Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nat. Commun.* **2018**, *9*, 310. [CrossRef] [PubMed]
36. Mangot, J.-F.; Logares, R.; Sánchez, P.; Latorre, F.; Seeleuthner, Y.; Mondy, S.; Sieracki, M.E.; Jaillon, O.; Wincker, P.; de Vargas, C.; et al. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci. Rep.* **2017**, *7*, 1–12. [CrossRef] [PubMed]
37. Delmont, T.O.; Kiefl, E.; Kilinc, O.; Esen, O.C.; Uysal, I.; Rappé, M.S.; Giovannoni, S.; Eren, A.M. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife* **2019**, *8*, e46497. [CrossRef]

38. Karsenti, E.; Acinas, S.G.; Bork, P.; Bowler, C.; Vargas, C.D.; Raes, J.; Sullivan, M.; Arendt, D.; Benzoni, F.; Claverie, J.-M.; et al. A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biol.* **2011**, *9*, e1001177. [CrossRef]
39. Langmead, B.; Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
40. Morgulis, A.; Gertz, E.; Schaffer, A.; Agarwala, R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *J. Comput. Biol. A J. Comput. Mol. Cell Biol.* **2006**, *13*, 1028–1040. [CrossRef]
41. Pesant, S.; Not, F.; Picheral, M.; Kandels-Lewis, S.; Bescot, N.L.; Gorsky, G.; Iudicone, D.; Karsenti, E.; Speich, S.; Troublé, R.; et al. Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2015**, *2*, 1–16. [CrossRef]
42. Clayton, S.; Dutkiewicz, S.; Jahn, O.; Hill, C.; Heimbach, P.; Follows, M.J. Biogeochemical versus ecological consequences of modeled ocean physics. *Biogeosciences* **2017**, *14*, 2877–2889. [CrossRef]
43. Benites, L.F.; Poulton, N.; Labadie, K.; Sieracki, M.E.; Grimsley, N.; Piganeau, G. Single cell ecogenomics reveals mating types of individual cells and ssDNA viral infections in the smallest photosynthetic eukaryotes. *Philos. Trans. R. Soc. B Biol. Sci.* **2019**, *374*, 20190089. [CrossRef] [PubMed]
44. Simmons, M.P.; Sudek, S.; Monier, A.; Limardo, A.J.; Jimenez, V.; Perle, C.R.; Elrod, V.A.; Pennington, J.T.; Worden, A.Z. Abundance and Biogeography of Picoprasinophyte Ecotypes and Other Phytoplankton in the Eastern North Pacific Ocean. *Appl. Environ. Microbiol.* **2016**, *82*, 1693–1705. [CrossRef] [PubMed]
45. Monier, A.; Worden, A.Z.; Richards, T.A. Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ. Microbiol. Rep.* **2016**, *8*, 461–469. [CrossRef] [PubMed]
46. De Vargas, C.; Audic, S.; Henry, N.; Decelle, J.; Mahé, F.; Logares, R.; Lara, E.; Berney, C.; Bescot, N.L.; Probert, I.; et al. Eukaryotic plankton diversity in the sunlit ocean. *Science* **2015**, *348*, 1261605. [CrossRef]
47. Ibarbalz, F.M.; Henry, N.; Brandão, M.C.; Martini, S.; Busseni, G.; Byrne, H.; Coelho, L.P.; Endo, H.; Gasol, J.M.; Gregory, A.C.; et al. Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell* **2019**, *179*, 1084–1097.e21. [CrossRef]
48. Caputi, L.; Carradec, Q.; Eveillard, D.; Kirilovsky, A.; Pelletier, E.; Karlusich, J.J.P.; Vieira, F.R.J.; Villar, E.; Chaffron, S.; Malviya, S.; et al. Community-Level Responses to Iron Availability in Open Ocean Plankton Ecosystems. *Glob. Biogeochem. Cycles* **2019**, *33*, 391–419. [CrossRef]
49. Pierella Karlusich, J.J.; Ceccoli, R.D.; Graña, M.; Romero, H.; Carrillo, N. Environmental Selection Pressures Related to Iron Utilization Are Involved in the Loss of the Flavodoxin Gene from the Plant Genome. *Genome Biol. Evol.* **2015**, *7*, 750–767. [CrossRef]
50. Umen, J.; Coelho, S. Algal Sex Determination and the Evolution of Anisogamy. *Ann. Rev. Microbiol.* **2019**, *73*, 267–291. [CrossRef]

III. Etude du mating type

Suite à la publication du second manuscrit, nous avons souhaité étendre l'analyse des chromosomes outliers aux autres espèces de Mamiellales afin de voir si la différence de couverture entre les régions outlier et non-outlier du BOC pouvait également servir de proxy aux ratios des mating-types. Les séquences de gènes correspondant à MT+ et MT- n'ayant pas été décrites à l'heure actuelle, nous nous sommes donc uniquement basés sur les couvertures moyennes à l'échelle de l'ensemble du chromosome en utilisant une méthodologie identique à celle utilisée pour *O. lucimarinus*.

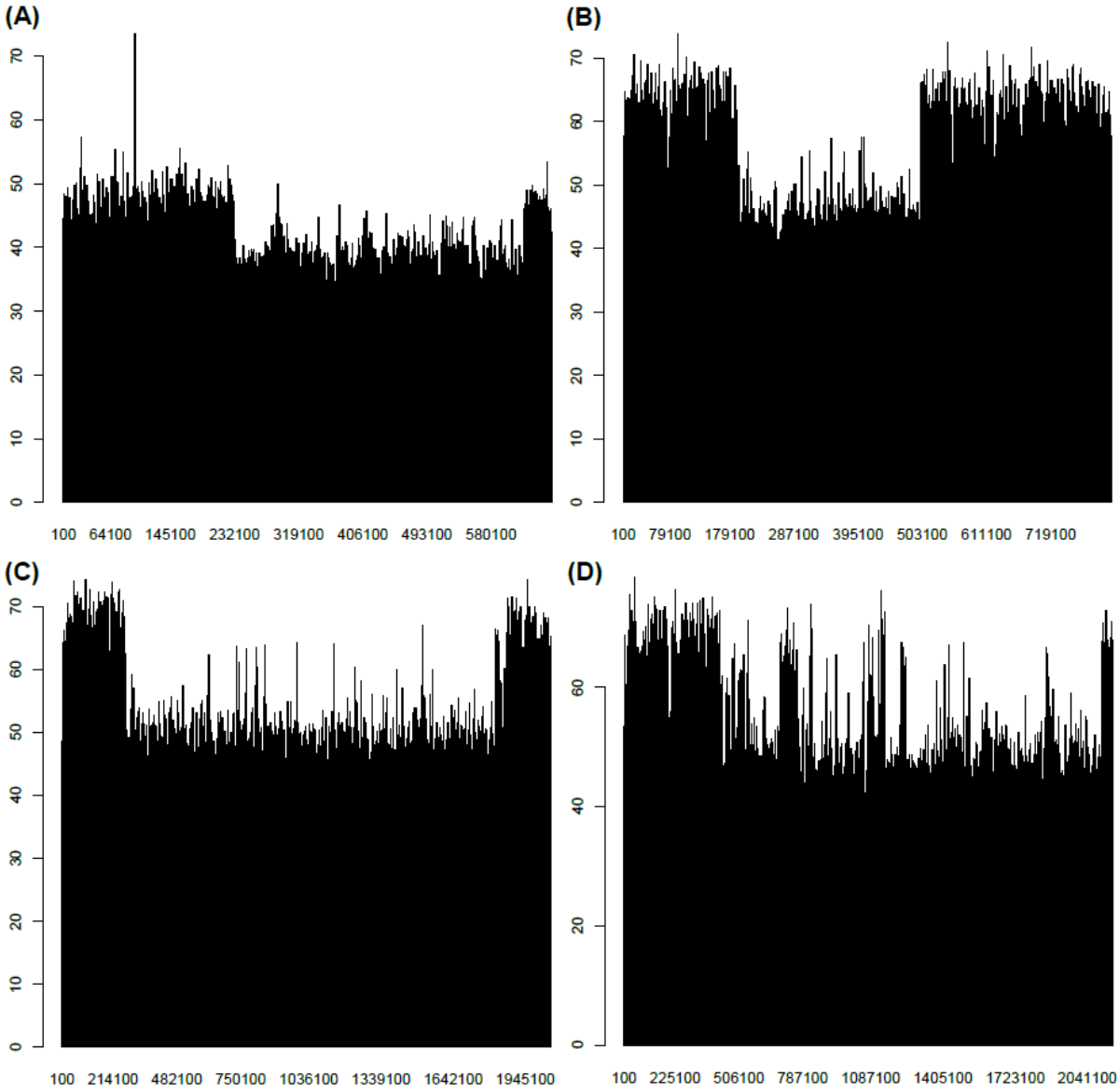


Figure 11 : Contenu en GC le long du (A) chromosome 14 de *B. prasinos*, (B) chromosome 2 de *O. RCC809*, (C) chromosome 1 de *M. commoda* et (D) chromosome 2 de *M. pusilla*.

Afin de connaître la position exacte de la zone outlier à partir de nos génomes d'intérêt, nous avons étudié les différences de contenu en GC, cette zone étant bien connue pour avoir un contenu nettement différent du reste du génome (Figure 11).

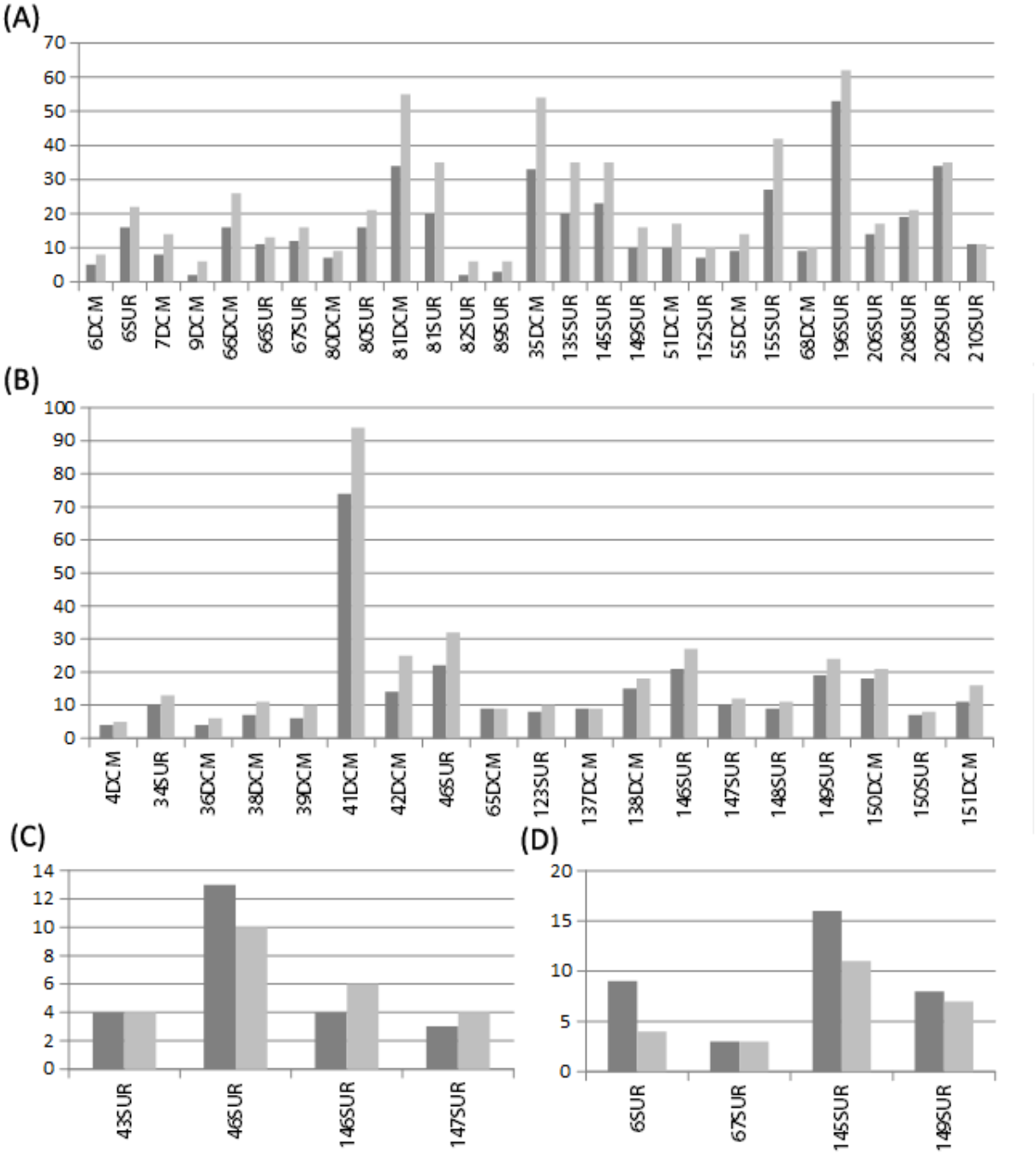


Figure 12 : Couverture médiane des régions à faible GC (gris clair) et haut GC (gris foncé) les échantillons métagénomiques *Tara* Oceans du (A) chromosome 14 de *B. prasinos*, (B) chromosome 2 de *O. RCC809*, (C) chromosome 1 de *M. commoda* et (D) chromosome 2 de *M. pusilla*. Seuls les échantillons où l'organisme présente une couverture de la région à haut contenu en GC d'au moins 4X ont été conservés.

A partir de ces informations, nous avons observé les variations de couvertures entre les stations où les différents organismes sont abondants, étudiant donc les profils de *B. prasinos*, *M. commoda*, *M. pusilla* et *O. RCC809*. (Figure 12).

Deux modèles se sont ainsi distingués parmi les Mamiellales. Le premier correspond à *O. lucimarinus*, étudié dans le manuscrit précédent, pour lequel la région outlier est moins couverte que le reste du génome. C'est ainsi que semble également se comporter *M. pusilla*. Par exemple dans la station 6 méditerranéenne, la zone outlier est deux fois moins couverte que le reste du chromosome (Figure 12D). En revanche, pour *M. commoda*, si ce pattern apparait dans la station 46, les stations 146 et 147 du Gulf Stream présentent une inversion avec la zone à faible contenu en GC plus couverte que le reste du chromosome (Figure 12C). Cela pourrait être un biais des échantillons à faible couverture, mais de la même manière pour *Bathycoccus* et *Ostreococcus* la région outlier est soit autant couverte, soit plus couverte que le reste du chromosome dans tous les échantillons (Figure 12A-B).

Il est difficile d'expliquer ce second pattern par la différence d'abondance des gènes du mating-type, qui n'ont selon notre hypothèse pas de raison de dépasser la couverture de la région ayant un taux de GC qui correspond au reste du génome. Les situations de plus grande couverture peuvent s'expliquer aussi par la présence d'un génome apparenté qui a une forte similarité dans cette région mais pas dans les autres (cross-mapping), ou par une duplication de cette région chez certains individus. Ces résultats appellent donc à des analyses complémentaires afin de mieux comprendre les éléments différenciant *Bathycoccus prasinos* et *Ostreoccocus* RCC809 des autres Mamiellales au niveau de leurs BOCs et de la région des gènes liés à la sexualité.

Au niveau des ratios entre les deux mating-types pour les espèces ou l'on peut observer une différence, ils semblent se comporter différemment selon les échantillons. Par exemple, *O. lucimarinus* présente une écrasante majorité de *MT+* dans tous les échantillons sauf dans l'océan austral et l'upwelling du Chili, qui comportent un mélange avec le *MT-*. Pour *M. pusilla*, c'est en Méditerranée qu'on observe un mélange, tandis que les échantillons provenant de l'upwelling du Benguela ou du Gulf Stream semblent moins divers. *M. commoda* est abondant dans un certain nombre d'échantillons, mais étonnamment son BOC a une couverture faible avec de nom-

breuses chutes dans la majorité d'entre eux, il est donc complexe de déterminer si la différence de couverture est significative même pour les quatre échantillons les plus couverts présentés ici. Pour *B. prasinos*, on peut noter que cette différence semble très faible dans l'arctique en particulier alors qu'elle est plus marquée ailleurs dans les eaux tempérées. Chaque espèce affiche donc une répartition qui lui est propre.

IV. Conclusion

Les deux espèces de *Bathycoccus* sont, en plus d'être distinctes d'un point de vue génomique, clairement séparées par les niches écologiques dans lesquelles elles se trouvent, typiquement des environnements présentant différentes températures. Bien qu'elles puissent parfois se trouver dans un même échantillon, l'une des deux semble toujours être plus abondante que l'autre et ce schéma n'est visiblement pas réservé à *Bathycoccus*, les mêmes facteurs semblant influer la distribution de nombreuses espèces de Mamiellales. En revanche, il n'est pas exclu pour *Bathycoccus*, *Ostreococcus* et *Micromonas* de se trouver dans le même échantillon. Au contraire, on peut même observer un pattern de co-abondance entre ces organismes, car parmi ceux étudiés, un groupe comprenant un membre de chaque correspond aux milieux froids tandis qu'un autre correspond aux milieux chauds, avec une coprésence dans de nombreux échantillons.

D'après ces résultats, la répartition des Mamiellales subit donc un fort impact des paramètres physico-chimiques. Les proportions entre les deux mating-types, basées sur l'analyse de la couverture du grand chromosome outlier, ne sont pas identiques entre les différentes espèces et ne semblent pas pour la plupart suivre une répartition géographique particulière.

Chapitre 2 : Diversité et biogéographie des Mamiellales arctiques à partir de métabarcodes

I. Introduction

L'étude des Mamiellales en milieu tempéré nous a permis de déterminer leur biogéographie, incluant les préférences environnementales. Mais certaines espèces comme celles appartenant au genre *Bathycoccus* ou *Micromonas* sont bien connues pour être abondantes dans les eaux arctiques, c'est pourquoi la mise à disposition des données provenant de la seconde partie de l'expédition *Tara* Oceans avec de nombreux échantillons du cercle polaire a été l'occasion de poursuivre et compléter nos analyses.

Afin d'étudier plus en détails l'océan Arctique, un environnement aux conditions particulières, une collaboration entre plusieurs laboratoires en particulier du consortium Tara Oceans visant à étudier la biodiversité du plancton dans ce milieu à partir des données de métabarcodes est en cours de mise en place. Les Mamiellophyceae, particulièrement abondants, y ont leur place en tant que représentants majeurs du phytoplancton.

Ce chapitre présentera donc en premier lieu les analyses que nous avons pu mener dans ce cadre en nous basant sur les données V9 pour obtenir une première approche de leur diversité et de leur répartition. Dans un second temps, afin d'affiner cette étude, nous avons également étendu les analyses se basant sur des génomes de référence de Mamiellales aux échantillons métagénomiques arctiques, afin d'étudier les potentielles différences avec le signal des barcodes et d'avoir une vue plus précise de la taxonomie de ces organismes.

II. Distribution des V9 de Mamiellophyceae

Nous avons dans un premier temps sélectionné les fichiers de métabarcodes obtenus à partir du séquençage des régions V9 pour les fractions 0.8µm-5µm provenant des échantillons tempérés, et 0.8µm-2000µm pour les échantillons arctiques ne disposant pas de la fraction précédente. Malheureusement, l'utilisation de fractions différentes complexifie la comparaison entre les deux environnements, rendant difficilement comparable l'abondance d'une espèce. En analysant l'abondance V9 de *Ba-*

thycoccus dans les rares stations présentant les deux fractions, nous avons cependant pu observer que l'abondance relative de l'espèce était deux fois plus élevée dans les échantillons 0.8µm-5µm par rapport aux 0.8µm-2000µm. Les autres espèces de Mamiellales n'étaient pas suffisamment abondantes dans ces stations pour pouvoir effectuer une comparaison similaire. Cela nous permet tout de même d'avoir une approximation du biais causé ici par la variation de fractions de tailles.

Les métabarcodes sont préalablement regroupés sous la forme d'unités taxonomiques opérationnelles (OTU) similairement au précédant jeu de données incomplet étudié et décrit en détails par Vargas et al[2]. Etant assignés taxonomiquement, nous avons pu extraire les OTUs correspondant aux Mamiellophyceae à partir de 194 échantillons (certains prélevés en surface et d'autres à la profondeur DCM), 164 provenant de la première partie de l'expédition et 30 de la partie arctique. Bien que l'objectif de ces analyses soit d'étudier et de comparer la diversité des échantillons polaires, il reste important d'observer la répartition des OTUs attribués à nos espèces d'intérêt dans un environnement plus connu.

1. Abondance relative par bassin océanique

Afin d'avoir une première idée de la proportion des Mamiellophyceae dans l'environnement, nous avons étudié leur abondance relative cumulée dans l'ensemble des échantillons, avant de la comparer à l'abondance relative cumulée de l'ensemble des organismes photosynthétiques dans ces mêmes échantillons (Figure 13).

Nous pouvons d'abord noter qu'en observant l'ensemble des organismes photosynthétiques, les océans Arctique et Austral, nos deux milieux polaires, sont ceux présentant les plus hautes abondances relatives avec des médianes à 37 et 38% respectivement dans les eaux de profondeur. Les autres milieux semblent se situer à des valeurs relativement équivalentes, la majorité étant entre 5 et 15% d'abondance. Ces proportions n'impactent pas les Mamiellophyceae qui n'ont pas une distribution divergente en milieu polaire. En revanche, ils sont très peu présents avec une médiale presque nulle dans trois régions distinctes : la mer Méditerranée (en surface et une partie de la DCM), le Pacifique Nord (en surface) et le Pacifique Sud (en surface et en DCM). Ces valeurs dans le Pacifique, notamment sud, ne sont pas surprenantes en raison des faibles taux de fer dans ces eaux. En Méditerranée, les organismes semblent donc se situer un peu en dessous de la surface, la DCM n'étant généralement

pas très profonde dans cette région (autour de 50m en moyenne). Bien qu'il n'y ait pas de différence majeure de distribution entre les océans, c'est dans l'Atlantique nord que les Mamiellophyceae ont en moyenne la plus grande abondance cumulée.

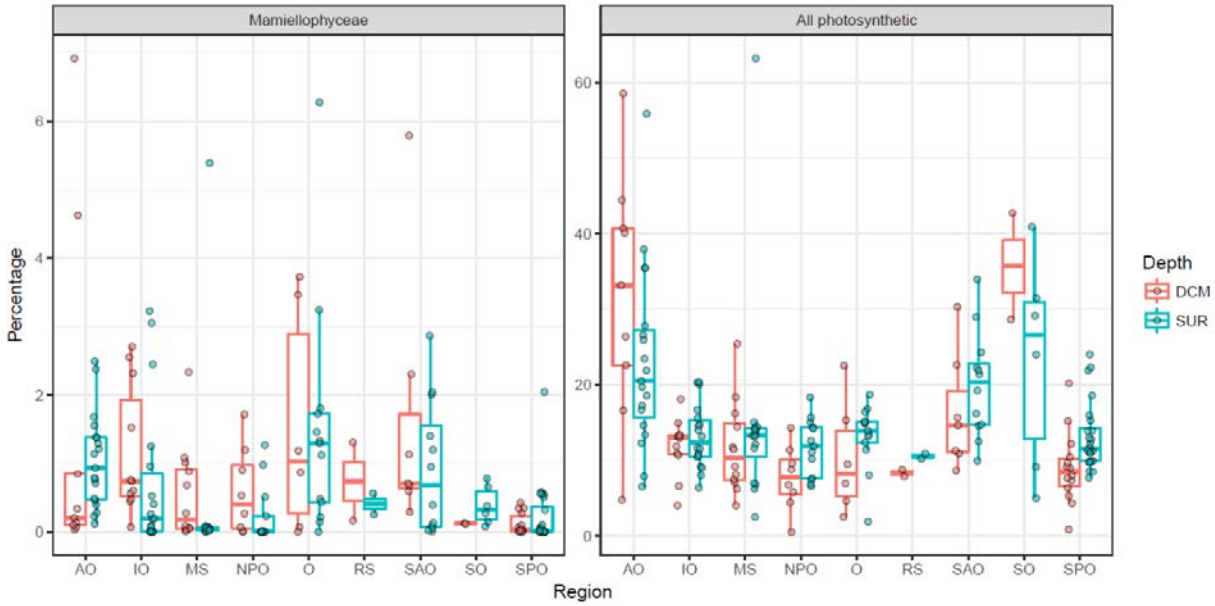


Figure 13 : Boxplot de l'abondance relative cumulée des Mamiellophyceae (à gauche) et de l'ensemble des organismes photosynthétiques (à droite) dans les échantillons *Tara* Oceans groupés par régions océaniques. Chaque point représente un échantillon, colorés selon s'ils ont été prélevés en surface (en bleu) ou en profondeur (en rouge).
(AO) Océan Arctique, (IO) Océan Indien, (MS) Mer Méditerranée, (NPO) Océan Pacifique Nord, (O) Océan Atlantique Nord, (RS) Mer Rouge, (SAO) Océan Atlantique Sud, (SO) Océan Austral, (SPO) Océan Pacifique Sud.

2. Répartition géographique globale

Dans un premier temps avant de nous concentrer sur l'Arctique, nous nous sommes intéressés à l'aspect global de la répartition de ces organismes. Pour cela, nous avons simplement cumulé l'abondance des OTUs à l'échelle du genre dans chaque échantillon, obtenant ainsi la biogéographie des Mamiellales *Bathycoccus*, *Micromonas* et *Ostreococcus* mais aussi *Mantoniella* et *Mamiella*. Nous avons également trouvé des organismes de l'ordre des Dolichomastigales, *Crustomastix* et *Dolichomastix*, mais pas de l'ordre des Monomastigales (Figure 14).

Comme attendu, les trois principaux Mamiellales se trouvent dans un grand nombre d'échantillons de manière abondante, dominant totalement les eaux tempérées, les différentes espèces qui composent chaque genre étant ici additionnées. En

revanche, les autres Mamiellales Mantoniella et Mamiella pour lesquels nous n'avons pas de génomes de référence ont surtout été retrouvés en Méditerranée à de faibles abondances. Concernant les Dolichomastigales, on retrouve surtout Crustomastix dans les eaux Pacifique sud et Atlantique sud aux plus basses latitudes, parfois associé à Dolichomastix en plus faible abondance.

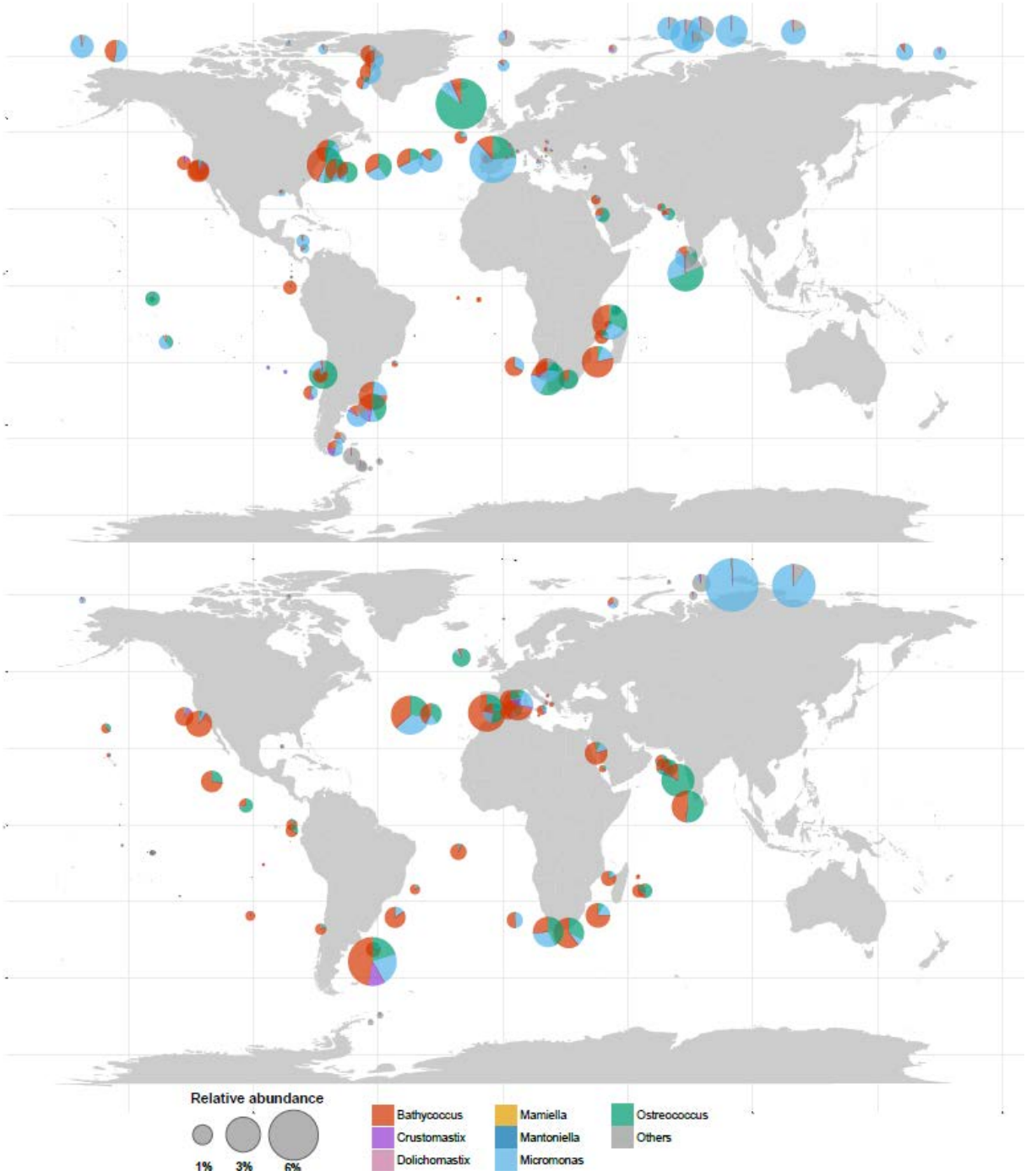


Figure 14 : Carte de répartition des Mamiellophyceae basée sur les données métabarcodes de *Tara* Oceans. Chaque point représente un échantillon, dont la taille correspond à l'abondance relative cumulée des Mamiellophyceae, découpé selon la proportion de chaque genre.

On les retrouve également dans les océans Austral et Arctique, accompagnés d'un autre genre de cet ordre n'ayant pu être identifié avec précision mais semblant appartenir à la famille Crustomastigaceae (la même que Crustomastix). Le groupe "autre" indiqué en gris sur la Figure 14 est en grande majorité constitué d'un ou plusieurs genres appartenant à cette famille.

Ces résultats dans les eaux tempérées semblent cohérents avec nos analyses basées sur les échantillons métagénomiques. Concernant l'Arctique, notre région d'intérêt, c'est *Micromonas* qui domine ces eaux, suivi par *Bathycoccus* à l'ouest, puis la famille Crustomastigaceae à l'est avec le genre non déterminé et dans quelques stations des traces de Crustomastix.

3. Analyse détaillée des taxons de *Micromonas*

Parmi notre groupe de Mamiellophyceae abondants, seul *Micromonas* possède une région V9 suffisamment variable pour obtenir des informations et une séparation à l'échelle de l'espèce. Nous nous sommes donc penchés sur la répartition des OTUs indiqués comme correspondant à des espèces distinctes de ce genre.

Nous avons ainsi pu distinguer quatre assignations différentes parmi nos OTUs: *M. commoda* et *M. pusilla*, dont nous avons déjà étudié la répartition dans les eaux tempérées sur la base des échantillons métagénomiques, *M. polaris*, espèce spécifique des eaux froides, et l'une des espèces candidates, *M.* candidate species 1. M. bravo, la quatrième espèce connue qui appartenait auparavant au même clade que *M. polaris* séparé en arctique et non-arctique[46], possède bien des OTUs lui correspondant mais seulement dans deux échantillons. De plus, son abondance s'avère quasiment nulle dans ces derniers et il n'a donc pas été pris en compte ici (Figure 15).

On peut observer que *M. pusilla* est présent dans très peu d'échantillons. C'est le cas également pour l'espèce candidate, les deux étant majoritairement présentes autour du Gulf Stream, et dans l'upwelling du Benguela pour *M. pusilla* ce qui correspond exactement à la répartition de son génome complet de référence étudié précédemment.

Concernant *M. commoda*, on retrouve l'espèce dans l'océan Indien et le Gulf Stream principalement, similairement aux analyses précédentes, mais aussi dans

l'Austral où il était détecté en très faible abondance avec l'approche par génome entier. Enfin il est détecté dans quelques échantillons arctiques en faible abondance.

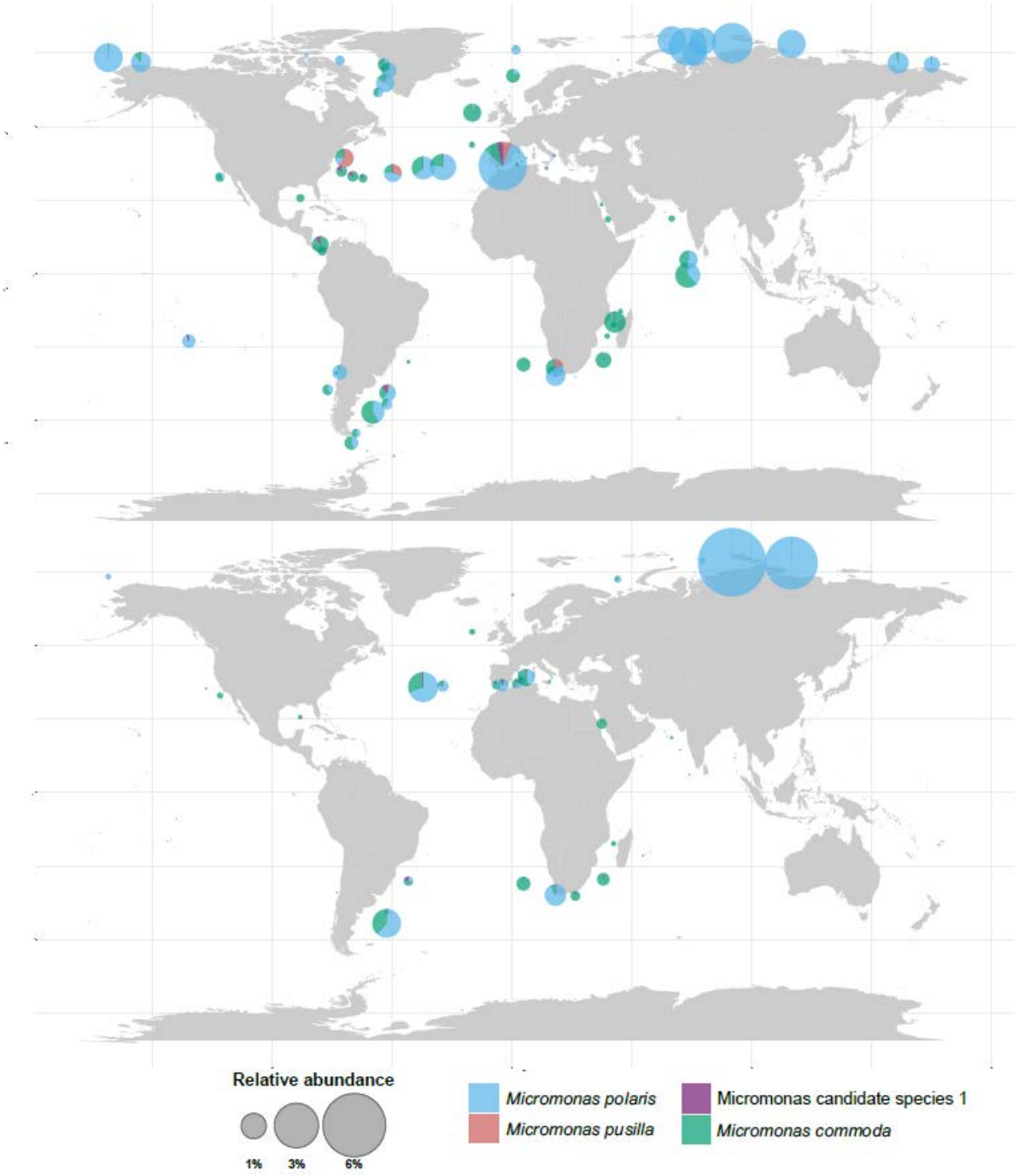


Figure 15 : Carte de répartition des espèces de Micromonas basée sur les données métabarcodes de *Tara* Oceans. Chaque point représente un échantillon, dont la taille correspond à l'abondance relative cumulée des quatre Micromonas, découpé selon la proportion de chacun.

M. polaris était attendu uniquement en Arctique, où il domine en effet les Mamiellales avec une forte abondance dans sa fraction de taille, mais il est étonnamment fortement détecté dans de très nombreux échantillons tempérés. Ce résultat n'était pas attendu, d'autres résultats tels que présentés par Simon et al[46] détectant cette espèce essentiellement en Arctique. Il est possible qu'il s'agisse ici d'une erreur de classification ou que la région V9 de *M. polaris* soit partagée avec une autre espèce de Micromonas.

4. Biodiversité du milieu Arctique

En observant les Mamiellales, groupe majeur des Mamiellophyceae, deux genres dominent l'océan Arctique : *Micromonas* et *Bathycoccus*. *Ostreoccocus*, dans les stations de l'expédition autour du cercle polaire, n'est trouvé que parmi celles faisant partie de l'Atlantique nord au début et à la fin du trajet. Au niveau du genre *Micromonas*, deux espèces distinctes sont détectées, *M. polaris* et *M. commoda*. D'autres OTUs sont également trouvés en très faible abondance dans ces eaux froides, associés au taxon "Mamiellales_X" sans famille ou genre particulier. Nous allons ici étudier plus en détails les variations d'abondances de l'ensemble de ces Mamiellales entre les échantillons (Figure 16).

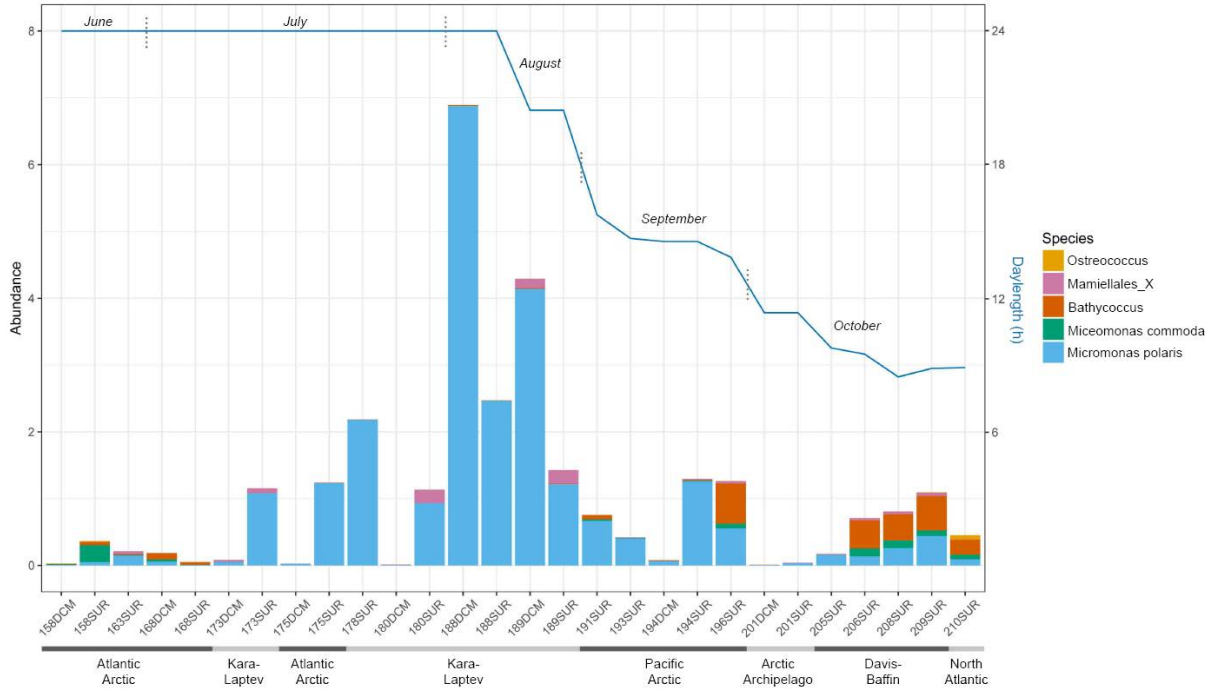


Figure 16 : Barplot de l'abondance relative des Mamiellales dans l'océan Arctique (échelle sur l'axe de gauche). Les bassins océaniques auxquels appartiennent les échantillons sont listés sous le graphique. La durée du jour, un paramètre important pour les organismes photosyn-

thétiques, est précisée pour l'ensemble des échantillons sous la forme d'une courbe bleue (échelle sur l'axe de droite). Les mois de prélèvement allant de Juin à Octobre sont également précisés.

Nous pouvons noter ici une forte dominance de *Micromonas polaris* des échantillons 173 à 194, avec la présence occasionnelle de Mamiellales_X. Au contraire, les échantillons 158 à 168 qui ont une faible proportion de Mamiellales présentent plutôt *Bathycoccus* ainsi que le second *Micromonas*. Finalement, à partir de l'échantillon 196, c'est *Bathycoccus* qui domine la majorité des stations de prélèvement, étant égal ou supérieur à *M. polaris*.

Il est difficile ici de distinguer un effet saisonnier d'un effet géographique, mais ces résultats semblent cohérents avec l'étude menée par Joli et al[56] dans la mer du Beaufort (Région "Pacific Arctic"). Celle-ci montre, en se basant sur des échantillons pris de Novembre à Juillet dans un même bassin, une dominance en été de *Micromonas* ensuite remplacé en hiver par *Bathycoccus*. Plusieurs hypothèses ont été présentées pour expliquer ce phénomène, telles qu'un potentiel avantage de *Bathycoccus* dans des conditions de faible luminosité, ou encore une plus haute proportion de virus ciblant *Micromonas* lorsque ce dernier est dominant, entrainant sa diminution et laissant *Bathycoccus* provisoirement plus abondant. Mais nous ne pouvons malheureusement pas conclure ici que l'absence de *Bathycoccus* dans les mers de Kara et de Laptev soit directement liée à un prélèvement estival, n'ayant pas d'échantillons au même endroit à une autre période pour le confirmer.

Enfin *Micromonas commoda* semble être présent, avec une abondance plus faible, aux mêmes sites que *Bathycoccus*, et *Ostreoccus* est comme établi précédemment uniquement trouvé dans l'échantillon Nord-Atlantique.

Pour conclure cette étude à partir des métabarcodes de Mamiellales, nous avons voulu étudier les conditions environnementales dans lesquelles les différentes espèces sont trouvées. Nous avons pour cela réalisé un NMDS basé sur l'abondance relative des Mamiellales similairement à Demory et al[132] prenant dans un premier temps l'ensemble des échantillons tempérés et polaires (Figure 17).

Ainsi les échantillons proches d'un nom d'espèce indiquent une forte abondance de celle-ci dans sa communauté. Nous avons ensuite ajusté des variables environnementales (Température, salinité, silicate, phosphate et NO2NO3) dans l'espace d'ordination avec une pvalue basée sur 999 permutations pour évaluer la significativité de cet ajustement. Seuls les paramètres ayant une pvalue inférieure à 0.05, et donc

représentant correctement le placement de ces derniers par rapport aux échantillons préalablement ordonnés par les valeurs d'abondance sont ici montrés. Nous pouvons noter que sur cette première figure, la majorité des paramètres sont significatifs, avec des échantillons de basse température et salinité et de hauts taux de nutriments en haut du NMDS, et des échantillons aux conditions opposées en bas. Cet effet est principalement causé par la séparation des océans Arctique et Austral du reste des régions océaniques.

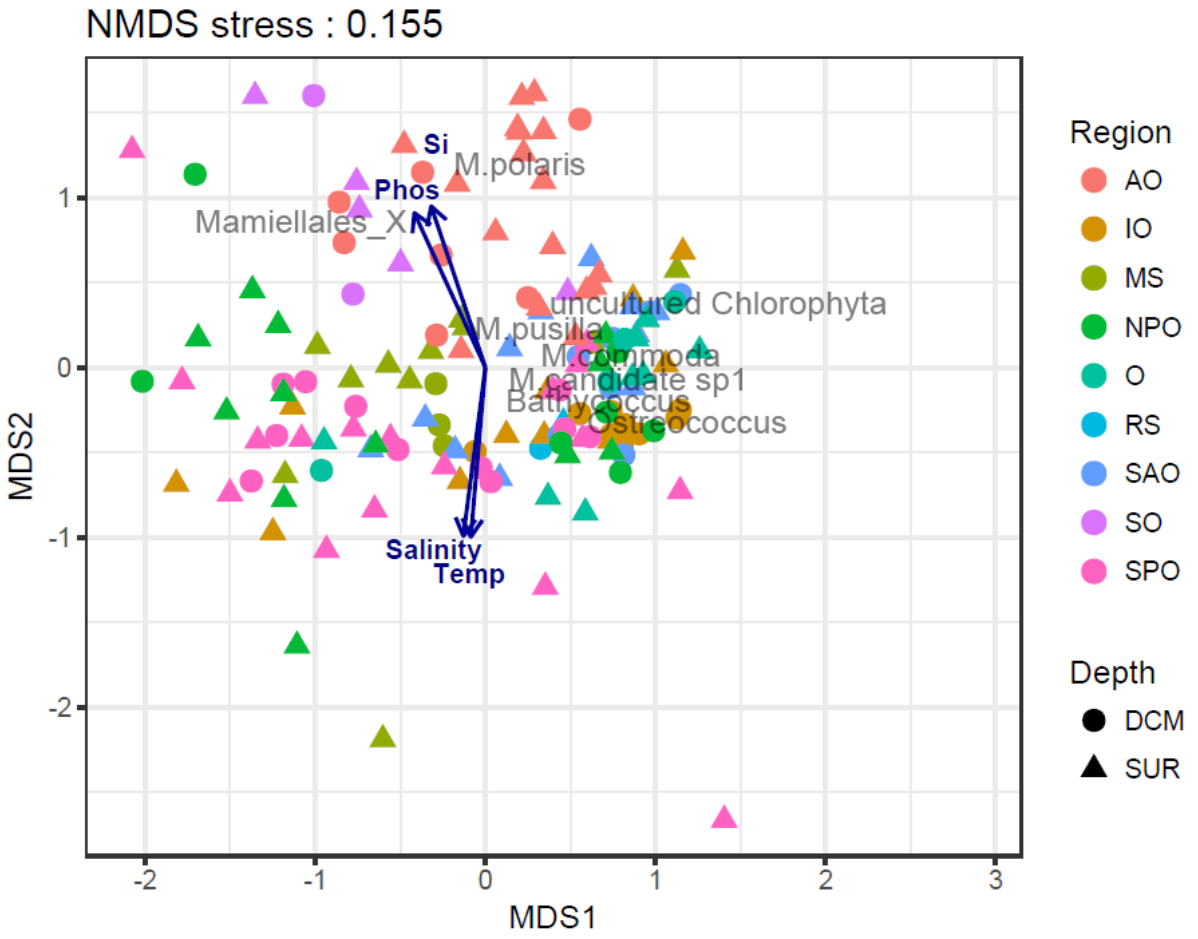


Figure 17 : NMDS basé sur l'abondance en métabarcodes des lignées de Mamiellales. Les noms d'espèces sont indiqués en gris, et chaque point représente un échantillon dont la forme dépend de la profondeur et la couleur de la région océanique. Plus un point est proche d'un nom d'espèce, plus l'abondance de l'espèce est forte dans cet échantillon. Les flèches représentent les paramètres environnementaux distinguant significativement les échantillons. Plus un échantillon est avancé sur l'axe de la flèche, plus la valeur du paramètre y est haute.
(AO) Océan Arctique, (IO) Océan Indien, (MS) Mer Méditerranée, (NPO) Océan Pacifique Nord, (O) Océan Atlantique Nord, (RS) Mer Rouge, (SAO) Océan Atlantique Sud, (SO) Océan Austral, (SPO) Océan Pacifique Sud.

A l'intérieur du genre Micromonas, se distinguent la présence de Mamiellales_X qui n'existe pas ailleurs et une forte abondance de *M. polaris*. Bien qu'il soit connu que les différentes espèces de *Bathycoccus* et *Ostreococcus* sont adaptées à différentes conditions de température[49,133], le manque de spécificité de la région V9 ne permet pas de faire cette distinction et les deux genres apparaissent simplement omniprésents.

Nous nous sommes donc ensuite concentrés sur la région Arctique uniquement, afin de voir si les différents bassins possédaient des caractéristiques environnementales différentes. Nous avons donc réalisé un second NMDS de la même manière en ne conservant que les échantillons présentés dans la Figure 16. Cette fois-ci, seul un paramètre s'est avéré séparer significativement les échantillons préalablement ordonnés selon l'abondance des espèces : la température de l'eau (Figure 18).

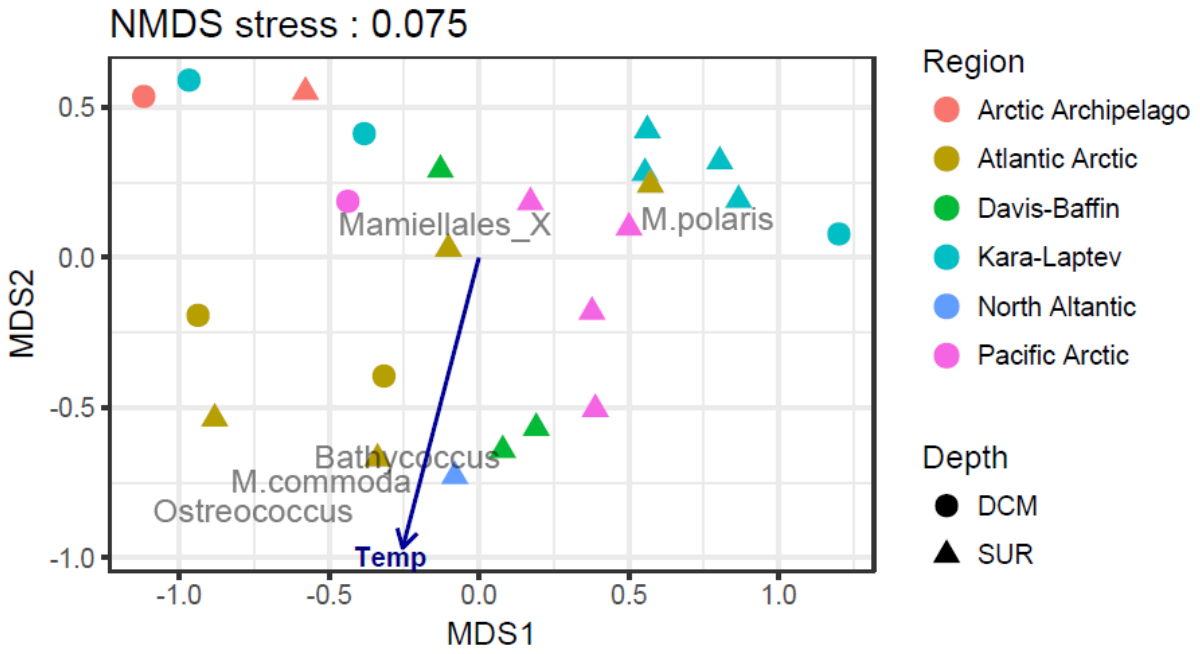


Figure 18 : NMDS basé sur l'abondance en métabarcodes des espèces de Mamiellales dans les échantillons arctiques. Les noms de phylum sont indiqués en gris, et chaque point représente un échantillon dont la forme dépend de la profondeur et la couleur de son bassin océanique.

La séparation des phyla réunit comme attendu *M. polaris* et Mamiellales_X d'un côté, présents majoritairement dans les mers de Kara et Laptev ainsi que dans l'Arctique pacifique, tandis que de l'autre s'associent *Bathycoccus*, *M. commoda* et *Ostreococcus* présents dans les mers de Davis et Baffin, ainsi que dans l'Arctique Atlantique et Pacifique. Le premier groupe d'échantillons constituerait donc un milieu

plus froid que le second bien que prélevés en été. Cette différence de température pourrait donc également être un facteur dans la répartition des différentes espèces.

Afin de confirmer cette corrélation entre l'abondance des Mamiellales et la température des eaux Arctiques, nous avons représenté ces deux facteurs pour chaque organisme dans les échantillons *Tara* et réalisé des régressions locales (Figure 19). La différence de température n'est pas très élevée et certains organismes se recoupent, mais nous pouvons tout de même observer que *Bathycoccus*, *M. commoda* et *Ostreococcus* sont abondants à de plus hautes températures (2°C et plus) que Mamiellales_X et *M. polaris* (inférieur à 2.5°C).

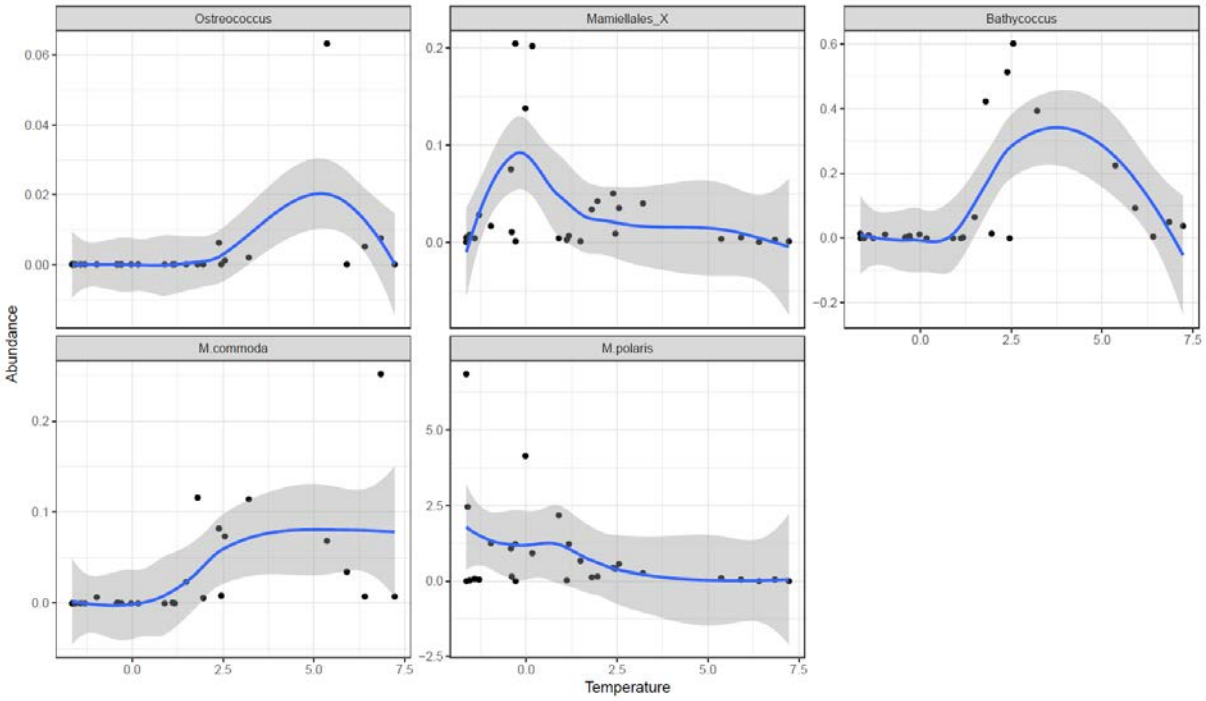


Figure 19 : Régressions locales basées sur l'abondance des Mamiellales à différentes températures. Chaque point représente un échantillon et la courbe bleue représente la courbe de régression lissée ajustée au nuage de points. Un pic correspond donc à une température de préférence pour l'espèce.

III. Comparaison avec les échantillons métagénomiques

Les analyses de biogéographies basées sur les métabarcodes permettent une première approche d'observation de la diversité globale des Mamiellales dans l'Arctique. Il est maintenant intéressant d'étudier les résultats d'analyses équivalentes basées sur les génomes de référence dont nous disposons et qui ont été utilisés dans l'article *Genome Resolved Biogeography of Mamiellales*. Afin d'obtenir des données

sur *Micromonas polaris*, nous avons ajouté à cela le transcriptome de la souche *Micromonas* sp. CCMP2099[45] provenant de l'océan Arctique, disponible depuis les données du MMETSP[134]. L'utilisation d'un transcriptome, bien que moins complet qu'un génome, nous permettra d'utiliser la même méthode sur les échantillons métagénomiques que pour les autres organismes.

Nous avons donc obtenu de la même manière que l'article présentant la biogéographie des eaux tempérés les abondances de ces sept espèces dans l'Arctique (Figure 20).

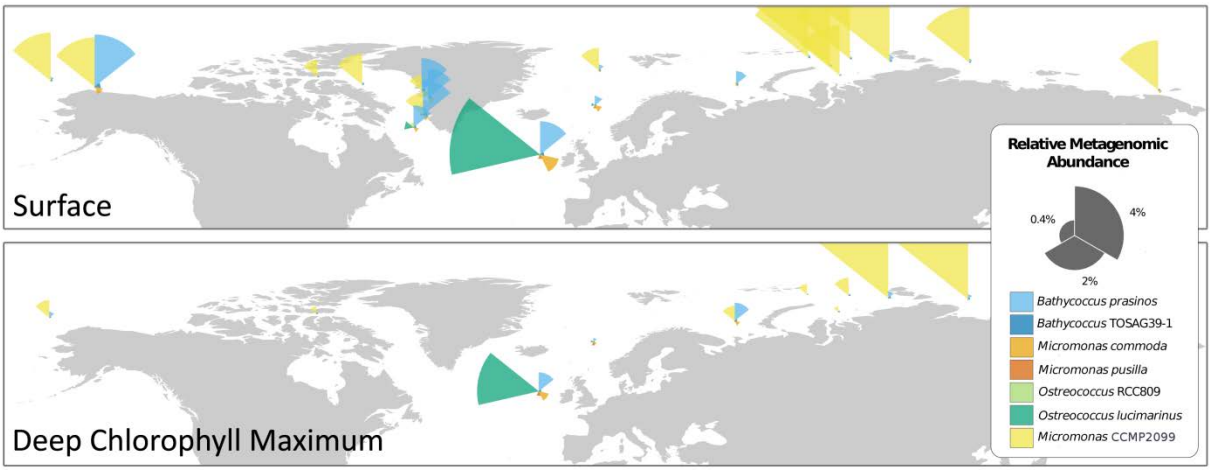


Figure 20 : Distribution géographique de sept génomes de Mamiellales dans les échantillons de la seconde partie de l'expédition *Tara* Oceans en surface et à la DCM. Chaque point représente un échantillon et la taille de chaque segment l'abondance relative du génome correspondant.

Nous pouvons observer que comme dans les échantillons de métabarcodes, *Micromonas* et *Bathycoccus* dominent ces eaux froides tandis qu'*Ostreococcus* n'est trouvable qu'à la limite de l'Atlantique Nord, où il est d'ailleurs extrêmement abondant dans le premier échantillon de l'expédition. Il est aussi en faible abondance dans le dernier échantillon, toujours dans le même océan. Pour *Micromonas*, c'est *M.* CCMP2099, appartenant à l'espèce *M. polaris* qui comme attendu a les plus fortes abondances sur la quasi-totalité des échantillons, et il est plus abondant sur la partie est que sur la partie ouest, où il est alors remplacé par *Bathycoccus*. *M. commoda* est également présent en très faible abondance dans quelques échantillons. Les résultats sont donc très cohérents entre les deux méthodes. Cependant, l'étude des génomes entiers nous informe que c'est uniquement l'écotype froid de *Bathycoccus*, *B. prasinos*, et pas du tout *B.* TOSAG39-1 qui est ici présent. De la même manière nous apprenons que c'est *O. lucimarinus* qui est trouvé au début et à la fin de l'expédition,

tandis qu'*O.* RCC809 est absent. Ceci semble logique au vu de leurs préférences écologiques déterminées auparavant, ces deux espèces étant celles trouvées dans des eaux typiquement plus froides. Il est à noter que *Micromonas* CCMP2099 n'a pas du tout été retrouvé dans les échantillons métagénomiques tempérés aux OTUs détectés précédemment.

IV. Conclusion

Au cours de cette étude, nous avons eu l'occasion d'analyser les Mamiellophyceae et plus particulièrement les Mamiellales dans un environnement très différent des eaux tempérées qui étaient jusqu'alors notre principal domaine d'intérêt. Les résultats de métabarcodes, bien que ne permettant pas d'avoir une vision détaillée de toutes les espèces présentes, nous donnent tout de même un point de vue intéressant sur la biodiversité dans ces échantillons. Nous avons notamment pu détecter un Crustomastigaceae très présent dans les eaux froides bien que n'ayant pas d'assignation plus détaillée, et pu avoir une bonne estimation de la biogéographie des Mamiellales qui pourra compléter une étude collaborative recoupant plusieurs ordres distincts.

La question de la répartition de *Micromonas* et *Bathycoccus* entre les différents bassins reste en suspens, pouvant potentiellement s'expliquer par un effet saisonnier ou par une simple différence de température. Une étude plus approfondie sera nécessaire afin de pallier au manque de connaissances sur cet environnement dans son ensemble, bien que d'autres analyses nous donnent déjà quelques pistes de réponse.

Enfin la métagénomique vient ici confirmer ce que nous avaient déjà montré les métabarcodes, en nous apportant grâce à quelques génomes de référence des détails à l'échelle des espèces, précisant notamment la présence de *B. prasinos* dans certains bassins, ce qui reste cohérent avec l'ensemble de nos analyses précédentes.

Chapitre 3 : Etude des variations génomiques de *Bathycoccus prasinos* dans les populations naturelles

I. Introduction

Nous avons pu déterminer au cours des chapitres précédents que *Bathycoccus prasinos* était une espèce cosmopolite, présente à la fois dans de nombreux océans tempérés et dans les eaux glacées des océans Arctique et Austral. Ce constat nous a encouragé à passer ici à une autre échelle d'analyse, celle des variations génomiques. En effet, si une seule espèce est présente dans des milieux aussi diversifiés, il est probable qu'il existe une stratégie d'adaptation passant par des différences à une échelle plus fine que celle de la spéciation, sous la forme de diverses populations de *Bathycoccus prasinos*.

Dans ce chapitre, nous étudierons donc les SNVs (*Single Nucleotide Variant*) de *Bathycoccus prasinos* mais aussi les SAAVs (*Single Amino-Acid Variant*) afin d'établir une structure des populations basée sur les mutations nucléotidiques mais également d'étudier l'impact de celles-ci au niveau protéique. Ces résultats sont ici présentés sous la forme d'un manuscrit en préparation pour lequel j'ai effectué la majorité des analyses et qui sera très prochainement soumis pour publication.

II. Article 3 : Equatorial to Polar genomic description of cosmopolitan *Bathycoccus prasinos* populations

Equatorial to Polar genomic description of cosmopolitan *Bathycoccus prasinos* populations

Jade Leconte[1,2], Tom Delmont[1,2], Gwenael Piganeau[3], Patrick Wincker[1,2], Olivier Jaillon[1,2]

1 Génomique Métabolique, Genoscope, Institut de biologie François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, 91057 Evry, France
2 Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, 3 rue Michel-Ange, 75016 Paris, France.
3 CNRS UMR7232 BIOM (Biologie Intégrative des Organismes Marin) Sorbonne University, 66650 Banyuls sur Mer, France

Abstract

Marine micro-eukaryotes are critical actors in the biogeochemical cycles and trophic webs of the oceans. Despite their importance, the enormous diversity of these organisms remains largely unexplored at the genomic level. Mamiellales is a group of prevalent photosynthetic picoeukaryotes that includes a haploid species of *Bathycoccus* found at almost all latitudes. The presence of a single species in negative temperature waters and in temperate waters up to twenty-five degrees raises the question of local adaptation and led us to study the population structure of this species at the SNV level. We analyzed 27 plankton metagenomic communities sampled by *Tara* Oceans across temperate and arctic oceans where the genome of *Bathycoccus prasinos* is detected as abundant. We found that these populations of the haploid *Bathycoccus* are composed of co-existing variants, most in uneven proportions, at 2% of nucleotide positions. We compared population diversity among sampled sites, taking into account environmental conditions. We found a clear differentiation between temperate and cold waters, with striking variant reversals at certain non-synonymous positions that change the amino acid composition of some proteins, while samples from transition waters displayed polymorphism at these positions. This study sheds light on key processes for the dynamics of genomes of unicellular eukaryotes that thrive in highly contrasted connected environments.

Introduction

Among phytoplankton, Mamiellales are the most prevalent photosynthetic picoeukaryotes[1,2]. Particularly abundant in coastal waters[3], they are widely distributed in open ocean as well and their geographical distribution has been studied in recent years on the basis of metabarcoding[4,5] and metagenomics datasets[6,7]. The three main genera of this order, *Bathycoccus*, *Micromonas* and *Ostreococcus*, are distributed over most latitudes and are therefore found in a wide range of environmental conditions. However, these environmental preferences seem to be reduced when analyzing the phyla. In temperate waters, *Bathycoccus prasinos*, *Ostreococcus lucimarinus* and *Micromonas pusilla* have been found at colder temperatures than *Bathycoccus* TOSAG39-1, *Ostreococcus* RCC809 and *Micromonas commoda*[7].

In the Arctic Ocean, which is colder and richer in nutrients than temperate waters, *Micromonas polaris* is largely dominant and seems restricted to this environment[8,9], however *Bathyccocus prasinos,* which is found in most oceans, is also abundant in those cold waters[10,11], making it the most cosmopolitan species of the order. The third genera, *Ostreococcus*, has never been reported in the Arctic despite being present in adjacent seasonally ice-covered waters such as the Baltic sea[12] or the White sea[13]. In Antarctic waters, only *Micromonas* has been detected, with populations defined as highly similar to the Arctic[14] ones.

The presence of the *Bathycoccus prasinos* genome at a stringency of more than 95% nucleotidic similarity in environmental metagenomes sampled from negative temperature waters and temperate waters up to twenty-five degrees argues against an ancient separation in two distinct species. This observation raises the question of putative related local adaptations that might have a signature in Single Nucleotide Variations (SNVs) within populations.

Experimental and environmental population genomic studies that have been conducted at gene level on Mamiellales markers[15,16] suggested an adaptation to climate changes. Recently, a complete genome analysis was conducted on *Ostreoccocus tauri* isolates from the Mediterranean sea, revealing high nucleotidic diversity within intergenic regions[17]. However, population genomic information at large geographical scale on Mamiellales is still lacking.
Using metabarcoding and metagenomics datasets from the *Tara* Oceans expedition[18], many studies focusing on different species were able to determine the distribution of eukaryotes in samples from all temperate oceans using reference genomes or markers[7,19,20]. Following these analyses, population structure studies based on genomic data started to emerge in order to assess species diversity, for example of crustaceans[21] or marine bacteria[22].
Here we leverage metagenomic data from Arctic Oceans that have been added to the *Tara* Oceans collection[23,24], to compare the genomic diversity of the populations of *Bathycoccus prasinos* in polar and temperate environments, exploiting a total of 27 selected samples from surface and deep chlorophyll maximum waters.

Materials and Methods

Genomic resources

Bathycoccus RCC1105 was isolated in the bay of Banyuls-sur-mer at the SOLA station at a depth of 3m in January 2006[25]. Sequences were downloaded from the Online Resource for Community Annotation of Eukaryotes[26].

Metagenomics reads from *Tara* Oceans samples[18,27] corresponding to the 0.8 to 5µm organism size fraction[28] collected at surface and deep chlorophyll maximum layers of the water column were used to assess the diversity of Bathycoccus. For the arctic samples, from TARA_155 to TARA_210, as this size fraction was not available the 0.8 to 2000 µm size fraction was used instead. In stations where both 0.8-5µm and 0.8-2000µm size fraction samples were available we obtained similar *Bathycoccus* relative abundance values (Supplementary Figure 1) probably due to the higher abundance of smaller organisms in plankton.

Environmental parameters

To assess the potential correlation between genomic variations and local environmental conditions, we used the physicochemical parameter values related to the Tara Oceans expedition sampling sites available in the PANGAEA database[28]. Those contextual data tables can be downloaded at the following link: https://doi.pangaea.de/10.1594/PANGAEA.875582.

Abundance counts

We mapped metagenomics reads on RCC1105 genome sequences using the Bowtie2 2.1.0 aligner with default parameters[29]. We then filtered out alignments corresponding to low complexity regions with the DUST algorithm[30] and selected reads with at least 95% identity and more than 30% high complexity bases.

Some gene sequences might be highly similar to orthologous genes from other organisms, in particular, *Bathycoccus* TOSAG39-1[6] co-occurring with *Bathycoccus prasinos* in some samples, and thus recruit additional metagenomic reads. To prevent bias from this putative artifact, we used a statistical approach to discriminate genes with atypical mapping counts. This analysis is based on the assumption that the values of the metagenomics RPKM (number of mapped reads per gene per kb per million of mapped reads) follow a normal distribution. We conducted Grubbs' test for outliers to provide for each sample a list of genes with RPKM distant from this distribution then merged all lists to have a global outliers set[19]. We finally computed relative genomic abundances as the number of reads mapped onto non-outlier genes normalized by the total number of reads sequenced for each sample.

Filtering steps

Using the previously filtered set of reads, we discarded those with MAPQ scores < 2 in order to remove reads mapping at multiple locations with the same score, which are randomly assigned at

either position by Bowtie2 and could cause errors in the variant detection. We then calculated genome coverage at each position in the coding regions using BEDTools 2.26.1[31] and kept samples having a mean coverage above 4x. On the initial set of 162 samples, 27 passed this filter. Among them, 4 samples considered to have very good coverage (more than 30x) were selected for a first in-depth variant analysis. The larger set of 27 samples was subsequently used for a global biogeography study.

For each sample, the "callable sites" used in variant analysis were selected from samtools mpileup results[32] as genomic positions covered by a number of reads comprised between 4 and a maximum corresponding to the average coverage in the sample plus twice the standard deviation.

Variant calling

We detected variable genomic sites using Anvi'o[33] on the two sets of *Tara* Oceans samples: a set of four samples above 30x and a set of 27 samples above 4x coverage. For that we created two Anvi'o databases then performed two separate SNV (Single Nucleotide Variant) and SAAV (Single Amino Acid Variant) calls. Quince mode was used in order to retrieve information for each variant locus in all samples. This method takes into account multiple variants in a codon by indicating all amino-acids present at a given position rather than independently projecting SNV results. Only positions callable in every sample of interest were kept, to be able to compare them. For the 4-sample set, a total of 10 585 350 positions (86% of *Bathycoccus* coding regions) were studied, while only 1 715 482 positions (14%) were kept for the 27-sample set. We considered an allele at a detected locus if it was confirmed by at least 4 reads. Variants were then considered either fixed in a sample (called fixed mutation) if presenting a single allele different from at least one other sample, or polymorphic (called SNV) if presenting two or more alleles in the sample. Only amino-acid variants for which a corresponding nucleotide variant passed those filters were kept.

Genomic distance computation

We computed a genomic distance for each pair of samples based on allele content at each SNV position. An allele can be considered either present or absent, without taking its frequency into account. The distance thus corresponds to the number of common alleles between the two samples (one nucleotide present in both samples would then count for two, while a nucleotide in a single sample would count for one) divided by the total number of alleles (for example three if two samples were sharing an allele and one had an exclusive allele). Identical allelic content would give a score of 1, no allele in common would give a score of 0.

Based on this distance metric, we computed a phylogenetic tree of all samples plus the reference genome RCC1105 using the core R function hclust with default parameters. We obtained bootstrapped values using the pvclust function with 9999 permutations, from the pvclust 2.0-0 R-package. Dendrograms were plotted using the dendextend 1.3.0 package.

Finally, in order to better visualize the information, we used the same values to assign colors to each sample, with the distance between colors reflecting the genomic distance between samples. To achieve this, we carried out Principal Component Analysis (PCA) with package vegan 2.4-1, and translated position values from the three first axes to a Red Green Blue (RGB) color-code for each sample. The resulting color circles were plotted on a map using R-packages ggplot2_2.2.1, scales_0.4.1 and maps_3.1.1.

Statistical approaches

Multiple statistical analyses were performed for this manuscript based on SNV and SAAV results. First, we computed pairwise water temperature distances in order to run a Mantel test against the genomic distances of all 27 samples, using R-package vegan 2.4-1.

Another experiment focused on those variants significantly responsible for the large distance between two main groups from the previously computed hierarchical clusters. We thus gathered pairwise distances between samples for each position. Finally, we selected loci with a distance above 0.6 between samples from different clusters and plotted the density curves of their frequencies for each sample independently using ggplot2_2.2.1.

The last statistical phase was the correlation study between amino-acid frequencies and sample temperatures in order to assess a potential swap of major and minor alleles between cold and temperate samples. For each SAAV, we took the amino-acid with the highest frequency at the position for each sample, and only kept the positions with at least two different alleles among samples. We then computed a Wilcoxon test for each position using as a first group the temperatures for which the first amino acid was found and as a second group the temperatures for which the other amino acid was found, and applied a Bonferroni correction to the resulting p-values. We kept positions with p-values smaller than 0.05, for a total of 13 variants, and plotted their amino-acid frequencies in our samples using R-packages gridExtra 2.2.1 and ggplot2_2.2.1.

Results

Bathycoccus genomic diversity

We analysed the diversity of *Bathycoccus* RCC1105 within natural populations by mapping metagenomics reads from the *Tara* Oceans expedition[18,27] on the reference genome[25] as previously reported[6,7]. For further analysis, we selected 27 samples providing a mean coverage of recruited reads higher than a threshold of 4X.

Among them, we considered a set of 4 samples where *Bathycoccus* RCC1105 appears relatively abundant with a mean coverage of recruited reads above 30x. Among these four samples, the coverage of recruited reads appeared to be consistent for all genes except those on chromosome 19 (Figure 1). Within each sample, most genes also present similar coverage. Two chromosomes, described as outliers[25], have different coverage and GC content behaviour.

Chromosome 19, the SOC or Small Outlier Chromosome, has previously been described as very variable, accordingly it presents a major coverage drop along most of its length. Chromosome 14, the BOC or Big Outlier Chromosome, possesses a large region considered as outlier. In our study, unlike the SOC, the coverage is higher in the outlier area of BOC, maybe corresponding to a signature of copy number variation in *Bathycoccus* populations or to the presence of a genome from another phylum in the sample that shares high genomic similarity with this region. The first third of the chromosome which is not an outlier and has a higher GC content presents a slight coverage drop in most samples.

Looking in detail at the Single Nucleotide Variant (SNV) and Single Amino-Acid Variant (SAAV) densities per sample, we can observe a positive correlation between densities and coverage but there is no clear impact of temperature or geographical clusters presenting similar densities. Coverage itself does not appear to be linked to the water temperature.

To analyse *Bathycoccus* diversity in detail we first worked on the 4 samples with highest coverage as it allowed us to have numerous reads confirming the presence of different alleles at a locus and thus lowering our error rate. As a second step, we extended our analysis to less-covered samples while verifying the coherence of both sets for comparison of different environments.

Using the reduced set of four well-covered samples, a total of 350 478 positions present variation in one or more samples, corresponding to 3.31% of our initial set of callable positions (Table 1). The mean coverage goes from 38.41 in one arctic sample to 63.87 in the other one, with two temperate samples falling in between. The total variant density reaches a maximum of 1.96% in the most covered sample, and the majority of those variants correspond to biallelic SNVs.

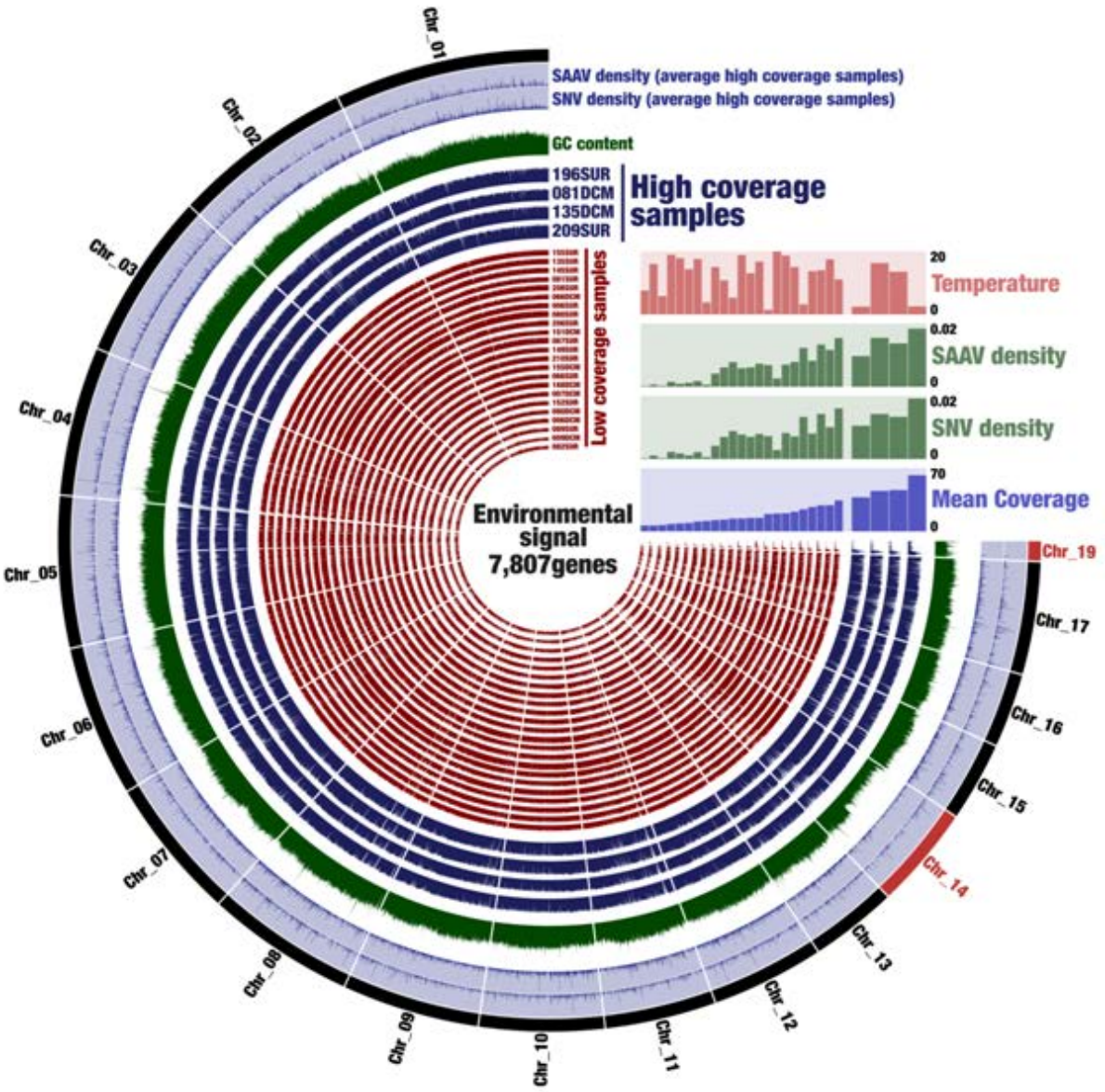


Figure 1: *Bathycoccus* whole genome diversity in 27 Tara Oceans samples. Description of layers starting from the exterior: chromosome number (in black), SAAV and SNV average density among the four best covered samples (in blue), GC content in percentage (in green), each layer then corresponds to the coverage of one sample, ordered by mean coverage. Figure computed with Anvi'o

Samples	Callable sites	Mean coverage on callable sites	Total number of variants	Number of fixed mutations	Number of biallelic SNVs	Number of triallelic SNVs	Number of quadriallelic SNVs
81DCM	10 972 751	45.73	146 793 *(1.38%)*	15 943	130 186	662	2
135DCM	10 983 591	44.80	155 249 *(1.46%)*	14 310	140 083	854	2
196SUR	11 181 235	62.13	208 320 *(1.96%)*	7 557	199 024	1732	7
209SUR	11 023 554	37.77	108 967 *(1.03%)*	14 303	94 463	201	0

Table 1: Number of variants for the four best-covered samples, total and separation according to the number of alleles found at the loci in each sample.

The coverage and the number of callable sites are consistent between chromosomes for all samples, except for chromosome 19 (79 to 91% against 23% for callable sites) which has a very low horizontal and vertical coverage (Supplementary Table 1). The variant density appears homogeneous among chromosomes, except for chromosome 14, where the density drops to half of that found for the other chromosomes. Variant density also appears homogeneous across the four different samples (Supplementary Figure 2).
At codon level, we detected SNVs leading to amino-acid changes (SAAVs, as previously reported[22]) to approximate the ratio of non-synonymous versus synonymous variants by dividing the number of SAAVs by the number of SNVs. About 6% of codons containing SNVs have more than one nucleotide variant. Here, the SAAV/SNV ratios per sample range from 0.33 to 0.40 without apparent geographical or environmental patterns for the four samples.
We simplified our dataset by associating each SAAV with an amino acid substitution type (AAST), defined as the two most frequent amino acids in a given SAAV. Among 210 070 SAAVs, we observed 206 out of 210 theoretically possible AASTs and very similar frequencies among samples (Supplementary Figure 3). As expected, the most common and prevalent amino-acid substitutions would have low impact at the protein level given their positive BLOSUM90[34] substitution matrix scores (Supplementary Figure 4). The ten most common AASTs in all samples have an average score of 0.7, while the 10 rarest AASTs have an average score of -4.5. The global distribution of all SAAV associated scores also shows a majority of positive values (Supplementary Figure 5). Similar results were found with the same rationale for bacteria[22].

We then added to this experiment 23 metagenomic samples that provide between 4x and 30x mean coverage and reselected positions with at least 4x coverage in all 27 samples. Variant density is positively linked among the sets with an almost perfect linear correlation (Supplementary Figure 6) and SAAV/SNV ratios are also similar, validating use of this larger set of samples on a reduced portion of the genome. In these 27 samples we obtained a total of 80 284 SNVs and fixed mutations which correspond to 4.68% of callable positions (Supplementary Table 2). This higher number of variant alleles compared to the previous value on four samples is probably due to the greater diversity of sampling environments. As previously observed, most samples present more polymorphic positions than fixed allele positions among our set of callable sites. But in a few samples, populations seem to hold as many SNVs as fixed alleles. Noticeably, some populations such as those in austral samples, have a very low polymorphism rate but no lack of coverage of recruited reads. At the SAAV level, we cannot see any particular correlation between the SAAV/SNV ratio and either environmental conditions or geographical patterns at oceanic basin scale. AASTs and BLOSUM90 score distributions are almost identical using the 4 and 27 samples sets, with only minor inversions in AAST prevalence.

Population structure analysis

To compare and analyze proximity among populations at SNV level we computed pairwise genomic distances between all samples considering fixed and polymorphic positions within the set of nucleotide variations found in the 27 samples. We constructed a dendrogram representing these distances among samples, including the RCC1105 reference sequence (Figure 2). This tree clearly separates Arctic and temperate samples. Austral samples (number 82 and 89) are the farthest from any other clusters as is the case for the Mediterranean sample 9DCM which, like the austral ones, presents mostly fixed mutations and low polymorphism.

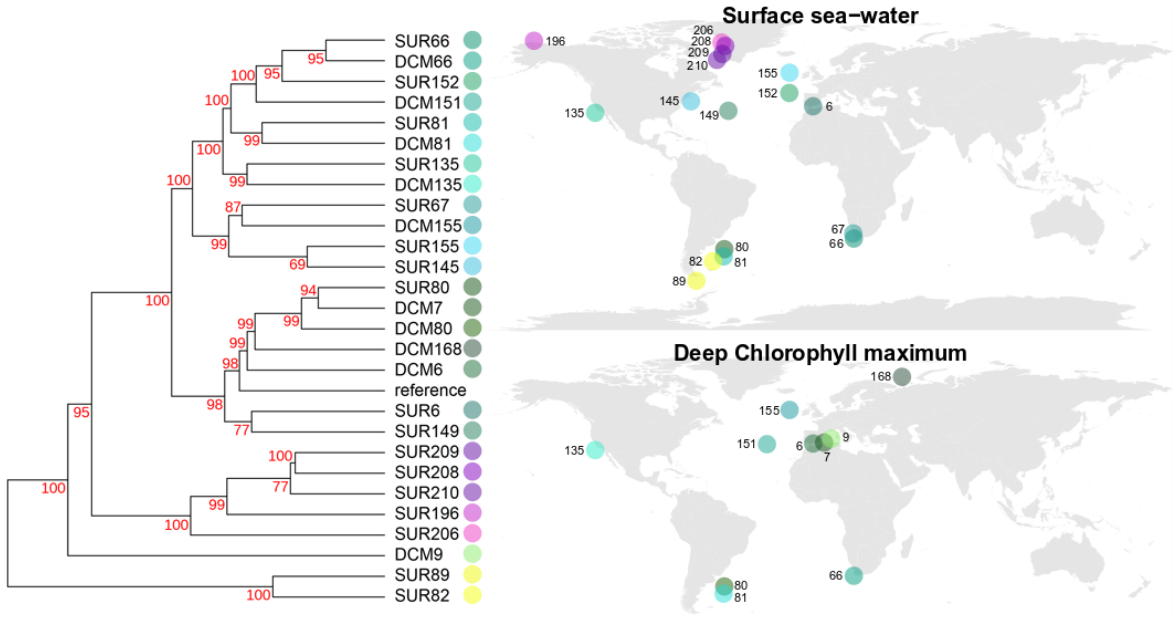


Figure 2: *Bathycoccus prasinos* genomic distance among 27 sampled populations based on nucleotide variant patterns, displayed as a phylogenetic tree (left panel) and a biogeography map (right panel for surface (top) or Deep Chlorophyll Maximum (bottom)). The similarities of sample colors (right) reflect the genomic similarities of populations (left) (Methods).

Each sample was attributed a color based on pairwise distances, the color difference representing the genomic distance between samples (Methods), and was plotted on a world map (Figure 2). Multiple biogeographical patterns emerge, for example a clear separation of Arctic samples in purple and austral ones in yellow. Most temperate samples are represented in a gradient of green, but sample 145SUR, near the end of the Labrador current thus under the influence of Arctic waters, and 155SUR near the end of the Gulf Stream just before Arctic waters, are both represented in blue. Sample 135DCM, located off the coast of California near a site of cold and rich upwelling and samples 81SUR and 81DCM situated close to austral waters are represented in blue-green colors, therefore in between temperate and warm-cold transition points. The Mediterranean sample 9DCM mentioned above appears in a yellow-green color, which is coherent with its genomic patterns similar to the austral samples.

A Mantel test between matrices of genomic and temperature differences for 27 samples confirms a positive correlation (0.4818 statistic at p-value=0.001, Supplementary Figure 7).

Genomic differences between temperate and cold waters populations

Since populations from temperate and cold waters have different SNV patterns, we decided to further analyze the genomic positions responsible for their differentiation. We identified a set of 2742 SNVs that significantly distinguish the two groups of populations (Methods) as illustrated by very different distributions of allele frequencies (Figure 3). In temperate samples, a majority of alleles are fixed as observed frequencies are close to 0% or to 100%. We cannot rule out the possibility that other variants are present in populations at rates below our detection capacity. In contrast, in all but one of our arctic samples (sample 206), these genomic positions mostly present local polymorphisms between two principal alleles, present in different ratios depending on the sample (80/20, 70/30 or 50/50 ratios). The high genomic distances between arctic and temperate populations would thus be mainly related to this group of loci that appear biallelic in the Arctic but present a single allele elsewhere.

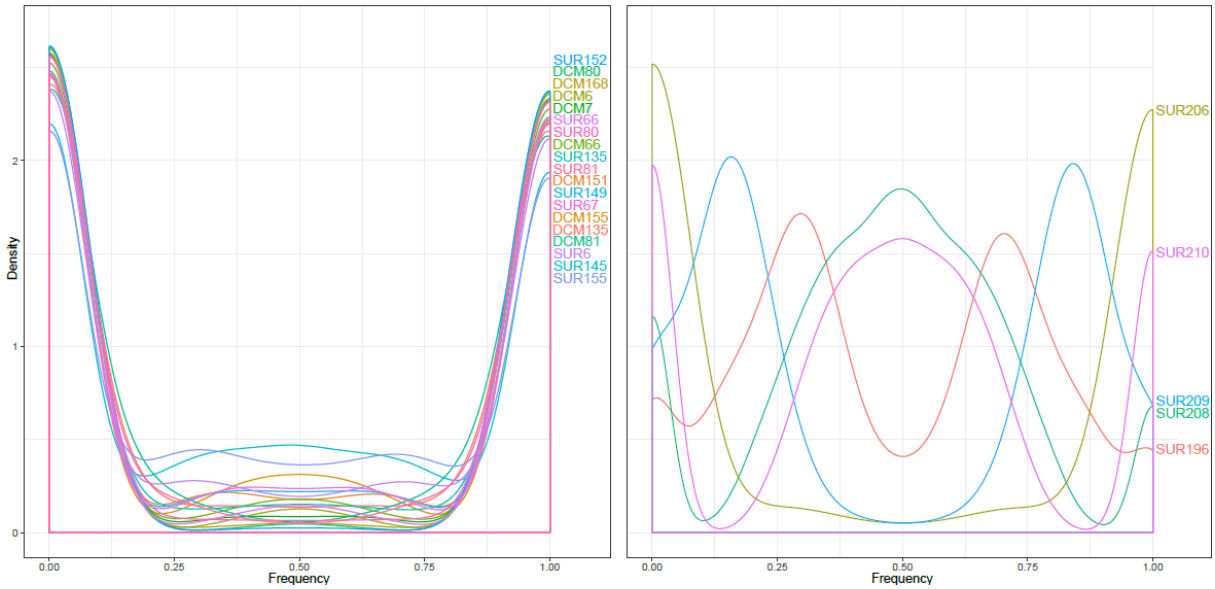


Figure 3: Allele frequency distributions for 2742 selected SNVs for different samples belonging to the temperate (left panel) or the arctic (right panel) cluster.

Finally, we examined amino-acid changes between cold and temperate populations, as such variations might have a functional impact. We selected 13 positions in the genome with allele variations highly correlated to water temperature and analyzed their frequencies in our 27 samples with respect to the genes in which they were found (Figure 4).

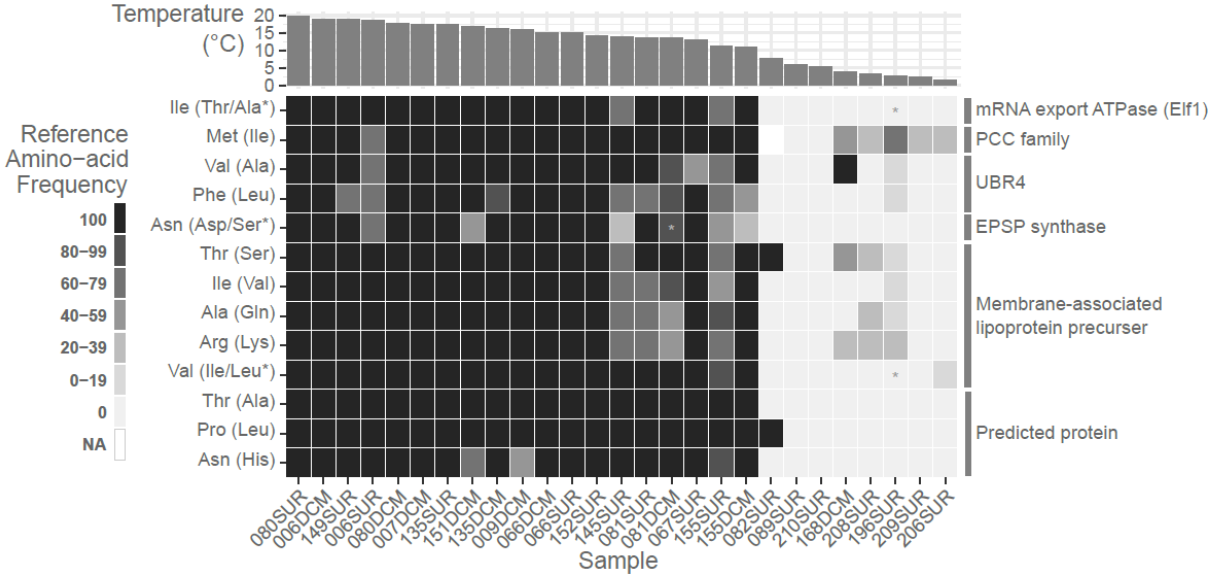


Figure 4: Heatmap presenting the frequency of amino-acids among 27 samples in 13 genomic positions that segregate populations according to temperature. Alternate amino-acids found are indicated in parentheses. Samples are sorted by their temperature indicated at the top, and the six proteins in which variants are found are indicated on the right.

We have a clear pattern of swap of major amino-acids between temperate and cold samples, for which arctic and austral populations have similar phenotypes for those positions. In most samples, the amino-acid frequencies of these positions correspond to fixation or near-fixation. Interestingly, populations presenting a protein polymorphism were sampled at intermediate temperatures and also at transition points between different oceanographic systems. For example, station 155 is located on the Northeast Passage, close to the Arctic ocean, and station 145 is at the end of the Northwest Passage in a cold current while station 81 on the other side is close to the Austral ocean. Station 67

populations clearly have a pattern similar to that of temperate samples despite expectation of similarity to polar samples given a location close to upwelling colder and richer waters.

The 13 discriminating positions are located in a total of six different genes. One of the genes has three amino-acids involved but no functional annotation. Annotations of other genes appear related to stability of protein structure or functions related to low or high temperature. The most segregated gene has six amino-acids involved and is a membrane-associated lipoprotein precursor. Structure and concentration of lipids has been shown to be important for cold adaptation in many organisms[35,36]. We can also notice an impact on transport proteins, such as the mRNA export or the PCC (Polycystin Cation Channel) family, which might need specific adaptations in order to function at low temperatures[37,38]. Other interesting proteins are impacted, such as UBR4, involved in the N-end rule pathway[39], marking the protein for degradation, or EPSP synthase, a highly important protein that participates in the biosynthesis of the aromatic amino acids phenylalanine, tyrosine, and tryptophan via the shikimate pathway in bacteria, fungi, and plants[40,41].

While the main pattern between cold and temperate samples, according to the analysis of allele frequencies, was high polymorphism in the first group and fixed alleles in the second, these 13 selected positions reveal a sub-pattern. Though austral populations are different from others, with very few polymorphic positions, they share with arctic populations similar major alleles at the 13 impactful positions.

Discussion

Using metagenomics samples from the *Tara* Oceans expedition and the genome sequence of the picoeukaryote alga *Bathycoccus prasinos* RCC1105, we assessed the genomic diversity of a cosmopolitan species model in temperate and polar marine biomes. With 27 metagenomic samples where the genome of *Bathycoccus prasinos* RCC1105 presents a significant coverage of recruited reads, we estimated that single nucleotide variations are mostly present in billalelic forms with a maximum density reaching 2% of the coding regions. Polar and temperate populations appear to be very similar genomically, although they can be segregated by patterns of single nucleotide variations in 0.16% of the genomic positions and at the protein level, 6 genes whose amino acid composition appears biome dependent.

Due to currents, a large part of the waters from the Arctic Ocean originate from the North Atlantic. Plankton is passively transported along this path and encounters the polar front; *Bathycoccus prasinos* RCC1105 seems to cross it with success as indicated by the very high genomic similarity between polar and temperate abundant populations. Therefore, in light of our results and this oceanographic context, multiple hypotheses can be raised concerning the evolutionary strategies that have shaped the genomic properties of *Bathycoccus prasinos*. Among these, the existence of alleles that would be restricted to each biome appears highly unlikely. Indeed, the polymorphic genomic loci of *Bathycoccus prasinos* populations consist mainly of two alleles whose proportions vary along the path of the currents connecting arctic and temperate waters. We favor the hypothesis that a relatively short life cycle combined with environmental selection occurring along the path would permit rapid recombination of dominant alleles and rapid swap of their relative proportions in populations transported by currents.

Genes that display a link between their amino-acid composition and temperate versus polar biomes also present interesting putative functions that seem environmentally coherent, potentially in relation to optimized structure or synthesis of lipids, transport-proteins, or the efficiency of post-translational processes. In line with this proposition, populations sampled in marine areas located between cold and temperate oceans present polymorphisms at those amino-acid positions. Further studies, from cultures or from natural populations, are required to better characterize the functional impact of these amino-acid variations. For example, the probable adaptation patterns would benefit from a gene expression study to test patterns of acclimation, as recently exemplified in fish[42] complementing patterns observed for bacterial communities in the same samples[24]. Such data would feed into efforts

to better understand and predict the impact of global warming, which could have a major impact on the polar biome community. It has recently been suggested that advection by North Atlantic currents to the Arctic Ocean, combined with warming, will shift the distribution of phytoplankton poleward, leading to a restructuring of biogeography and complete communities[43]. Such rapid and significant changes challenge adaptation and acclimatization strategies that have evolved over millions of years, especially for cosmopolitan organisms such as *Bathycoccus prasinos*.

References

1. Vaulot, D., Eikrem, W., Viprey, M. & Moreau, H. The diversity of small eukaryotic phytoplankton (< or =3 microm) in marine ecosystems. *FEMS Microbiol. Rev.* **32**, 795–820 (2008).
2. Massana, R. Eukaryotic Picoplankton in Surface Oceans. *Annu. Rev. Microbiol.* **65**, 91–110 (2011).
3. Worden, A., Nolan, J. & Palenik, B. Assessing the Dynamics and Ecology of Marine Picophytoplankton: The Importance of the Eukaryotic Component. *Limnology and Oceanography* **49**, 168–179 (2004).
4. Monier, A., Worden, A. Z. & Richards, T. A. Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ Microbiol Rep* **8**, 461–469 (2016).
5. Tragin, M. & Vaulot, D. Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Sci Rep* **9**, 1–14 (2019).
6. Vannier, T. *et al.* Survey of the green picoalga Bathycoccus genomes in the global ocean. *Sci Rep* **6**, 1–11 (2016).
7. Leconte, J. *et al.* Genome Resolved Biogeography of Mamiellales. *Genes* **11–1**, (2020).
8. Lovejoy, C. *et al.* Distribution, phylogeny, and growth of cold−adapted picoprasinophytes in Arctic SeaS1. *Journal of Phycology* **43**, 78–89 (2007).
9. Simon, N. *et al.* Revision of the Genus Micromonas Manton et Parke (Chlorophyta, Mamiellophyceae), of the Type Species M. pusilla (Butcher) Manton & Parke and of the Species M. commoda van Baren, Bachy and Worden and Description of Two New Species Based on the Genetic and Phenotypic Characterization of Cultured Isolates. *Protist* **168**, 612–635 (2017).
10. Joli, N., Monier, A., Logares, R. & Lovejoy, C. Seasonal patterns in Arctic prasinophytes and inferred ecology of Bathycoccus unveiled in an Arctic winter metagenome. *ISME J* **11**, 1372–1385 (2017).
11. Lovejoy, C. & Potvin, M. Microbial eukaryotic distribution in a dynamic Beaufort Sea and the Arctic Ocean. *J Plankton Res* **33**, 431–444 (2011).
12. Majaneva, M., Enberg, S., Autio, R., Blomster, J. & Rintala, J. m. Mamiellophyceae shift in seasonal predominance in the Baltic Sea. *Aquatic Microbial Ecology* **83**, 181–187 (2019).
13. Belevich, T. A. *et al.* Photosynthetic Picoeukaryotes in the Land-Fast Ice of the White Sea, Russia. *Microb Ecol* **75**, 582–597 (2018).
14. Simmons, M. P. *et al.* Intron Invasions Trace Algal Speciation and Reveal Nearly Identical Arctic and Antarctic Micromonas Populations. *Mol. Biol. Evol.* **32**, 2219–2235 (2015).
15. Schaum, C.-E., Rost, B. & Collins, S. Environmental stability affects phenotypic evolution in a globally distributed marine picoplankton. *The ISME Journal* **10**, 75–84 (2016).
16. Li, W. K. W., McLaughlin, F. A., Lovejoy, C. & Carmack, E. C. Smallest Algae Thrive As the Arctic Ocean Freshens. *Science* **326**, 539–539 (2009).
17. Blanc-Mathieu, R. *et al.* Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Science Advances* **3**, e1700239 (2017).
18. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *PLOS Biology* **9**, e1001177 (2011).
19. Seeleuthner, Y. *et al.* Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nat Commun* **9**, 1–10 (2018).
20. Biard, T. *et al.* Biogeography and diversity of Collodaria (Radiolaria) in the global ocean. *The ISME Journal* **11**, 1331–1344 (2017).
21. Madoui, M.-A. *et al.* New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. *Mol. Ecol.* **26**, 4467–4482 (2017).
22. Delmont, T. O. *et al.* Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *eLife* **8**, e46497 (2019).
23. Ibarbalz, F. M. *et al.* Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell* **179**, 1084-1097.e21 (2019).
24. Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083.e21 (2019).
25. Moreau, H. *et al.* Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biology* **13**, R74 (2012).

26. Sterck, L., Billiau, K., Abeel, T., Rouzé, P. & Van de Peer, Y. ORCAE: online resource for community annotation of eukaryotes. *Nature Methods* **9**, 1041–1041 (2012).

27. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific Data* **4**, 170093 (2017).

28. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* **2**, 1–16 (2015).

29. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357–9 (2012).

30. Morgulis, A., Gertz, E., Schaffer, A. & Agarwala, R. A Fast and Symmetric DUST Implementation to Mask Low-Complexity DNA Sequences. *Journal of computational biology : a journal of computational molecular cell biology* **13**, 1028–40 (2006).

31. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current Protocols in Bioinformatics* **47**, 11.12.1-11.12.34 (2014).

32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

33. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).

34. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915–10919 (1992).

35. Upchurch, R. G. Fatty acid unsaturation, mobilization, and regulation in the response of plants to stress. *Biotechnol Lett* **30**, 967–977 (2008).

36. van Dooremalen, C., Suring, W. & Ellers, J. Fatty acid composition and extreme temperature tolerance following exposure to fluctuating temperatures in a soil arthropod. *Journal of Insect Physiology* **57**, 1267–1273 (2011).

37. Gerday, C. & Glansdorff, N. *EXTREMOPHILES - Volume II*. (EOLSS Publications, 2009).

38. Ting, L. *et al.* Cold adaptation in the marine bacterium, Sphingopyxis alaskensis, assessed using quantitative proteomics. *Environmental Microbiology* **12**, 2658–2676 (2010).

39. Tasaki, T. *et al.* The Substrate Recognition Domains of the N-end Rule Pathway. *J. Biol. Chem.* **284**, 1884–1895 (2009).

40. Bentley, R. & Haslam, E. The Shikimate Pathway — A Metabolic Tree with Many Branche. *Critical Reviews in Biochemistry and Molecular Biology* **25**, 307–384 (1990).

41. Kishore, G. M. & Shah, D. M. Amino acid biosynthesis inhibitors as herbicides. *Annu. Rev. Biochem.* **57**, 627–663 (1988).

42. Sandoval-Castillo, J. *et al.* Adaptation of plasticity to projected maximum temperatures and across climatically defined bioregions. *PNAS* (2020) doi:10.1073/pnas.1921124117.

43. Oziel, L. *et al.* Faster Atlantic currents drive poleward expansion of temperate phytoplankton in the Arctic Ocean. *Nature Communications* **11**, 1705 (2020).

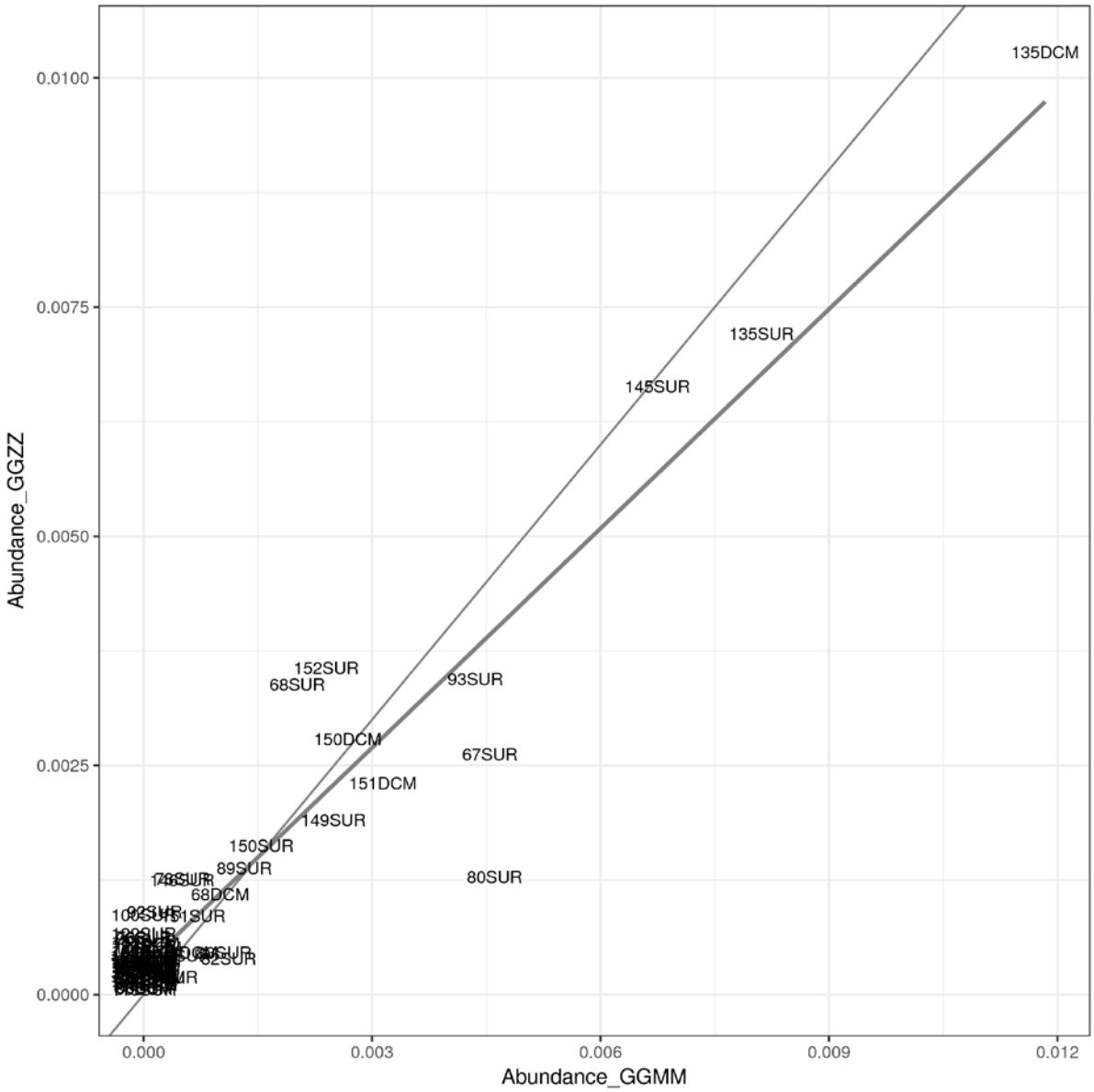
Supplementary Materials

Chr.	Total size	Coding size	Callable sites %	Mean cov. 81DCM	Mean cov. 135DCM	Mean cov. 196SUR	Mean cov. 209SUR	Variant density 81DCM	Variant density 135DCM	Variant density 196SUR	Variant density 209SUR
01	1 352 724	1 133 982	91%	45.7	45.63	65.14	38.57	1.31%	1.41%	1.95%	0.97%
02	1 122 692	922 092	90%	46.16	45.54	63.86	38.31	1.38%	1.48%	2.02%	1.03%
03	1 091 008	909 405	89%	46.65	45.65	64.02	38.29	1.36%	1.43%	2.02%	1.02%
04	1 037 991	863 091	87%	47.82	46.54	65.36	39.65	1.35%	1.38%	1.91%	1.04%
05	1 019 276	850 563	89%	46.95	45.81	65.17	38.78	1.39%	1.48%	1.91%	1.03%
06	989 707	821 307	89%	45.76	45.03	63.75	38.77	1.37%	1.49%	2.05%	1.07%
07	955 652	791 748	87%	47.16	46.82	65.2	39.91	1.31%	1.42%	1.75%	0.92%
08	937 610	767 367	89%	47.08	46.31	64.27	38.95	1.30%	1.39%	1.91%	1.02%
09	895 536	740 394	89%	46.5	45.39	64.35	38.87	1.30%	1.38%	1.95%	1.00%
10	794 368	667 857	88%	46	45.38	62.7	37.78	1.40%	1.45%	2.05%	1.09%
11	741 603	610 035	86%	45.59	45.42	63.02	38.92	1.43%	1.51%	2.08%	0.99%
12	712 459	597 036	90%	46.33	45.89	63.17	38.55	1.42%	1.50%	2.11%	1.08%
13	708 035	607 836	79%	46.79	44.95	64.57	37.37	1.52%	1.51%	1.97%	1.19%
14	663 424	478 431	90%	47.43	46.5	60.31	34.77	0.98%	1.04%	1.00%	0.53%
15	519 835	431 610	87%	46.53	45.38	64.31	38.46	1.40%	1.49%	1.94%	1.05%
16	494 108	407 472	84%	47.33	46.14	62.55	38.19	1.43%	1.50%	1.98%	0.99%
17	465 570	386 298	78%	46.02	44.99	61.94	37.25	1.57%	1.67%	2.05%	1.15%
18	310 170	251 943	84%	44.99	43.89	60.16	36.87	1.78%	1.86%	2.30%	1.24%
19	146 238	84 381	23%	26.46	27.17	35.01	19.29	1.41%	1.50%	1.16%	0.62%
Total	14 958 006	12 322 848	87%	45.43	44.65	62.05	37.24	1.37%	1.44%	1.94%	1.01%

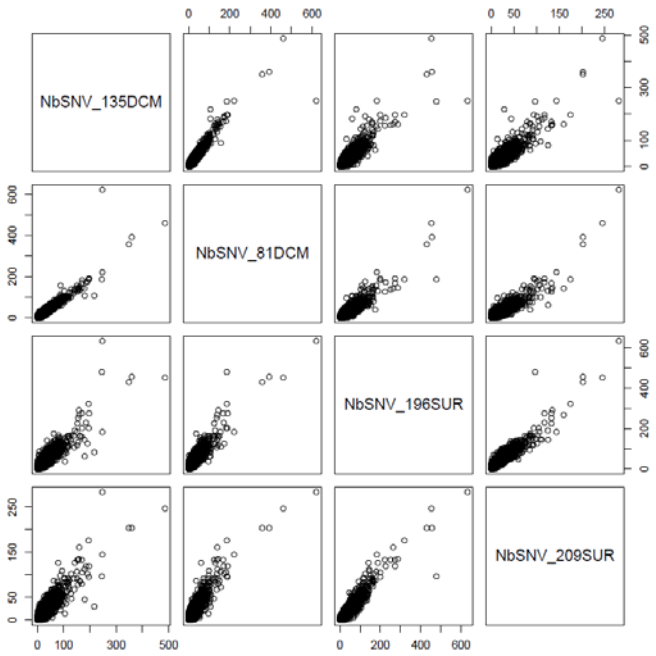
Supplementary Table 1: Main statistics of genomic diversity among populations of *Bathycoccus prasinos* from the four best-covered samples. For each chromosome or the whole genome, the columns indicate total chromosome size, cumulative length of coding regions, percentage of the chromosome considered as callable, coverage and variant density for each of the samples.

Samples	Callable sites	Mean coverage on callables	Total number of variants	Number of fixed mutations	Number of biallelic SNVs	Number of triallelic SNVs	Number of quadriallelic SNVs
6DCM	8 804 921	8.22	8 662 *(0.50%)*	4 786	3 876	0	0
6SUR	10 698 809	20.01	16 627 *(0.97%)*	2 491	14 109	27	0
7DCM	10 053 998	10.94	7 624 *(0.44%)*	3 873	3 749	2	0
9DCM	6 355 421	6.53	13 171 *(0.77%)*	11 293	1 878	0	0
66DCM	10 646 963	21.67	15 940 *(0.93%)*	3 662	12 261	17	0
66SUR	10 030 196	12.45	11 829 *(0.69%)*	4 613	7 213	3	0
67SUR	10 172 546	14.89	16 710 *(0.97%)*	4 673	12 027	10	0
80DCM	9 324 737	8.84	7 098 *(0.41%)*	4 084	3 012	2	0
80SUR	10 778 634	19.82	8 005 *(0.47%)*	3 087	4 915	3	0
81DCM	10 972 751	45.73	26 111 *(1.52%)*	3 184	22 843	84	0
81SUR	10 688 629	26.93	18 300 *(1.07%)*	4 431	13 846	23	0
82SUR	6 644 362	6.48	14 599 *(0.85%)*	14 220	379	0	0
89SUR	7 104 067	6.98	14 322 *(0.83%)*	13 818	504	0	0
135DCM	10 983 591	44.8	27 574 *(1.61%)*	3 119	24 344	111	0
135SUR	10 597 732	28.99	20 978 *(1.22%)*	3 674	17 245	59	0
145SUR	10 894 102	28.8	26 989 *(1.57%)*	2 461	24 462	66	0
149SUR	10 272 274	14.3	16 645 *(0.97%)*	3 340	13 294	11	0
151DCM	10 197 027	15.04	17 967 *(1.05%)*	4 345	13 596	26	0
152SUR	9 172 532	9.35	9 276 *(0.54%)*	7 231	2 044	1	0
155DCM	10 072 636	12.96	16 263 *(0.95%)*	4 255	11 999	9	0
155SUR	10 983 560	34.28	29 959 *(1.75%)*	2 447	27 399	113	0
168DCM	10 084 235	11.53	6 399 *(0.37%)*	5 364	1 035	0	0
196SUR	11 181 235	62.13	34 715 *(2.02%)*	1 877	32 586	251	1
206SUR	10 122 199	19.22	23 435 *(1.37%)*	11 020	12 402	13	0
208SUR	10 803 020	24.33	25 493 *(1.49%)*	2 650	22 793	50	0
209SUR	11 023 554	37.77	21 038 *(1.23%)*	2 915	18 088	35	0
210SUR	10 253 900	14.15	19 307 *(1.13%)*	4 099	15 202	6	0

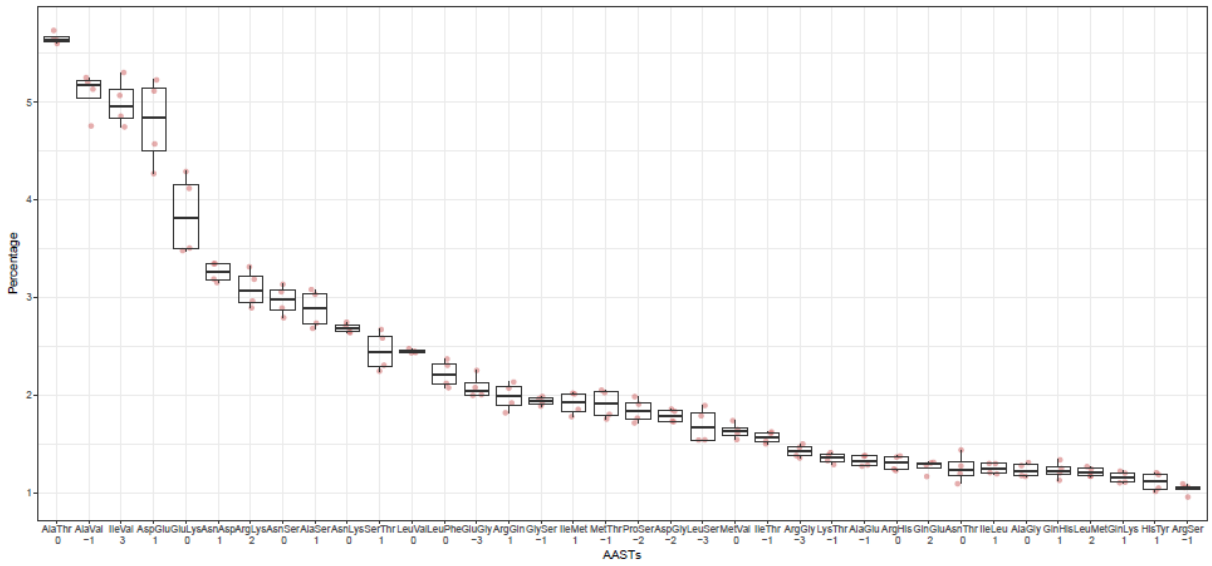
Supplementary Table 2: Number of variants for the 27 samples, total and separation according to the number of alleles found at the loci in each sample.



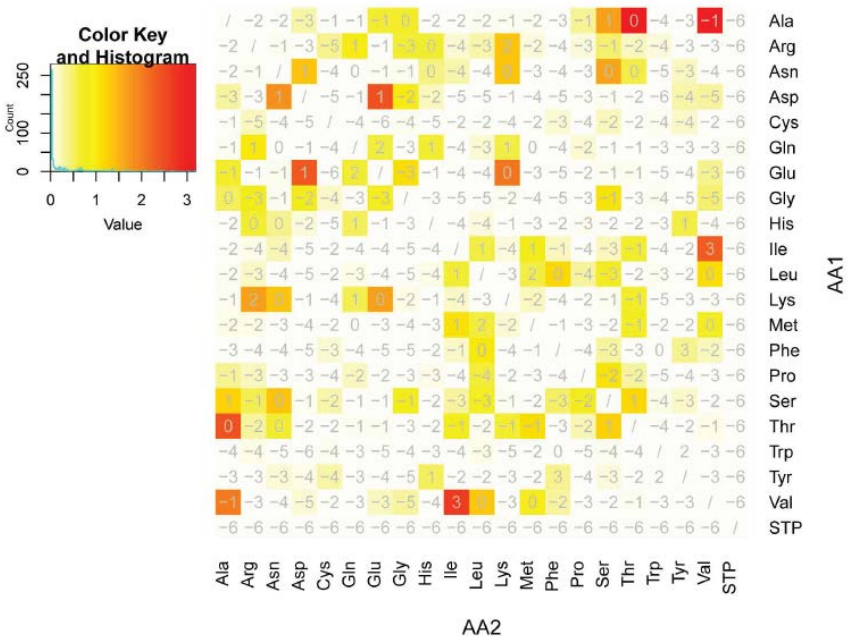
Supplementary Figure 1: Comparison of *Bathycoccus prasinos* RCC105 relative abundance in samples from the 0.8-5µm (X axis) and 0.8-2000µm (Y axis) size fractions in the same stations. The two lines correspond to the identity line (1:1) and to a linear regression from relative abundances values.



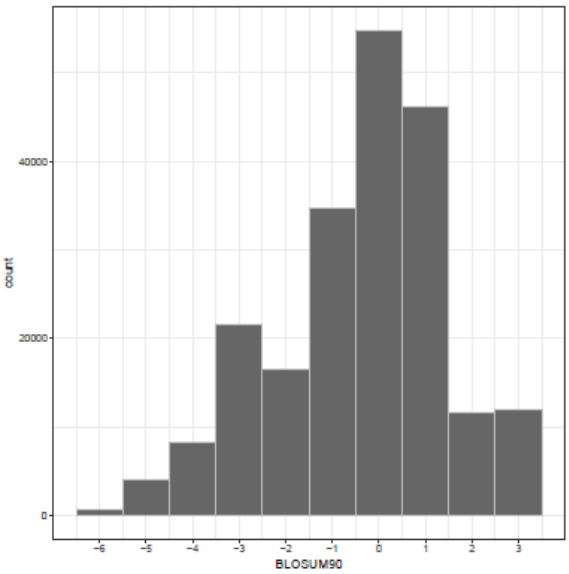
Supplementary Figure 2: Pairwise SNV number per gene in the four best-covered samples.



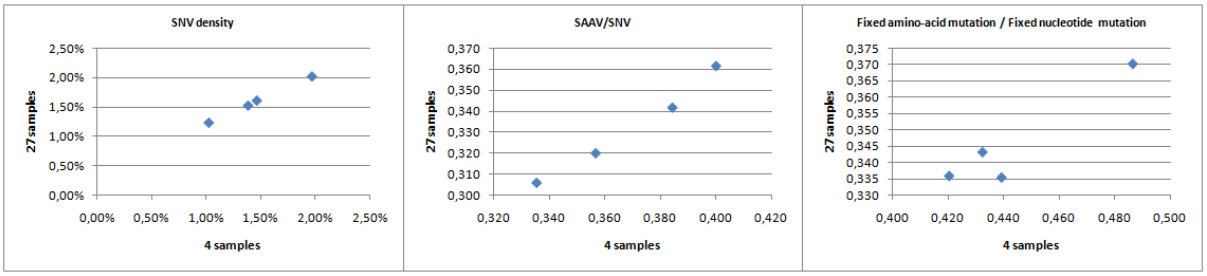
Supplementary Figure 3: Frequencies of AASTs with more than 1% prevalence for the four best-covered samples. Associated BLOSUM90 scores are indicated under the amino-acid pair on the x-axis.



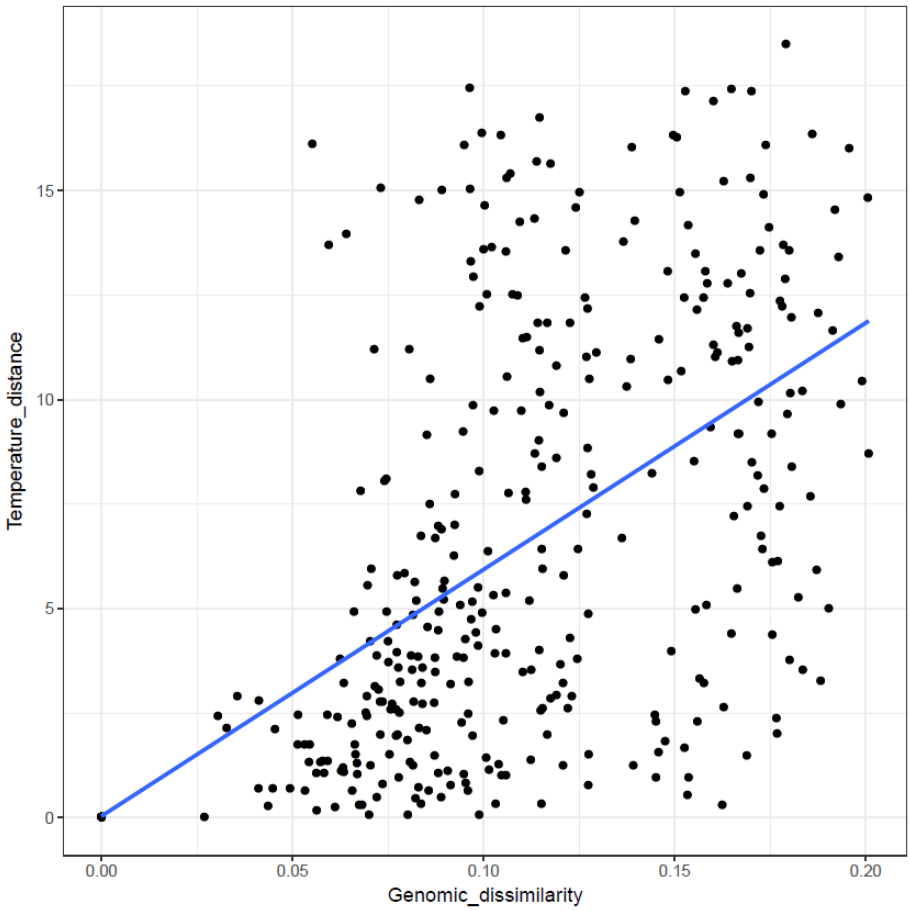
Supplementary Figure 4: Percentage of pairs of amino-acids (AASTs) found in SAAVs. Color corresponds to the average percentage between the 4 best-covered samples, ranging from white (absence) to red (high percentage).



Supplementary Figure 5: BLOSUM90 score distribution including all SAAVs, considering the transition between the two major amino-acids.



Supplementary Figure 6: Comparison of the SNV density, SAAV/SNV ratio, fixed amino-acid variant/fixed nucleotide variant ratio in the four samples when analysing them in the 4-sample set and the 27-sample set.



Supplementary Figure 7: Genomic dissimilarity (1 - genomic distance computed according to Methods) versus delta temperature (difference between two samples) for each pair of samples. The linear regression line is indicated in blue. Mantel test between the two parameters indicates a 0.4818 statistic at p-value=0.001.

III. Conclusion

Partant d'un set d'échantillons de surface et de profondeur métagénomiques dans lesquels *Bathycoccus prasinos* a une couverture moyenne supérieure à 4X, nous avons pu réaliser un appel de variants afin d'étudier la structure de ses populations. Pour cela, nous avons établi un certain nombre de filtres de qualité dans le but d'éviter tout cross-mapping avec des espèces proches, évitant ainsi de confondre des sous-populations potentielles avec par exemple la seconde espèce de *Bathycoccus.*

Dans un premier temps, nous avons vérifié que les résultats obtenus sur nos 27 échantillons d'intérêt étaient cohérent avec un sous-set de quatre échantillons couverts à plus de 30X en moyenne afin d'être certains de la validité de nos analyses. Nous avons ensuite observé le paysage global de la diversité de B. *prasinos*, établissant notamment un nombre de positions variables allant jusqu'à 2% des régions codantes.

Une fois cette étape préliminaire validée, nous avons pu nous intéresser à la structure des populations sur la base d'une distance génomique calculée entre chaque paire de stations, à partir des allèles présents aux positions variables. Ces résultats montrent très clairement une séparation des échantillons en trois groupes distincts : ceux provenant de l'océan Austral, ceux de l'océan Arctique, et les autres, majoritairement des échantillons tempérés.

Les échantillons austraux sont clairement distincts des autres par leur très faible nombre de positions polymorphiques, contre beaucoup plus de positions avec une mutation fixe que les autres échantillons (un seul allèle localement qui est différent des autres échantillons). Nous nous sommes donc ensuite intéressés en particulier aux positions responsables de la plus grande distance entre le cluster arctique et le cluster tempéré afin d'étudier ce qui les différencie.

La distribution des fréquences alléliques à ces 2472 positions statistiquement sélectionnées montre encore une fois un pattern clair entre les eaux arctiques et les autres échantillons. En effet, ces variants sont détectés dans les eaux froides essentiellement biallélіques, tandis que dans les eaux plus chaudes ces mêmes positions

présentent un seul allèle fixé. Il s'agit donc de la différence la plus marquée entre ces deux populations.

Enfin, nous nous sommes intéressés aux variations d'acides aminés afin d'étudier l'impact au niveau protéique des différences entre les eaux chaudes et froides. Pour cela, nous avons sélectionné les variants les plus significatifs, dont les changements d'acides aminés corrèlent au mieux avec la température des échantillons. Nous avons ainsi pu étudier 13 positions correspondant à nos critères statistiques, toutes inclues dans les 2472 SNVs précédents. Mais contrairement au schéma général observé, ces dernières montrent plutôt un seul acide aminé localement fixé dans chacun des environnements (froid, incluant arctique et austral, ou tempéré), avec un mélange de ces deux acides aminés dans les stations de transition à température intermédiaire situées géographiquement entre les deux milieux, ou encore dans un upwelling. Ces positions sont réunies dans 6 gènes aux fonctions variées et cohérentes avec la forte différence de milieu.

Nous avons donc ici présenté une structure des populations de *Bathycoccus*, séparant les biomes tempérés et polaires (arctique et sub-austral) qui présentent quelques similarités. Cependant la majorité des variants ne sont pas significativement corrélés à un paramètre environnemental ou géographique. Le milieu austral est caractérisé par un très faible polymorphisme et un grand nombre de variants fixés, tandis que les populations des régions tempérées et arctiques se séparent par un groupe de variants polymorphiques dans le froid et fixés dans les eaux plus chaudes. Cependant, quelques positions très significatives ont un signal différent, étant fixées avec des acides aminés différents entre les deux environnements.

Chapitre 4 : Analyses de biogéographie au niveau des communautés

I. Application de ces méthodes à d'autres espèces : exemple des Straménopiles

L'étude des Mamiellales à différents niveaux nous a permis une première approche de l'étude de la diversité du phytoplancton et de l'impact de l'environnement sur ce dernier. Néanmoins, de nombreux autres micro-organismes peuplent le milieu marin et un grand nombre d'entre eux ne sont pas cultivables contrairement aux Mamiellales, rendant leur étude d'autant plus complexe.

Nous nous sommes donc ici intéressés à des organismes très différents, une famille de protistes hétérotrophes, les straménopiles, en nous concentrant sur sept lignées prédominantes non-cultivées grâce aux données de séquençage en cellule unique de *Tara* Oceans. Ces résultats sont compilés dans le manuscrit *Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans* publié en 2018 dans *Nature communication* [135] (Annexe 3). Dans le cadre de cette étude, j'ai principalement contribué à l'analyse des données environnementales et leur corrélation avec la répartition de nos espèces d'intérêt, ce qui m'a donné l'occasion de voir si leur schéma de répartition était influencé de la même manière que le phytoplancton étudié jusqu'alors.

Nous avons donc pris en compte sept espèces distinctes et peu étudiées réunies à partir de quarante cellules appartenant à trois différents groupes. Les straménopiles marins du groupe 4 (MAST-4) tout d'abord, sont de petites cellules bactérivores possédant un flagelle, abondants dans les milieux tempérés et tropicaux et possédant plusieurs clades distincts[136,137]. Nous avons ici étudié les trois clades A (dont nous avons deux représentants), C et E. Ensuite, le groupe 3 (MAST-3) est particulièrement divers de petits organismes flagellés[136], dont nous avons pu obtenir des représentants des clades A et F. Enfin les Chrysophytes du clade H, selon des études environnementales[138], apparaissent abondants dans les océans et en font donc un troisième groupe d'intérêt pour lequel nous avons étudié deux génomes distincts. A cela s'ajoute pour certaines analyses un huitième génome partiel de MAST-4 clade D précédemment caractérisé par Single-cell également[139].

Les abondances métagénomiques ont donc été calculées similairement aux analyses précédentes sur les Mamiellales, et la distribution géographique de ces sept SAGs (Single Amplified Genomes) a été établie, confirmant pour la plupart leur abondance dans les océans (Figure 21).

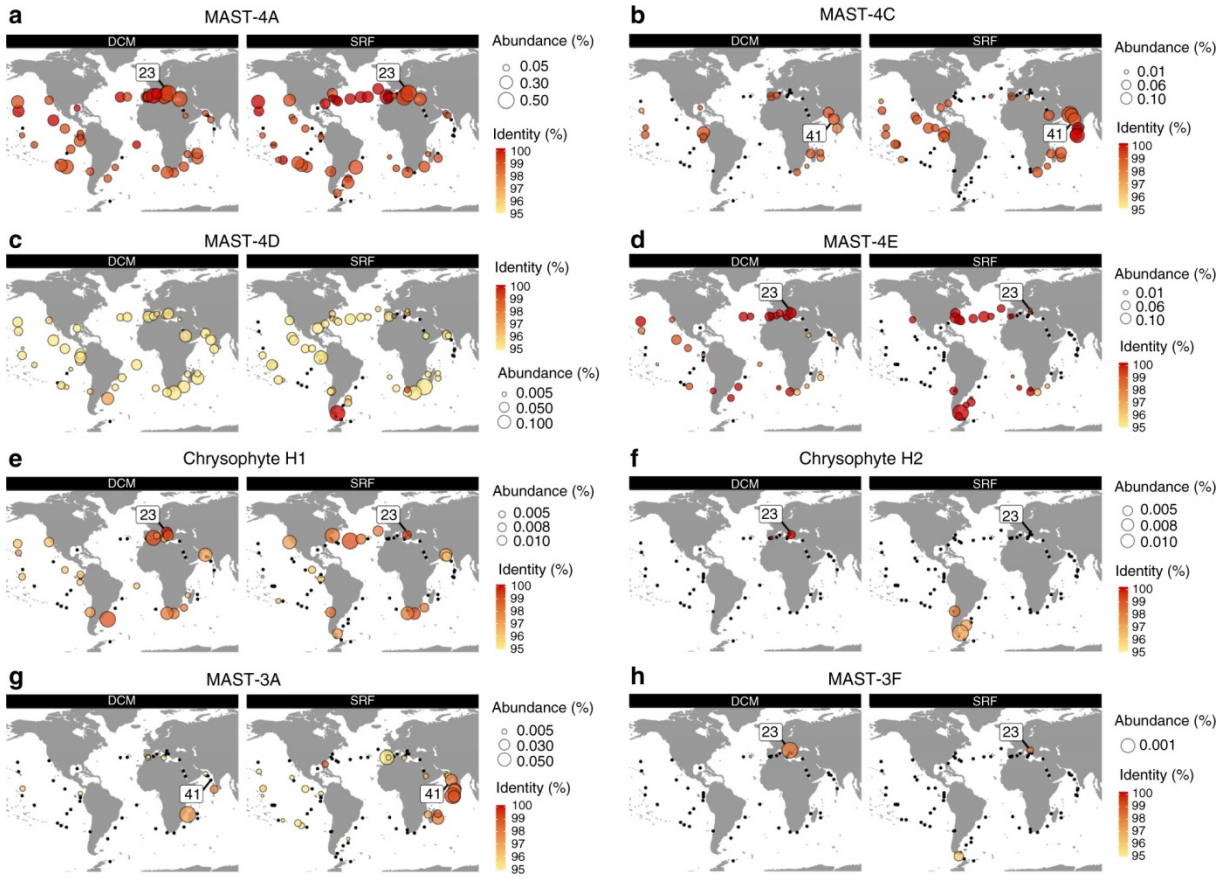


Figure 21 : Distributions géographiques des SAGs séparées selon la profondeur (surface à droite et DCM à gauche) incluant a MAST-4A ; b MAST-4C ; c MAST-4D ; d MAST-4E ; e Chrysophyte H1 ; f Chrysophyte H2 ; g MAST-3A ; et h MAST-3F. Chaque échantillon est représenté par un point noir (pas de signal détecté) ou un cercle de diamètre correspondant à l'abondance relative de l'espèce. La couleur des cercles représente le pourcentage de similarité médian des séquences mappées par rapport à la référence. Le numéro de la station de prélèvement de chaque SAG est indiqué par une étiquette sur chaque carte.

A partir de ces résultats, nous avons pu nous intéresser à la distribution des espèces selon les différents paramètres environnementaux des échantillons en réalisant un test de Kruskal-Wallis pour définir lesquels discriminaient au mieux les sept phyla. Le paramètre le plus significatif au niveau des distributions s'avère de loin être la température de l'eau (p-value = 2.2×10^{-16}), suggérant que certaines de ces lignées

ont probablement des plages de température préférentielles dans lesquelles elles sont les plus abondantes (Figure 22). Des distributions dépendantes de la profondeur sont également fréquentes, MAST-4C et MAST-3A étant typiquement localisés en sub-surface tandis que MAST-4E et Chrysophyte H1 sont essentiellement à la DCM, à part dans les colonnes d'eau très mélangées (Figure 20). Ces résultats montrent bien que ces protistes hétérotrophes ont différentes préférences environnementales au sein d'une même famille, de la même manière que le phytoplancton déjà étudiés.

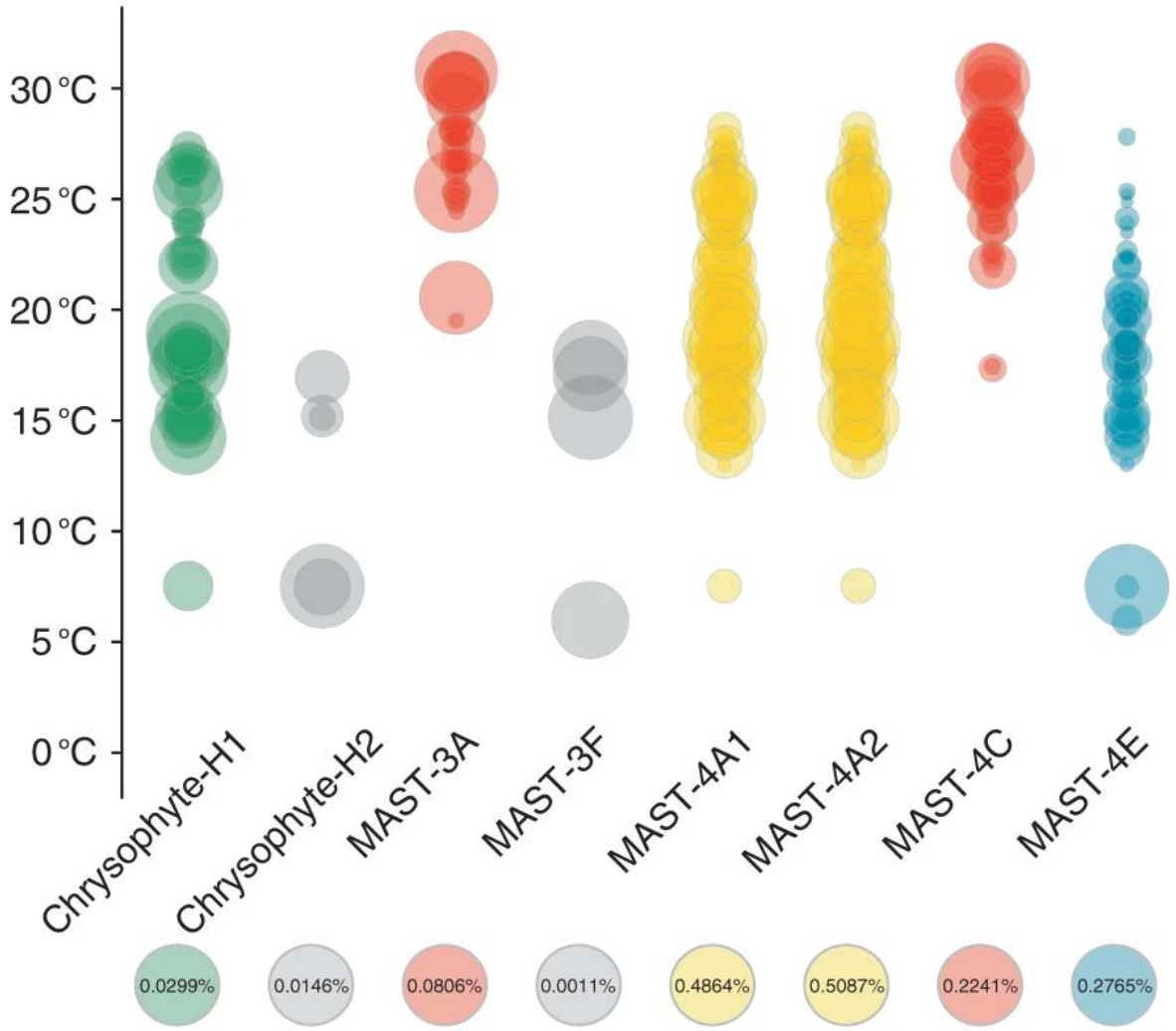


Figure 22 : Distribution de l'abondance des SAGs selon la température des échantillons correspondants. Chaque cercle est un échantillon dont la taille représente l'abondance de l'organisme indiqué sur l'axe horizontal, placé à sa température sur l'axe vertical. L'échelle de chaque colonne est indiquée sous le nom de l'espèce, la taille du cercle correspondant à l'abondance maximum de cette dernière parmi tous les échantillons.

II. Article 4 : Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems

Si nous pouvons observer des schémas de répartition similaires et fortement influencés par les mêmes paramètres environnementaux chez des organismes abondants aussi bien autotrophes qu'hétérotrophes, il est maintenant intéressant de passer à une plus large échelle d'étude, celle des communautés. Il s'agit ici de prendre en compte non pas une famille d'espèces, mais un ensemble d'organismes vivants dans le but d'étudier encore une fois leur répartition et les paramètres qui influent sur cette dernière.

Jusqu'alors, la biogéographie globale de plus grande envergure du plancton avait été réalisée par Longhurst[140], basée essentiellement sur des paramètres associés au plancton photosynthétique. Nous évaluons ici dans un manuscrit collaboratif, disponible depuis 2019 dans *bioRxiv*, la biogéographie mondiale du plancton et sa relation avec le contexte biologique, chimique et physique de l'océan (le "paysage marin"). Pour cela, nous avons utilisé les données de métabarcodes et de métagénomique du projet *Tara* Oceans ainsi que les nombreux paramètres environnementaux qui leur sont associés, utilisant toutes les fractions de taille afin de prendre en compte les communautés planctoniques virales, procaryotes et eucaryotes.

Ma contribution à cette publication a été dans un premier temps une participation au clustering des provinces génomiques basées sur la dissimilarité génomique entre les échantillons, puis majoritairement une analyse des paramètres environnementaux distinguant significativement ces provinces.

La première partie de cette analyse a été le calcul de distances entre chaque paire d'échantillons basé sur leur contenu génomique, et ce pour chaque fraction de taille. A partir de ces distances, nous avons pu réaliser un clustering hiérarchique, et définir statistiquement un seuil permettant de grouper les échantillons en provinces (Figure 1a et Figure supplémentaire 4 du manuscrit).

A partir de ces résultats, nous avons pu appliquer des tests de Kruskal-wallis afin de déterminer si certains facteurs séparaient significativement les différentes provinces d'une fraction de taille, montrant ainsi que la température est toujours le

paramètre dominant, impactant pour la distribution des six fractions étudiées. Parmi les autres facteurs les plus significatifs viennent ensuite les phosphates en moindre mesure pour toutes les fractions également, l'ammonium pour les grandes fractions, les nitrates pour les bactéries, protistes et petits métazoaires, puis la chlorophylle pour les protistes. Ces tests sont accompagnés d'analyses en composantes principales montrant plus en détails la relation entre les provinces et les paramètres physico-chimiques, puis ont été suivis de tests post-hoc de Tukey permettant d'étudier pour les facteurs significatifs les paires de provinces présentant des compositions environnementales significativement différentes. Nous pouvons ainsi voir par exemple que si la province australe est très distincte des autres par sa température et ses nutriments, certaines des autres provinces se distinguent les unes des autres par ces paramètres également (Figure supplémentaire 7 du manuscrit).

Le reste des analyses menées par mes collaborateurs ont notamment permis de montrer que les corrélations entre le temps de transport du plancton et les dissimilitudes métagénomiques révèlent des continuités biologiques et environnementales à l'échelle des bassins. La modulation des communautés planctoniques durant le transport à travers les courants marins varie notamment en fonction de la taille des organismes, de sorte que la répartition des plus petites fractions du plancton correspond le mieux aux provinces biogéochimiques établies par Longhurst, tandis que les plus grandes fractions se regroupent dans les provinces les plus larges. L'ensemble de ces résultats pourrait donc représenter une référence pour l'interprétation de l'organisation des communautés dans leur environnement, ouvrant la voie à une meilleure compréhension du fonctionnement des écosystèmes océaniques.

Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems

Daniel J. Richter[1,2*], Romain Watteaux[3*], Thomas Vannier[4,5*], Jade Leconte[5], Paul Frémont[5], Gabriel Reygondeau[6,7], Nicolas Maillet[8], Nicolas Henry[1], Gaëtan Benoit[9], Antonio Fernàndez-Guerra[10,11,12], Samir Suweis[13], Romain Narci[14], Cédric Berney[1], Damien Eveillard[15,16], Frederick Gavory[17], Lionel Guidi[18,19], Karine Labadie[17], Eric Mahieu[17], Julie Poulain[5], Sarah Romac[1], Simon Roux[20], Céline Dimier[1,21], Stefanie Kandels[22,23], Marc Picheral[24,25], Sarah Searson[24,25], *Tara* Oceans Coordinators, Stéphane Pesant[26,27], Jean-Marc Aury[17], Jennifer R. Brum[20,28], Claire Lemaitre[9], Eric Pelletier[5], Peer Bork[22,29,30], Shinichi Sunagawa[22,31], Lee Karp-Boss[32], Chris Bowler[21], Matthew B. Sullivan[20,33], Eric Karsenti[21,23], Mahendra Mariadassou[14], Ian Probert[1], Pierre Peterlongo[9], Patrick Wincker[5], Colomban de Vargas[1**], Maurizio Ribera d'Alcalà[3**], Daniele Iudicone[3**$], Olivier Jaillon[5**$]

* and $: equal contributions
**: corresponding authors

***Tara* Oceans Coordinators**: Silvia G. Acinas[34], Peer Bork[22,29,30], Emmanuel Boss[32], Chris Bowler[21], Guy Cochrane[35], Colomban de Vargas[1], Gabriel Gorsky[36], Nigel Grimsley[37,38], Lionel Guidi[18,19], Pascal Hingamp[39], Daniele Iudicone[3], Olivier Jaillon[5], Stefanie Kandels[22,23], Lee Karp-Boss[32], Eric Karsenti[21,23], Fabrice Not[1], Hiroyuki Ogata[40], Stéphane Pesant[26,27], Jeroen Raes[41,42], Christian Sardet[18,43], Mike Sieracki[44,45], Sabrina Speich[46,47], Lars Stemmann[18], Matthew B. Sullivan[20,33], Shinichi Sunagawa[22,31], Patrick Wincker[5]

Data availability: http://doi.org/10.6084/m9.figshare.11303177
Supplemental Tables 1-19 (including DDBJ/ENA/GenBank short read archive identifiers for *Tara* Oceans metagenomic & 18S V9 sequence reads, and distance matrices), Datasets 1-3 (18S V9 metabarcode and OTU tables, and reference database).

1 Sorbonne Université, CNRS, Station Biologique de Roscoff, AD2M, UMR 7144, 29680 Roscoff, France
2 Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain
3 Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.
4 Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110 , 13288, Marseille, France
5 Génomique Métabolique, Genoscope, Institut de biologie François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, Evry, France
6 Changing Ocean Research Unit, Institute for the Oceans and Fisheries, University of British Columbia. Aquatic Ecosystems Research Lab. 2202 Main Mall. Vancouver, BC V6T 1Z4. Canada.
7 Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA.
8 Institut Pasteur - Bioinformatics and Biostatistics Hub - C3BI, USR 3756 IP CNRS - Paris, France.
9 INRIA/IRISA, Genscale team, UMR6074 IRISA CNRS/INRIA/Université de Rennes 1, Campus de Beaulieu, 35042, Rennes, France.
10 Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark
11 MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany
12 Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany
13 Dipartimento di Fisica e Astronomia 'G. Galilei' & CNISM, INFN, Università di Padova, Via Marzolo 8, 35131 Padova, Italy.
14 MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France
15 Université de Nantes, Centrale Nantes, CNRS, LS2N, F-44000 Nantes, France
16 Research Federation (FR2022) Tara Oceans GO-SEE, 3 rue Michel-Ange, 75016 Paris, France
17 Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France

18 Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France.
19 Department of Oceanography, University of Hawaii, Honolulu, Hawaii 96822, USA.
20 Department of Microbiology, The Ohio State University, Columbus, OH 43214, USA.
21 Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France.
22 Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany.
23 Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg Germany.
24 Sorbonne Universités, UPMC Univ Paris 06, UMR 7093 LOV, F-75005, Paris, France.
25 CNRS, UMR 7093 LOV, F-75005, Paris, France.
26 MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany.
27 PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany.
28 Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA, 70808, USA
29 Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany.
30 Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany.
31 Institute of Microbiology, Department of Biology, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.
32 School of Marine Sciences, University of Maine, Orono, Maine 04469, USA.
33 Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus OH 43214 USA.
34 Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM), CSIC, Barcelona, Spain.
35 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
36 Sorbonne Universités, CNRS, Laboratoire d'Oceanographie de Villefranche, LOV, F-06230 Villefranche-sur-Mer, France.
37 CNRS, UMR 7232, BIOM, Avenue Pierre Fabre, 66650 Banyuls-sur-Mer, France.
38 Sorbonne Universités Paris 06, OOB UPMC, Avenue Pierre Fabre, 66650 Banyuls-sur-Mer, France.
39 Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110, 13288, Marseille, France
40 Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan.
41 Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.
42 VIB Center for Microbiology, Herestraat 49, 3000 Leuven, Belgium.
43 CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France.
44 National Science Foundation, Arlington, VA 22230, USA.
45 Bigelow Laboratory for Ocean Sciences East Boothbay, ME, USA.
46 Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France.
47 Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France.

Abstract

Biogeographical studies have traditionally focused on readily visible organisms, but recent technological advances are enabling analyses of the large-scale distribution of microscopic organisms, whose biogeographical patterns have long been debated[1,2]. The most prominent global biogeography of marine plankton was derived by Longhurst[3] based on parameters principally associated with photosynthetic plankton. Localized studies of selected plankton taxa or specific organismal sizes[1,4–7] have mapped community structure and begun to assess the roles of environment and ocean current transport in shaping these patterns[2,8]. Here we assess global plankton biogeography and its relation to the biological, chemical and physical context of the ocean (the 'seascape') by analyzing 24 terabases of metagenomic sequence data and 739 million metabarcodes from the *Tara* Oceans expedition in light of environmental data and simulated ocean current transport. In addition to significant local heterogeneity, viral, prokaryotic and eukaryotic plankton communities all display near steady-state, large-scale, size-dependent biogeographical patterns. Correlation analyses between plankton transport time and metagenomic or environmental dissimilarity reveal the existence of basin-scale biological and environmental continua emerging within the main current systems. Across oceans, there is a measurable, continuous change within communities and environmental factors up to an average of 1.5 years of travel time. Modulation of plankton communities during transport varies with organismal size, such that the distribution of smaller plankton best matches Longhurst biogeochemical provinces, whereas larger plankton group into larger provinces. Together these findings provide an

integrated framework to interpret plankton community organization in its physico-chemical context, paving the way to a better understanding of oceanic ecosystem functioning in a changing global environment.

Main Text

Plankton communities are constantly on the move, transported by ocean currents[9]. Transport involves both advection and mixing. While being advected by currents, plankton are influenced by multiple processes, both physico-chemical (fluxes of heat, light and nutrients[10]) and biological (species interactions, life cycles, behavior, acclimation/adaptation[11,12]), which act across various spatio-temporal scales. In turn, plankton impact seawater physico-chemistry while they are being advected[10]. The community composition and biogeochemical properties of a water mass are also partially dependent on its history of mixing with neighboring water masses during transport. These intertwined processes form the pelagic seascape[13] (Supplementary Fig. 1a). Previous studies on plankton distribution have tended to focus on individual factors, such as nutrient or light availability[3,14], or have investigated the role of transport for specific nutrients[15] or types of planktonic organisms[8,16]. Here, instead, we integrated uniformly collected metagenomic data across multiple size fractions with large-scale ocean circulation simulations in the context of the seascape.

We assessed global patterns of plankton biogeography in the context of the seascape using samples collected at 113 stations during the *Tara* Oceans expedition[17], including DNA sequence data from six organismal size fractions: one virus-enriched (0-0.22 µm)[5], one prokaryote-enriched (either 0.22-1.6 or 0.22-3 µm)[18], and four eukaryote-enriched (0.8-5 µm, 5-20 µm, 20-180 µm and 180-2000 µm[19]; Supplementary Fig. 1b). We analyzed 24.2 terabases of metagenomic sequence reads and 320 million new eukaryotic 18S V9 ribosomal DNA marker sequences (Supplementary Table 1), complementing previously described *Tara* Oceans data[5,18,19]. We used metagenomic data and Operational Taxonomic Units (OTUs, representing groups of genetically related organisms) as independent proxies to compute pairwise comparisons of plankton community dissimilarity (β-diversity). Metagenomic dissimilarity highlighted, at species and sub-species resolution, differences in the genomic identity of organisms between stations. Our metagenomic sampling resulted in pairwise metagenomic dissimilarities that likely represent an overestimate of true β-diversity (Supplementary Information 1). However, since we applied an identical procedure to compute dissimilarity between all pairs of samples, these values nevertheless provide an accurate picture of β-diversity variation among samples. The more deeply sampled OTU dissimilarity, in contrast, incorporated the numerous rare taxa within the plankton, but at genus or higher-level taxonomic resolution[19]. Metagenomic and OTU dissimilarities were correlated for all size fractions (Spearman's ρ 0.53 to 0.97, $p \leq 10^{-4}$, Supplementary Fig. 2), indicating that both proxies, although characterized by different sampling depth and taxonomic resolution, provided coherent and complementary estimates of β-diversity (Supplementary Information 1). We performed subsequent analyses using both measures, which produced consistent results. We focus on analyses of metagenomic dissimilarity here, with accompanying results for OTU dissimilarity presented in Supplementary Figures.

Globally, we observed significant dissimilarities at both the metagenomic and OTU level between sampled stations (including adjacent sites) across all size fractions (Supplementary Fig. 3a, Supplementary Information 1). The resulting portrait is of a locally heterogeneous oceanic ecosystem dominated by a small number of abundant and cosmopolitan taxa, with a much larger number of less abundant taxa found at fewer sampling sites (Supplementary Fig. 3b-e), corroborating previous studies[19].

Underlying this local heterogeneity, we found robust evidence for the existence of large-scale biogeographical patterns within all plankton size classes using two complementary analyses of dissimilarity among samples (Fig. 1a, Supplementary Fig. 4a-f, Supplementary Fig. 5, Supplementary

Information 2). First, we grouped metagenomic samples within each size fraction into 'genomic provinces' via hierarchical clustering (Supplementary Fig. 6). Second, we derived colors for each sample based on a principal coordinates analysis (PCoA-RGB; see Methods) in order to visualize transitions in community composition within and between genomic provinces. Most genomic provinces were composed of large-scale geographically contiguous stations (consistent with previous studies documenting patterns in plankton biogeography[1,2,5,6]) with some independent distant samples (Fig. 1a, Supplementary Fig. 4a-f). Genomic provinces of smaller plankton (viruses, bacteria and eukaryotes <20 µm) tended to be limited to a single ocean basin and to approximately correspond to Longhurst biogeochemical provinces[3] (Supplementary Fig. 4a-d; Supplementary Information 3). In contrast, provinces of larger plankton (micro- and meso-plankton, >20 µm) spanned multiple basins (Supplementary Fig. 4e-f, Supplementary Information 4).

These large-scale biogeographical patterns derived from metagenomes were linked to environmental parameters including nutrients, temperature and trophic level. Seawater temperature was significantly different among genomic provinces for all plankton size classes (Kruskal-Wallis test, $p < 10^{-5}$), corroborating previous results for prokaryotes[18], whereas other environmental conditions were significantly different only with respect to specific size classes (Supplementary Fig. 7). The geography of combined nutrient and temperature variations resembled the biogeography of smaller plankton size classes (Fig. 1a-b, Supplementary Fig. 4a-d,g), whereas temperature alone more closely matched the distribution of larger plankton (Supplementary Fig. 4e,f,h), reflecting different potential ecological constraints. Many genomic provinces were spatially consistent with ocean basin-scale circulation patterns, such as western boundary currents or major subtropical gyres[20] (Fig. 1a, Supplementary Fig. 4a-f), suggesting a particular role for large-scale surface transport (a core component of the seascape) in the emergence of spatial patterns of plankton community composition, as previously proposed[21]. We therefore investigated community composition differences between sampled stations in light of the corresponding transit time. We inferred the time of mean transport between stations from trajectories computed with the physically well-constrained MITgcm ocean model (see Methods), which takes into account directionalities[9] and meso- to large-scale circulation, potential dispersal barriers and mixing effects[22,23]. We quantified transport using the minimum travel time[24] (T_{min}) between pairs of *Tara* stations. These trajectories corresponded to the dominant paths that transport the majority of water volume and its contents (e.g., heat, nutrients and plankton; Fig. 1c). For all plankton size classes, community composition differences between stations were correlated to travel time (Supplementary Fig. 8). Cumulative correlation values (correlations between metagenomic dissimilarity and T_{min} computed for an increasing range of T_{min}) were maximal for pairs of stations separated by T_{min} <~1.5 years for all size classes ($p \leq 10^{-4}$; Spearman's ρ 0.45 to 0.71 depending on size class, Fig. 2a, Supplementary Fig. 9a-e), hence revealing measurable plankton community dynamics on time scales far longer than typical plankton growth rates or life cycles. In contrast, no such unimodal pattern was found for correlations between metagenomic dissimilarity and geographic distance (without traversing land; Supplementary Fig. 9f). Over the timescale <~1.5 years, which corresponds well with the average time to travel across a basin or gyre, large-scale transport is therefore an appropriate framework for studying differences in plankton community composition (Fig. 2b). The fact that simulated transport times and metagenomic dissimilarity were correlated despite a 3 year pan-season sampling campaign highlights the overall stability of plankton dynamics along the main ocean currents.

Transit time also covaried (although less strongly) with differences in environmental conditions for pairs of stations for which T_{min} <~1.5 years (Fig. 3). This indicates that along large-scale oceanic current systems, changes in environmental conditions and plankton community composition are concurrent. In our data, beyond ~1.5 years of transport, correlations of T_{min} with metagenomic dissimilarity decreased (Fig. 2a, Fig. 3, Supplementary Fig. 9a-e), meaning the signature of transport in generating large-scale diversity changes weakened and travel time therefore becomes a less appropriate framework to study β-diversity. A similar trend was observed for the correlation between T_{min} and

nutrient concentrations whereas temperature was better correlated when considering larger transit times (Fig. 3).

Together, these analyses suggest the existence in the seascape of stable biogeochemical continua induced by basin-scale currents with predictable, interlinked changes in environmental conditions and plankton community composition (Supplementary Information 5). It has previously been posited that transport could generate continuous transitions between niches[25], but it was not anticipated that this would occur on the scale of ocean basins. Beyond ~1.5 years, the correlation of metagenomic dissimilarity with differences in temperature increased while that with differences in nutrients decreased (Fig. 3, Supplementary Fig. 9a-e). However, both of these correlations with metagenomic dissimilarity remained strong on these time scales. This might be related to distant *Tara* Oceans stations experiencing similar oceanographic phenomena (notably temperature), for example upwelling zones, producing generally similar environmental conditions.

The existence of a size-class dependent (smaller or larger than 20 µm) plankton biogeography indicates that organisms contribute differently to the basin-scale biogeochemical continua present in the seascape. In the case of the North Atlantic current system (including the Mediterranean Sea), a simple exponential fit of metagenomic dissimilarity along T_{min} for T_{min} <~1.5 years (Fig. 2c) revealed that the smaller size classes (<20 µm) had a shorter metagenomic turnover time (ca. 1y) than larger plankton (ca. 2y) (Supplementary Fig. 10, Supplementary Information 6). At global geographical scales, the genomic provinces of small size classes, which are enriched in phytoplankton[18,19], corresponded with differences in environmental parameters such as nutrient levels (Fig. 1b, Supplementary Fig. 7) that are often constrained by regional oceanographic processes[26], as shown in our data. On the other hand, genomic provinces of larger plankton, dominated by heterotrophic and symbiotic organisms[19], often crossed biogeochemical boundaries and were more related to global scale gradients and circulation patterns, notably major latitudinal temperature zones or the separation between Atlantic and Indo-Pacific large-scale surface circulations (Supplementary Fig. 4e,f,h). These divergent effects were also evident in comparisons of metagenomic dissimilarity with variations in environmental conditions (Supplementary Fig. 9b). For smaller plankton, correlations with differences in nutrient concentrations were stronger for T_{min} up to ~1.5 years, but for larger plankton, correlations were stronger with temperature variations for T_{min} beyond ~1.5 years. These results indicate a significant size-based decoupling within planktonic food webs (see Supplementary Information 4).

In this study, we provide genomic evidence for an organism-size-dependent global plankton biogeography shaped by currents at the scale of ocean basins. We measured, using metagenomes, the underlying plankton dynamics driven by seascape processes such as intrinsic biological dynamics, variation in environmental conditions, and/or long-range transport. Our analyses reveal that global plankton communities include components that are in a near steady-state that emerges from the integration of the seascape. This behavior resembles self-organizing systems within reaction-advection-diffusion contexts[27]. This work shows that studies of the dynamics of plankton communities must consider the critical influence of ocean currents in stretching and altering, on the scale of basins, the distribution of both planktonic organisms and the physico-chemical nature of the water mass in which they reside. In this context, our study confirms that the combination of ocean circulation modelling with the use of metagenomic DNA as a tracer of plankton communities is a key tool for unravelling the regulation of plankton dynamics. The planktonic ecosystem is fundamentally different in many ways from other major planetary ecosystems and this study provides a framework to understand and predict the structuring of the ocean ecosystem in a scenario of rapid environmental and current system changes[28,29].

Methods

Sampling, sequencing and environmental parameters

Sampling, size fractionation, measurement of environmental parameters and associated metadata, DNA extraction and metagenomic sequencing were conducted as described previously[30,31]. Samples were collected at 113 *Tara* Oceans stations for six size fractions (0-0.2, 0.22-1.6/3, 0.8-5, 5-20, 20-180, 180-2000 µm; Supplementary Fig. 1b; Supplementary Table 1) and two depths (subsurface and deep chlorophyll maximum (DCM)). The prokaryote-enriched size fraction was collected either a 0.22-1.6 µm or 0.22-3 µm filter[18,30].

We used physico-chemical data measured *in situ* during the *Tara* Oceans expedition (depth of sampling, temperature, chlorophyll, phosphate, nitrate and nitrite concentrations), supplemented with simulated values for iron and ammonium (using the MITgcm Darwin model described below in "Ocean circulation simulations"), day length, and 8-day averages calculated for photosynthetically active radiation (PAR) in surface waters (AMODIS, https://modis.gsfc.nasa.gov). In order to obtain PAR values at the deep chlorophyll maximum, we used the following formula[32]:

$$PAR(Z) = PAR(0)*\exp(-k*Z)$$

$$x=\log(Chl)$$

$$\log(Z)=1.524-0.426x-0.0145x^2+0.0186x^3$$

$$k=-\ln(0.01)/Z$$

in which k is the attenuation coefficient, and Z is the depth of the DCM (in meters). Other data, such as silicate and the nitrate/phosphate ratio, were extracted from the World Ocean Atlas 2013 (WOA13 version 2, https://www.nodc.noaa.gov/OC5/woa13/), by retrieving the annual mean values at the closest available geographical coordinates and depths to *Tara* sampling stations. For temperature and nitrate, we calculated seasonality indexes (SI) from monthly WOA13 data. For each sample, the index is the annual variation of the parameter (max - min) at this location divided by the highest variation value among all samples.

A list of samples, metagenomic and metabarcode sequencing information and associated environmental data is available in Supplementary Tables 1-2.

Calculation of metagenomic community dissimilarity

Metagenomic community distance between pairs of samples was estimated using whole shotgun metagenomes for all six size fractions. We used a metagenomic comparison method (Simka[33]) that computes standard ecological distances by replacing species counts by counts of DNA sequence k-mers (segments of length *k*). K-mers of 31 base pairs (bp) derived from the first 100 million reads sequenced in each sample (or the first 30 million reads for the 0-0.2 µm size fraction) were used to compute a similarity measure between all pairs of samples within each organismal size fraction. Based on a benchmark of Simka, we selected 100 million reads per sample (or 30 million for the 0-0.2 µm fraction) because increasing this number did not produce a qualitatively different set of results, and to ensure that the same number of reads were used in each pairwise comparison within a size fraction. Nearly all samples in our data set had at least 100 million reads (or at least 30 million for the 0-0.2 µm fraction; Supplementary Table 1).

We estimated β-diversity for metagenomic reads with the following equation within Simka:

$$\text{Metagenomic } \beta\text{-diversity} = (b + c) / (2a + b + c)$$

Where a is the number of distinct k-mers shared between two samples, and b and c are the number of distinct k-mers specific to each sample. We represented the distance between each pair of samples on a heatmap using the heatmap.2 function of the R-package[34] gplots_2.17.0[35]. The dissimilarity matrices we produced for each plankton size fraction (on a scale of 0 = identical to 100 = completely dissimilar) are available as Supplementary Tables 3-8.

Calculation of OTU-based community dissimilarity

Within the 0-0.2 µm size fraction, we used previously published viral populations (equivalent to OTUs)[36] and viral clusters (analogous to higher taxonomic levels)[5] based on clustering of protein content. For the 0.22-1.6/3 µm size fraction, we used previously derived miTAGs based on metagenomic matches to 16S ribosomal DNA loci and processed them as described[18]. For the four

eukaryotic size fractions, we added additional samples to a previously published *Tara* Oceans metabarcoding data set and processed them using the same methods[19] (also described at DOI: 10.5281/zenodo.15600).

We calculated OTU-based community dissimilarity for all size fractions as the Jaccard index based on presence/absence data using the vegdist function implemented in vegan 2.4-0[37] in the software package R. The dissimilarity matrices we produced for each plankton size fraction (on a scale of 0 = identical to 100 = completely dissimilar) are available as Supplementary Tables 9-14.

Calculating distances of environmental parameters

We calculated Euclidean distances[38] for physico-chemical parameters. Each were scaled individually to have a mean of 0 and a variance of 1 and thus to contribute equally to the distances. Then the Euclidean distance between two stations i and j for parameters P was computed as follows:

$$ED(i,j,P) = \sqrt{\sum_{p \in P} \left(x_{ip} - x_{jp}\right)^2}$$

RGB encoding of environmental positions

We color-coded the position of stations in environmental space for Fig. 1b and Supplementary Fig. 4g as follows. First, environmental variables were power-transformed using the Box-Cox transformation to have Gaussian-like distributions to mitigate the effect of outliers and scaled to have zero mean and unit variance. We then performed a principal component analysis (PCA) with the R command prcomp from the package stats 3.2.1[34] on the matrix of transformed environmental variables and kept only the first 3 principal components. Finally, we rescaled the scores in each component to have unit variance and decorrelated them using the Mahalanobis transformation. Each component was mapped to a color channel (red, green or blue) and the channels were combined to attribute a single composite color to each station. The components (x, y, z) were mapped to color channel values (r, g, b) between 0 and 255 as r = 128 * (1 + x / max(abs(x)), and similarly for g and b. This map ensures that the global dispersion is equally distributed across the three components and composite colors span the whole color space.

Definition of genomic provinces

We used a hierarchical clustering method on the metagenomic pairwise dissimilarities produced by Simka for all surface and DCM samples, and multiscale bootstrap resampling for assessing the uncertainty in hierarchical cluster analysis. We focused on metagenomic dissimilarity due to its higher resolution, and confirmed that the patterns found in metagenomic data were consistent when using OTU data (Supplementary Fig. 5). We used UPGMA (Unweighted Pair-Group Method using Arithmetic averages) clustering, as it has been shown to have the best performance to describe clustering of regions for organismal biogeography[39]. The R-package pvclust_1.3-2[40], with average linkage clustering and 1,000 bootstrap replications, was used to construct dendrograms with the approximately unbiased p-value for each cluster (Supplementary Fig. 6). Because the number of genomic provinces by size fraction was not known *a priori*, we applied a combination of visualization and statistical methods to compare and determine the consistency within clusters of samples. First, the silhouette method[41] was used to measure how similar a sample was within its own cluster compared to other clusters using the R package cluster_2.0.1[42]. The Silhouette Coefficient *s* for a single sample is given as:

$$s = (b - a) / \max(a,b)$$

Where *a* is the mean distance between *a* sample and all other points in the same class and *b* is the mean distance between *a* sample and all other points in the next nearest cluster. We used the value of *s*, in addition to bootstrap values, to partition each tree into genomic provinces (see Supplementary Information 2 for further details on statistical validation of genomic provinces). Additionally, we used the Radial Reingold-Tilford Tree representation from the JavaScript library D3.js (https://d3js.org/)[43]

to visualize sample partitions from the dendrogram. Single samples were not considered as genomic provinces.

In a complementary approach, we performed a principal coordinates analysis (PCoA) with the R command cmdscale (eig = TRUE, add = TRUE) from the package stats 3.2.1[34] on the matrices of pairwise metagenomic dissimilarities calculated by Simka (or OTU dissimilarity measured with the Jaccard index) within each size fraction and kept only the first 3 principal coordinates. We then converted those coordinates to a color using the RGB encoding described above, with one modification: scaling factors λ_r, λ_g and λ_b were calculated as the ratios of the second and third eigenvalues to the first (dominant) eigenvalue to ensure that the dispersion of stations along each color channel reproduced the dispersion of the stations along the corresponding principal component (the ratio for the color corresponding to the dominant eigenvalue is 1). The components (x, y, z) were then mapped to color channel values (r, g, b) between 0 and 255 as $r = 128 * (1 + \lambda_c x / \max(\text{abs}(x))$, where λ_c is the ratio of the eigenvalue of color c to the dominant eigenvalue.

We represented number and PCoA-RGB color of genomic provinces for each sample on a world map (Fig. 1, Supplementary Fig. 4a-f) generated with the R packages maps_3.0.0.2[44], mapproj 1.2-4[45], gplots_2.17.0[35] and mapplots_1.5[46]. We also plotted phosphate and temperature (Supplementary Fig. 4a-f) obtained from the *Csiro Atlas of Regional Seas* (CARS2009, http://www.cmar.csiro.au/cars) using the phosphate_cars2009.nc and temprerature_cars2009a.nc files and the R package RNetCDF[47].

Comparison of genomic provinces to previous ocean divisions

To evaluate the spatial similarity between the clusters obtained in our study for each size fraction and previous biogeographic divisions, we performed an analysis of similarity (ANOSIM, Fathom toolbox, matlab®). First, we collected coordinates for three spatial divisions at a resolution of 0.5° x 0.5°: biomes, biogeochemical provinces (BGCPs)[3,48] and objective global ocean biogeographic provinces (OGOBPs)[49]. Second, we assigned *Tara* Oceans stations to biomes, BGCPs, and OGOBPs based on their GPS coordinates. Third, for each size fraction we performed an ANOSIM with the metagenomic dissimilarity matrix calculated by Simka, using biogeographic clusters (biome, BGCP, OGOBP) as group membership for each station. Each ANOSIM was bootstrapped 1,000 times to evaluate the interval of confidence around the strength of the relationships we detected (Supplementary Fig. 4a-f).

Environmental differences among genomic provinces

For each size fraction, we tested which environmental parameters significantly discriminated among genomic provinces (Supplementary Fig. 7). A total of 12 parameters characterizing each sample, grouped by genomic provinces, were evaluated with a Kruskal-Wallis test within each size fraction with a significance threshold of $p < 10^{-5}$. Selected parameters for each size fraction were then used to perform a principal components analysis of the samples using the R package vegan_1.17-11[37]. Samples were plotted with the same PCoA-RGB colors used in the genomic province maps above and each genomic province surrounded by a grey polygon. In analyses where Southern Ocean (including Antarctic) stations were considered independently from other stations, the following were considered Southern Ocean stations: 82, 83, 84, 85, 86, 87, 88, 89.

Ocean circulation simulations

We derived travel times from the MITgcm Darwin simulation[50] based on an optimized global ocean circulation model from the ECCO2 group[51]. The horizontal resolution of the model was approximately 18 km, with 1,103,735 total ocean cells. We ran the model for six continuous years in order to smooth anomalies that might occur during any single year. We used surface velocity simulation data to compute trajectories of floats originating in ocean cells containing all *Tara* Oceans stations, and applied the following stitching procedure to generate a large number of trajectories for each initial position. (The use of surface velocity data implies that Ekman transport also influences trajectories within the simulation.)

First, we precomputed a set of monthly trajectories: for each of the 72 months in the dataset, we released floats in every ocean cell of the model grid and simulated transport for one month. We used a fourth-order Runge-Kutta method with trilinearly interpolated velocities and a diffusion of 100 m^2/s. Second, following previous studies[4], we stitched together monthly trajectories to create 10,000 year trajectories: for each float released within a 200 km radius of a *Tara* station, we constructed 1,000 trajectories, each 10,000 years long. To avoid seasonal effects, we began by selecting a random starting month. We followed the trajectory of a float released within that month to the grid cell containing its end point at the end of the month. Next, we randomly selected a trajectory starting on the following month (e.g., February would follow January) from that grid cell, and repeated until reaching a 10,000 year trajectory.

We searched the resulting 50.8 million trajectories for those that connected pairs of *Tara* Oceans stations. To ensure robustness of our results, we only included pairs of stations that were connected by more than 1,000 trajectories. For each pair of stations, T_{min} was defined as the minimum travel time of all trajectories (if any) connecting the two stations. The travel time matrix we produced (measured in years) is available as Supplementary Table 15. Standard minimum geographic distance without traversing land[52] is available as Supplementary Table 16.

Correlations of β-diversity, T_{min} and environmental parameters

We excluded stations that were not from open ocean locations from correlation analyses to avoid sites impacted by coastal processes (those numbered 54, 61, 62, 79, 113, 114, 115, 116, 117, 118, 119, 120, and 121). In analyses where Southern Ocean (including Antarctic) stations were considered independently from other stations, the following were considered Southern Ocean (including Antarctic) stations: 82, 83, 84, 85, 86, 87, 88, 89. We calculated rank-based Spearman correlations between β-diversity, T_{min} and environmental parameters (either differences in temperature or the Euclidean distance composed of differences in NO_2NO_3, PO_4 and Fe, see above) for surface samples with a Mantel test with 1,000 permutations and a nominal significance threshold of $p < 0.01$. For the correlations presented in Fig. 2a, Fig. 3 and Supplementary Fig. 9 correlation values were derived from pairs of stations connected by T_{min} up to the value on the x-axis. We calculated partial correlations of metagenomic and OTU dissimilarity and T_{min} by controlling for differences in temperature and for differences in nutrient concentrations, and partial correlations of dissimilarity with temperature or nutrient variation by controlling for T_{min}.

Community turnover in the North Atlantic

Tara Oceans stations numbered 72, 76, 142, 143, 144, and all stations from 146 to 151 were located along the main current system connecting South Atlantic and North Atlantic oceans and continuing to the strait of Gibraltar. In addition, we included stations 4, 7, 18, and 30 located on the main current system in the Mediterranean Sea (Supplementary Fig. 10). As the *Tara* Oceans samples within the subtropical gyre of the North Atlantic and in the Mediterranean Sea were all collected in winter, seasonal variations should not play a role in the variability in community composition that we observed (see Supplementary Table 2). We calculated genomic e-folding times (the time after which the detected genomic similarity between plankton communities changes by 63%) over scales from months to years based on an exponential fit of metagenomic dissimilarity to T_{min} with the form $y = C_0 e^{-x/\tau}$ (where C_0 is a constant and τ the folding time). Exponential fits for size fractions 0-0.2 µm and 5-20 µm were not calculated due to an insufficient number of sampled stations in the North Atlantic (Supplementary Information 6).

The synthetic map (Supplementary Fig. 10a) was generated with the R packages maps_3.0.0.2, mapproj 1.2.4, gplots_2.17.0 and mapplots_1.5. We derived dynamic sea surface height from the *Csiro Atlas of Regional Seas* (CARS2009, http://www.cmar.csiro.au/cars) using the hgt2000_cars2009a.nc file and plotted with the R package RNetCDF.

References

1. Martiny, J. B. H. et al. Microbial biogeography: putting microorganisms on the map. Nat. Rev. Microbiol. 4, 102–112 (2006).
2. Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C. & Martiny, J. B. H. Beyond biogeographic patterns: processes shaping the microbial landscape. Nat. Rev. Microbiol. (2012). doi:10.1038/nrmicro2795
3. Longhurst, A. Ecological Geography of the Sea. (Academic Press, 2006).
4. Hellweger, F. L., van Sebille, E. & Fredrick, N. D. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. Science 345, 1346–1349 (2014).
5. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537, 689–693 (2016).
6. McGowan, J. A. & Walker, P. W. Structure in the Copepod Community of the North Pacific Central Gyre. Ecol. Monogr. 49, 195–226 (1979).
7. Reygondeau, G. & Dunn, D. Pelagic Biogeography. in Encyclopedia of Ocean Sciences 588–598 (Elsevier, 2019). doi:10.1016/B978-0-12-409548-9.11633-1
8. Villarino, E. et al. Large-scale ocean connectivity and planktonic body size. Nat. Commun. 9, 142 (2018).
9. Watson, J. R. et al. Realized and potential larval connectivity in the Southern California Bight. Mar. Ecol. Prog. Ser. 401, 31–48 (2010).
10. Moore, C. M. et al. Processes and patterns of oceanic nutrient limitation. Nat. Geosci. 6, 701–710 (2013).
11. Flynn, K. J. et al. Acclimation, adaptation, traits and trade-offs in plankton functional type models: reconciling terminology for biology and modelling. J. Plankton Res. 37, 683–691 (2015).
12. Armbrust, E. V. The life of diatoms in the world's oceans. Nature 459, 185–192 (2009).
13. Pittman, S.J. (ed.). Seascape Ecology. (Wiley-Blackwell, 2017).
14. Tagliabue, A. et al. The integral role of iron in ocean biogeochemistry. Nature 543, 51–59 (2017).
15. Letscher, R. T., Primeau, F. & Moore, J. K. Nutrient budgets in the subtropical ocean gyres dominated by lateral transport. Nat. Geosci. 9, 815–819 (2016).
16. Wilkins, D., van Sebille, E., Rintoul, S. R., Lauro, F. M. & Cavicchioli, R. Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. Nat. Commun. 4, 2457 (2013).
17. Karsenti, E. et al. A Holistic Approach to Marine Eco-Systems Biology. PLoS Biol. 9, e1001177 (2011).
18. Sunagawa, S. et al. Structure and function of the global ocean microbiome. Science 348, 1261359 (2015).
19. de Vargas, C. et al. Eukaryotic plankton diversity in the sunlit ocean. Science 348, 1261605–1261605 (2015).
20. Talley, L. D., Pickard, G. L., Emery, W. J. & Swift, J. H. Descriptive Physical Oceanography: An Introduction. (Elsevier, 2011).
21. Clayton, S., Dutkiewicz, S., Jahn, O. & Follows, M. J. Dispersal, eddies, and the diversity of marine phytoplankton. Limnol. Oceanogr. Fluids Environ. 3, 182–197 (2013).
22. Goetze, E. et al. Ecological dispersal barrier across the equatorial Atlantic in a migratory planktonic copepod. Prog. Oceanogr. (2016). doi:10.1016/j.pocean.2016.07.001
23. Mousing, E. A., Richardson, K., Bendtsen, J., Cetinić, I. & Perry, M. J. Evidence of small-scale spatial structuring of phytoplankton alpha- and beta-diversity in the open ocean. J. Ecol. 104, 1682–1695 (2016).
24. Jönsson, B. F. & Watson, J. R. The timescales of global surface-ocean connectivity. Nat. Commun. 7, 11239 (2016).
25. Lévy, M., Jahn, O., Dutkiewicz, S. & Follows, M. J. Phytoplankton diversity and community structure affected by oceanic dispersal and mesoscale turbulence. Limnol. Oceanogr. Fluids Environ. 4, 67–84 (2014).
26. Sarmiento, J. L. & Gruber, N. Ocean Biogeochemical Dynamics. (Princeton University Press, 2006).
27. Feudel, U. Pattern Formation in Marine Systems. in Complexity and Synergetics 179–196 (Springer International Publishing, 2018). doi:10.1007/978-3-319-64334-2_15
28. Beaugrand, G. Reorganization of North Atlantic Marine Copepod Biodiversity and Climate. Science 296, 1692–1694 (2002).
29. Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G. & Saba, V. Observed fingerprint of a weakening Atlantic Ocean overturning circulation. Nature 556, 191–196 (2018).
30. Pesant, S. et al. Open science resources for the discovery and analysis of Tara Oceans data. Sci. Data 2, 150023 (2015).

31. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci. Data 4, 170093 (2017).
32. Morel, A. et al. Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. Remote Sens. Environ. 111, 69–88 (2007).
33. Benoit, G. et al. Multiple comparative metagenomics using multiset k -mer counting. PeerJ Comput. Sci. 2, e94 (2016).
34. R Core Team, T. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, 2017).
35. Warnes, G. R. et al. R package gplots: Various R Programming Tools for Plotting Data. (2015).
36. Brum, J. R. et al. Patterns and ecological drivers of ocean viral communities. Science 348, 1261498 (2015).
37. Oksanen, J. et al. R package vegan: Community Ecology Package. (2019).
38. Legendre, P. & Legendre, L. Numerical Ecology. (Elsevier, 2012).
39. Kreft, H. & Jetz, W. A framework for delineating biogeographical regions based on species distributions. J. Biogeogr. 37, 2029–2053 (2010).
40. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22, 1540–1542 (2006).
41. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65 (1987).
42. Maechler, M., Rousseeuw, P. J., Struyf, A., Hubert, M. & Hornik, K. R package cluster: Cluster Analysis Basics and Extensions. (2015).
43. Bostock, M., Ogievetsky, V. & Heer, J. D3 Data-Driven Documents. IEEE Trans. Vis. Comput. Graph. 17, 2301–2309 (2011).
44. Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P. & Deckmyn, A. R package maps: Draw Geographical Maps. (2018).
45. McIlroy, D., Brownrigg, R., Minka, T. P. & Bivand, R. R package mapproj: Map Projections. (2015).
46. Gerritsen, H. R package mapplots: Data Visualization on Maps. (2014).
47. Ridgway, K. R., Dunn, J. R. & Wilkin, J. L. Ocean Interpolation by Four-Dimensional Weighted Least Squares—Application to the Waters around Australasia. J. Atmospheric Ocean. Technol. 19, 1357–1375 (2002).
48. Reygondeau, G. et al. Dynamic biogeochemical provinces in the global ocean. Glob. Biogeochem. Cycles 27, 1046–1058 (2013).
49. Oliver, M. J. & Irwin, A. J. Objective global ocean biogeographic provinces. Geophys. Res. Lett. 35, L15601 (2008).
50. Clayton, S. et al. Biogeochemical versus ecological consequences of modeled ocean physics. Biogeosciences Discuss. 1–20 (2016). doi:10.5194/bg-2016-337
51. Menemenlis, D. et al. ECCO2: High resolution global ocean and sea ice data synthesis. Mercat. Ocean Q. Newsl. 31, 13–21 (2008).
52. Rattray, A. et al. Geographic distance, water circulation and environmental conditions shape the biodiversity of Mediterranean rocky coasts. Mar. Ecol. Prog. Ser. 553, 1–11 (2016).
53. Carradec, Q. et al. A global ocean atlas of eukaryotic genes. Nat. Commun. 9, 373 (2018).
54. Wu, S., Xiong, J. & Yu, Y. Taxonomic Resolutions Based on 18S rRNA Genes: A Case Study of Subclass Copepoda. PLoS ONE 10, e0131498 (2015).
55. Vannier, T. et al. Survey of the green picoalga Bathycoccus genomes in the global ocean. Sci. Rep. 6, 37900 (2016).
56. Piganeau, G., Eyre-Walker, A., Grimsley, N. & Moreau, H. How and Why DNA Barcodes Underestimate the Diversity of Microbial Eukaryotes. PLoS ONE 6, e16342 (2011).
57. Worden, A. Z. et al. Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. Science 324, 268–272 (2009).
58. Seeleuthner, Y. et al. Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. Nat. Commun. 9, 310 (2018).
59. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 31, 3718–3720 (2015).
60. Sokal, R. R. & Rohlf, F. J. The Comparison of Dendrograms by Objective Methods. Taxon 11, 33–40 (1962).
61. Sneath, P. H. A. & Sokal, R. R. Numerical taxonomy. The principles and practice of numerical classification. (W.H. Freeman and Company, 1973).

62. Baker, F. B. & Hubert, L. J. Measuring the Power of Hierarchical Cluster Analysis. J. Am. Stat. Assoc. 70, 31–38 (1975).
63. Wei, T. & Simko, V. R package corrplot: Visualization of a Correlation Matrix. (2016).
64. Terada, Y. & von Luxburg, U. R package loe: Local Ordinal Embedding. (2016).
65. Speich, S., Blanke, B. & Cai, W. Atlantic meridional overturning circulation and the Southern Hemisphere supergyre. Geophys. Res. Lett. 34, n/a–n/a (2007).
66. Madoui, M.-A. et al. New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. Mol. Ecol. 26, 4467–4482 (2017).
67. Eppley, R. W. Temperature and phytoplankton growth in the sea. Fish Bull 70, 1063–1085 (1972).
68. Reygondeau, G. et al. Biogeography of tuna and billfish communities. J. Biogeogr. 39, 114–129 (2012).
69. Fofonoff, N. P. The Gulf Stream system. in Evolution of Physical Oceanography: Scientific Surveys in Honor of Henry Stommel (eds. Warren, B. A. & Wunsch, C.) 112–139 (MIT Press, 1980).
70. Dornelas, M. et al. Assemblage Time Series Reveal Biodiversity Change but Not Systematic Loss. Science 344, 296–299 (2014).
71. Franklin, B. A Letter from Dr. Benjamin Franklin, to Mr. Alphonsus le Roy, Member of Several Academies, at Paris. Containing Sundry Maritime Observations. Trans. Am. Philos. Soc. 2, 294–329 (1786).

Acknowledgements

We acknowledge Oliver Jahn and M. J. Follows for providing numerical simulations of particle trajectories from *Tara* Oceans stations. We thank the commitment of the following people and sponsors who made this expedition possible: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, the French Government 'Investissement d'Avenir' programs OCEANOMICS (ANR-11-BTBR-0008) and FRANCE GENOMIQUE (ANR-10-INBS-09), Fund for Scientific Research – Flanders, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, MEMO LIFE (ANR-10-LABX-54), Paris Sciences et Lettres (PSL) Research University (ANR-11-IDEX-0001- 02),
ANR (projects POSEIDON/ANR-09-BLAN-0348, PROMETHEUS/ANR-09-PCS-GENM-217, MAPPI/ANR-2010-COSI-004, TARA-GIRUS/ANR-09-PCS-GENM-218, HYDROGEN/ANR-14-CE23-0001), EU FP7 MicroB3/No. 287589, US NSF grant DEB-1031049, FWO, BIO5, Biosphere 2, Agnès b., the Veolia Environment Foundation, Région Bretagne, World Courier, Illumina, Cap L'Orient, the EDF Foundation EDF Diversiterre, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, the *Tara* schooner and its captain and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. The bulk of genomic computations were performed using the Airain HPC machine provided through GENCI- [TGCC/CINES/IDRIS] (grants t2011076389, t2012076389, t2013036389, t2014036389, t2015036389 and t2016036389). We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who granted us sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org).
DJR was supported by postdoctoral fellowships from the Conseil Régional de Bretagne, the Beatriu de Pinós programme of the Government of Catalonia's Secretariat for Universities and Research of the Ministry of Economy and Knowledge, and a fellowship from "la Caixa" Foundation (ID 100010434) with the fellowship code LCF/BQ/PI19/11690008. RW, DI and MRd'A were supported by the Italian Flagship Project RITMARE and Premiale MIUR NEMO. MBS was supported by US NSF grants OCE-1536989 and OCE-1829831, grant #3709 from the Gordon and Betty Moore Foundation, and HPC support from the Ohio Super Computer.
We also acknowledge Stéphane Audic for assistance with metabarcoding analyses, C. Scarpelli for support in high-performance computing, Mathieu Raffinot and Dominique Lavenier for discussions on sequence comparison algorithms, Samuel Chaffron for help with sample contextual data, Noan Le Bescot (Ternog Design) for assistance in preparing figures, and Marion Gehlen. We thank all members of the *Tara* Oceans consortium for maintaining a creative environment and for their constructive criticism.

Author Contributions

DI, OJ, CdV, and PW designed and directed the study. IP, DJR, RW, OJ, DI, MRd'A, TV and CdV wrote the manuscript .TV, GB, NM, PP, CL and OJ designed and computed pairwise metagenomic comparisons. TV, DJR, RW, JL and PF performed the analyses of genomic data with substantial input from MRd'A, DI, OJ and PW. RW, DI, TV, PF and DJR analyzed ocean circulation simulations. GR, NH, AF-G, S Suweis, RN, J-MA, MM and EP contributed additional analysis. S Sunagawa, LG, PB, CB, MBS and EK provided additional interpretation of results. KL, EM and JP coordinated the genomic sequencing with the informatics assistance of CD, FG and J-MA. S Roux, JRB and MBS contributed viral data, PB and S Sunagawa contributed bacterial data. CB, S Romac, NH, CdV and DJR analyzed eukaryotic metabarcoding data. CD, SK, MP, S Searson and JP coordinated collection and management of *Tara* Oceans samples. *Tara* Oceans Coordinators provided support and guidance throughout the study. All authors discussed the results and commented on the manuscript.

Author Information

The authors declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations in whose waters the *Tara* Oceans expedition sampled. Metagenomic and metabarcoding sequencing reads have been deposited at the European Nucleotide Archive under accession numbers provided in Supplementary Table 1. Contextual metadata of *Tara* Oceans stations are available in Supplementary Table 2. Metagenomic dissimilarity, OTU community dissimilarity, simulated travel times and geographic distances are provided in Supplementary Tables 3-16. All Supplementary Tables, in addition to tables of 18S V9 barcodes and OTUs and the V9 reference database are available on FigShare at the following URL: http://doi.org/10.6084/m9.figshare.11303177
The authors declare no competing financial interests.
Correspondence and requests for materials should be addressed to Olivier Jaillon, Daniele Iudicone, Maurizio Ribero d'Alcalà, Colomban de Vargas.

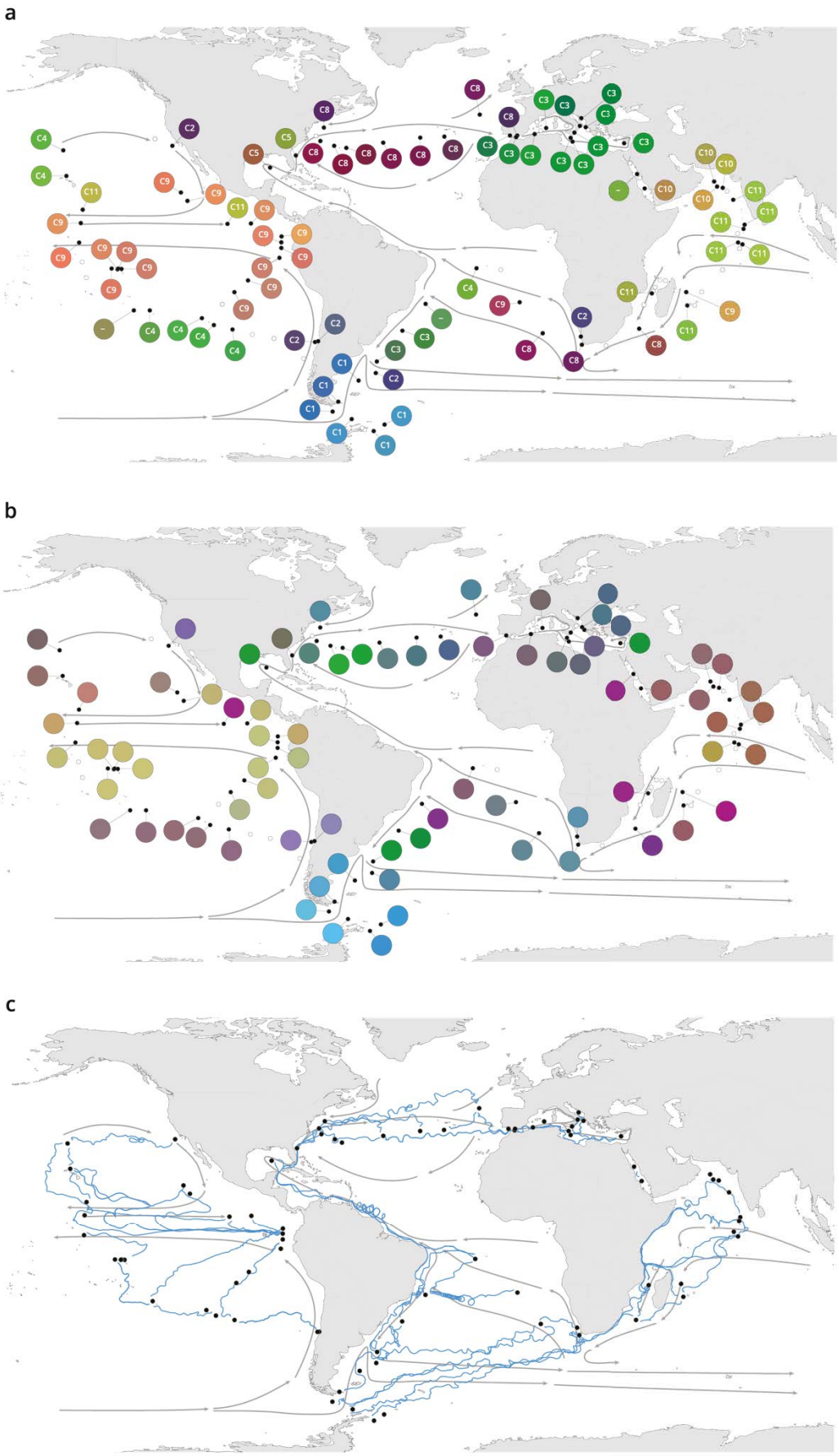


Figure 1 | Plankton biogeography, environmental variation and ocean transport among *Tara* Oceans stations. Major currents are represented by solid arrows. **a**, Genomic provinces of *Tara* Oceans surface

655 samples for the 0.8-5 µm size fraction, each labeled with a letter prefix ('C' represents the 0.8-5 µm size
656 fraction) and a number; samples not assigned to a genomic province are labeled with '-'. Maps of all six size
657 fractions and including DCM samples are available in Supplementary Fig. 4. Station colors are derived from an
658 ordination of metagenomic dissimilarities; more dissimilar colors indicate more dissimilar communities (see
659 Methods). **b**, Stations colored based on an ordination of temperature and the ratio of NO_2NO_3 to PO_4 (replaced
660 by 10^{-6} for 3 stations where the measurement of PO_4 was 0) and of NO_2NO_3 to Fe. Colors do not correspond
661 directly between maps; however, the geographical partitioning among stations is similar between the two
662 maps. **c**, Simulated trajectories corresponding to the minimum travel time (T_{min}) for pairs of stations (black
663 dots) connected by $T_{min} < 1.5$ years. Directionality of trajectories is not represented.

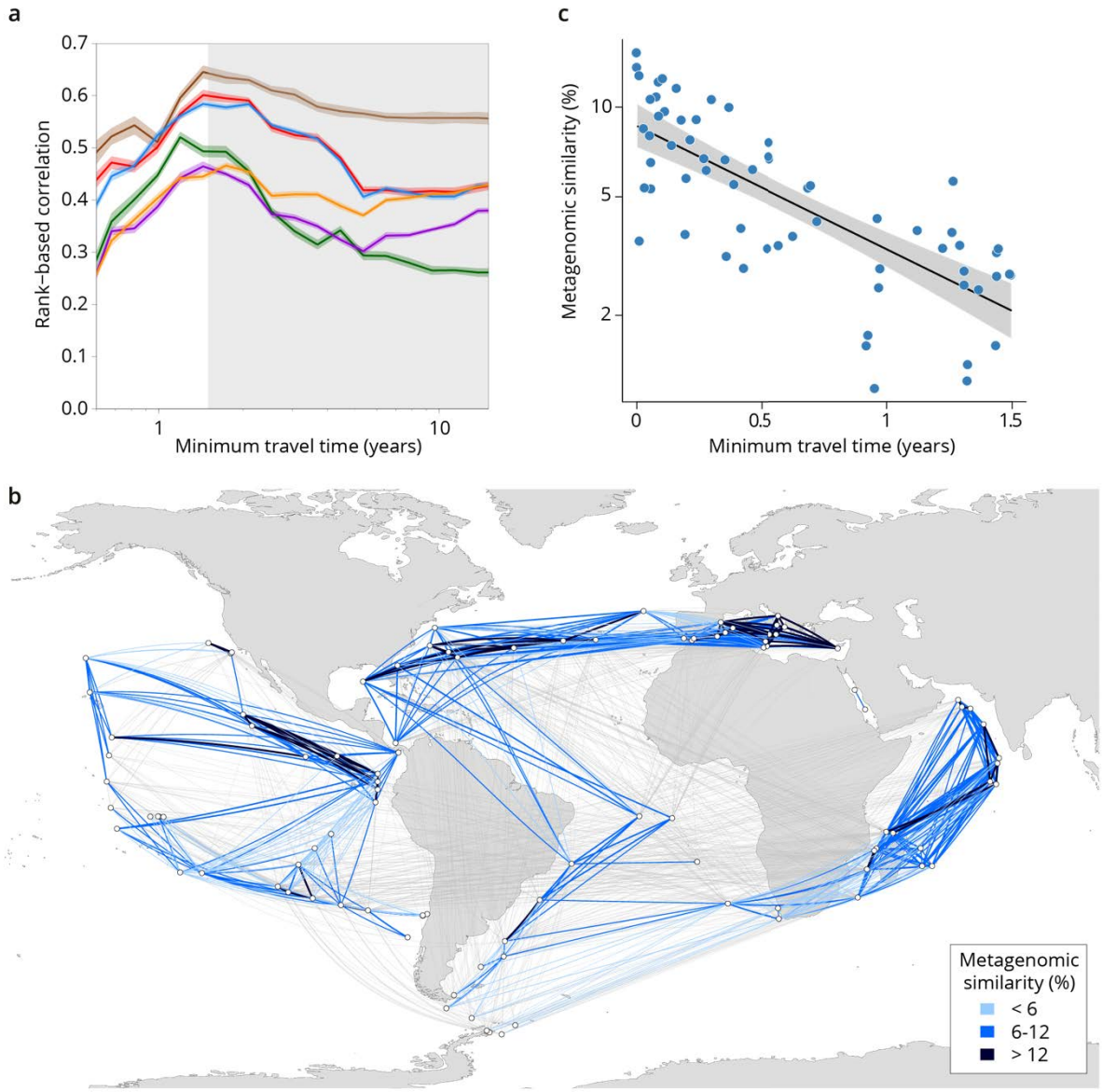


Figure 2 | Metagenomic dissimilarity and travel time of plankton are maximally correlated up to ~1.5 years. **a**, Spearman rank-based correlation by size fraction between metagenomic dissimilarity and minimum travel time along ocean currents (T_{min}) for pairs of *Tara* Oceans samples separated by a minimum travel time less than the value of T_{min} on the x axis. Brown line: 0-0.2 µm size fraction, red: 0.22-1.6/3 µm, blue: 0.8-5 µm, green: 5-20 µm, purple: 20-180 µm, orange: 180-2000 µm. Shaded colored areas represent 95% confidence intervals. T_{min} >1.5 years is shaded in grey. See plots for OTU dissimilarity in Supplementary Fig. 9. **b**, Pairs of *Tara* stations connected by T_{min} <1.5 years in blue/black and >1.5 years in grey. Shading reflects metagenomic similarity from the 0.8-5 µm size fraction. **c**, The relationship of metagenomic similarity to T_{min} with an exponential fit (black line, grey 95% CI), for pairs of surface samples in the 0.8-5 µm size fraction within the North Atlantic and Mediterranean current system (see map and plots for other size fractions and OTUs in Supplementary Fig. 10, and Supplementary Information 1 for a discussion of metagenomic similarity).

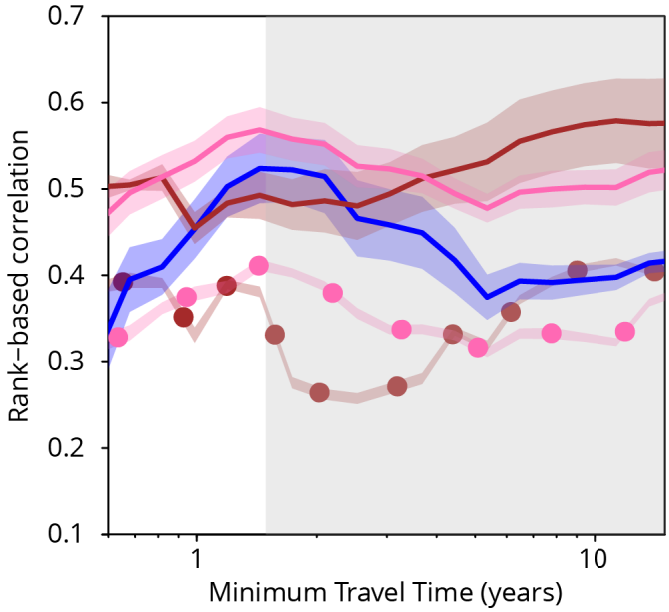
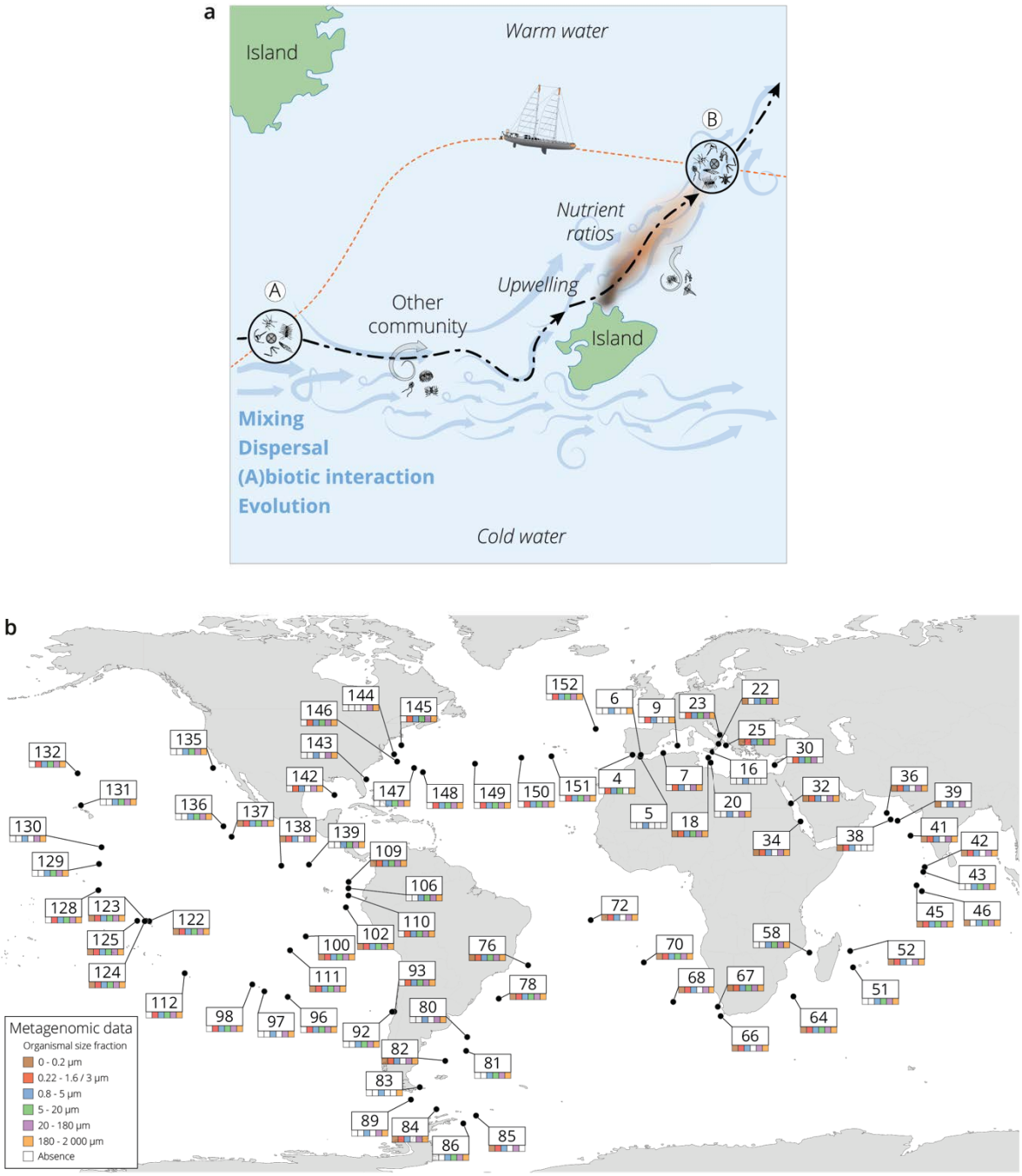
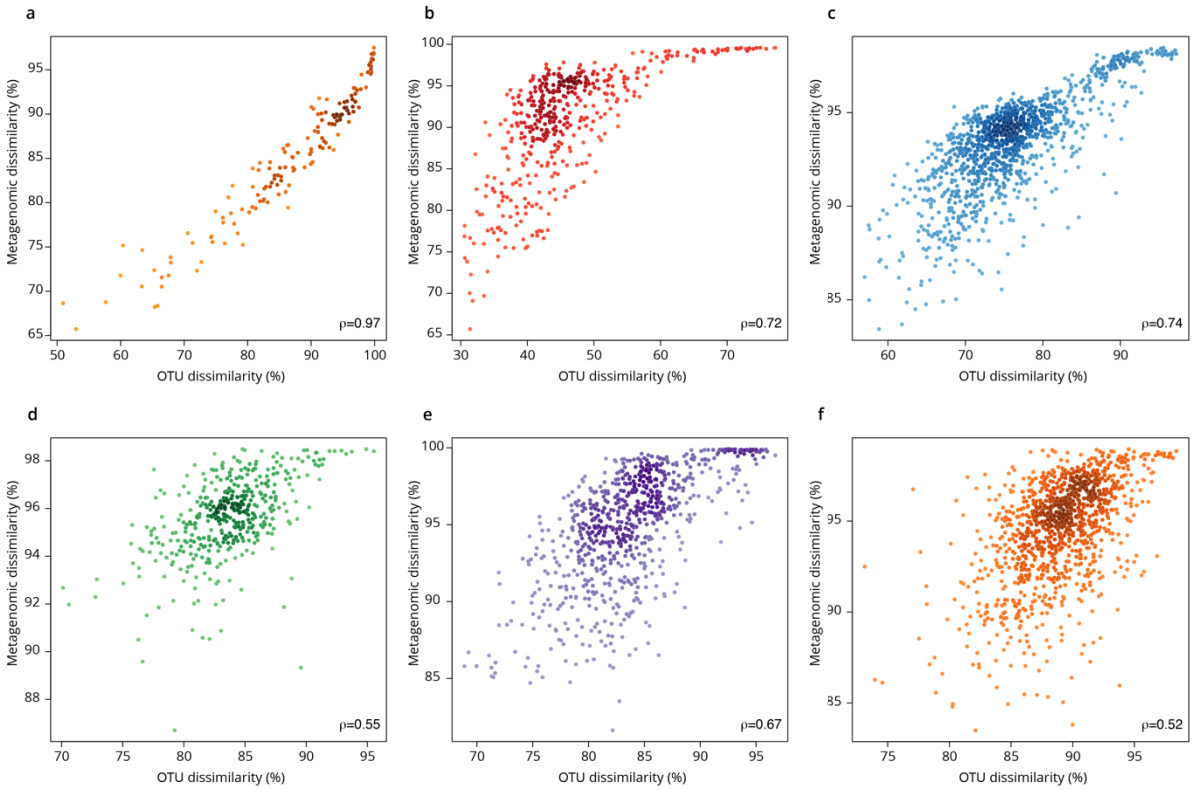


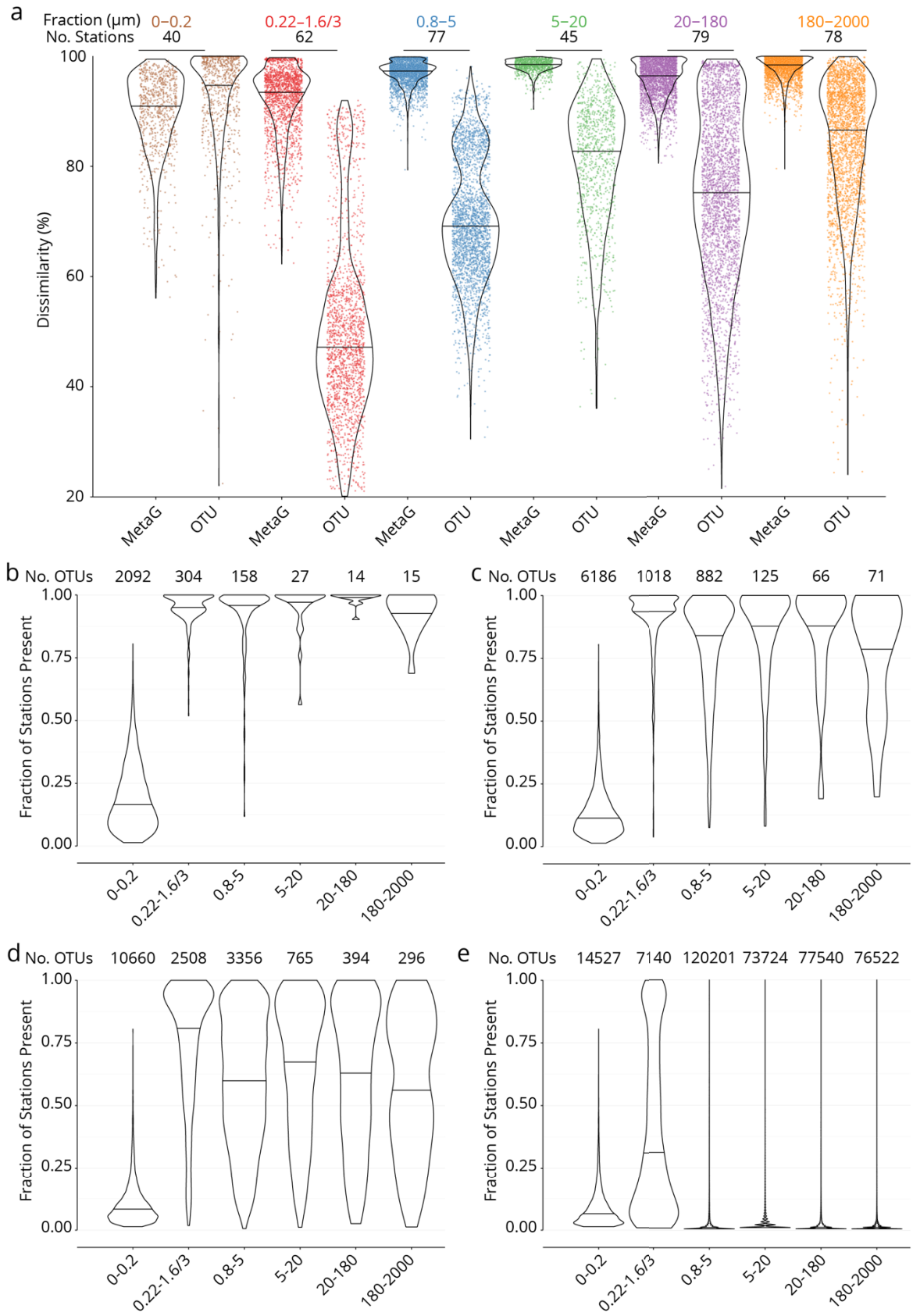
Figure 3 | Plankton travel time, metagenomic dissimilarity and environmental differences show different temporal patterns of pairwise correlation. Spearman rank-based correlations between metagenomic dissimilarity and minimum travel time (T_{min}, blue), metagenomic dissimilarity and differences in NO_2NO_3, PO_4 and Fe (pink), metagenomic dissimilarity and differences in temperature (red), T_{min} and differences in NO_2NO_3, PO_4 and Fe (pink, dashed), and T_{min} and differences in temperature (red, dashed) for pairs of *Tara* Oceans samples separated by a minimum travel time less than the value of T_{min} on the x axis. Shaded regions represent standard error of the mean. Correlations represent averages across four of six size fractions represented in Fig. 2a; the 0-0.2 µm and 5-20 µm size fractions are excluded due to a lack of samples at the global level. Individual size fractions, partial correlations, and correlations with OTU data are in Supplementary Fig. 9.



Supplementary Figure 1 | The seascape, plankton transport and community metagenomic samples of *Tara* Oceans stations. a, A community sampled at a given location (A) changes over time as it travels along ocean currents (dashed bold line) to a second location (B). It is affected by numerous external processes, including mixing with water containing other communities and changes in local nutrient concentration, and by internal processes, such as biotic interactions. In this study, the *Tara* schooner followed a sampling route (orange dashed line) leading to an elapsed time between the 2 sampling sites A and B that was independent of plankton travel time. **b**, Location, station number, and sequenced surface metagenomic samples.

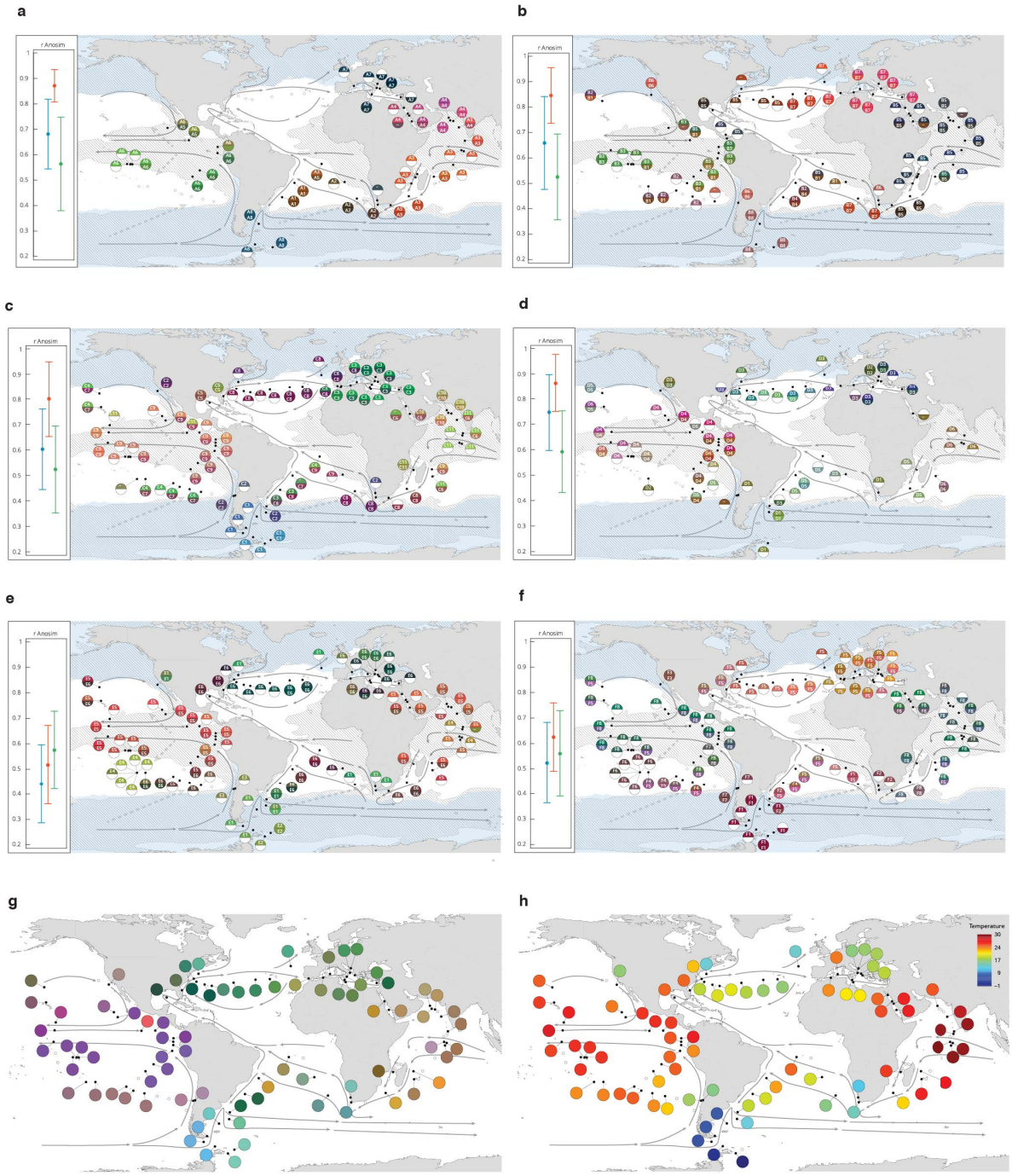


Supplementary Figure 2 | β-diversity estimates from metagenomic and OTU-based dissimilarity are correlated. Scatter plots of metagenomic dissimilarity versus OTU community dissimilarity for six organismal size fractions. Each point represents a pairwise comparison between two samples. **a**, 0-0.2 µm size fraction. **b**, 0.22-1.6/3 µm size fraction. **c**, 0.8-5 µm size fraction. **d**, 5-20 µm size fraction. **e**, 20-180 µm size fraction. **f**, 180-2000 µm size fraction. Global rank-based correlations (Spearman, $p \leq 10^{-4}$) are indicated in the bottom right of each plot.



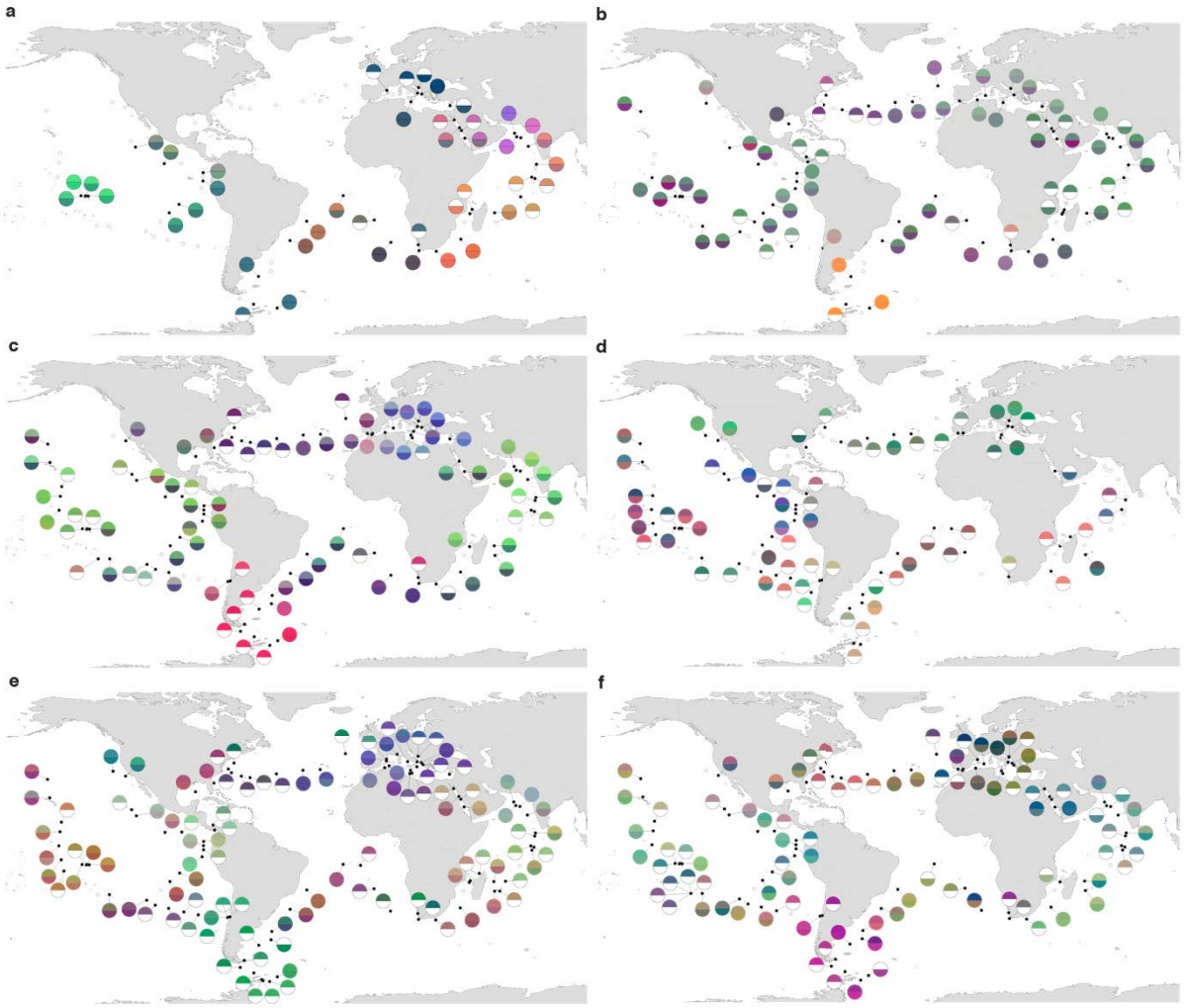
Supplementary Figure 3 | **Global dissimilarity and OTU occupancy. a**, Distributions of dissimilarity for six organismal size fractions (measured either as metagenomic or OTU dissimilarity; see Supplementary Information 1). One colored point represents one pair of stations. Violin plots (horizontal line: median) summarize each distribution. The number of stations in common between the metagenomic/OTU data sets

706 within each size fraction is indicated above. **b-e**, **OTU occupancy for different proportions of total abundance.**
707 Fraction of stations present (occupancy) for the minimum number of OTUs (indicated above) necessary to
708 represent different proportions of the total abundance within each organismal size fraction. A relatively small
709 number of abundant and cosmopolitan taxa represents the majority of the abundance within each size
710 fraction; this effect is more pronounced with increasing organismal size. **b**, OTUs representing 50% of the total
711 abundance within each size fraction. **c**, 80%. **d**, 95%. **e**, 100% (all OTUs).

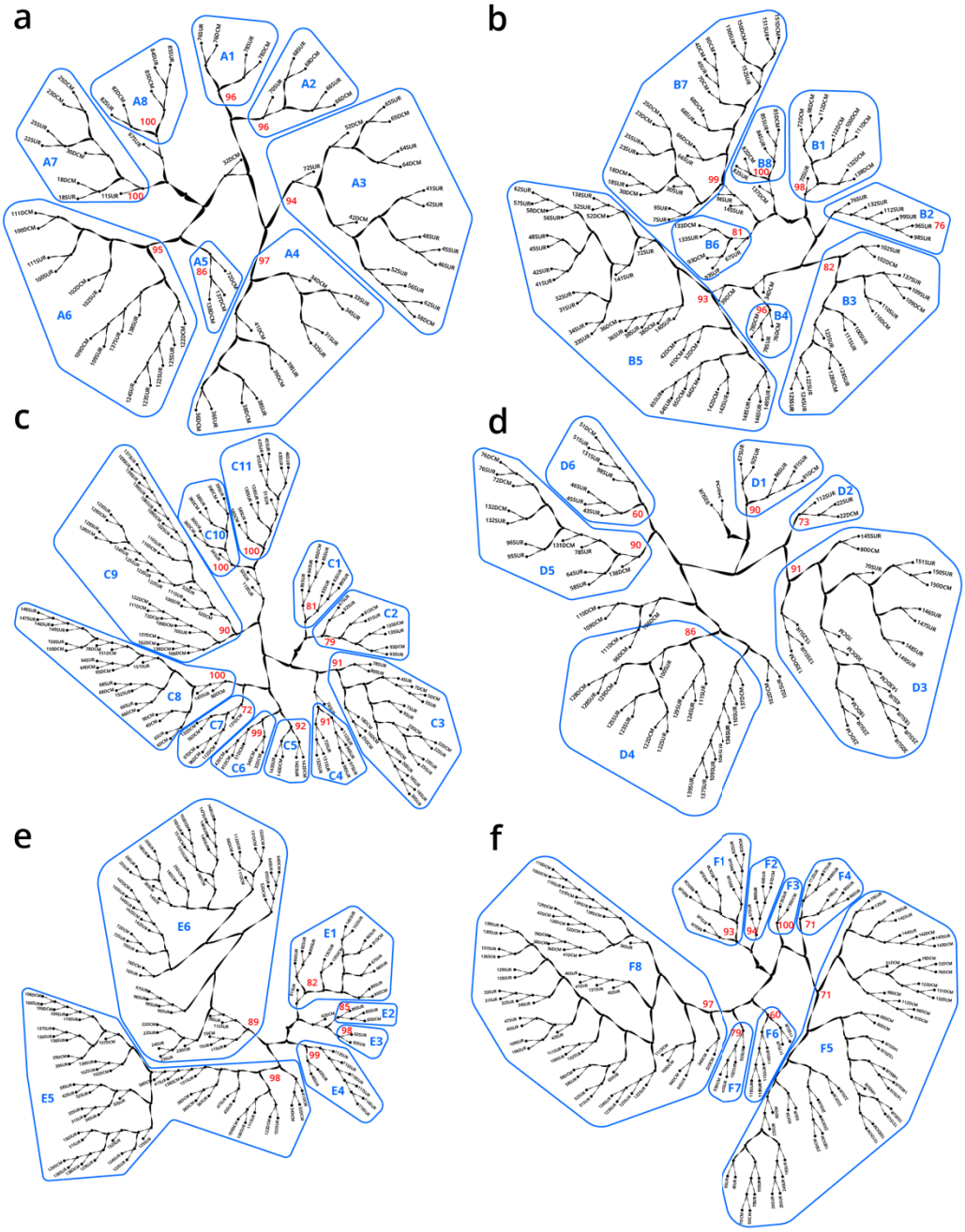


Supplementary Figure 4 | Genomic provinces in comparison to previous ocean divisions, and ordination maps of environmental parameters. a-f, Geographical maps of genomic provinces by organismal size fraction (see Supplementary Information 2). Circles denote stations with data available for the size fraction and contain the corresponding genomic province identifiers (one letter prefix per size fraction (A-F); stations not assigned to genomic provinces are shown as '-'). The top portion of each circle represents samples collected at the surface and the bottom portion represents the deep chlorophyll maximum (stations missing metagenomic data for one of the two depths are drawn as half circles). Colors are based on PCoA-RGB (Methods) and do not correspond among size fractions. Major currents are shown with solid black arrows, wind transport with dashed grey arrows. Blue zones indicate temperature < 14 °C. Hashed zones indicate phosphate concentration > 0.4 mmol. Hierarchical dendrograms that were used to build genomic provinces are shown in Supplementary Fig. 6. Maps with colors based on OTU dissimilarity are shown in Supplementary Fig. 5. **a**, 'A' prefix, 0-0.2 µm size fraction. **b**, 'B' prefix, 0.22-1.6/3 µm. **c**, 'C' prefix, 0.8-5 µm. **d**, 'D' prefix, 5-20 µm. **e**, 'E' prefix, 20-180 µm. **f**, 'F' prefix, 180-2000. **Insets**, Results of ANOSIM to determine, independently for each size fraction, the ability of three nested levels of ocean partitioning to explain metagenomic dissimilarities among stations (blue,

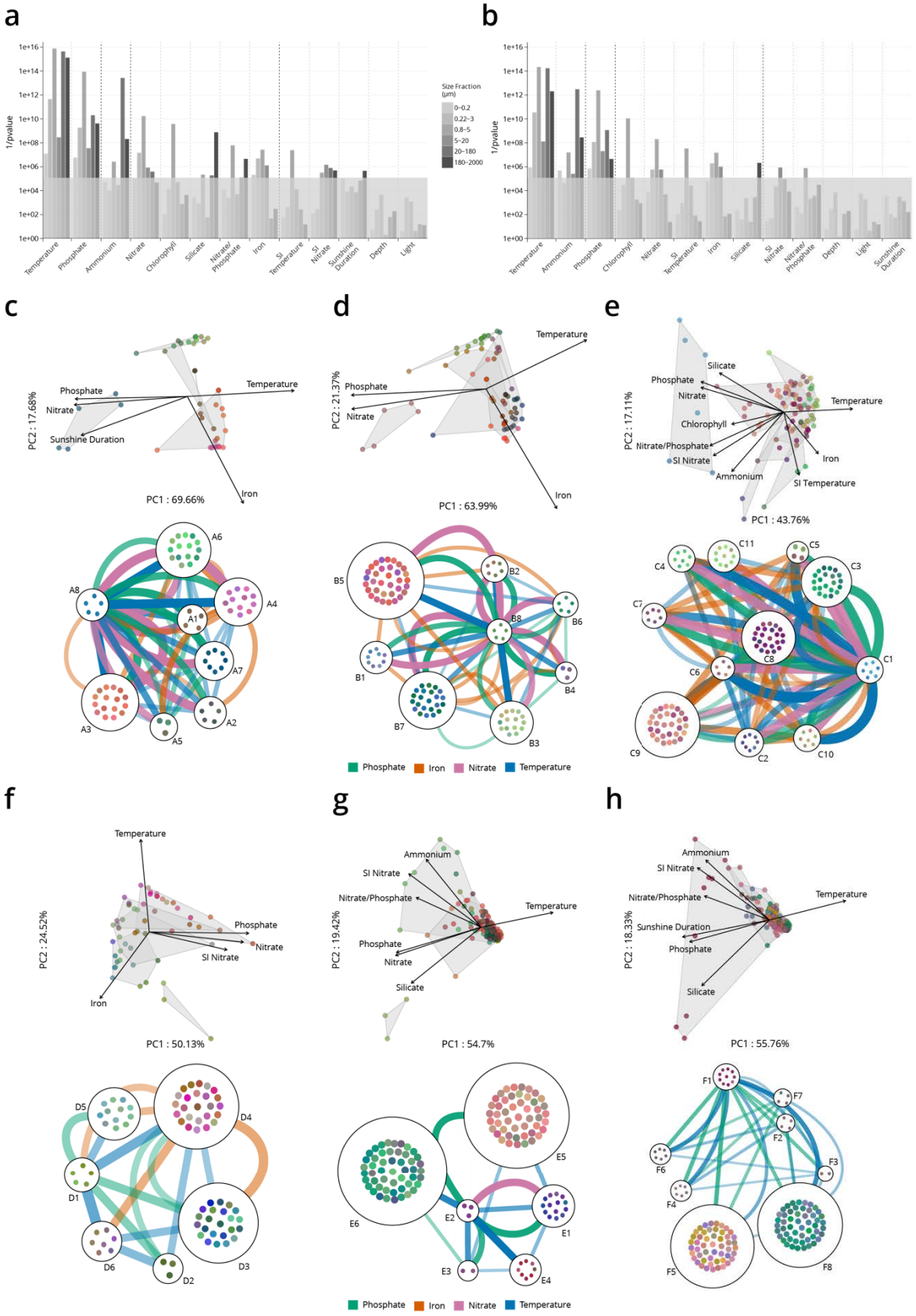
Longhurst biomes; red, Longhurst biogeochemical provinces; green, Oliver and Irwin objective provinces; see Methods and Supplementary Information 3). **g**, The distribution of temperature and nutrient variations matches the biogeography of small plankton (< 20 μm). Stations are colored based on an ordination of Euclidean distances in temperature, NO_2NO_3, PO_4 and Fe. **h**, The distribution of temperature matches the biogeography of large plankton (> 20 μm). Stations are colored following a Box-Cox transformation (Methods).



Supplementary Figure 5 | Biogeography based on an ordination of OTU dissimilarity. a-f, Principal coordinates analysis (PCoA)-RGB color maps for OTUs (see Methods). The top of each half circle represents samples collected at the surface and the bottom portion represents the deep chlorophyll maximum (stations missing OTU data for one of the two depths are drawn as half circles). Station colors do not correspond among size fractions. **a**, 0-0.2 µm size fraction. **b**, 0.22-1.6/3 µm. **c**, 0.8-5 µm. **d**, 5-20 µm. **e**, 20-180 µm. **f**, 180-2000 µm.

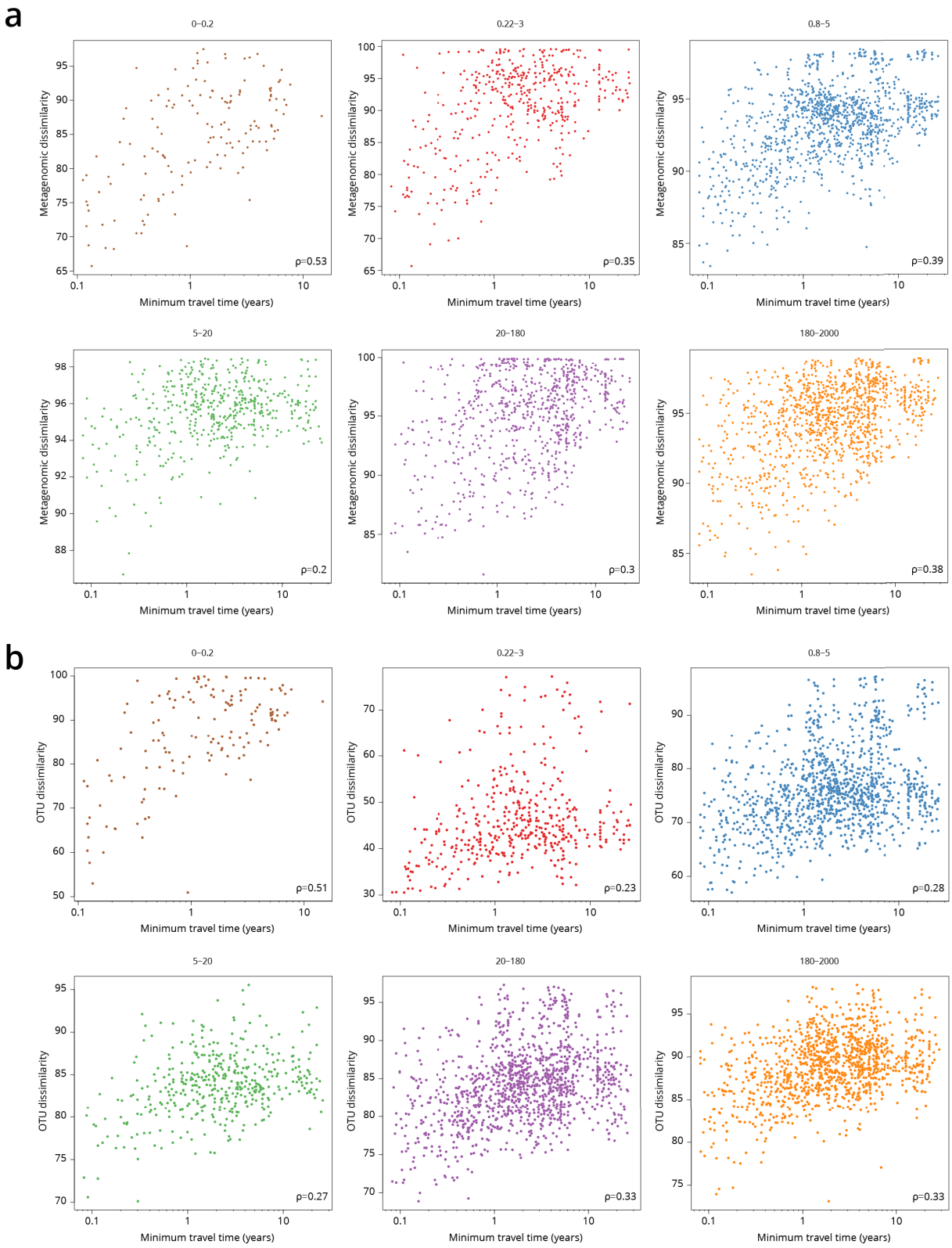


Supplementary Figure 6 | Hierarchical trees illustrating how samples were partitioned into genomic provinces. Dendrograms resulted from UPGMA clustering. Each sample (SUR: surface, DCM: deep chlorophyll maximum) is shown as a leaf. Genomic provinces are shown with their identifiers in blue polygons; identifiers are composed of one letter prefix per size fraction (A-F) and a number. Bootstrap values in red show the support at the key nodes that separate genomic provinces from one another. See also Supplementaryt Information 2 on the robustness of genomic provinces. **a**, 'A' prefix, 0-0.2 µm size fraction. **b**, 'B' prefix, 0.22-1.6/3 µm. **c**, 'C' prefix, 0.8-5 µm. **d**, 'D' prefix, 5-20 µm. **e**, 'E' prefix, 20-180 µm. **f**, 'F' prefix, 180-2000 µm.

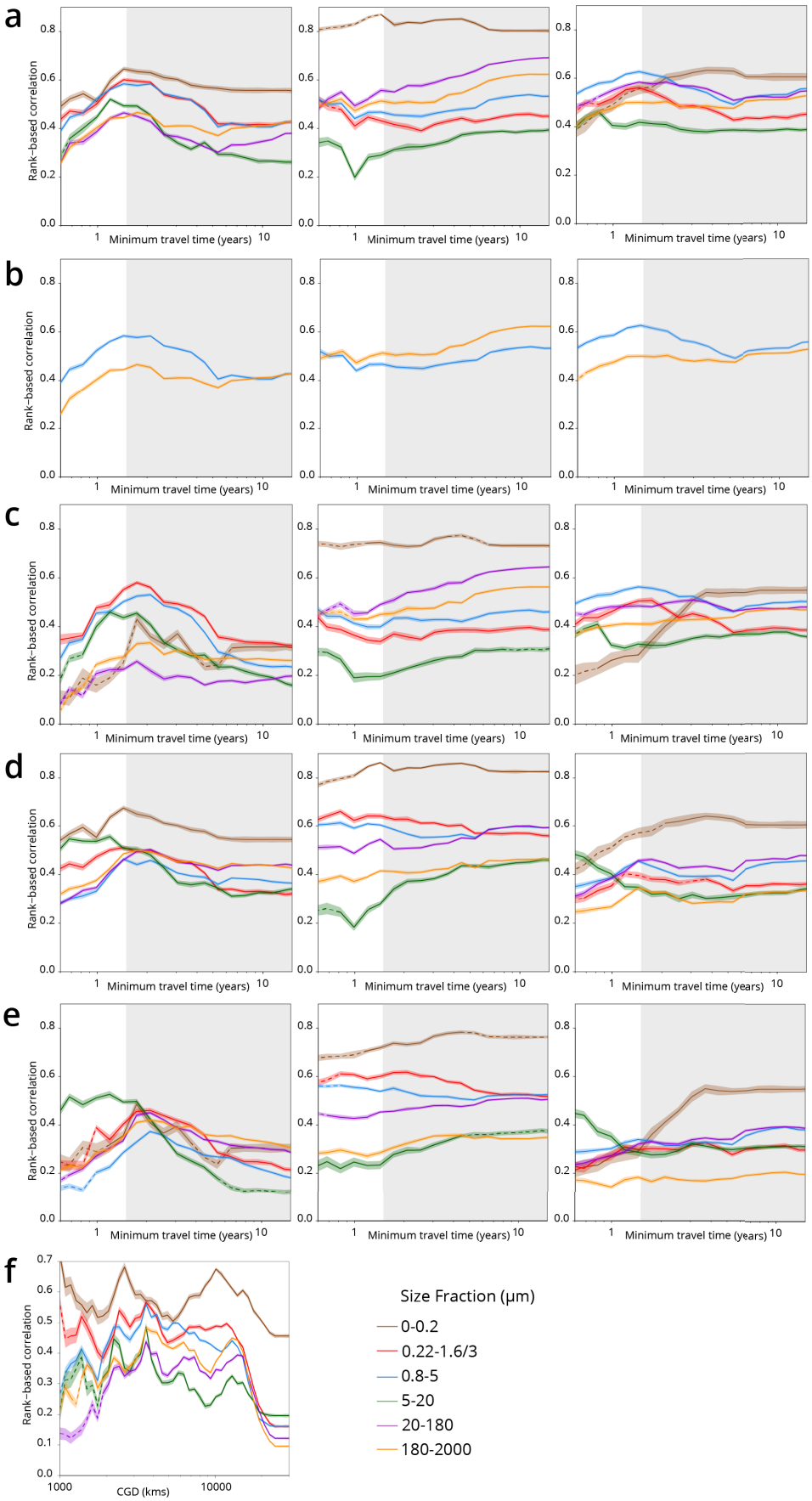


Supplementary Figure 7 | Environmental parameters that distinguish genomic provinces. a-b, Environmental parameters that significantly differentiate among genomic provinces (Kruskal-Wallis test, grey box indicates p values > 10^{-5}). SI = Seasonality Index. **a**, all stations. **b**, Antarctic stations removed (see Methods). Eliminating Antarctic stations does not result in a large change in the parameters that significantly differentiate among

provinces. **c-h**, Two types of visualizations of the relationships between genomic provinces and environmental parameters. Sample colors are those from Supplementary Fig. 4. **Top plots within panels c-h**: principal components analysis-based visualization. Samples, and environmental parameters differing significantly ($p \leq 10^{-5}$) among genomic provinces, are projected onto the first two axes of variation. Grey polygons enclose different genomic provinces. **Bottom plots within panels c-h**: network-based visualization. Each genomic province is represented as a node, with the individual samples composing the province within the node. Edges between nodes represent differences in temperature, nitrate, phosphate and iron that significantly differentiate ($p \leq 10^{-5}$) among genomic provinces, that are statistically significantly different between individual pairs of genomic provinces (*post hoc* Tukey test, $p < 0.01$) and whose difference in median parameter values is ≥ 1 standard deviation (calculated from the parameter values of all samples in the size fraction). Thicker edges represent larger differences. **c**, 0-0.2 µm size fraction. **d**, 0.22-1.6/3 µm. **e**, 0.8-5 µm. **f**, 5-20 µm. **g**, 20-180 µm. **h**, 180-2000 µm.

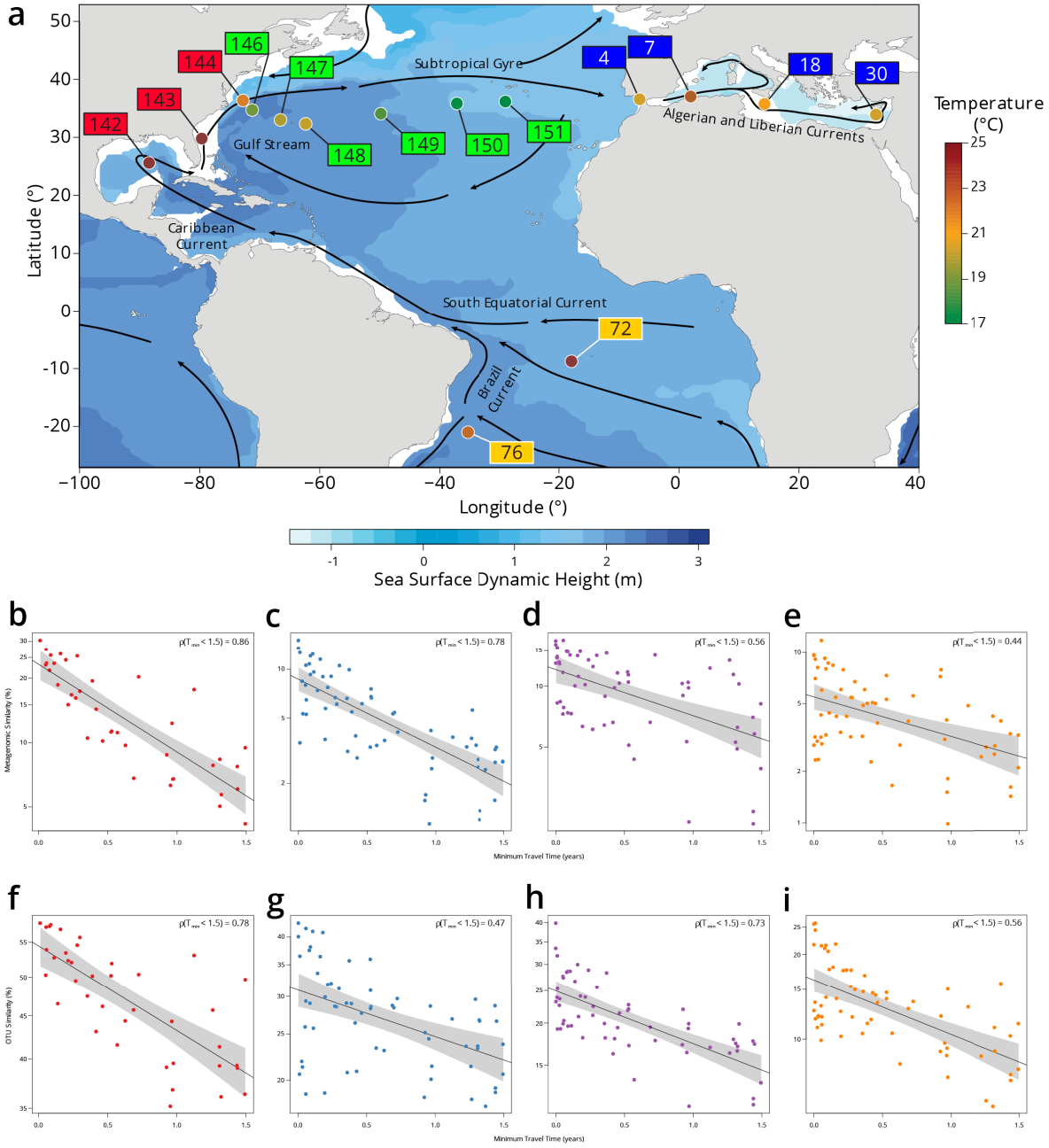


Supplementary Figure 8 | Global correlations of dissimilarity with minimum travel time (T_{min}). Scatter plots of dissimilarity versus T_{min}. One point represents a pair of samples. **a**, metagenomic dissimilarity. **b**, OTU dissimilarity. Global Spearman correlation values are indicated within each panel.



Supplementary Figure 9 | Plankton travel time, dissimilarity, environmental distance and geographic distance show different temporal patterns of pairwise correlation. Spearman correlation values are shown

separately by organismal size fraction. Non-significant correlations ($p > 0.01$) are shown with dashed lines. **a-e**, Correlations for pairs of *Tara* Oceans samples separated by a minimum travel time less than the value of T_{min} on the x axis. T_{min} >1.5 years is shaded in grey. Left panels: correlation of dissimilarity with T_{min}; middle panels, dissimilarity with temperature; right panels: dissimilarity with differences in NO_2NO_3, PO_4 and Fe. **a-c**, metagenomic dissimilarity. **d-e**, OTU dissimilarity. There is a maximum correlation of dissimilarity with T_{min} (and, for most size fractions, of dissimilarity with nutrients) for T_{min} <~1.5 years, but the correlation between dissimilarity and temperature does not display a similar maximum. **b** displays only the 0.8-5 µm (blue) and 180-2000 µm (orange) size fractions from **a**, to highlight that for smaller plankton, correlations with differences in nutrient concentrations were stronger for T_{min} up to ~1.5 years, but for larger plankton, correlations were stronger with temperature variations for T_{min} beyond ~1.5 years. **c** and **e**, Partial correlations to estimate the independent effects of T_{min} and environmental distances on β-diversity. Left panels: controlling for differences in temperature and for differences in NO_2NO_3, PO_4 and Fe; middle and right panels: controlling for T_{min}. Partial correlations do not affect the maximum correlation of dissimilarity with T_{min} for T_{min} <~1.5 years. **f**, Correlation of geographic distance (without traversing land) with metagenomic dissimilarity for pairs of *Tara* Oceans samples separated by a geographic distance less than the value on the x axis.



Supplementary Figure 10 | Plankton community composition turnover through the North Atlantic. a, Map of *Tara* Oceans stations, currents (solid lines), temperature by station (colored circles) and sea surface climatological dynamic height from CARS2009 (http://www.cmar.csiro.au/cars). Each station label has a color corresponding to a sub-region: South Atlantic in orange, Gulf Stream in red, Recirculation/Gyre in green and Mediterranean Sea in blue. **b-e**, Scatter plots of metagenomic similarity versus minimum travel time (T_{min}) for these stations in the **b**, 0.22-3 µm; **c**, 0.8-5 µm; **d**, 20-180 µm; and **e**, 180-2000 µm size fractions. **f-i**, Scatter plots of OTU community similarity for the **f**, 0.22-3 µm; **g**, 0.8-5 µm; **h**, 20-180 µm; and **i**, 180-2000 µm size fractions. The black line represents an exponential fit, with a light grey shaded 95% confidence interval. The resulting turnover times using metagenomic similarity are τ = 0.91 y for 0.22-3 µm, τ = 0.91 y for 0.8-5 µm, τ = 2.22 y for 20-180 µm and τ = 1.99 y for 180-2000 µm. Turnover times using the OTU community similarity are τ = 4.23 y for 0.22-3 µm, τ = 4.08 y for 0.8-5 µm, τ = 2.6 y for 20-180 µm and τ = 2.1 y for 180-2000 µm. The viral-enriched 0-0.2 µm and the nanoplanktonic 5-20 µm size fractions are not shown due to insufficient sampling of these stations.

Supplementary Information

Supplementary Information 1. Comparison of metagenomes and OTUs

Metagenomic comparisons reflect fine-scale differences in genome content at the community level as a function of diversity, genome size and organismal abundance, and also depend on the rate of evolution of each specific lineage. With exhaustive sampling, metagenomic dissimilarity could theoretically distinguish among genomes in a sample separated by a single mutation. However, our metagenomic sequencing depth was likely not able to reach saturation due to the number of genomes per sample and their putative large size (metatranscriptomes, which contain fewer sequences per species than do metagenomes, did not reach saturation within *Tara* Oceans samples[53]). For example, if for a pair of samples we sequence 50% of the total amount of the unique genomic DNA present, we expect the maximum similarity of the two samples to be roughly 25% (0.5 x 0.5). Therefore, the pairwise metagenomic dissimilarities we calculated between samples probably reflected a combination of genomic differences weighted towards more abundant organisms. In contrast, OTUs, obtained by sequencing single marker genes, approach biodiversity saturation[5,18,19]. However, OTU resolution depends on the choice of the marker to be used, the threshold of similarity for the marker, and its lineage-specific substitution rate, and may therefore confound evolutionarily and/or ecologically distant organisms[54–58]. We observed a significant agreement between the two proxies (Supplementary Fig. 2), although dissimilarities based on OTUs were generally lower than those computed from metagenomic data (Supplementary Fig. 3a).

Analyses of plankton biogeography produced consistent results based on metagenomic and OTU data (Supplementary Fig. 4, Supplementary Fig. 5, Supplementary Fig. 8, Supplementary Fig. 9). For simplicity, in the main text, we chose to highlight results based on metagenomes rather than on OTUs for three reasons. First, the metagenomic sequencing protocol and subsequent measurement of dissimilarity was uniform across size fractions, whereas OTUs were defined differently for the viral-enriched, bacterial-enriched and eukaryote-enriched size fractions (Methods). Second, the biogeographical patterns we obtained (see below) may be more evident in comparisons among metagenomic sequences (our data source in identifying genomic provinces), as genomes, accumulate single-base changes and other variants more quickly than a single ribosomal gene marker. Third, β-diversity estimated by metagenomic dissimilarity generally displayed higher correlation values with minimum travel time (T_{min}; Supplementary Fig. 8).

Supplementary Information 2. Robustness of genomic provinces

We assessed the robustness of genomic provinces in five separate ways. First, we tested 5 different hierarchical clustering algorithms from R-package pvclust_1.3-2[40] (UPGMA - Unweighted Pair Group Method with Arithmetic mean; McQuitty's method; Complete linkage; Ward's method; Single linkage) on the metagenomic pairwise dissimilarities produced by Simka separately for the six organismal size fractions, followed by multiscale bootstrap resampling. We used the cophenetic correlation coefficient from the R-package dendextend_1.5.2[59] to measure how accurately the dendrograms produced by each method preserved the pairwise distances within the input dissimilarity matrices[60,61]. The ranking of the cophenetic correlation coefficient for different clustering methods within each size fraction was consistent with a published large-scale methodological comparison of clustering methods for biogeography (Supplementary Table 17), which considered UPGMA agglomerative hierarchical clustering to have consistently the best performance[39]. Second, we compared clustering results among all size fractions using Baker's Gamma Index[62] from the R-package corrplot_0.77[63], which is a measure of association (similarity) between two trees based on hierarchical clustering (dendrograms). The Baker's Gamma Index is defined as the rank correlation between the stages at which pairs of objects combine in each of the two trees. For each type of correlation, the UPGMA was consistently the most correlated with other clustering methods (Supplementary Table 18). This allowed us to conclude, in

agreement with previous results[39], that the UPGMA method is likely more robust than the other methods we tested.

Third, we compared the genomic provinces found by our UPGMA hierarchical clustering approach to those found by two different non-hierarchical methods: K-means on the positions found by multidimensional scaling and spectral clustering on the nearest-neighbor graph. Both methods rely on (i) a dissimilarity matrix and (ii) a tuning parameter (dimension of the projection space for K-means, and number of neighbors for spectral clustering). K-means uses the numeric values of the dissimilarities, whereas spectral relies only on their ordering (e.g., community A is closer to B than to C). We compared the genomic provinces to clusters found by K-means and spectral clustering for all values of the tuning parameter using the Rand Index (RI; from the GARI function of the loe R package version 1.1[64]), a score of agreement between partitions. Results are reported as mean +/- s.d. of the RI: 1 means perfect agreement and 0 complete disagreement. Fourth, in order to assess the significance of the genomic provinces, we performed a multivariate ANOVA to partition metagenomic dissimilarity across regions, using the adonis function of the vegan R package version 2.5-4[37]. Note, however, that since the same data were used both to construct the genomic provinces and to assess their significance, the p-values estimated by ADONIS might be anti-conservative. The results of the third and fourth analyses are presented in Supplementary Table 19.

Fifth, we found that clustering of samples in genomic provinces was consistent with a complementary visualization based on the same data: RGB colors derived from the first three axes of a principal coordinates analysis (PCoA-RGB) of β-diversity, in which similar colors represent similar communities (Supplementary Fig. 4; see Methods). Samples within the same genomic province generally shared the same range of PCoA-RGB colors. Because the clustering approach was hierarchical, samples sharing some similarity could have been assigned to different genomic provinces due to binary decisions during the clustering process. This was also reflected in the PCoA-RGB colors, where the boundaries of genomic provinces did not indicate a complete change of communities among genomic provinces (and, conversely, belonging to the same genomic province did not imply identical community). Nonetheless, samples with similar PCoA-RGB colors were generally situated in closely-related branches in the UPGMA tree (Supplementary Fig. 6). An illustrative example is genomic province F5 (of the 180-2000 µm size fraction; Supplementary Fig. 4f), which encompassed stations in the Atlantic, Mediterranean Sea and some subtropical stations in the Indo-Pacific. In this wide region, the PCoA-RGB colors indicate the variation in community composition within the genomic province, and also reflect the relatedness of F5 to its adjacent samples, in particular those in the subtropical Atlantic/Pacific region F4, its neighbor in the UPGMA tree (Supplementary Fig. 6f).

Supplementary Information 3. Comparison of genomic provinces to previous biogeographical divisions

Current approaches in biogeographic theory divide the ocean into regions based either on expert knowledge applied to satellite data, as in the hierarchical nesting by Longhurst[3] into biomes (macro-scale, essentially representing a division of the world's oceans into cold and warm waters, and coastal upwelling zones) and biogeochemical provinces (BGCPs, areas within biomes defined by observable boundaries and predicted ecological characteristics), or, alternatively, into the objective provinces of Oliver and Irwin[49], which are based solely on statistical analyses. Longhurst BGCPs are based upon, primarily, monthly variations of chlorophyll a, the geography of the seasonal cycle of physical factors (such as the depth of the upper ocean mixed layer) and surface temperatures. In turn, these ocean properties are strongly modulated by oceanic currents (for example, moderate to large mixed layer depths are observed generally on the poleward side of the subtropical gyres). In contrast, the objective global ocean biogeographic provinces proposed by Oliver and Irwin[49] were based upon clustering temporal variability of chlorophyll concentration and surface temperatures, both measured from satellite data. They combined a proxy for the intensity of primary productivity with water temperature, therefore emphasizing regions similar in their temporal variability for both properties

(which essentially corresponds to the seasonal cycle). None of these ocean partitionings directly considered organismal community composition.

We tested whether genomic provinces were comparable with these partitionings by performing an analysis of similarity (ANOSIM; Supplementary Fig. 4, insets; Methods). The four small size classes, 0-0.2 µm, 0.22-1.6/3 µm, 0.8-5 µm, and 5-20 µm (Supplementary Fig. 4a-d) were more consistent with Longhurst BGCPs. In contrast, for the two larger size fractions 20-180 µm and 180-2000 µm, the three biogeographical divisions were not strongly different within the ANOSIM (Supplementary Fig. 4e-f).

From an oceanographic point of view, plankton should be quasi-neutrally redistributed (i.e., homogenized) by currents and their biogeography should follow the structure of the main recirculations, within a range of physiologically compatible temperatures. In this point of view, our results are consistent with the large-scale geographic distributions found by Hellweger *et al*.[4] using a neutral model.

Supplementary Information 4. Differences in genomic province sizes among organismal size fractions

Globally, we obtained more numerous, smaller genomic provinces in the smaller size fractions and fewer, larger genomic provinces in the larger size fractions (Supplementary Fig. 4, Supplementary Fig. 7). We observed a similar pattern using OTU data (Supplementary Fig. 5). Whereas smaller size fractions generally lacked geographically widespread genomic provinces containing numerous *Tara* Oceans samples, the two largest size fractions were both characterized by two very widespread genomic provinces: F5 and F8 for the 180-2000 µm size fraction, and E5 and E6 for the 20-180 µm size fraction. These large genomic provinces were latitudinally limited by the boundary between the subtropics and subpolar regions, and spanned different oceanic basins. Notably, in the Southern Hemisphere the subtropical gyres actually form a single supergyre[65] and there are almost no metabolic (mainly temperature) barriers between the northern and southern subtropical gyres (see Supplementary Fig. 4), potentially explaining genomic provinces in the 20-180 µm and 180-2000 µm size fraction that contain samples from the North and South Atlantic. For example, in the 180-2000 µm size fraction, F5 mostly covered the North and South Atlantic Oceans and adjacent systems, and F8 covered the Indo-Pacific low- and mid-latitudes. No clear correspondence existed with biogeochemical patterns (e.g., nutrient ratios), except for the clusters coinciding with upwelling systems (F3 for the California upwelling, F7 for the Chile-Peru upwelling and F2 for the Benguela upwelling system) and for the samples collected at the deep chlorophyll maximum (DCM) in the Pacific subtropical gyres (F5); this is consistent with the comparison of genomic provinces to previous biographical divisions, in which the genomic provinces of smaller size fractions were more consistent with Longhurst BGCPs, but those of larger size fractions were not (Supplementary Information 3). A bimodal zooplankton species distribution (split into subtropical and subpolar communities, with ubiquitous warm water species) was also detected by a recent study on copepod population dynamics that used alternative approaches to analyze the same metagenomic dataset[66] (see their Fig. 2). More locally, within the North Atlantic (see also Supplementary Information 6), along the northern boundary of the subtropical gyre, cold and warm copepod species overlapped because of cross-current dispersal. Nonetheless, although both cold and warm species appeared to be able to travel long distances, mixing among them was not sufficient to create a local genomic province in our data.

We interpret the difference in genomic province sizes between smaller and larger size fractions as the result of various factors. Plankton smaller than 20 µm (femto-, pico- and nanoplankton), which represent most of the prokaryotic and eukaryotic phototrophs[18,19], are sensitive to a suite of environmental factors (i.e., temperature[67], nutrients and trace elements[10]; see also Supplementary Fig. 7) and generally have a shorter life cycle, together leading to faster fluctuations in their relative abundance in the communities we sampled. In contrast, larger plankton have longer life cycles and, if they are predators that are not strongly selective in their feeding, or are photosymbiotic hosts capable of partnering with multiple different symbionts, may cope with local fluctuations in environmental

conditions. Therefore, they should be affected primarily by large scale, mostly latitudinal, variations in the environment, leading to larger genomic provinces, whereas smaller plankton are grouped into smaller provinces more influenced by local environmental conditions. Overall, this difference in biogeography suggests a size-based decoupling between smaller and larger plankton (which may also extend to nekton such as tuna and billfish[68]), with implications for the structure and function of oceanic food webs and other types of biotic interactions.

Supplementary Information 5. Genomic provinces as stable ecological continua

As plankton communities are transported by ocean currents, they change over time due to the various processes that occur in the context of the seascape: variations in temperature, light and nutrients (where changes in the latter may also be induced by plankton communities), intra- and inter-individual and species biological interactions, and mixing with neighboring water masses. Thus, a continuum of composition among nearby samples is expected as a natural consequence of community turnover within the seascape over time. We observed the effects of continuous turnover in our biogeographical analyses (Fig. 1a, Supplementary Fig. 4, Supplementary Fig. 5, Supplementary Information 2) in which nearby samples often reflected gradual, but not complete changes in community composition.

We measured the time window of transport by currents separating two samples during which the changes in their community composition were maximally correlated with travel time, resulting in a global average of T_{min} < roughly 1.5 years. This represents the travel time during which predictable continuous turnover occurs in our dataset. Notably, T_{min} does not necessarily define the turnover rate itself which depends on how strongly different seascape processes affect communities with differing biological characteristics (see Supplementary Information 6).

The global ocean current system is composed of a series of large-scale main currents and associated recirculations (which are also referred to as gyres). Therefore, we present the following hypothesis as a potential explanation of our results: the average global timescale of 1.5 years is comparable to the crossing time of an ocean gyre (i.e., the amount of time it takes a water parcel to travel from one side of a gyre to the other), e.g., to cross the North Atlantic basin while riding the Gulf Stream system. This time scale of 1.5 years is probably an underestimate, since our sparse sampling did not cover all current systems. Within different systems, the transport by main currents leads to stable, continuous patterns of changes in community structure and nutrient concentrations, and also explains how temporally stable genomic provinces can exist in the face of ocean circulation. Within each system we have thus to expect that a community turnover is long enough to allow for this long range predictability due to smooth, continuous changes. Significant heterogeneity in environmental conditions among different circulation patterns means that moving from system to another (and therefore, in our case here, beyond the 1.5 year timescale; Supplementary Fig. 9c-f) disrupts the interlinked relationship among local seascape processes, leading to a global delimitation into separate ecological continua among different gyre-scale current systems.

Supplementary Information 6. Community turnover in the North Atlantic

In order to characterize the impact of physical and biological processes on changes in metagenomic composition during travel along currents, we focused on the well-known current systems crossing the North Atlantic into the Mediterranean Sea (the Gulf Stream and other currents around the subtropical gyre[20,69–71]; Supplementary Fig. 10a). Across this region, the piconanoplankton (0.8-5 µm) were split into three genomic provinces, C5, C8 and C3, each less than 5,000 km wide (~11 months of travel time; Supplementary Fig. 4c). In contrast, mesoplankton (180-2000 µm) biogeography corresponded to a single province, F5, spanning from the Caribbean to Cyprus (> 9,700 km or ~18 months of travel time; Supplementary Fig. 4f; see also Supplementary Information 4). Metagenomic dissimilarity and T_{min} were strongly correlated within the region (Spearman's ρ between 0.44 and 0.86 depending on size

fraction, Supplementary Fig. 10b-e), which allowed us to explore the relationship of genomic province size, ocean transport and plankton community turnover over scales from months to years. We calculated metagenomic turnover times as e-folding times based on an exponential fit of metagenomic dissimilarity to T_{min} (ranging from a few months to a few years, Methods). The metagenomic turnover time of smaller plankton (< 20 µm) was approximately one year. In contrast, for the larger size fractions, the metagenomic turnover time was approximately two years, suggesting that a lower turnover rate for larger plankton may explain their geographically larger genomic provinces.

We note that our results on metagenomic turnover time appear different from a recently published study that also calculated turnover rates for plankton, which found faster rates for larger organisms[8]. This may be explained by two significant differences between our approach and theirs: first, their measurements of β-diversity were based on presence/absence (Jaccard) comparisons among either morphological species or OTUs, whereas our calculations of turnover time above were based on metagenomic sequences. As described above (Supplementary Information 1), there are significant differences in resolution between OTU-based and metagenomic data, and we would expect similar differences in resolution between organismal observation data and metagenomic sequences. In fact, due to these differences in resolution, our estimates of metagenomic time based on OTU rather than metagenomic data show a similar trend to those of Villarino *et al*.[8] (Supplementary Fig. 10f-i). Second, their turnover rates were calculated separately for individual plankton groups (the 9 main groups were prokaryotes, coccolithophores, dinoflagellates, diatoms, all microbial eukaryotes, gelatinous zooplankton, mesozooplankton, macrozooplankton and myctophids), whereas our metagenomic data represent samples of the full plankton community within each size fraction. Among these, several groups (e.g., dinoflagellates or mesozooplankton) would be expected to be found across multiple *Tara* Oceans size fractions, blurring potential comparisons. Thus, our study and Villarino *et al*. calculated rates of change using broadly similar approaches, but based on very different underlying biological substrates.

III. Etude de l'impact du changement climatique sur les communautés planctoniques

En nous basant sur les résultats du manuscrit présenté dans la partie II de ce chapitre, nous avons mis en place un projet collaboratif avec le LSCE (Laboratoire des Sciences du Climat et de l'Environnement) visant à analyser l'impact des changements climatiques à venir sur les communautés planctoniques.

J'ai ainsi réalisé une étude préliminaire étudiant les différences de températures marines entre la période actuelle et la fin du siècle sur la base de modèles fournis par le LSCE, avant de mettre en relation ces informations avec les paramètres environnementaux de *Tara* Oceans et les provinces génomiques déterminées préalablement.

Les valeurs utilisées pour cette étude sont des moyennes annuelles multi-modèles (provenant de dix simulations) de températures de surface correspondant aux données de Bopp et al, 2013[141]. Nous avons ainsi deux sets de températures : les valeurs historiques, moyenne des années 1990 à 1999, et les valeurs de la fin du 21ième siècle cent ans plus tard, moyenne de 2090 à 2099 (Figure 23). Ces données prévisionnelles correspondent au scénario climatique RCP 8.5 (*Representative Concentration Pathway*), le scénario le plus pessimiste dans lequel les émissions de carbone atmosphériques continuent d'augmenter avec les années plutôt que de se stabiliser ou de diminuer. Les montées de température observées avec ce scénario sont celles qui seraient obtenues si aucune mesure particulière n'était prise et que les émissions de carbone continuaient de progresser de la même manière que ces dernières années.

Nous avons ainsi commencé par comparer les valeurs moyennes "historiques" de la fin des années 1990 avec les valeurs *in situ* relevées par l'expédition *Tara* Oceans, quelques années plus tard afin d'étudier la cohérence de nos données théoriques avec celles environnementales et donc avec nos communautés planctoniques (Figure 24). Pour cela nous avons récupéré les températures aux positions géographiques des échantillons, comparant ainsi les valeurs pour 99 échantillons de surface. D'après un test de Wilcoxon comparant la distribution de ces deux sets, il n'y a pas de différence significative entre ces valeurs (p-value = 0.52). En revanche, si l'on compare les températures de la fin du 21ème siècle avec celles de 1990 ou celles de *Tara*, on

observe une distribution significativement plus élevée (p-value de $3.95x10^{-18}$ et $3.5x10^{-15}$ respectivement).

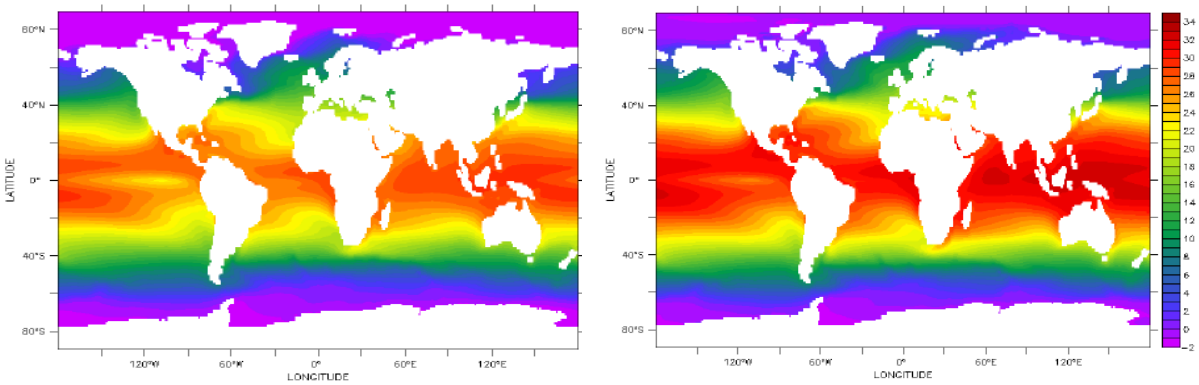


Figure 23 : Cartes des valeurs de température de surface moyennes en degrés Celsius passées (1990-1999) et futures (2090-2099) provenant de Bopp et al, 2013. Figure réalisée avec Ferret[142].

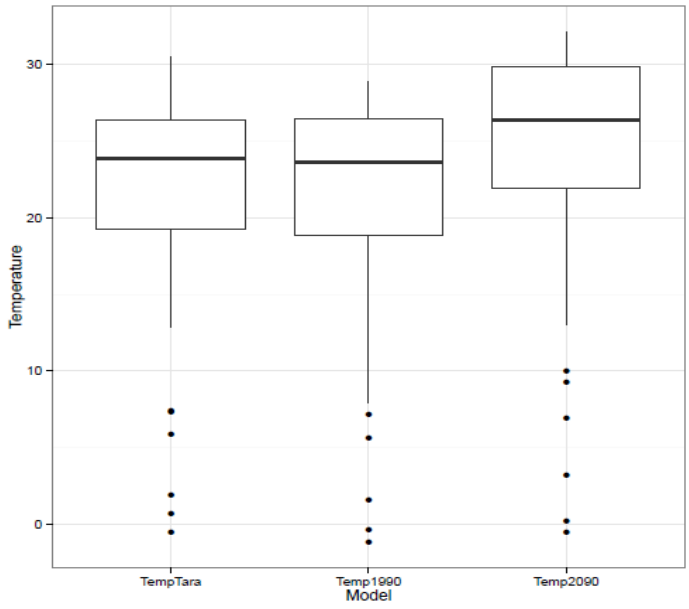


Figure 24 : Boxplot de la distribution des températures de surface provenant des échantillons *Tara* Oceans (à gauche), des températures passées (au centre) et futures (à droite) provenant du modèle de Bopp et al.

Afin d'avoir une meilleure idée des variations de température à cent ans sur les points d'échantillonnage, nous avons représenté sur une carte la différence de température entre 2090 et 1990 pour chacun de ces points (Figure 25). Nous pouvons ainsi observer que l'ensemble des lieux étudiés subit une augmentation de température allant de 0.5°C dans les stations australes à 4°C en Méditerranée ou à l'est de l'Argentine. Globalement, les augmentations de température semblent plus intenses

sur les échantillons à l'équateur et dans l'hémisphère nord que dans l'hémisphère sud. Cet effet est bien connu, il est dû à la plus forte densité en continent sur l'hémisphère nord ce qui diminue l'effet tampon des volumes d'eau océaniques.

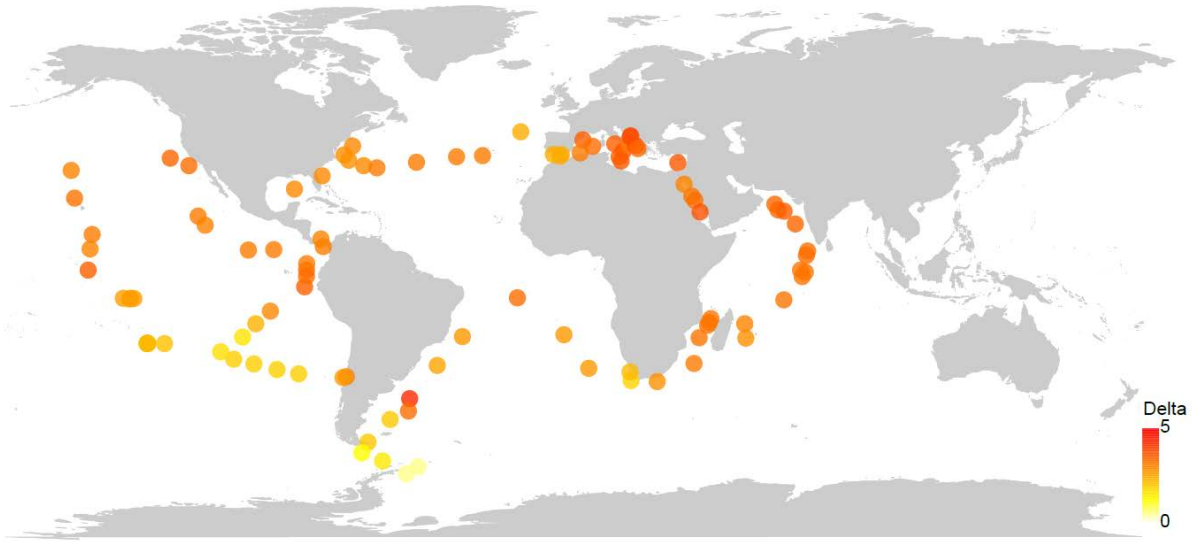


Figure 25 : Carte des stations *Tara* Oceans représentées par des points de couleurs. Le gradient de couleur correspond à la variation de température subie entre 1990 et 2090.

Nous pouvons maintenant nous intéresser à l'échelle des communautés. En effet, on peut faire la supposition qu'une province génomique étant composée d'organismes génétiquement proches, ceux d'un échantillon pourraient s'adapter et survivre dans les conditions environnementales d'un autre. Partant de ce principe, nous avons voulu examiner les variations de température qui seront subies par les espèces en prenant en compte l'ensemble de chaque province comme référence.

Pour cela, nous avons noté la température maximum actuelle de chaque province, tous échantillons confondus, et déterminé la différence avec la température future de chaque site qui a été échantillonné. Nous avons ensuite refait le calcul avec la température moyenne au lieu de maximum (Figure 26). Les résultats montrent qu'en 2100, considérant les valeurs maximums, entre 59 et 75% des sites de prélèvements seront à des températures plus chaudes que le maximum connu actuellement dans l'ensemble de leur province génomique correspondante. Considérant la moyenne connue par la province, entre 79 et 85% des sites seront à des températures plus chaudes. Ceci signifie que les communautés seront exposées à des températures

qu'elles n'ont encore jamais atteintes, même en prenant en compte l'ensemble de leur province génomique.

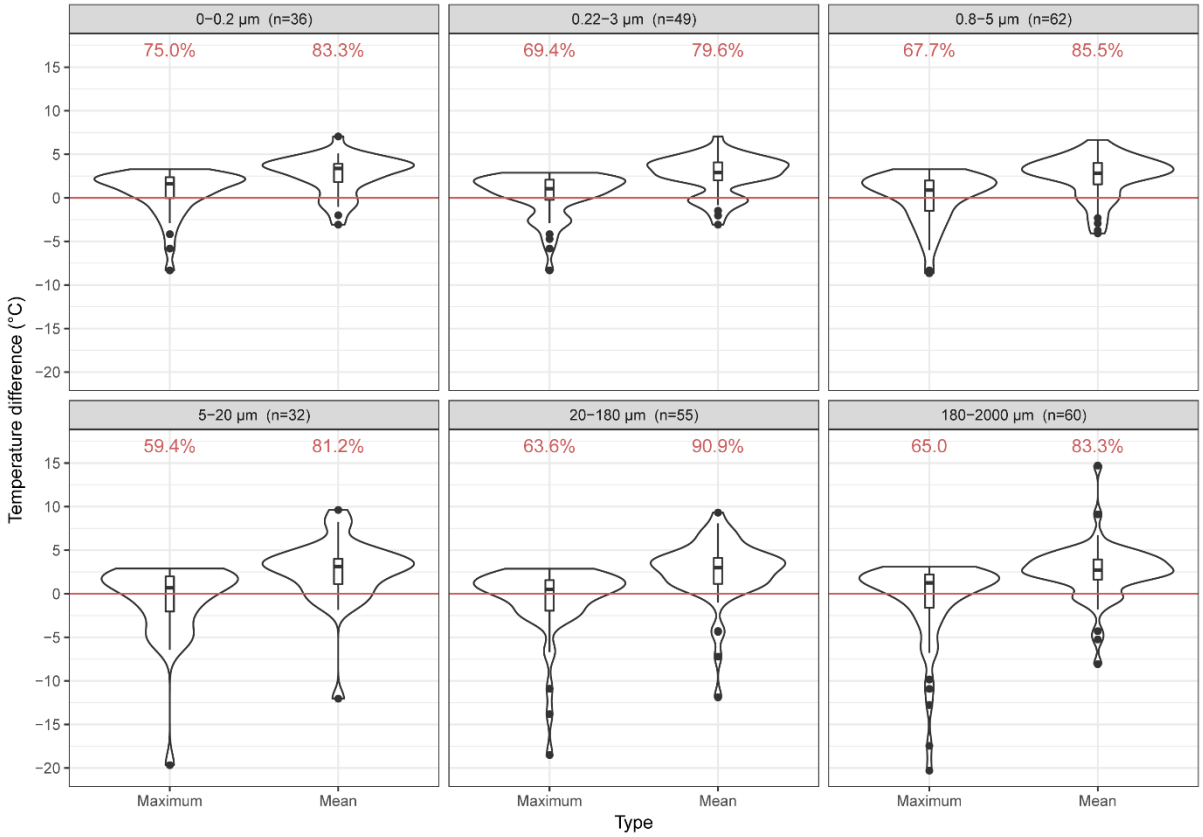


Figure 26 : Violin plot des différences de température, entre la valeur de température future trouvée dans un échantillon en 2100 et le maximum (gauche de chaque case) ou la moyenne (droite de chaque case) actuelle trouvée dans une province génomique. Chaque case correspond à une fraction de taille, avec le nombre d'échantillons pris en compte (n). La ligne rouge correspond à une valeur de zéro, c'est à dire le seuil où la température future est égale à celle actuelle. Le chiffre en rouge correspond au pourcentage d'échantillons qui sont au-dessus de ce seuil, et donc qui atteindront des températures plus hautes que ce que connait actuellement l'ensemble de leur province.

IV. Conclusion

Après l'étude du phytoplancton, nous avons pu nous intéresser à une large et diverse famille d'hétérotrophes, les straménopiles, qui ont révélé des distributions notamment liées à la température des océans, de la même manière que les Mamiellales, ce qui a confirmé que ce schéma pouvait être valable pour des espèces très variées.

Afin d'élargir encore notre analyse, nous sommes donc passés à l'échelle des communautés, observant ainsi aussi bien les communautés virales que les grands organismes planctoniques. Encore une fois, nous avons pu constater que la température de l'eau était un facteur majeur de leur distribution, suivie par les nutriments. Les résultats obtenus sont cohérents avec les analyses métagénomiques et de métabarcodes précédemment réalisées à l'échelle des espèces.

Nous avons enfin pu utiliser les provinces génomiques définies dans cette étude pour pousser l'analyse des températures vers l'avenir, dans un contexte de changement climatique. En observant l'évolution prévue des températures de surface, nous avons pu conclure que malgré l'hétérogénéité géographique du réchauffement climatique, l'ensemble des communautés actuelles allaient devoir faire face à un réchauffement de l'eau.

Conclusions générales et perspectives

Durant ce projet de thèse, nous avons étudié différentes lignées de plancton, surtout des picoalgues eucaryotes, afin d'appréhender à la résolution des génomes et au travers d'études de métagénomique l'impact de l'environnement. Pour cela, nous sommes passés par des analyses à plusieurs échelles très différentes et complémentaires, allant de l'étude des variants génomiques chez *Bathycoccus prasinos* à une étude globale des communautés en passant par la distribution géographique de quelques espèces d'intérêt.

Dans les deux premiers chapitres de ce manuscrit, nous avons tout d'abord réalisé un inventaire relativement complet de la biogéographie des Mamiellales, des organismes phytoplanctoniques ayant un fort impact écologique notamment par leur abondance et leur aspect cosmopolite. Nous nous sommes basés sur les données du projet *Tara* Oceans comportant un grand nombre d'échantillons de métabarcodes et métagénomiques provenant de tous les océans, et avons ainsi pu montrer une séparation entre les espèces de ce groupe fortement liée à la température de l'eau, ainsi que pour certaines à la profondeur et à la luminosité. Dans l'ensemble, ces organismes sont en effet abondants et trouvés dans de nombreux bassins mais restent peu présents dans les milieux oligotrophes tels que l'océan Pacifique.

Afin de compléter cette étude, il serait évidemment avantageux de pouvoir analyser un plus grand nombre de points géographiques sur plusieurs mois, par exemple l'expédition a peu échantillonné près des côtes et n'a pas effectué de séries temporelles sur une même station. Il est en effet connu que certains Mamiellales peuvent être impactés par les variations saisonnières, ce qui est difficile à estimer ici avec le type de données utilisé. Cependant l'étude *Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems* à laquelle j'ai contribué suggère fortement qu'à l'échelle des bassins, la structure de la biogéographie n'est pas dépendante de la saisonnalité. Cette dernière, combinée à d'autres effets qui seraient locaux contribuerait surtout à faire varier les abondances relatives mais de façon insuffisante pour modifier des structures en provinces qui sont surtout dépendantes des courants et de la température.

Il serait également intéressant de travailler sur la corrélation entre les Mamiellales et les prasinovirus, un des groupes de virus marins très abondants[143] qui

ciblent en particulier nos organismes d'intérêt. Ces virus sont espèce-spécifiques[144] et trois groupes majeurs sont connus pour infecter en particulier les très abondants *Ostreococcus*, *Micromonas* et *Bathycoccus*, pouvant ainsi avoir un impact sur leur biogéographie. Il a par exemple été suggéré que ces virus seraient responsables des transitions saisonnières entre *Micromonas* et *Bathyoccus* dans l'océan Arctique[56], et il pourrait donc être enrichissant pour cette étude de prendre la présence des prasinovirus en compte dans la description de l'environnement des Mamiellales.

Le chapitre suivant portant sur la génomique des populations de *Bathycoccus prasinos* a très clairement permis d'observer la séparation de cette espèce en trois populations distinctes, celle des eaux tempérées et deux populations des eaux froides dans les océans Arctique et Austral. On a ici le cas d'une adaptation à des milieux de températures drastiquement différentes qui a structuré ces populations. Il existe des variants liés aux deux environnements pouvant coexister mais à des fréquences différentes, un allèle prenant le dessus sur l'autre selon la température jusqu'à éventuellement faire parfois disparaitre l'allèle mineur. La population froide ayant plus de polymorphisme sur ces positions clés, il est possible que les allèles adaptés au milieu froid ne soient par exemple pas conservés dans les eaux tempérées, ou qu'à l'inverse le passage vers une eau plus froide entraine des mutations. Il serait intéressant d'effectuer des expériences en laboratoire en plaçant des souches adaptées à un milieu dans l'environnement opposé afin d'observer les variations de fréquences alléliques de ces positions d'intérêt.

En étudiant les variations d'acides aminés les plus marquées en fonction de la température, nous avons sélectionné six gènes présentant treize mutations, avec une version "froide" trouvée dans les eaux arctiques et australes et une version "chaude" trouvée dans les autres échantillons. Si nous avons une idée de la fonction de certains de ces gènes, nous n'avons que peu d'informations sur l'impact de ces changements sur les protéines associées. Nous pourrions notamment nous intéresser plus en détail à la structure tridimensionnelle de ces deux versions, et encore une fois mener des expériences en culture en ciblant ces gènes d'intérêt et leurs variations dans différents milieux.

Nous avons ici concentré nos analyses sur *Bathycoccus prasinos* en raison de sa forte abondance dans des environnements très variés, mais nous pourrions également étudier la structure des populations des autres Mamiellales, *Bathycoccus*

TOSAG39-1 ou *Micromonas commoda* étant par exemple présents dans un grand nombre de stations tempérées. Il est possible que des espèces telles que *Micromonas polaris* présentent également différents patterns selon les bassins Arctiques et ces structures pourraient ensuite être comparées entre les Mamiellales.

Le dernier chapitre de cette thèse s'éloigne de notre groupe phytoplanctonique de référence pour d'abord inclure des organismes hétérotrophes, les straménopiles, et reproduire les analyses de corrélation à l'environnement. Nous avons ainsi pu observer un modèle similaire aux Mamiellales, avec des espèces d'une même famille trouvées à différentes températures.

Finalement, nous sommes passés de l'échelle des espèces à celle des communautés, nous basant toujours sur nos échantillons métagénomiques mais incluant ici des tailles d'organisme beaucoup plus diverses, et prenant en compte l'ensemble de chaque échantillon environnemental. Nous avons ainsi pu confirmer qu'une fois de plus, la température de l'eau était un facteur majeur influant sur la distribution du plancton, et ce pour toutes les fractions de taille sans exception. Cependant nous avons pu observer que les communautés ne présentaient pas exactement la même répartition géographique selon les fractions, ce qui montre l'importance d'étudier ici les organismes selon leur taille. Il serait comme nous l'avons évoqué plus tôt pour les Mamiellales intéressant d'analyser les interactions entre les différentes espèces, les communautés virales par exemple étant typiquement connues pour avoir un impact sur les autres[145,146]. Des analyses supplémentaires étudiant l'impact potentiel des fractions les unes sur les autres pourraient donc compléter la description environnementale présentée.

Pour compléter cette étude, nous avons brièvement étudié l'évolution prévue dans les cent prochaines années de la température de l'eau aux points de prélèvement, prenant en compte l'ensemble des températures connues par les communautés, et avons pu observer que chacune d'entre elles rencontrerait des températures de surface plus élevées que celles trouvées aujourd'hui. Pour poursuivre ces analyses se pose évidemment la question de la manière dont les espèces vont réagir, s'adaptant à leur nouveau milieu ou présentant une nouvelle répartition géographique. Il serait donc intéressant d'envisager des analyses plus poussées sur les niches écologiques, visant à prévoir les futures provinces génomiques et les comparant à celles connues à l'heure actuelle.

Globalement, cette thèse montre le très fort lien entre le plancton et son environnement par des approches de génomique, principalement chez les Mamiellales analysés à plusieurs échelles, mais aussi en s'intéressant à d'autres espèces pour avoir un point de comparaison, puis aux communautés dans leur ensemble. La température étant systématiquement un facteur clé, il est important dans les recherches à venir de s'intéresser à l'impact du changement climatique sur ces organismes notamment à l'aide de modèles prédictifs et d'expériences complémentaires en laboratoire. Enfin de nombreuses questions restent à explorer, telles que les points de similarité entre l'océan arctique et le bassin austral qui bien que très distants montrent chez *Bathycoccus* quelques traits d'évolution communs. D'autres espèces telles que le copépode *Oithona similis* sont également connues pour leurs ressemblance entre les deux milieux[147]. L'étude des données métatranscriptomiques pourraient également très bien compléter nos analyses, ajoutant les variations d'expression des gènes aux mutations comme réponse des espèces aux différents environnements.

Références

1. Lalli, C. & Parsons, T. R. *Biological Oceanography: An Introduction*. (Elsevier, 1997).
2. Vargas, C. de *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, (2015).
3. Burki, F. The Eukaryotic Tree of Life from a Global Phylogenomic Perspective. *Cold Spring Harb. Perspect. Biol.* **6**, (2014).
4. Worden, A. Z. *et al.* Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science* **347**, (2015).
5. Ochoa, J., Maske, H., Sheinbaumc, J. & Candela, J. Diel and lunar cycles of vertical migration extending to below 1000 m in the ocean and the vertical connectivity of depth-tiered populations. *Limnol. Oceanogr.* **58**, 1207–1214 (2013).
6. Fenchel, T. Marine Plankton Food Chains. *Annu Rev Ecol Syst 19* 19–38 (1988).
7. Johnson, M. D. Inducible Mixotrophy in the Dinoflagellate Prorocentrum minimum. *J. Eukaryot. Microbiol.* **62**, 431–443 (2015).
8. Rottberger, J., Gruber, A., Boenigk, J. & Kroth, P. G. Influence of nutrients and light on autotrophic, mixotrophic and heterotrophic freshwater chrysophytes. *Aquat. Microb. Ecol.* **71**, 179–191 (2013).
9. Weber, M. X. & Medina, M. Chapter Four - The Role of Microalgal Symbionts (Symbiodinium) in Holobiont Physiology. in *Advances in Botanical Research* (ed. Piganeau, G.) vol. 64 119–140 (Academic Press, 2012).
10. Bidle, K. D. & Falkowski, P. G. Cell death in planktonic, photosynthetic microorganisms. *Nat. Rev. Microbiol.* **2**, 643–655 (2004).
11. Le Quéré, C. *et al.* Global carbon budget 2014. *Earth Syst. Sci. Data* **7**, 47–85 (2015).
12. Stone, L. & Weisburd, R. S. Positive feedback in aquatic ecosystems. *Trends Ecol. Evol.* **7**, 263–267 (1992).
13. Fenchel, T. & Finlay, B. Oxygen and the spatial structure of microbial communities. *Biol. Rev. Camb. Philos. Soc.* **83**, 553–569 (2008).
14. Petrou, K. *et al.* Acidification diminishes diatom silica production in the Southern Ocean. *Nat. Clim. Change* **9**, 781–786 (2019).
15. Pittman, S. J. *Seascape Ecology*. (Wiley-Blackwell, 2017).
16. Arteaga, L., Pahlow, M. & Oschlies, A. Global patterns of phytoplankton nutrient and light colimitation inferred from an optimality-based model. *Glob. Biogeochem. Cycles* **28**, 648–661 (2014).
17. Lewandowska, A. M., Hillebrand, H., Lengfellner, K. & Sommer, U. Temperature effects on phytoplankton diversity — The zooplankton link. *J. Sea Res.* **85**, 359–364 (2014).
18. Toseland, A. *et al.* The impact of temperature on marine phytoplankton resource allocation and metabolism. *Nat. Clim. Change* **3**, 979–984 (2013).
19. Martin, J. H. & Fitzwater, S. E. Iron deficiency limits phytoplankton growth in the north-east Pacific subarctic. *Nature* **331**, 341–343 (1988).
20. Herndl, G. J. & Reinthaler, T. Microbial control of the dark end of the biological pump. *Nat. Geosci.* **6**, 718–724 (2013).

21. von Dassow, P. & Montresor, M. Unveiling the mysteries of phytoplankton life cycles: patterns and opportunities behind complexity. *J. Plankton Res.* **33**, 3–12 (2011).

22. Smith, J. M. Optimization Theory in Evolution. *Annu. Rev. Ecol. Syst.* **9**, 31–56 (1978).

23. Tibayrenc, M., Kjellberg, F. & Ayala, F. J. A clonal theory of parasitic protozoa: the population structures of Entamoeba, Giardia, Leishmania, Naegleria, Plasmodium, Trichomonas, and Trypanosoma and their medical and taxonomical consequences. *Proc. Natl. Acad. Sci.* **87**, 2414–2418 (1990).

24. Dunthorn, M. & Katz, L. A. Secretive ciliates and putative asexuality in microbial eukaryotes. *Trends Microbiol.* **18**, 183–188 (2010).

25. Derelle, E. *et al.* Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11647–11652 (2006).

26. Dunthorn, M. *et al.* Meiotic Genes in Colpodean Ciliates Support Secretive Sexuality. *Genome Biol. Evol.* **9**, 1781–1787 (2017).

27. Hofstatter, P. G., Brown, M. W. & Lahr, D. J. G. Comparative Genomics Supports Sex and Meiosis in Diverse Amoebozoa. *Genome Biol. Evol.* **10**, 3118–3128 (2018).

28. Kraus, D. *et al.* Putatively asexual chrysophytes have meiotic genes: evidence from transcriptomic data. *PeerJ* **6**, (2019).

29. Patil, S. *et al.* Identification of the meiotic toolkit in diatoms and exploration of meiosis-specific SPO11 and RAD51 homologs in the sexual species Pseudo-nitzschia multistriata and Seminavis robusta. *BMC Genomics* **16**, (2015).

30. Weedall, G. D. & Hall, N. Sexual reproduction and genetic exchange in parasitic protists. *Parasitology* **142**, S120–S127 (2015).

31. Chi, J., Mahé, F., Loidl, J., Logsdon, J. & Dunthorn, M. Meiosis Gene Inventory of Four Ciliates Reveals the Prevalence of a Synaptonemal Complex-Independent Crossover Pathway. *Mol. Biol. Evol.* **31**, 660–672 (2014).

32. Ramesh, M. A., Malik, S.-B. & Logsdon, J. M. A Phylogenomic Inventory of Meiotic Genes: Evidence for Sex in Giardia and an Early Eukaryotic Origin of Meiosis. *Curr. Biol.* **15**, 185–191 (2005).

33. Blanc-Mathieu, R. *et al.* Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Sci. Adv.* **3**, e1700239 (2017).

34. Grimsley, N., Péquin, B., Bachy, C., Moreau, H. & Piganeau, G. Cryptic Sex in the Smallest Eukaryotic Marine Green Alga. *Mol. Biol. Evol.* **27**, 47–54 (2010).

35. Speijer, D., Lukeš, J. & Eliáš, M. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proc. Natl. Acad. Sci.* (2015) doi:10.1073/pnas.1501725112.

36. Goodenough, U. & Heitman, J. Origins of Eukaryotic Sexual Reproduction. *Cold Spring Harb. Perspect. Biol.* **6**, a016154 (2014).

37. Dacks, J. & Roger, A. J. The First Sexual Lineage and the Relevance of Facultative Sex. *J. Mol. Evol.* **48**, 779–783 (1999).

38. Marin, B. & Melkonian, M. Molecular Phylogeny and Classification of the Mamiellophyceae class. nov. (Chlorophyta) based on Sequence Comparisons of the Nuclear- and Plastid-encoded rRNA Operons. *Protist* **161**, 304–336 (2010).

39. Monier, A., Worden, A. Z. & Richards, T. A. Phylogenetic diversity and biogeography of

the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ. Microbiol. Rep.* **8**, 461–469 (2016).

40. Vaulot, D. *et al.* Metagenomes of the Picoalga Bathycoccus from the Chile Coastal Upwelling. *PLOS ONE* **7**, e39648 (2012).
41. Simmons, M. P. *et al.* Abundance and Biogeography of Picoprasinophyte Ecotypes and Other Phytoplankton in the Eastern North Pacific Ocean. *Appl. Environ. Microbiol.* **82**, 1693–1705 (2016).
42. Not, F. *et al.* A single species, Micromonas pusilla (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. *Appl. Environ. Microbiol.* **70**, 4064–4072 (2004).
43. Treusch, A. H. *et al.* Phytoplankton distribution patterns in the northwestern Sargasso Sea revealed by small subunit rRNA genes from plastids. *ISME J.* **6**, 481–492 (2012).
44. Countway, P. D. & Caron, D. A. Abundance and distribution of Ostreococcus sp. in the San Pedro Channel, California, as revealed by quantitative PCR. *Appl. Environ. Microbiol.* **72**, 2496–2506 (2006).
45. Lovejoy, C. *et al.* Distribution, phylogeny, and growth of cold−adapted picoprasinophytes in Arctic SeaS1. *J. Phycol.* **43**, 78–89 (2007).
46. Simon, N. *et al.* Revision of the Genus Micromonas Manton et Parke (Chlorophyta, Mamiellophyceae), of the Type Species M. pusilla (Butcher) Manton & Parke and of the Species M. commoda van Baren, Bachy and Worden and Description of Two New Species Based on the Genetic and Phenotypic Characterization of Cultured Isolates. *Protist* **168**, 612–635 (2017).
47. Chrétiennot-Dinet, M.-J. *et al.* A new marine picoeucaryote: Ostreococcus tauri gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia* **34**, 285–292 (1995).
48. Subirana, L. *et al.* Morphology, genome plasticity, and phylogeny in the genus ostreococcus reveal a cryptic species, O. mediterraneus sp. nov. (Mamiellales, Mamiellophyceae). *Protist* **164**, 643–659 (2013).
49. Vannier, T. *et al.* Survey of the green picoalga Bathycoccus genomes in the global ocean. *Sci. Rep.* **6**, 1–11 (2016).
50. Leliaert, F., Verbruggen, H. & Zechman, F. W. Into the deep: New discoveries at the base of the green plant phylogeny. *BioEssays* **33**, 683–692 (2011).
51. Foulon, E. *et al.* Ecological niche partitioning in the picoplanktonic green alga Micromonas pusilla: evidence from environmental surveys using phylogenetic probes. *Environ. Microbiol.* **10**, 2433–2443 (2008).
52. Guillou, L. *et al.* Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. *Protist* **155**, 193–214 (2004).
53. Rodríguez, F. *et al.* Ecotype diversity in the marine picoeukaryote Ostreococcus (Chlorophyta, Prasinophyceae). *Environ. Microbiol.* **7**, 853–859 (2005).
54. Demir-Hilton, E. *et al.* Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *ISME J.* **5**, 1095–1107 (2011).
55. Tragin, M. & Vaulot, D. Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Sci. Rep.* **9**, 1–14 (2019).

56. Joli, N., Monier, A., Logares, R. & Lovejoy, C. Seasonal patterns in Arctic prasinophytes and inferred ecology of Bathycoccus unveiled in an Arctic winter metagenome. *ISME J.* **11**, 1372–1385 (2017).

57. Limardo, A. J. *et al.* Quantitative biogeography of picoprasinophytes establishes ecotype distributions and significant contributions to marine phytoplankton. *Environ. Microbiol.* **19**, 3219–3234 (2017).

58. Majaneva, M., Enberg, S., Autio, R., Blomster, J. & Rintala, J. -m. Mamiellophyceae shift in seasonal predominance in the Baltic Sea. *Aquat. Microb. Ecol.* **83**, 181–187 (2019).

59. Belevich, T. A. *et al.* Photosynthetic Picoeukaryotes in the Land-Fast Ice of the White Sea, Russia. *Microb. Ecol.* **75**, 582–597 (2018).

60. Vaulot, D., Eikrem, W., Viprey, M. & Moreau, H. The diversity of small eukaryotic phytoplankton (< or =3 microm) in marine ecosystems. *FEMS Microbiol. Rev.* **32**, 795–820 (2008).

61. Worden, A. Z. & Not, F. Ecology and Diversity of Picoeukaryotes. in *Microbial Ecology of the Oceans* (ed. Kirchman, D. L.) 159–205 (John Wiley & Sons, Inc., 2008). doi:10.1002/9780470281840.ch6.

62. Worden, A., Nolan, J. & Palenik, B. Assessing the Dynamics and Ecology of Marine Picophytoplankton: The Importance of the Eukaryotic Component. *Limnol. Oceanogr.* **49**, 168–179 (2004).

63. Li, W. K. W., McLaughlin, F. A., Lovejoy, C. & Carmack, E. C. Smallest Algae Thrive As the Arctic Ocean Freshens. *Science* **326**, 539–539 (2009).

64. Monier, A. *et al.* Oceanographic structure drives the assembly processes of microbial eukaryotic communities. *ISME J.* **9**, 990–1002 (2015).

65. Moreau, H. *et al.* Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol.* **13**, R74 (2012).

66. Worden, A. Z. *et al.* Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science* **324**, 268–272 (2009).

67. Yau, S. *et al.* A Viral Immunity Chromosome in the Marine Picoeukaryote, Ostreococcus tauri. *PLOS Pathog.* **12**, e1005965 (2016).

68. Jancek, S., Gourbière, S., Moreau, H. & Piganeau, G. Clues about the genetic basis of adaptation emerge from comparing the proteomes of two Ostreococcus ecotypes (Chlorophyta, Prasinophyceae). *Mol. Biol. Evol.* **25**, 2293–2300 (2008).

69. Lee, S. C., Ni, M., Li, W., Shertz, C. & Heitman, J. The Evolution of Sex: a Perspective from the Fungal Kingdom. *Microbiol. Mol. Biol. Rev. MMBR* **74**, 298–340 (2010).

70. Grimsley, N., Yau, S., Piganeau, G. & Moreau, H. Typical Features of Genomes in the Mamiellophyceae. in *Marine Protists: Diversity and Dynamics* (eds. Ohtsuka, S., Suzaki, T., Horiguchi, T., Suzuki, N. & Not, F.) 107–127 (Springer Japan, 2015). doi:10.1007/978-4-431-55130-0_6.

71. Okie, J. G. General models for the spectra of surface area scaling strategies of cells and organisms: fractality, geometric dissimilitude, and internalization. *Am. Nat.* **181**, 421–439 (2013).

72. Dupuy, C. *et al.* Feeding rate of the oyster Crassostrea gigas in a natural planktonic

community of the Mediterranean Thau Lagoon. *Mar. Ecol. Prog. Ser.* **205**, 171–184 (2000).

73. Šlapeta, J., López-García, P. & Moreira, D. Global Dispersal and Ancient Cryptic Species in the Smallest Marine Eukaryotes. *Mol. Biol. Evol.* **23**, 23–29 (2006).
74. Valentini, A., Pompanon, F. & Taberlet, P. DNA barcoding for ecologists. *Trends Ecol. Evol.* **24**, 110–117 (2009).
75. Bucklin, A., Lindeque, P. K., Rodriguez-Ezpeleta, N., Albaina, A. & Lehtiniemi, M. Metabarcoding of marine zooplankton: prospects, progress and pitfalls. *J. Plankton Res.* **38**, 393–400 (2016).
76. Piganeau, G., Eyre-Walker, A., Grimsley, N. & Moreau, H. How and Why DNA Barcodes Underestimate the Diversity of Microbial Eukaryotes. *PLOS ONE* **6**, e16342 (2011).
77. Chase, M. W. & Fay, M. F. Barcoding of Plants and Fungi. *Science* **325**, 682–683 (2009).
78. Guillou, L. *et al.* The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597-604 (2013).
79. Massana, R., Guillou, L., Terrado, R., Forn, I. & Pedrós-Alió, C. Growth of uncultured heterotrophic flagellates in unamended seawater incubations. *Aquat. Microb. Ecol.* **45**, 171–180 (2006).
80. Lindner, M. S. & Renard, B. Y. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Res.* **41**, e10–e10 (2013).
81. Hou, Y. & Lin, S. Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes. *PLOS ONE* **4**, e6978 (2009).
82. Marcy, Y. *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci.* **104**, 11889–11894 (2007).
83. Saadatpour, A., Lai, S., Guo, G. & Yuan, G.-C. Single-Cell Analysis in Cancer Genomics. *Trends Genet.* **31**, 576–586 (2015).
84. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci.* **99**, 5261–5266 (2002).
85. Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).
86. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.* **7**, 19 (2007).
87. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science* **338**, 1622–1626 (2012).
88. Gulcher, J. & Stefansson, K. Population Genomics: Laying the Groundwork for Genetic Disease Modeling and Targeting. *Clin. Chem. Lab. Med. CCLM* **36**, 523–527 (1998).
89. Laso-Jadart, R., Ambroise, C., Peterlongo, P. & Madoui, M.-A. metaVaR: introducing metavariant species models for reference-free metagenomic-based population genomics. *bioRxiv* 2020.01.30.924381 (2020) doi:10.1101/2020.01.30.924381.

90. Gruber, B., Unmack, P. J., Berry, O. F. & Georges, A. dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Mol. Ecol. Resour.* **18**, 691–699 (2018).

91. Cooke, N. P. & Nakagome, S. Fine-tuning of Approximate Bayesian Computation for human population genomics. *Curr. Opin. Genet. Dev.* **53**, 60–69 (2018).

92. Steinthorsdottir, V. *et al.* A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat. Genet.* **39**, 770–775 (2007).

93. Cao, K. *et al.* Genome-wide association study of 12 agronomic traits in peach. *Nat. Commun.* **7**, 13246 (2016).

94. Yin, L. *et al.* KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters. *Genome Biol.* **21**, 146 (2020).

95. Black IV, W. C., Baer, C. F., Antolin, M. F. & DuTeau, N. M. POPULATION GENOMICS: Genome-Wide Sampling of Insect Populations. *Annu. Rev. Entomol.* **46**, 441–469 (2001).

96. Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* **4**, 981–994 (2003).

97. Lynch, M., Gabriel, W. & Wood, A. M. Adaptive and demographic responses of plankton populations to environmental change. *Limnol. Oceanogr.* **36**, 1301–1312 (1991).

98. Lohbeck, K. T., Riebesell, U. & Reusch, T. B. H. Adaptive evolution of a key phytoplankton species to ocean acidification. *Nat. Geosci.* **5**, 346–351 (2012).

99. Reusch, T. B. H. & Boyd, P. W. Experimental Evolution Meets Marine Phytoplankton. *Evolution* **67**, 1849–1859 (2013).

100. Kashtan, N. *et al.* Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science* **344**, 416–420 (2014).

101. Sohm, J. A. *et al.* Co-occurring Synechococcus ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J.* **10**, 333–345 (2016).

102. Zwirglmaier, K. *et al.* Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. *Environ. Microbiol.* **10**, 147–161 (2008).

103. Masseret, E. *et al.* Unexpected Genetic Diversity among and within Populations of the Toxic Dinoflagellate Alexandrium catenella as Revealed by Nuclear Microsatellite Markers. *Appl. Environ. Microbiol.* **75**, 2037–2045 (2009).

104. Stuart, R. K., Brahamsha, B., Busby, K. & Palenik, B. Genomic island genes in a coastal marine Synechococcus strain confer enhanced tolerance to copper and oxidative stress. *ISME J.* **7**, 1139–1149 (2013).

105. Bibby, T. S., Mary, I., Nield, J., Partensky, F. & Barber, J. Low-light-adapted Prochlorococcus species possess specific antennae for each photosystem. *Nature* **424**, 1051–1054 (2003).

106. Blanco-Bercial, L. & Bucklin, A. New view of population genetics of zooplankton: RAD-seq analysis reveals population structure of the North Atlantic planktonic copepod Centropages typicus. *Mol. Ecol.* **25**, 1566–1580 (2016).

107. Perry, E. B. *et al.* Tumor diversity and evolution revealed through RADseq.

Oncotarget **8**, 41792–41805 (2017).

108. Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G. & Hohenlohe, P. A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **17**, 81–92 (2016).

109. Costea, P. I. *et al.* metaSNV: A tool for metagenomic strain level analysis. *PLOS ONE* **12**, e0182392 (2017).

110. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).

111. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).

112. Madoui, M.-A. *et al.* New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. *Mol. Ecol.* **26**, 4467–4482 (2017).

113. Wright, S. The Genetical Structure of Populations. *Ann. Eugen.* **15**, 323–354 (1949).

114. Nei, M. Analysis of Gene Diversity in Subdivided Populations. *Proc. Natl. Acad. Sci.* **70**, 3321–3323 (1973).

115. Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462 (1995).

116. Bonhomme, M. *et al.* Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. *Genetics* **186**, 241–262 (2010).

117. Ma, L., Ji, Y.-J. & Zhang, D.-X. Statistical measures of genetic differentiation of populations: Rationales, history and current states. *Curr. Zool.* **61**, 886–897 (2015).

118. Nielsen, R. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641–647 (2001).

119. Krasovec, M., Eyre-Walker, A., Sanchez-Ferandin, S. & Piganeau, G. Spontaneous Mutation Rate in the Smallest Photosynthetic Eukaryotes. *Mol. Biol. Evol.* **34**, 1770–1779 (2017).

120. Krasovec, M. *et al.* Fitness Effects of Spontaneous Mutations in Picoeukaryotic Marine Green Algae. *G3 Genes Genomes Genet.* **6**, 2063–2071 (2016).

121. Evans, D. W., Roberts, D. J. & Thomas, D. N. *Ernst Haeckel: Art Forms from the Abyss : Images from the HMS Challenger Expedition*. (Prestel, 2015).

122. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLOS Biol.* **5**, e77 (2007).

123. Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLOS Biol.* **5**, e16 (2007).

124. Barberán, A., Fernández-Guerra, A., Bohannan, B. J. M. & Casamayor, E. O. Exploration of community traits as ecological markers in microbial metagenomes. *Mol. Ecol.* **21**, 1909–1917 (2012).

125. Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl. Acad. Sci.* **106**, 1374–1379 (2009).

126. Falcou-Préfol, M. *et al.* Statistical Methodology for Identifying Microplastic Samples

Collected During TARA Mediterranean Campaign. in *Proceedings of the International Conference on Microplastic Pollution in the Mediterranean Sea* (eds. Cocca, M. et al.) 31–35 (Springer International Publishing, 2018). doi:10.1007/978-3-319-71279-6_5.

127. Gorsky, G. *et al.* Expanding Tara Oceans Protocols for Underway, Ecosystemic Sampling of the Ocean-Atmosphere Interface During Tara Pacific Expedition (2016–2018). *Front. Mar. Sci.* **6**, (2019).

128. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**, 1–16 (2015).

129. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *PLOS Biol.* **9**, e1001177 (2011).

130. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4**, 170093 (2017).

131. Benites, L. F., Bucchini, F., Sanchez-Brosseau, F., Grimsley, N. & Piganeau, G. Evolutionary dynamics of sex-related chromosomes at the base of the green lineage. *submitted* (2019).

132. Demory, D. *et al.* Picoeukaryotes of the Micromonas genus: sentinels of a warming ocean. *ISME J.* **13**, 132–146 (2019).

133. Leconte, J. *et al.* Genome Resolved Biogeography of Mamiellales. *Genes* **11–1**, (2020).

134. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biol.* **12**, e1001889 (2014).

135. Seeleuthner, Y. *et al.* Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nat. Commun.* **9**, 1–10 (2018).

136. Massana, R. *et al.* Phylogenetic and Ecological Analysis of Novel Marine Stramenopiles. *Appl. Environ. Microbiol.* **70**, 3528–3534 (2004).

137. Massana, R., del Campo, J., Sieracki, M. E., Audic, S. & Logares, R. Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J.* **8**, 854–866 (2014).

138. del Campo, J. & Massana, R. Emerging Diversity within Chrysophytes, Choanoflagellates and Bicosoecids Based on Molecular Surveys. *Protist* **162**, 435–448 (2011).

139. Roy, R. S. *et al.* Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci. Rep.* **4**, 4780 (2014).

140. Longhurst, A. R. *Ecological Geography of the Sea*. (Academic Press, 2007).

141. Bopp, L. *et al.* Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models. *Biogeosciences* **10**, 6225–6245 (2013).

142. NOAA's Pacific Marine Environmental Laboratory. Ferret. https://ferret.pmel.noaa.gov/Ferret/.

143. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013).

144. Clerissi, C., Desdevises, Y. & Grimsley, N. Prasinoviruses of the Marine Green Alga

Ostreococcus tauri Are Mainly Species Specific. *J. Virol.* **86**, 4611–4619 (2012).

145. Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).

146. Mojica, K. D. A. & Brussaard, C. P. D. Factors affecting virus dynamics and microbial host–virus interactions in marine environments. *FEMS Microbiol. Ecol.* **89**, 495–515 (2014).

147. Cornils, A., Wend-Heckmann, B. & Held, C. Global phylogeography of Oithona similis s.l. (Crustacea, Copepoda, Oithonidae) - A cosmopolitan plankton species or a complex of cryptic lineages? *Mol. Phylogenet. Evol.* **107**, 473–485 (2017).

Annexes

Annexe 1: Informations supplémentaires de l'article "Survey of the green picoalga Bathycoccus genomes in the global ocean"

Supplementary Information for

Survey of the green picoalga *Bathycoccus* genomes in the global ocean

Thomas Vannier[1,2,3], Jade Leconte[1,2,3], Yoann Seeleuthner[1,2,3], Samuel Mondy[1,2,3], Eric Pelletier[1,2,3], Jean-Marc Aury[1], Colomban de Vargas[4], Michael Sieracki[5], Daniele Iudicone[6], Daniel Vaulot[4], Patrick Wincker*[1,2,3] & Olivier Jaillon*[1,2,3]

[1]CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry France.

[2]CNRS, UMR 8030, CP5706, Evry France.

[3]Université d'Evry, UMR 8030, CP5706, Evry France.

[4]Sorbonne Universités, UPMC Université Paris 06, CNRS, UMR7144, Station Biologique de Roscoff, 29680 Roscoff, France.

[5]National Science Foundation, 4201 Wilson Blvd., Arlington, VA 22230, USA.

[6]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy

*Correspondence: Olivier Jaillon (ojaillon@genoscope.cns.fr) and Patrick Wincker (pwincker@genoscope.cns.fr).

Genomic data

Bathycoccus RCC1105[1] was isolated in the bay of Banyuls-sur-mer at the SOLA station at a depth of 3 m in January 2006. Sequences were downloaded from the Online Resource for Community Annotation of Eukaryotes[2]. Two metagenomes of uncultured *Bathycoccus* sorted by flow cytometry[3] were obtained from samples taken in the Eastern South Pacific Ocean at depths of 5 and 30 m (33°59′46″S, 73°22′10″W and 33°51′37″S, 73°20′24″W). Their accession numbers are CAFX01000000 and CAFY01000000. A third flow cytometry sorted metagenome[4] originated from the Deep Chlorophyll Maximum layer (DCM) at station OLIGO in the Atlantic Ocean (12°22′40″N, 27°14′27″W) with accession number AFUW01000000.

Single-cell isolation and amplification

The four cells composing the final genome sequence assembly of TOSAG39-1 (for *Tara* Oceans Single Amplified Genome from Station 39 numbered 1) originated from a sample of the *Tara* Oceans expedition, obtained in December 2009 in the Arabian Sea (18°34′52.3″N, 66°33′43.7″E) at station TARA_039 in surface (Supplementary Figure S13). Samples were preserved in 6% glycine betaine final and frozen quickly in liquid nitrogen. Samples were shipped to the Bigelow Laboratory Single Cell Genomics Center where they were thawed. Single cells were sorted into a lysis buffer by flow cytometry based on their cell size and chlorophyll content. The DNA content of each cell was amplified separately using Multiple Displacement Amplification (MDA), following previously described protocols[5]. The identification of cells was based on the 18S rRNA gene sequence. After multiple alignments using MUSCLE[6], it appeared that the 18S rRNA sequence of TOSAG39-1 was strictly identical to that of *Bathycoccus prasinos* (GenBank: AY425315, FN562453).

DNA sequencing and assembly

The four cells, A, B, C and D were sequenced independently on 1⁄8th Illumina HiSeq lane, producing a total of 96 million 101-bp paired-end reads. For the combined-SAG assembly, we pooled the reads from

the different cells to increase the completion of the final assembly. To ensure that genomes of these cells could be correctly co-assembled, we first analyzed the contribution of each cell to a global assembly using the HyDA assembler[7]. HyDA produced a colored de Bruijn graph in which most contigs were covered by reads from at least three different cells, suggesting that the genomes were close enough to be successfully co-assembled. We used SPAdes 2.4[8] using parameter k = 21, 33 and 55 to obtain the final assembly, and we scaffolded contigs using the SSPACE program[9]. We used GapCloser (v 1.12-6 from SOAPdeNovo2 package[10]) with default settings to perform gap filling on the resulting scaffolds. Scaffolds shorter than 500 bp were discarded from the assembly.

We obtained individual assemblies for each cell, A, B, C and D separately using the same versions of SPAdes, SSPace and GapCloser. We computed a merged-assembly by pooling all scaffolds from the four individual assemblies and removing the redundancy using CD-HIT[11,12] v 4.6.1. Scaffolds with $\geq 95\%$ identity and $\geq 80\%$ overlapping (considering the shortest sequence) were clustered together and the longest scaffold of each cluster was kept as representative. The combined-SAGs assembly is the longest and appears as the most complete (Table 1).

Gene prediction on the TOSAG39-1 assembly

To predict different structures or specific genes that would be absent from the RCC1105 genome, we performed a *de novo* gene prediction using three different resources: protein mapping from a custom database enriched in marine protists transcripts, including the RCC1105 proteome; *ab initio* gene predictions; and transcriptional evidence from *Tara* Oceans metatranscriptomic data. Before this process, we masked the TOSAG39-1 assembly against repeated sequences using RepeatMasker version open-3.3.0[13].

We then mapped all proteins with BLAST+ 2.2.27[14] (e-value < 10^{-2}). The reference database was built with Uniref100[15] (version July 25th 2013) and the MMETSP transcriptomes[16] (version August 2013). We obtained a total of 6 560 distinct matches. For *ab initio* predictions, we used the SNAP predictor[17] after

calibration on *Bathycoccus prasinos* RCC1105 gene models. This resulted in the prediction of 6 797 gene models. Biological evidence was also provided by *Tara* Oceans metatranscriptomes. After mapping metatranscriptomic reads from all *Tara* Oceans samples of the 0.8-5 µm size fraction, we used the Gmorse pipeline[18] to define the gene structures from vertical coverage. We applied a minimum read coverage threshold of 32 because of the large abundance of *Bathycoccus* in *Tara* Ocean samples. We detected 6 112 genes. We finally integrated protein mapping, SNAP *ab initio* predictions and metatranscriptome derived gene models using a combiner process modified from the Gmorse software[16] and obtained 6 444 gene models. Further quality control filtering on putative non-*Bathycoccus* nuclear DNA reduced the final gene set to 6 157 (see below). Comparisons of TOSAG39-1 and RCC1105 gene sets are given in Supplementary Table 1.

TOSAG39-1 and RCC1105 genomic comparison

Best reciprocal hits (BRH)

We identified orthologous genes between RCC1105 and TOSAG39-1. We aligned each pair of genes using the Smith-Waterman algorithm[19] and retained alignments having a score higher than 300 (BLOSUM62, gapo = 10, gape = 1). We defined 4 153 best reciprocal hits as orthologs. The distribution of the percent identities for these BRH between the two *Bathycoccus* genomes is shown in Supplementary Figure S3.

Synteny and collinear genes analysis

We aligned the RCC1105 genomic data against the twenty longest TOSAG39-1 scaffolds (containing 656 genes) using promer (default parameter) from the MUMmer 3.19 package[20]. We used mummerplot to select RCC1105 chromosomes that corresponded to TOSAG39-1 scaffolds. We identified 18 scaffolds having an alignment covering their entire length with 11 chromosomes. We identified 573 RCC1105 genes localized within these syntenic regions. One of the two remaining scaffolds had matches with one RCC1105 contig that is not mapped to any chromosome, and the other could not be aligned and had a

lower GC% (0.44 *vs*. 0.48 averages for the other scaffolds) suggesting a chromosome 19 origin. To identify genes that are shared between the two genomes, we compared TOSAG39-1 scaffolds and RCC1105 in the six translated frames using tblastx[14] (e-value < 10^{-3}). We visually inspected genomic alignment regions using Artemis[21] and identified 52 RCC1105 genes localized in syntenic regions that lacked any alignments. We further compared these 52 genes against the whole genome at the protein level with tblastx[14] (e-value < 10^{-3}) and identified a total of 24 exclusive genes.

Comparison between *Bathycoccus* genomes and MMETSP transcriptome

We compared the RCC1105 and TOSAG39-1 gene sets to the two *Bathycoccus* transcriptomes available in the MMETSP collection[16]. We computed the best reciprocal hit at the amino acid level, as defined previously, and distributed their percentage of identity. We identified unambiguously MMETSP1460 (culture strain RCC716) and MMETSP1399 (culture strain CCMP1898) as corresponding to TOSAG39-1 and RCC1105, respectively (Supplementary Figure S5)

Comparison between *Bathycoccus* genome assemblies and metagenomes containing *Bathycoccus*

We compared by tblastn[14] (selecting e-value lower than 10^{-3}) the gene sets of RCC1105 and TOSAG39-1 to the two metagenomes (T142 and T149) from the Chile upwelling[3] and to the metagenome from the Atlantic Ocean DCM[4,22]. We selected matches covering more than 80% of the genes. We identified that RCC1105 corresponds to the T142 and T149 metagenome and TOSAG39-1 corresponds to the Atlantic Ocean metagenome (Supplementary Figure S5).

Metagenomic fragment recruitment

In order to analyze the diversity of *Bathycoccus* genomes and of dispensable genes, metagenomic reads from the *Tara* Oceans 0.8–5-µm fraction samples were recruited to whole sequence assemblies. We used Bowtie2-2.1.0[23] to align reads longer than 80 bp. We retained matches having more than 80% identity and more than 30% of high-complexity bases. From the initial 122 samples, we further analyzed the 36

samples for which at least 98% of the genes of *Bathycoccus* were detected (more than one mapped read). Using R-package 'ggplot2'[24], we displayed the density of reads mapping along the genome in 5 000-bp bins and 1% identity height (Supplementary Figure S11). This representation reduces the granularity of the Y-axis, particularly for high identity levels, caused by the short length of reads.

Gene set filtering

Mitochondrial and plastid genes

tblastn (e-value < 10^{-20})[14] was used to compare the mitochondrial and chloroplast RCC1105 proteins against TOSAG39-1 scaffolds. To check the validity of these scaffolds, we compared these selected scaffold against the nr database[25] using blastn[14]. We identified 35 genes as putatively of chloroplast or mitochondrial origin. The corresponding scaffolds were not further considered in the analysis.

Foreign sequences in TOSAG39-1 assembly

To improve detection of non-*Bathycoccus* DNA sequences in the TOSAG39-1 assembly, we used the results of metagenomic fragment recruitments for *Tara* Oceans samples. We postulated that assembly contigs corresponding to *Bathycoccus vs.* to non-*Bathycoccus* would be mapped by metagenomic reads at different coverages in the various samples. Therefore, we analyzed the variations of coverage of each gene along *Tara* Oceans samples to retrieve the specific *Bathycoccus* coverage profile. We assumed that the coverage profile of the majority of genes was the signature of TOSAG39-1. Considering these profiles as a time series, we used the "diss.CORT" function of the "TSclust" R-package[26] to compute distances based on abundance values and spatial correlation between profiles. We tagged 533 genes having a profile quite different from that of TOSAG39-1. However, we untagged from this list genes having an ortholog in *Bathycoccus prasinos* RCC1105. Finally, we discarded scaffolds containing tagged genes only. The aim of this filter is to discard the maximum of contigs that have an outlier statistical signal on fragment recruitment to avoid any putative bias due to atypical genomic region. Using this approach, we removed 223 scaffolds from the assembly. We compared these scaffolds on public databases using blast[14]. Due to

the stringency of this filter, some of these scaffolds (37.8%) seem to correspond to *Bathycoccus*, but the majority doesn't have any match or match different other organisms (Supplementary Table 6).

We also followed this rationale to detect genes having "outlier" profiles. We identified 826 and 1 051 genes on RCC1105 and TOSAG39-1, respectively. Among these, 111 and 223 were identified as cross-mapped genes (see below).

Estimation of cross-species mapped genes

In order to analyze the abundance of the two *Bathycoccus* genomes in the *Tara* Oceans metagenomic samples, we checked the possibility that some genes could be cross-mapped, that is genes that could be mapped by metagenomic reads from both genotypes. These genes could lead to a background signal in species detection survey. We identified 1 057 and 1 020 genes from TOSAG39-1 and RCC1105, respectively, that could be aligned on the other genome using Bowtie[23]. In order to do this, we fragmented one genome into 100-bp fragments that we mapped on the second genome to simulate metagenomics fragment recruitment conditions. We retained results having more than 95% identity. Since TOSAG39-1 is 64% complete, we extrapolated the total number of cross-mapped genes to about 1 500.

Abundance counts

Relative genomic abundance

We mapped metagenomic reads on RCC1105 and TOSAG39-1 genome sequence using Bowtie2 2.1.0 aligner with default parameters[23]. We filtered out alignments corresponding to low complexity regions using the dust algorithm[27] and we discarded alignments with less than 95% mean identity or with less than 30% of high complexity bases. For each *Bathycoccus,* we computed relative genomic abundances as the number of reads mapped onto non-outlier genes normalized by the total number of reads sequenced for each sample. We took into account the estimated fraction of genome recovery of TOSAG39-1 assembly to extrapolate the number of reads mapped on non-outlier genes to a complete genome assembly. Cross mapped genes, organelles and outlier genes were dismissed for the calculation. We generated the world

maps and heatmaps with R-package maps_2.1-6, mapproj_1.1-8.3, gplots_2.8.0 and mapplots_1.4 (version R-2.13, https://cran.r-project.org/web/packages/maps/index.html).

$RPKM_{MG}$ and $RPKM_{MT}$

Metagenomic and metatranscriptomic read counts per gene ($RPKM_{MG}$ and $RPKM_{MT}$) correspond to the number of mapped reads per gene (intron plus exon for $RPKM_{MG}$) or per CDS (for $RPKM_{MT}$) divided by the total number of reads sequenced for each sample multiplied by gene length. We used the following formula for figures: $\frac{\log(1+(\mathrm{RPKM}*10^9))}{\log(2)}$. We investigated relative transcriptomic activity of genes by dividing $RPKM_{MT}$ by $RPKM_{MG}$. If $RPKM_{MT} > 0$ but $RPKM_{MG}$ is null, we used the median of the total $RPKM_{MG}$.

Metabarcoding

Metabarcoding abundance values (V9 region of 18S rRNA genes) were extracted from a previous study[28] and correspond to the proportion of all eukaryotic reads assigned to *Bathycoccus.*

Analyses of dispensable genes

Identification and characterization

To detect variations in gene content of the two *Bathycoccus* genomes in the different samples, in particular gene loss, we analyzed the coverage of metagenomic reads that were specifically mapped on each genome at high stringency. To avoid putative background signals, we restricted this analysis to samples where 98% of the genes were detected (metagenomic abundance > 0). We retained 34 samples for RCC1105 and 21 samples for TOSAG39-1. We then focused on genes that were detected in at least four samples, and not detected in at least five samples. We obtained 108 and 106 dispensable genes in RCC1105 and TOSAG39-1, respectively. We performed a Mann-Whitney-Wilcoxon test (using R function wilcox.test

with default parameters) to compare RPKM values and gene length between dispensable and non-dispensable genes. We considered a significant difference at a p-value < 10^{-3}.

Validation of dispensable cassette genes in metagenomes

We aimed to validate the genomic pattern of gain or loss of cassettes of dispensable genes on RCC1105 using long metagenomic contigs from the *Tara* Ocean expedition data. We selected *Tara* Oceans stations having a high abundance of RCC1105 and a negligible abundance of TOSAG39-1 (relative abundance < 0.05%). We assembled merged metagenomic reads using SOAPdenovo[10] and a kmer size of 31. Most of the metagenomics contigs were short (N50 sizes ranged from 804 to 836 nt in the different samples) because of the difficulty of assembling eukaryotic metagenomes. However, we identified by blastn[14] several long metagenomics contigs that covered two dispensable cassettes, including the longest one. These metagenomics contigs were from the following stations and depths: TARA_082 surface, TARA_093 surface, TARA_152 surface, TARA_089 surface, TARA_093 DCM and TARA_152 surface (Figure 4, Supplementary Figure 13). These alignments confirmed the total absence of these dispensable cassettes in these metagenomic contigs. Furthermore, the positions of the insertion or deletion of a given cassette were identical for several metagenomic contigs, indicating a common event and suggesting the existence of only two genomic forms at these genomic positions in these samples

Analysis of environmental parameters

We used physicochemical parameter values related to the expedition sampling sites and available in the Pangaea database[29]. We extrapolated PAR values (corresponding to weekly averages values of Photosynthetically Active Radiation) at sample depth using the following formula with *k* derived from surface chlorophyll concentration (Chl_{sur}) using the following published formulas[30].

$$PAR(Z) = PAR(0) * \exp(-k * z)$$

$$x = \log(Chl)$$

$$log(Z) = 1.524 - 0.426x - 0.0145x^2 + 0.0186x^3$$

$$k = \frac{-\ln(0.01)}{Z}$$

PAR values were only available for 59 out of 122 samples among which 21 out of the 36 samples contained abundant *Bathycoccus* genome. Consequently PAR was not included into the principal component analysis presented in figure 3, as it would have reduced the data set considerably. A principal component analysis including PAR values is presented in Supplementary Figure S9 and did not alter our conclusions.

We carried these analyses for stations for which at least 98% of genes from one of the two *Bathycoccus* were detected. For each parameter, we performed a Mann-Whitney-Wilcoxon test (using the R function wilcox.test with default parameters) between the TOSAG39-1 and RCC1105 sets of values.

rRNA operon comparison

The *Bathycoccus* RCC1105 rRNA operon was used as the reference sequence to align the rRNA operons of TOSAG39-1, of two metagenomes (T142 and T149) from the Chile upwelling[3], of a metagenome from the Atlantic Ocean DCM[4,22], and the ITS from strains RCC715 and 716 (Genbank accession KT809427, KT809428) that have been isolated from the Indian Ocean. The alignments were done with MAFFT, as implemented in Geneious 7.1 (http://www.geneious.com/).

Functional analysis of dispensable genes

We predicted functional annotations of protein domains using CDD database (version v3.11)[31].

Supplementary Figures

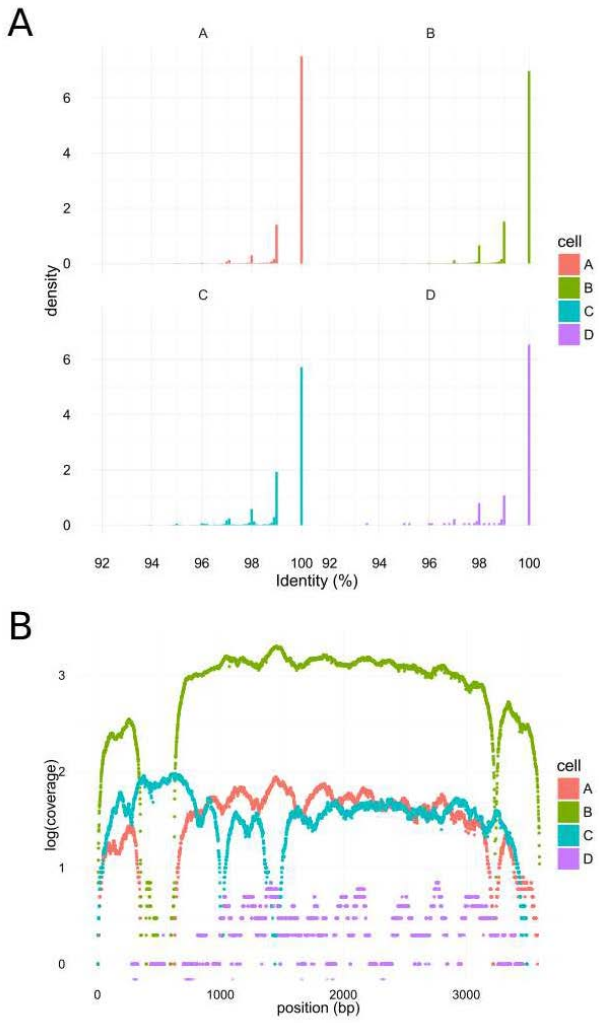


Figure S1. A. Distribution of identity percent of reads from each individual cell A (red), B (green), C (blue) and D (purple) once mapped onto the final combined SAG assembly. B. Example of the contributions of reads of each cell A (red), B (green), C (blue) and D (purple) along one contig of the final combined SAG assembly. X axis correspond to position and Y axis to coverage (log scale).

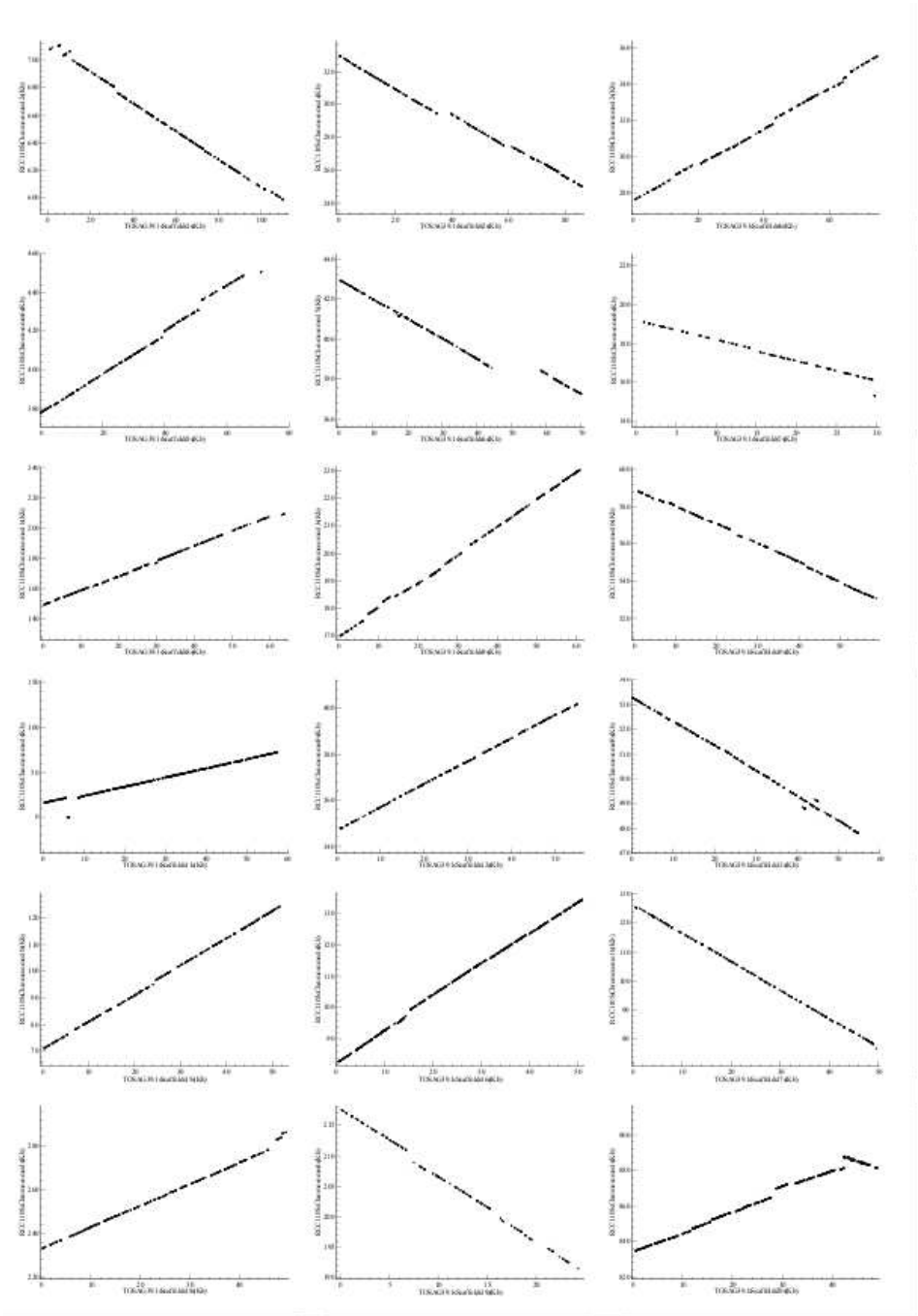


Figure S2. Synteny conservation between the two *B. prasinos* genomes. The RCC and TOSAG39-1 genomes are displayed on the X- and Y-axis, respectively. Dots correspond to regions conserved at the protein level (tblastx hits). Only the 18 longest scaffolds of TOSAG39-1 are represented. The two genomes are largely collinear and present only local and small rearrangements.

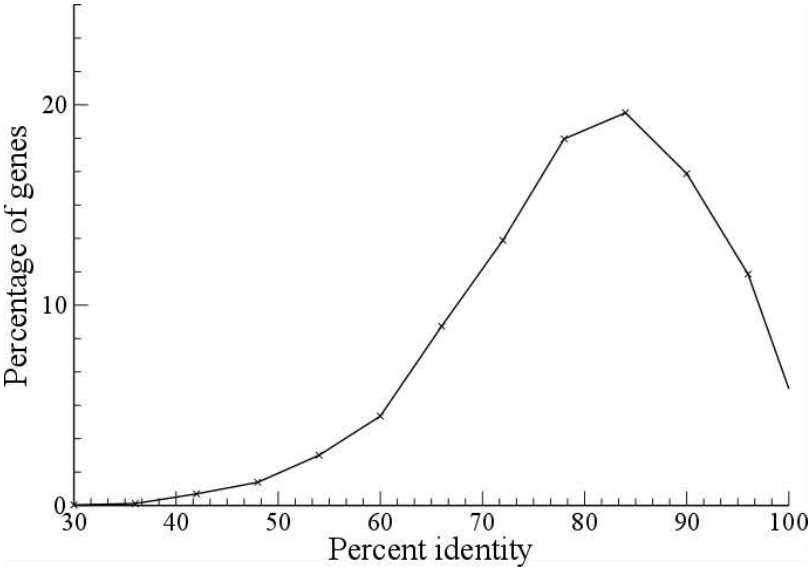


Figure S3. Distribution of orthologous gene divergence at the protein level between *Bathycoccus* RCC1105 and TOSAG39-1.

Figure S4. rRNA operon. Comparison of the rRNA ITS region between the two *Bathycoccus* genomes. RCC1105 and two metagenomes from the Chile upwelling[3] share identical ITS1 and ITS2, while TOSAG39-1 ITSs are identical to those of a metagenome from the Atlantic Ocean DCM[4] and to those from strains RCC715 and 716 that were isolated from the Indian Ocean.

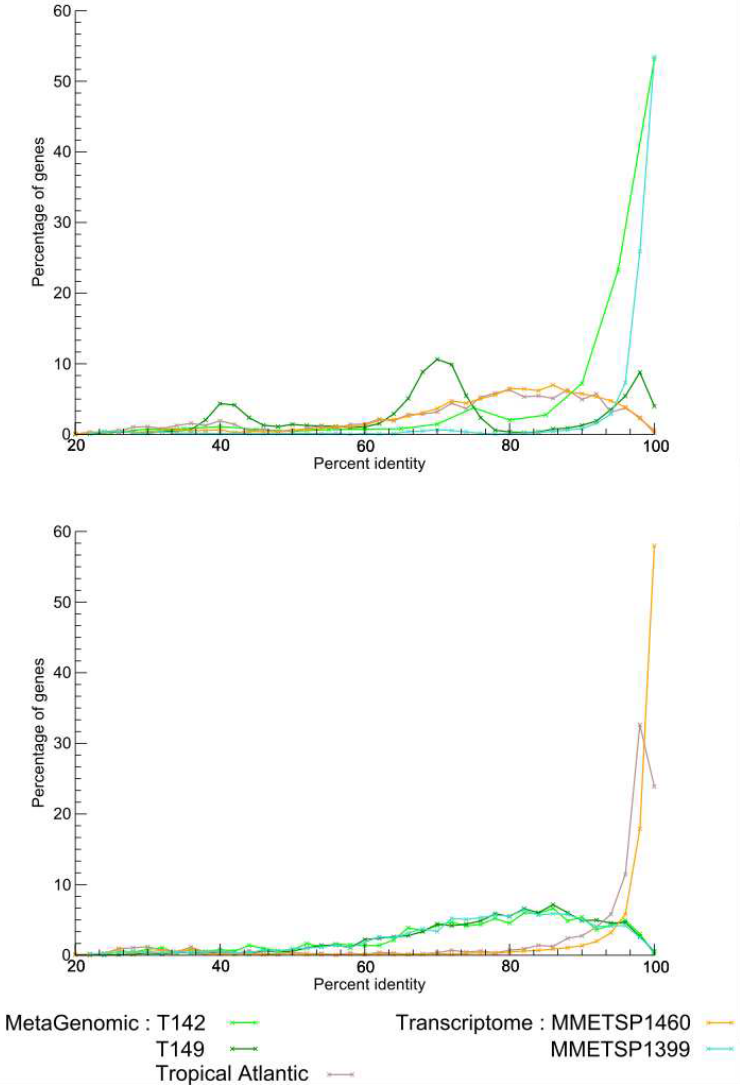


Figure S5. Affiliations of three metagenomes containing *Bathycoccus* and two *Bathycoccus* transcriptomes of the MMETSP database to the two genome assemblies. Distributions correspond to similarities at the amino acid level for one *Bathycoccus* genome assembly (top: RCC1105, bottom: TOSAG39-1) with two *Bathycoccus* transcriptomes (MMETSP1460 and MMETSP1399) and with three metagenomes containing *Bathycoccus*. MMETSP1399 transcriptome and T42 and T149 metagenomes correspond to RCC1105 genome, whereas MMETSP1460 and the tropical Atlantic metagenome correspond to TOSAG39-1.

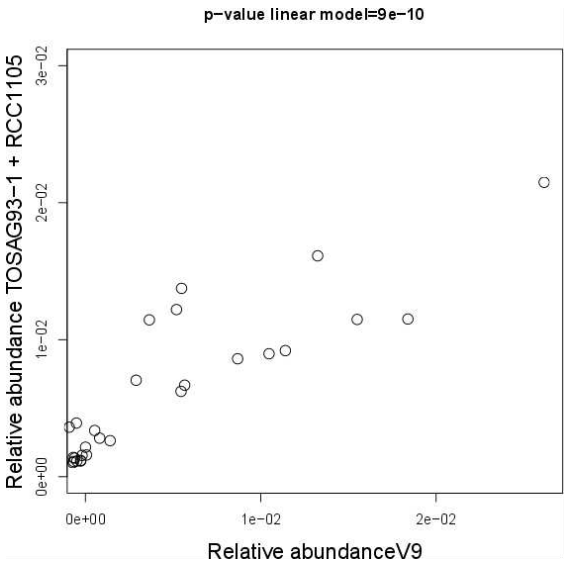


Figure S6. Correlation between the abundance of *Bathycoccus* estimated from whole metagenomes (two genomes summed) and V9 amplicons abundances.

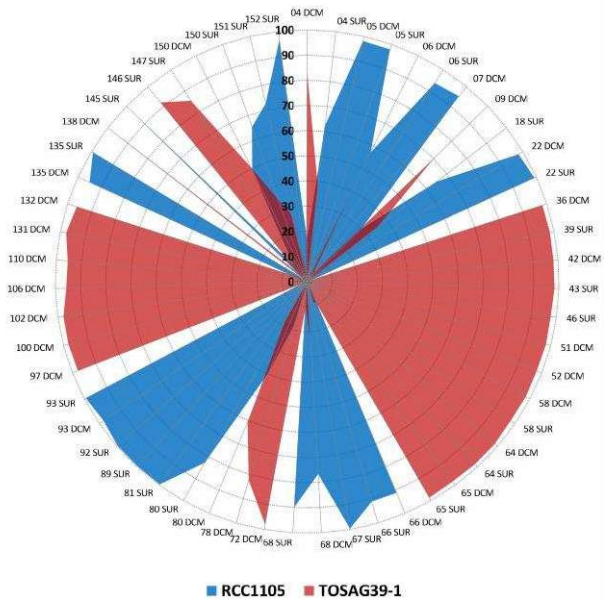


Figure S7. Relative contribution of each genome at *Bathycoccus*-rich stations. Within the 58 DCM and surface samples where *Bathycoccus* metagenomic abundance represents more than 0.01%, one of the two *Bathycoccus* genome was dominant (>70% of the *Bathycoccus* metagenomic reads) in 91% of the cases. The two genomes were measured in similar proportion (range 40% – 60%) in only two samples (stations 6 and 150 at the DCM).

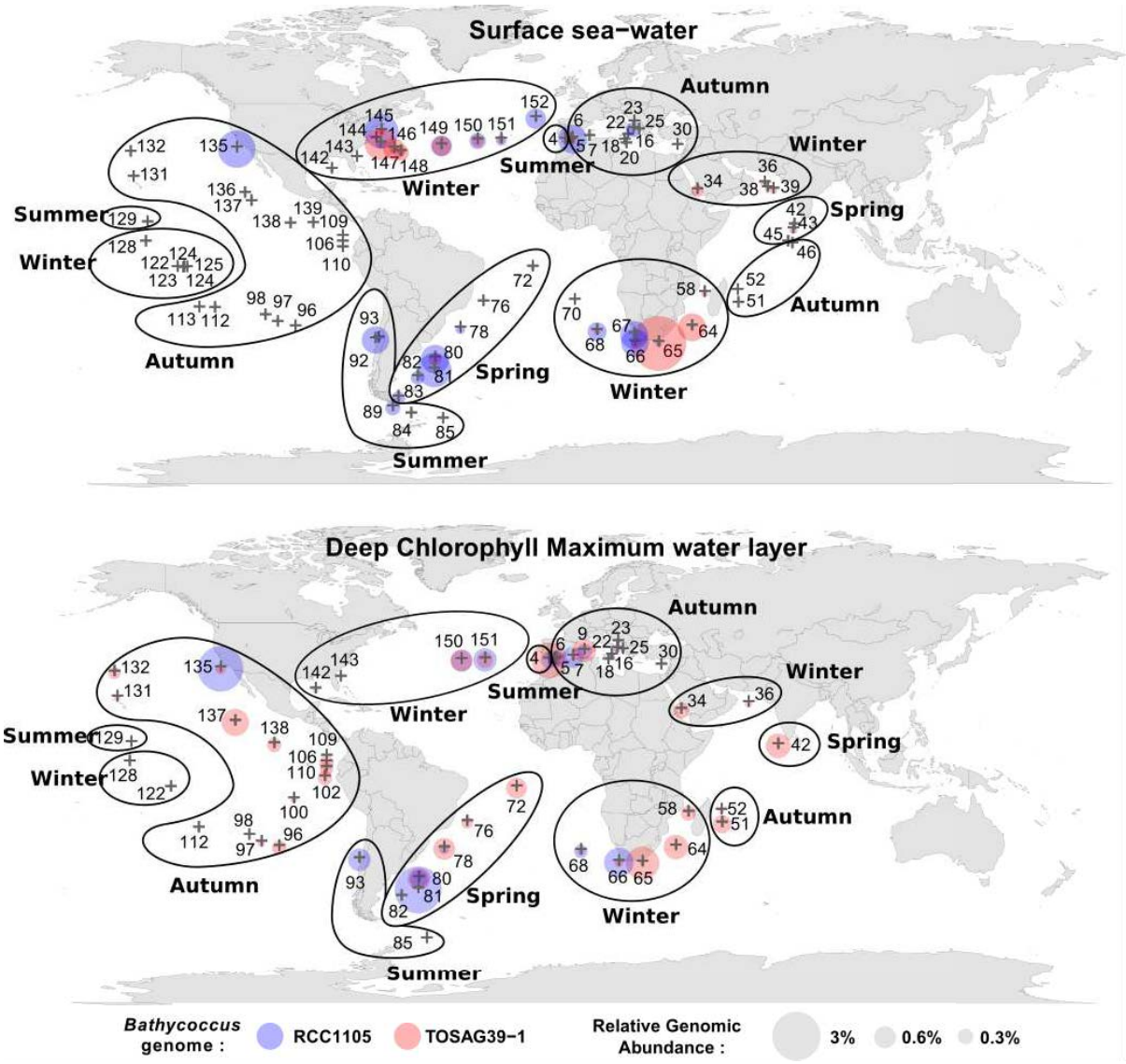


Figure S8. Map of relative metagenomic abundances of the two *Bathycoccus* in *Tara* Oceans stations with sampling season. This map was created using R-package maps_2.1-6, mapproj_1.1-8.3, gplots_2.8.0 and mapplots_1.4 (version R-2.13, https://cran.r-project.org/web/packages/maps/index.html).

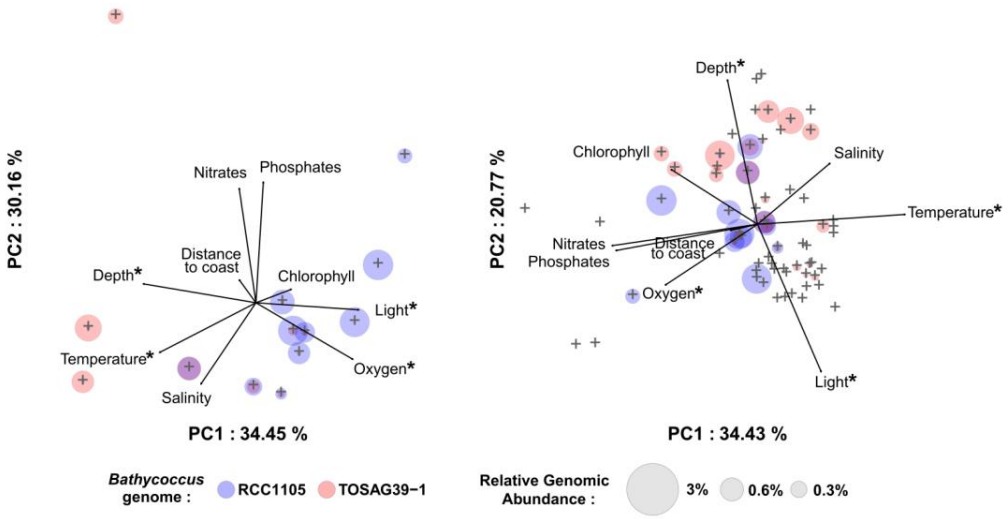


Figure S9. Principal Component Analysis including Photosynthetically Active Radiation (PAR). Left: Using only 13 samples for which we measured a large relative genomic abundance of *Bathycoccus* that have available PAR (indicated as light). Right: Idem but with all *Tara* Oceans samples that have available PAR values (indicated as light). Stars indicate parameters statistically discriminant.

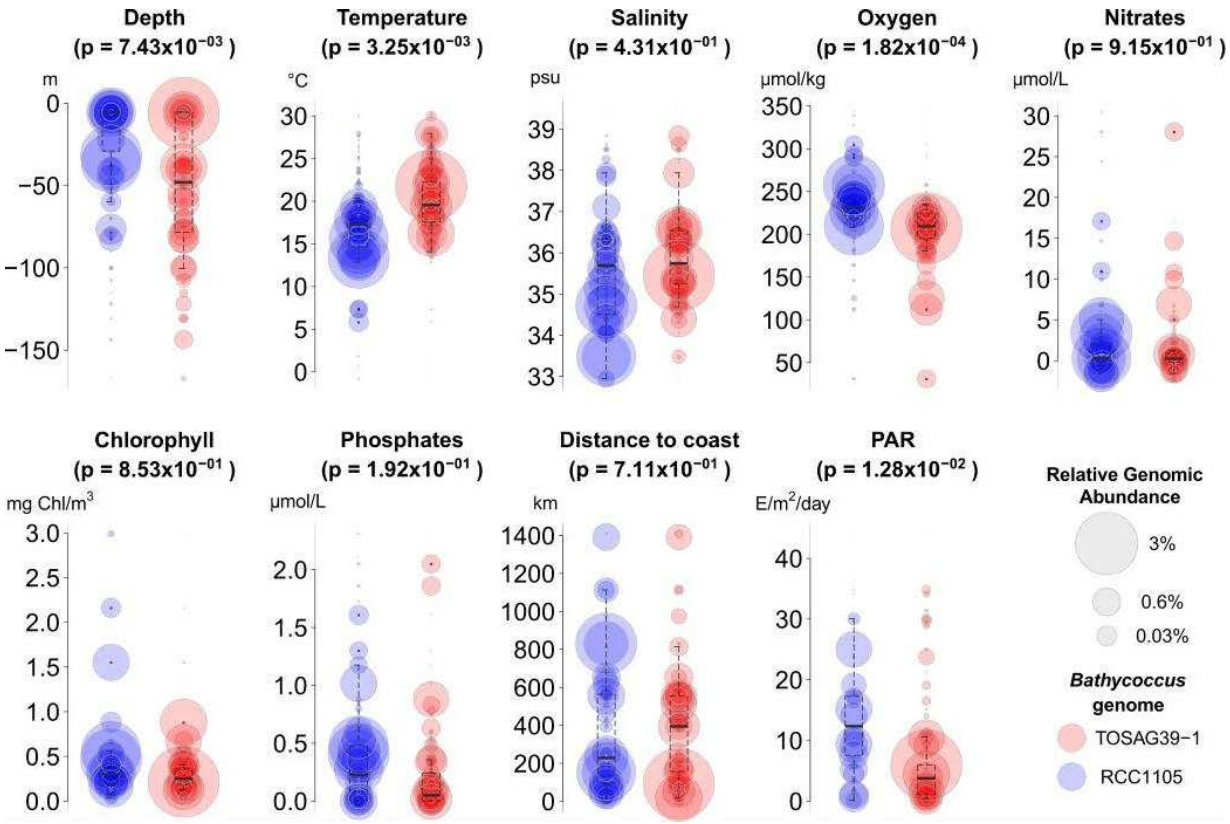


Figure S10. Environmental parameters and genomic abundances of *Bathycoccus*. PAR (Photosynthetically Active Radiation) corresponds to AMODIS satellite data for surface samples and to computed estimations for DCM samples. Temperature, oxygen, depth and light are parameters that gave significantly different distributions between the two *Bathycoccus* (Wilcoxon probability values). Sizes of circles are proportional to relative metagenomics abundance, according to the scale given in the legend. Boxplots over bubble plots indicate organism range distribution within samples containing high abundances of *Bathycoccus*, without taking in account abundance values.

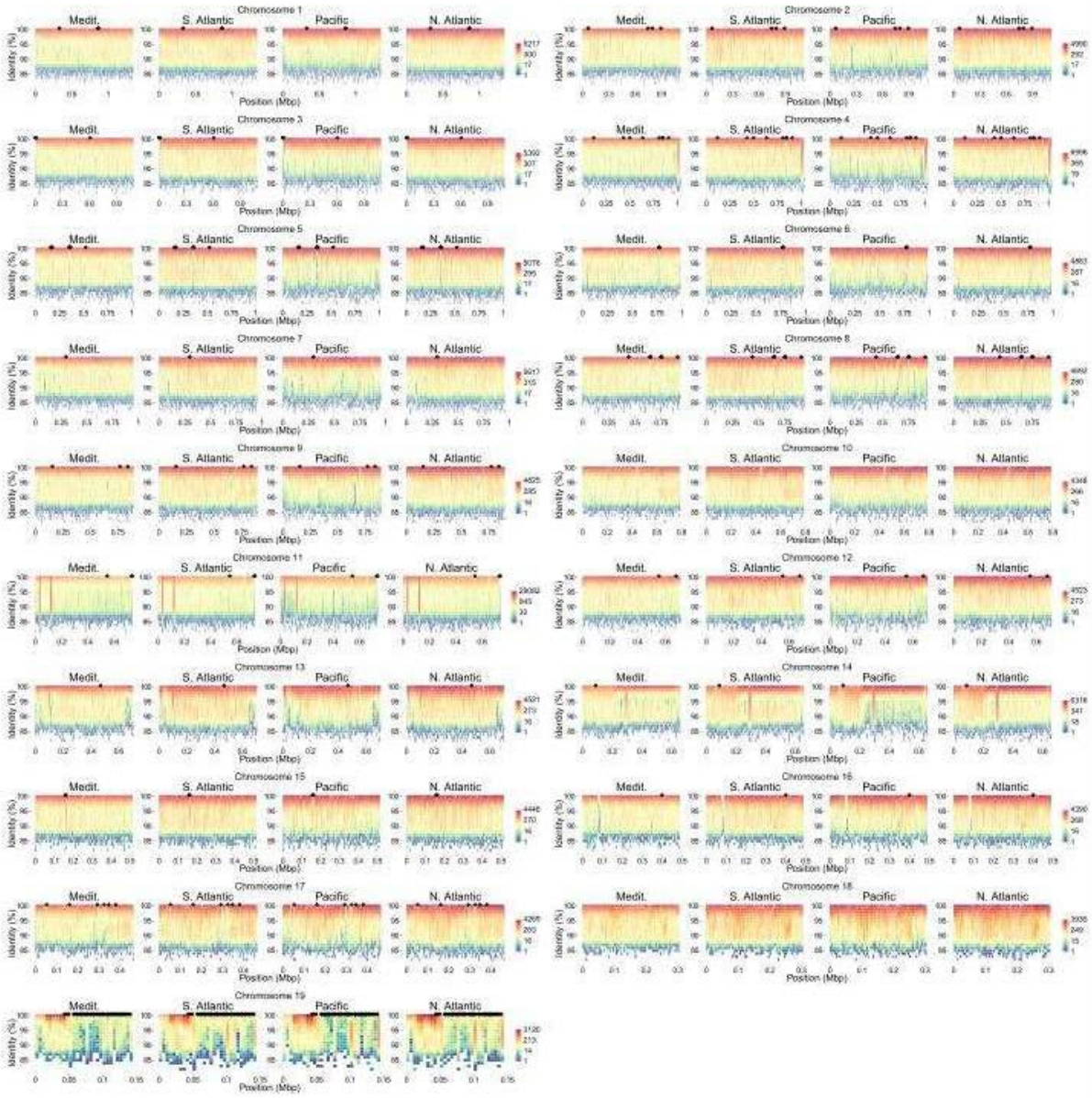


Figure S11. Metagenomic fragment recruitment plot on all chromosomes separated by large marine basins. Chromosome positions of dispensable genes are indicated by black dots. Gradient colors correspond to density of recruited metagenomic reads from low (blue) to high (red).

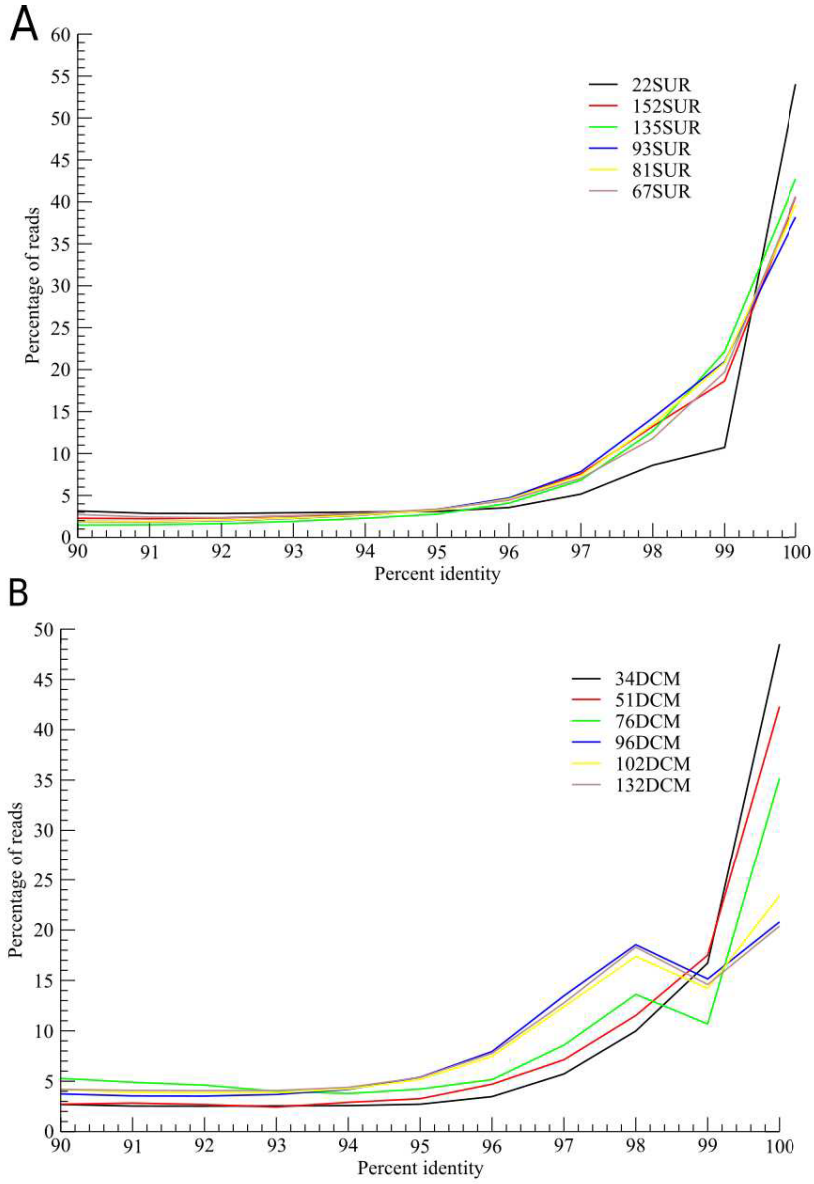


Figure S12. Distribution of identity percent of *Tara* Oceans metagenomic reads mapped onto RCC1105 genome (A) and TOSAG39-1 assembly (B). We only used *Tara* Oeans samples where the presence of only one *Bathycoccus* genome was detected.

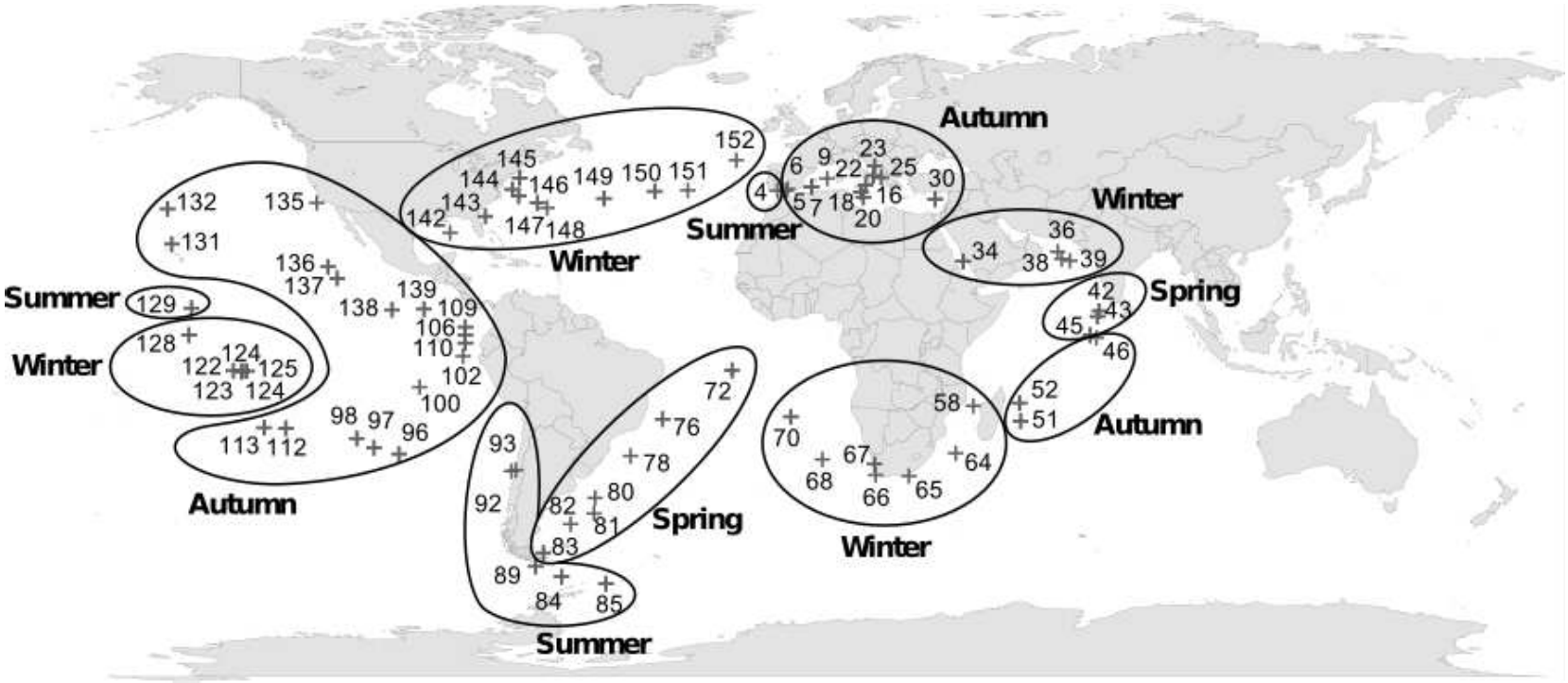


Figure S13. Map of the stations of the *Tara* Oceans expedition with seasons when sampled. This map was created using R-package maps_2.1-6, mapproj_1.1-8.3, gplots_2.8.0 and mapplots_1.4 (version R-2.13, https://cran.r-project.org/web/packages/maps/index.html).

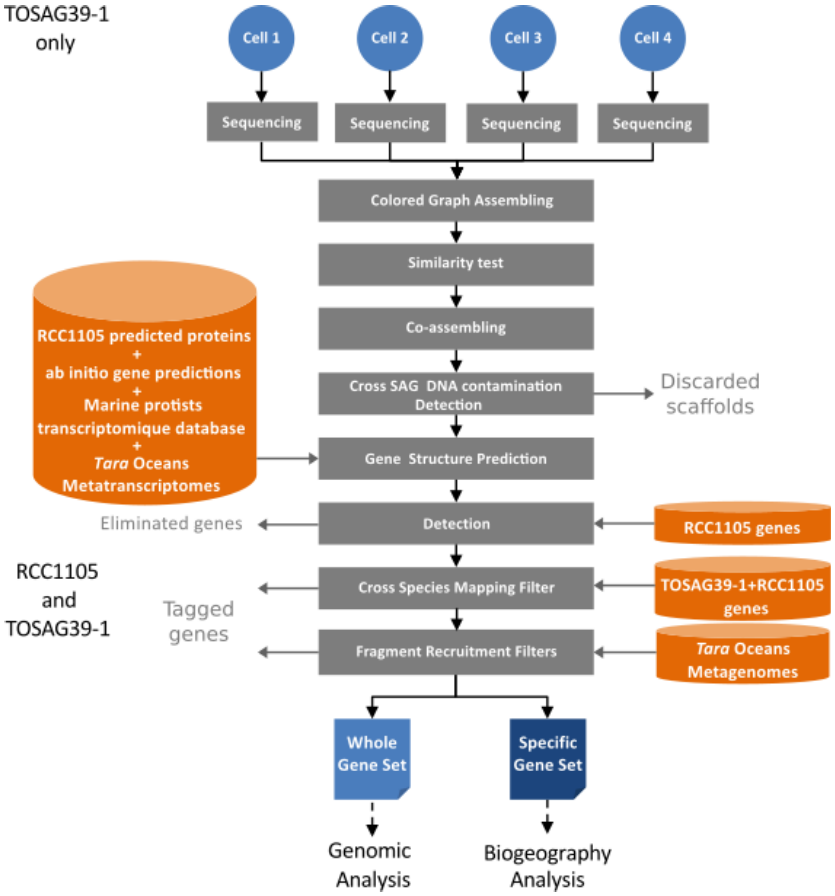


Figure S14. Pipeline for data acquisition and quality control.

Table S1. Comparisons of gene features of the two *Bathycoccus* gene sets.

Characterization		RCC1105 Genes			RCC1105 Genes (except chromosome 19)			TOSAG39-1 Predicted Genes		
		All	**Dispensable**	**Not Dispensable**	**All**	**Dispensable**	**Not Dispensable**	**All**	**Dispensable**	**Not Dispensable**
Gene number		7807	108	7699	7735	58	7677	6157	106	6051
Gene size (nt) mean. : sd		1609.36 : 1281	1014.19 : 942	1617.70 : 1287	1613.30 : 1287	1137.43 : 1049	1616.90 : 1287	1344.62 : 1074	511.89 : 392	1359.21 : 1087
Monoexonic Genes		6648 (85%)	100 (93%)	6548 (85%)	6585	52 (90%)	6533 (85%)	4596 (75%)	75 (71%)	4521 (75%)
Number of exons mean : sd		1.19 : 1	1.08 : 1	1.19 : 1	1.19 : 1	1.10 : 1	1.19 : 1	1.33 : 1	1.30 : 1	1.33 : 1
CDS length (nt) mean : sd		1578.44 : 1251	1006.78 : 939	1586.45 : 1257	1582.22 : 1257	1126.76 : 1026	1585.66 : 1257	1242.16 : 984	455.12 : 330	1255.95 : 999
Number of introns		1504	9	1495	1494	6	1488	2028	32	1996
Introns Size (nt) mean : sd		160.50 : 131	88.89 : 57	160.93 : 132	160.92 : 131	103.17 : 44	161.15 : 132	217.25 : 154	101.88 : 83	219.10 : 154
Metagenomic Abundance (*a*) (RPKM values)	All Samples. mean. : sd	2.47 : 1.16	0.44 : 0.69	2.50 : 1.14	2.49 : 1.14	0.56 : 0.82	2.51 : 1.13	3.28 : 1.34	0.50 : 0.73	3.33 : 1.30
	Samples with detected signal only. mean. : sd	2.49 : 1.14	0.75 : 0.76	2.51 : 1.13	2.50 : 1.13	0.92 : 0.88	2.51 : 1.13	3.31 : 1.31	0.82 : 0.78	3.34 : 1.29
Metatranscriptomic Abundance (*b*) (RPKM values)	All Samples. Mean. : sd	1.34 : 1.40	0.15 : 0.46	1.36 : 1.41	1.35 : 1.41	0.16 : 0.55	1.68 : 1.64	1.64 : 1.64	0.12 : 0.36	1.71 : 1.64
	Samples with detected signal only. Mean. : sd	1.58 : 1.39	0.58 : 0.76	1.58 : 1.39	1.58 : 1.40	0.70 : 0.96	2.04 : 1.59	2.03 : 1.59	0.67 : 0.59	2.05 : 1.59
Relative Transcriptomic Activity (a / b)	All Samples. mean. : sd	0.47 : 0.71	0.20 : 0.55	0.47 : 0.71	0.47 : 0.71	0.18 : 0.55	0.47 : 0.71	0.49 : 0.73	0.13 : 0.43	0.49 : 0.73
	Samples with detected signal only. mean. : sd	0.56 : 0.74	0.77 : 0.84	0.56 : 0.74	0.56 : 0.74	0.73 : 0.89	0.56 : 0.74	0.59 : 0.76	0.72 : 0.78	0.59 : 0.76

RPKM: reads per kilobase of transcript per million reads mapped.

Table S2. Depths of the Mixed Layer Depth (MLD) and of samples from the Deep Chlorophyll Maximum (DCM; italic red correspond to DCM samples taken above the MLD) for each *Tara* Ocean station used in this paper.

***Tara* Oceans Station**	**DCM sample depths (m)**	**MLD (m)**
4	39	4
7	42	18
8	45	3
9	55	21
18	62	39
22	31	9
23	55	9
25	52	29
30	69	41
34	60	26
36	17	7
38	25	11
39	25	9
42	79	21
51	80	40
52	79	47
58	67	17
64	*64*	*71*
65	*29*	*47*
66	*29*	*90*
68	*40*	*187*
72	95	75
76	148	34
78	118	34
80	83	12
81	38	29
82	42	29
85	87	38
93	34	22
96	153	42
97	174	50
98	183	53
100	58	35
102	46	18

Tara Oceans Station	**DCM sample depths (m)**	**MLD (m)**
106	47	12
109	30	9
110	49	22
112	154	131
122	113	71
125	138	95
128	42	35
129	85	76
131	109	36
132	114	41
135	30	13
137	44	17
138	58	24
142	*124*	*142*
143	*49*	*69*
150	*40*	*77*
151	78	36

Table S3. Annotations of the RCC1105 dispensable genes that have functional predictions.

Pfam	Note	Gene Identifier	Number of Dispensable Genes	
			Whole Genome	Chromosome 19
Pfam14312	FG-GAP repeat	Bathy02g04860	1	0
Pfam13465	Zinc-finger double domain	Bathy04g03240, Bathy04g03240, Bathy09g04110, Bathy09g04110	4	0
Pfam00808	Histone-like transcription factor (CBF/NF-Y) and archaeal histone	Bathy04g04090	1	0
Pfam06977	SdiA-regulated	Bathy04g04270	1	0
Pfam07727	Reverse transcriptase (RNA-dependent DNA polymerase)	Bathy04g04610, Bathy19g00670	2	1
Pfam01844	HNH endonuclease	Bathy05g02900	1	0
pfam12796	Ankyrin repeats (3 copies)	Bathy07g01420, Bathy12g03030	2	0
pfam14099	Polysaccharide lyase	Bathy08g04110	1	0
pfam01866	Putative diphthamide synthesis protein	Bathy08g04120	1	0
pfam03382	Mycoplasma protein of unknown function, DUF285	Bathy17g01470, Bathy17g01550	2	0
pfam11913	Protein of unknown function (DUF3431)	Bathy19g00310	1	1

pfam13383	Methyltransferase domain	Bathy19g00340, Bathy19g00540	2	2
pfam13578	Methyltransferase domain	Bathy19g00410	1	1
pfam00777	Glycosyltransferase family 29 (sialyltransferase)	Bathy19g00420	1	1
pfam04321	RmlD substrate binding domain	Bathy19g00510	1	1
pfam13489	Methyltransferase domain	Bathy19g00590	1	1

Table S4. Dispensable genes of RCC1105.

Bathy01g01790	Bathy08g04120	Bathy19g00350
Bathy01g04690	Bathy08g04130	Bathy19g00360
Bathy01g04700	Bathy08g04940	Bathy19g00370
Bathy02g00365	Bathy09g00830	Bathy19g00380
Bathy02g04020	Bathy09g04110	Bathy19g00390
Bathy02g04230	Bathy09g04450	Bathy19g00400
Bathy02g04860	Bathy11g02890	Bathy19g00410
Bathy03g00010	Bathy11g03900	Bathy19g00420
Bathy03g00030	Bathy11g03920	Bathy19g00430
Bathy03g00040	Bathy12g03030	Bathy19g00440
Bathy03g03150	Bathy12g03670	Bathy19g00450
Bathy04g00740	Bathy13g02130	Bathy19g00460
Bathy04g02210	Bathy14g00440	Bathy19g00470
Bathy04g02620	Bathy15g00910	Bathy19g00480
Bathy04g03240	Bathy16g02050	Bathy19g00490
Bathy04g04090	Bathy17g00250	Bathy19g00510
Bathy04g04270	Bathy17g00780	Bathy19g00520
Bathy04g04280	Bathy17g01470	Bathy19g00530
Bathy04g04610	Bathy17g01550	Bathy19g00540
Bathy05g00940	Bathy17g01690	Bathy19g00550
Bathy05g00970	Bathy17g01840	Bathy19g00560
Bathy05g00980	Bathy19g00160	Bathy19g00570
Bathy05g02010	Bathy19g00175	Bathy19g00580
Bathy05g02020	Bathy19g00200	Bathy19g00590

Bathy05g02030	Bathy19g00230	Bathy19g00600
Bathy05g02040	Bathy19g00240	Bathy19g00610
Bathy05g02900	Bathy19g00250	Bathy19g00620
Bathy06g04070	Bathy19g00260	Bathy19g00630
Bathy06g04080	Bathy19g00270	Bathy19g00640
Bathy06g04090	Bathy19g00280	Bathy19g00650
Bathy07g01420	Bathy19g00290	Bathy19g00660
Bathy08g02440	Bathy19g00300	Bathy19g00670
Bathy08g03500	Bathy19g00310	Bathy19g00680
Bathy08g03510	Bathy19g00320	Bathy19g00690
Bathy08g03520	Bathy19g00330	Bathy19g00700
Bathy08g04110	Bathy19g00340	

Table S5. Dispensable genes of TOSAG39-1.

TOSAG39-1_gene78	TOSAG39-1_gene2608	TOSAG39-1_gene4518
TOSAG39-1_gene145	TOSAG39-1_gene2703	TOSAG39-1_gene4704
TOSAG39-1_gene223	TOSAG39-1_gene2704	TOSAG39-1_gene4784
TOSAG39-1_gene226	TOSAG39-1_gene2878	TOSAG39-1_gene4883
TOSAG39-1_gene229	TOSAG39-1_gene2935	TOSAG39-1_gene5106
TOSAG39-1_gene278	TOSAG39-1_gene2982	TOSAG39-1_gene5107
TOSAG39-1_gene358	TOSAG39-1_gene2987	TOSAG39-1_gene5131
TOSAG39-1_gene382	TOSAG39-1_gene3033	TOSAG39-1_gene5174
TOSAG39-1_gene383	TOSAG39-1_gene3035	TOSAG39-1_gene5178
TOSAG39-1_gene394	TOSAG39-1_gene3051	TOSAG39-1_gene5189
TOSAG39-1_gene509	TOSAG39-1_gene3339	TOSAG39-1_gene5291
TOSAG39-1_gene521	TOSAG39-1_gene3340	TOSAG39-1_gene5327
TOSAG39-1_gene588	TOSAG39-1_gene3341	TOSAG39-1_gene5480
TOSAG39-1_gene615	TOSAG39-1_gene3361	TOSAG39-1_gene5523
TOSAG39-1_gene616	TOSAG39-1_gene3460	TOSAG39-1_gene5695
TOSAG39-1_gene791	TOSAG39-1_gene3505	TOSAG39-1_gene5721
TOSAG39-1_gene993	TOSAG39-1_gene3508	TOSAG39-1_gene5791
TOSAG39-1_gene997	TOSAG39-1_gene3562	TOSAG39-1_gene5792
TOSAG39-1_gene1003	TOSAG39-1_gene3690	TOSAG39-1_gene5901
TOSAG39-1_gene1004	TOSAG39-1_gene3830	TOSAG39-1_gene5902
TOSAG39-1_gene1048	TOSAG39-1_gene3846	TOSAG39-1_gene5986
TOSAG39-1_gene1113	TOSAG39-1_gene3880	TOSAG39-1_gene5987
TOSAG39-1_gene1178	TOSAG39-1_gene3915	TOSAG39-1_gene6023
TOSAG39-1_gene1388	TOSAG39-1_gene3958	TOSAG39-1_gene6026

TOSAG39-1_gene1392	TOSAG39-1_gene3959	TOSAG39-1_gene6027
TOSAG39-1_gene1403	TOSAG39-1_gene3966	TOSAG39-1_gene6079
TOSAG39-1_gene1416	TOSAG39-1_gene3967	TOSAG39-1_gene6104
TOSAG39-1_gene1417	TOSAG39-1_gene3972	TOSAG39-1_gene6187
TOSAG39-1_gene1483	TOSAG39-1_gene4016	TOSAG39-1_gene6188
TOSAG39-1_gene1694	TOSAG39-1_gene4042	TOSAG39-1_gene6222
TOSAG39-1_gene1740	TOSAG39-1_gene4043	TOSAG39-1_gene6362
TOSAG39-1_gene1751	TOSAG39-1_gene4060	TOSAG39-1_gene6376
TOSAG39-1_gene1765	TOSAG39-1_gene4062	TOSAG39-1_gene6422
TOSAG39-1_gene1818	TOSAG39-1_gene4273	TOSAG39-1_gene6426
TOSAG39-1_gene2202	TOSAG39-1_gene4303	TOSAG39-1_gene6440
TOSAG39-1_gene2203	TOSAG39-1_gene4517	

Table S6. Summary of the matches obtained with the discarded scaffolds of TOSAG39-1 assembly.

Match	Proportion
No match	42.4%
Bathycoccus prasinos	37.8%
Bacteria	10.8%
Mitochondrion	3.6%
Cyprinus carpio	0.6%
Chloroplast	0.5%
BpV2 virus	0.4%
Bacteriophage S13	0.2%
Other	3.8%

References

1. Moreau, H. *et al.* Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biol.* **13,** R74 (2012).
2. Sterck, L., Billiau, K., Abeel, T., Rouzé, P. & Van de Peer, Y. ORCAE: online resource for community annotation of eukaryotes. *Nat. Methods* **9,** 1041–1041 (2012).
3. Vaulot, D. *et al.* Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS ONE* **7,** e39648 (2012).
4. Monier, A. *et al.* Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. *Environ. Microbiol.* **14,** 162–176 (2012).
5. Stepanauskas, R. & Sieracki, M. E. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 9052–9057 (2007).
6. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
7. Movahedi, N. S., Forouzmand, E. & Chitsaz, H. De novo co-assembly of bacterial genomes from multiple single cells. in *2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 1–5 (2012). doi:10.1109/BIBM.2012.6392618
8. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **19,** 455–477 (2012).
9. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinforma. Oxf. Engl.* **27,** 578–579 (2011).
10. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1,** 18 (2012).

11. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinforma. Oxf. Engl.* **22,** 1658–1659 (2006).
12. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).
13. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0 at http://www.repeatmasker.org. (2013).
14. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
15. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinforma. Oxf. Engl.* **23,** 1282–1288 (2007).
16. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLOS Biol* **12,** e1001889 (2014).
17. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5,** 59 (2004).
18. Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9,** R175 (2008).
19. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).
20. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
21. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinforma. Oxf. Engl.* **16,** 944–945 (2000).
22. Monier, A., Sudek, S., Fast, N. M. & Worden, A. Z. Gene invasion in distant eukaryotic lineages: discovery of mutually exclusive genetic elements reveals marine biodiversity. *ISME J.* **7,** 1764–1774 (2013).

23. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
24. Wickham, H. *ggplot2*. (Springer New York, 2009).
25. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33,** D501–D504 (2005).
26. Montero Manso, P. & Vilar, A. TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software* 1–43 (2014).
27. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **13,** 1028–1040 (2006).
28. de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348,** 1261605 (2015).
29. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2,** 150023 (2015).
30. Morel, A. *et al.* Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. *Remote Sens. Environ.* **111,** 69–88 (2007).
31. Marchler-Bauer, A. *et al.* CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43,** D222–226 (2015).

Annexe 2: Informations supplémentaires de l'article "Genome Resolved Biogeography of Mamiellales"

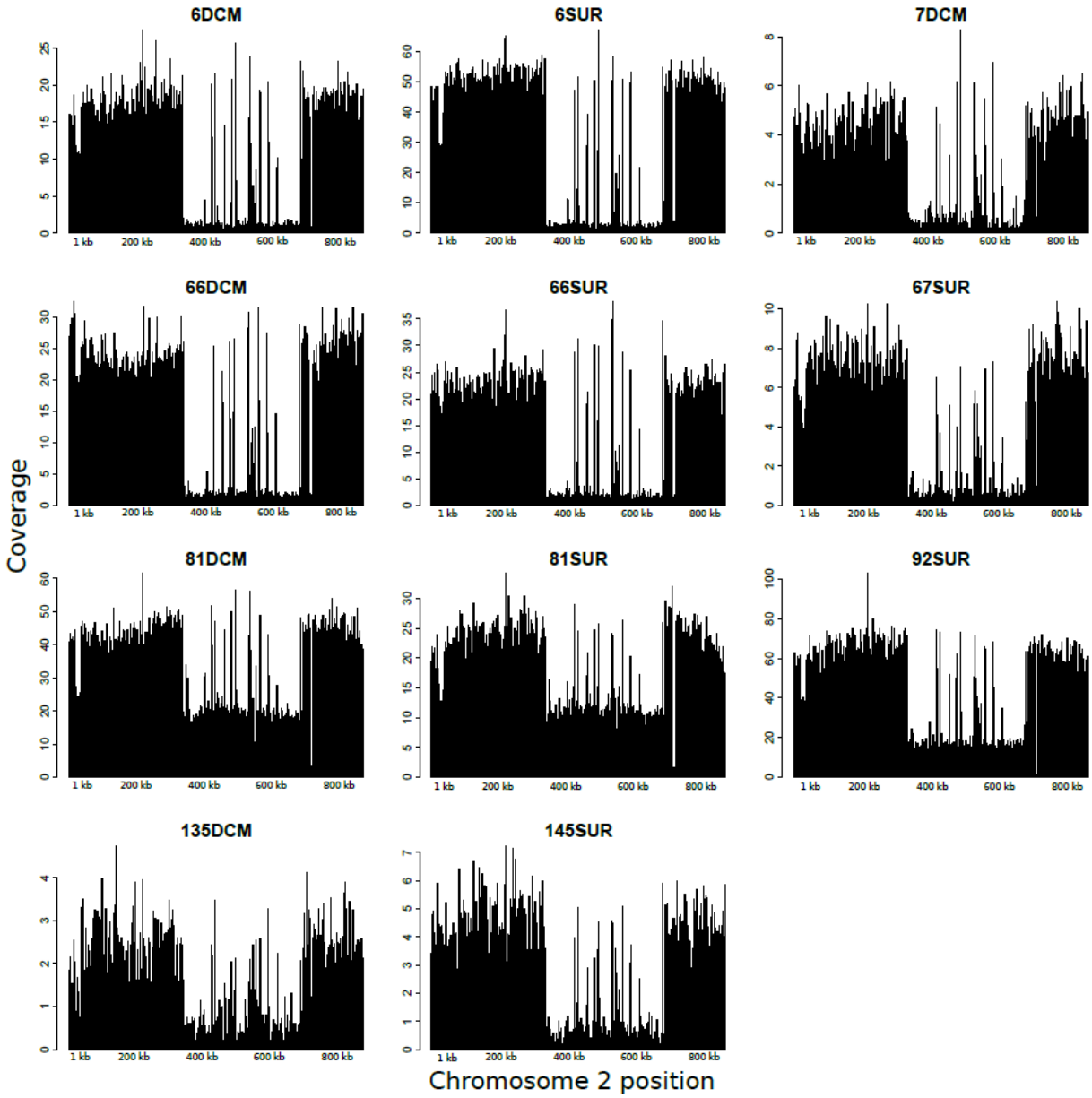


Figure S1. Metagenomic reads coverage along chromosome 2 of *O. lucimarinus* in the 11 samples where this genome recruits at least 0.1% of reads.

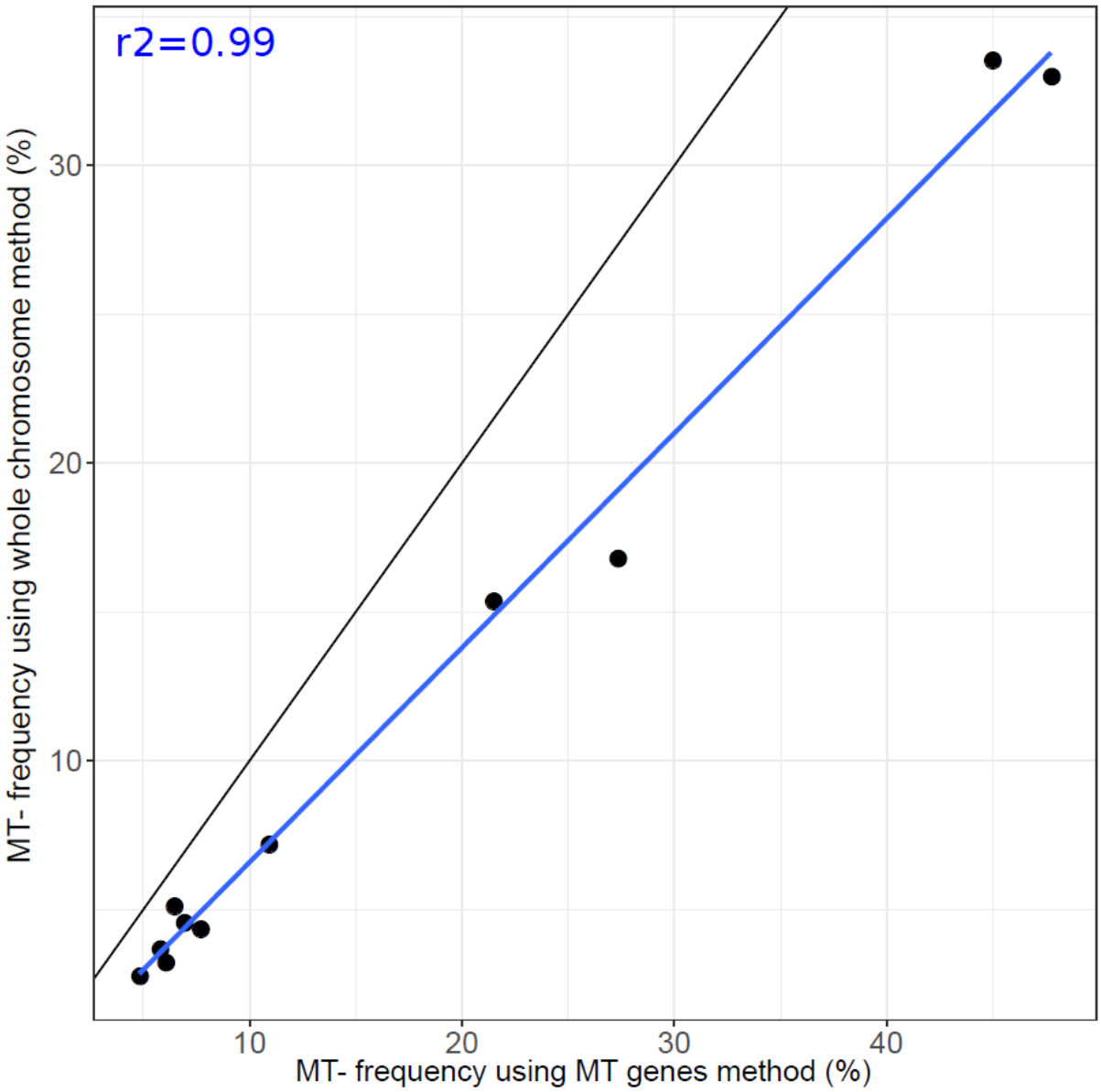


Figure S2. Scatterplot of *O. lucimarinus*chromosome 2 differential coverage based on median values againstthe ratio of relative metagenomic abundances of the *MT-* and *MT+* mating types genes.

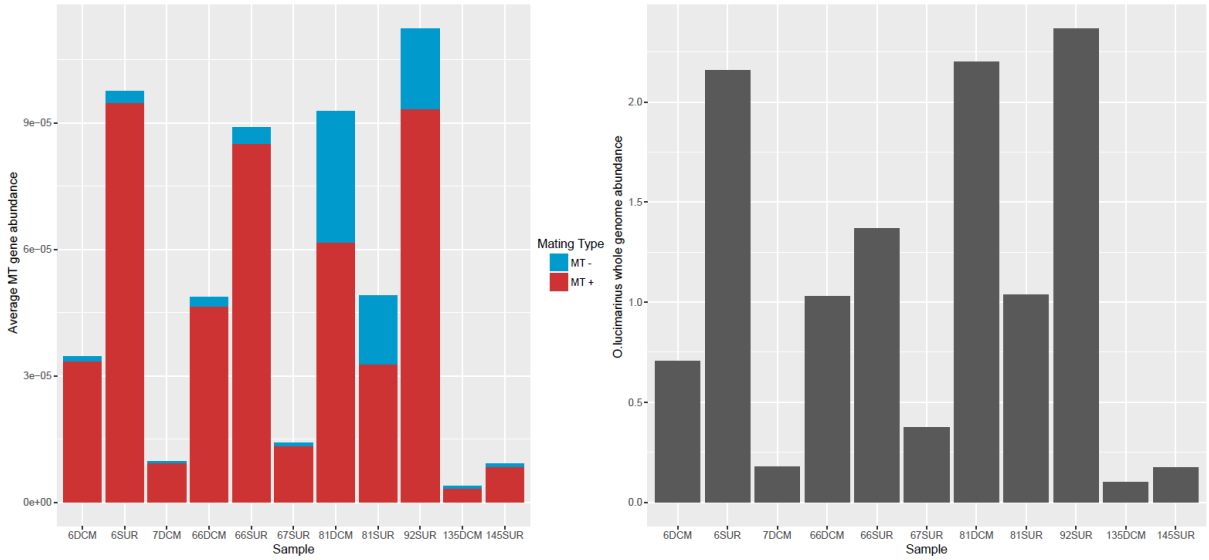


Figure S3. Barplot comparisons of relative metagenomic abundances of the *MT-* and *MT+* mating types (left panel) versus whole genome (right panel) for *O. lucimarinus*

Table S1. Metagenomic samples identifying information (133 Tara Oceans Samples ID to access genomic and contextual data).

Site	Depth	PANGAEA.ID	BioSamples.ID	ENA.ID
4	DCM	TARA_X000000408	SAMEA2619419	ERS487956
4	SUR	TARA_X000000323	SAMEA2619396	ERS487919
5	DCM	TARA_X000000567	SAMEA2619456	ERS488005
5	SUR	TARA_X000000488	SAMEA2619443	ERS487982
6	DCM	TARA_X000000706	SAMEA2619486	ERS488048
6	SUR	TARA_X000000608	SAMEA2619467	ERS488020
7	DCM	TARA_A200000158	SAMEA2591077	ERS477956
7	SUR	TARA_A200000123	SAMEA2591060	ERS477934
9	DCM	TARA_X000001040	SAMEA2619551	ERS488150
9	SUR	TARA_X000000954	SAMEA2619534	ERS488122
16	DCM	TARA_A100000035	SAMEA2619664	ERS488317
16	SUR	TARA_A100000032	SAMEA2619654	ERS488306
18	DCM	TARA_A100000604	SAMEA2619684	ERS488352
18	SUR	TARA_A100000595	SAMEA2619675	ERS488338
20	SUR	TARA_A100000758	SAMEA2619723	ERS488410
22	DCM	TARA_A100000698	SAMEA2619754	ERS488460
22	SUR	TARA_A100000534	SAMEA2619745	ERS488446
23	DCM	TARA_A100000559	SAMEA2591103	ERS478003
23	SUR	TARA_A100000551	SAMEA2591093	ERS477988
25	DCM	TARA_A100000396	SAMEA2619788	ERS488515
25	SUR	TARA_A100000393	SAMEA2619777	ERS488497
30	DCM	TARA_A100001333	SAMEA2591128	ERS478046
30	SUR	TARA_A100001640	SAMEA2591115	ERS478024
32	DCM	TARA_A100001145	SAMEA2619850	ERS488609
32	SUR	TARA_A100001141	SAMEA2619832	ERS488583
34	DCM	TARA_N000000323	SAMEA2619916	ERS488694
34	SUR	TARA_N000000327	SAMEA2619917	ERS488695
36	DCM	TARA_N000000317	SAMEA2619958	ERS488753
36	SUR	TARA_N000000316	SAMEA2619943	ERS488730
38	DCM	TARA_N000000030	SAMEA2620023	ERS488832
38	SUR	TARA_N000000029	SAMEA2620010	ERS488809
39	DCM	TARA_N000000007	SAMEA2620089	ERS488924
39	SUR	TARA_N000000006	SAMEA2620071	ERS488885
41	DCM	TARA_N000000072	SAMEA2620221	ERS489078
41	SUR	TARA_N000000071	SAMEA2620204	ERS489053
42	DCM	TARA_N000000086	SAMEA2620267	ERS489142
42	SUR	TARA_N000000085	SAMEA2620249	ERS489106
43	SUR	TARA_N000000292	SAMEA2620292	ERS489175
45	SUR	TARA_N000000255	SAMEA2620354	ERS489251
46	SUR	TARA_N000000267	SAMEA2620378	ERS489279
51	DCM	TARA_N000000217	SAMEA2620527	ERS489506
51	SUR	TARA_N000000214	SAMEA2620503	ERS489451

52	DCM	TARA_N000000600	SAMEA2620581	ERS489596
52	SUR	TARA_N000000598	SAMEA2620556	ERS489543
58	DCM	TARA_N000000419	SAMEA2620745	ERS489857
58	SUR	TARA_N000000469	SAMEA2620711	ERS489785
64	DCM	TARA_N000000524	SAMEA2620845	ERS490019
64	SUR	TARA_N000000522	SAMEA2620802	ERS489933
65	DCM	TARA_N000000961	SAMEA2620910	ERS490105
66	DCM	TARA_N000000807	SAMEA2620959	ERS490172
66	SUR	TARA_N000000805	SAMEA2620939	ERS490134
67	SUR	TARA_N000000756	SAMEA2620988	ERS490201
68	DCM	TARA_N000000728	SAMEA2621048	ERS490307
68	SUR	TARA_N000000722	SAMEA2621029	ERS490281
70	SUR	TARA_N000000678	SAMEA2621082	ERS490343
72	DCM	TARA_N000000833	SAMEA2621169	ERS490490
72	SUR	TARA_N000000831	SAMEA2621147	ERS490448
76	DCM	TARA_N000000860	SAMEA2621226	ERS490607
76	SUR	TARA_N000000855	SAMEA2621209	ERS490553
78	DCM	TARA_N000000622	SAMEA2621282	ERS490701
78	SUR	TARA_N000000620	SAMEA2621265	ERS490670
80	DCM	TARA_N000001485	SAMEA2621337	ERS490792
80	SUR	TARA_N000001491	SAMEA2621315	ERS490751
81	DCM	TARA_N000001434	SAMEA2621388	ERS490860
81	SUR	TARA_N000001436	SAMEA2621362	ERS490817
82	SUR	TARA_N000001386	SAMEA2621412	ERS490896
83	SUR	TARA_N000001374	SAMEA2621470	ERS490977
84	SUR	TARA_N000001438	SAMEA2621498	ERS491012
85	DCM	TARA_N000001030	SAMEA2621544	ERS491103
85	SUR	TARA_N000001028	SAMEA2621522	ERS491057
86	DCM	TARA_N000001071	SAMEA2621609	ERS491187
86	SUR	TARA_N000001069	SAMEA2621583	ERS491142
89	SUR	TARA_N000001217	SAMEA2621696	ERS491289
92	SUR	TARA_N000001299	SAMEA2621772	ERS491398
93	DCM	TARA_N000001111	SAMEA2621823	ERS491474
93	SUR	TARA_N000001296	SAMEA2621791	ERS491433
95	SUR	TARA_N000001266	SAMEA2621854	ERS491514
96	DCM	TARA_N000001236	SAMEA2621894	ERS491572
96	SUR	TARA_N000001256	SAMEA2621872	ERS491538
97	DCM	TARA_N000001525	SAMEA2621971	ERS491673
97	SUR	TARA_N000001522	SAMEA2621922	ERS491613
98	SUR	TARA_N000001574	SAMEA2622010	ERS491719
100	DCM	TARA_N000001610	SAMEA2622130	ERS491885
100	SUR	TARA_N000001608	SAMEA2622106	ERS491845
102	DCM	TARA_N000001652	SAMEA2622230	ERS492023
102	SUR	TARA_N000001650	SAMEA2622184	ERS491949
106	DCM	TARA_N000001692	SAMEA2622278	ERS492090
106	SUR	TARA_N000001690	SAMEA2622258	ERS492052

109	DCM	TARA_N000001732	SAMEA2622345	ERS492186
109	SUR	TARA_N000001730	SAMEA2622325	ERS492154
110	DCM	TARA_N000001752	SAMEA2622417	ERS492279
110	SUR	TARA_N000001750	SAMEA2622391	ERS492243
111	DCM	TARA_N000001816	SAMEA2622489	ERS492368
111	SUR	TARA_N000001812	SAMEA2622463	ERS492332
112	DCM	TARA_N000001872	SAMEA2622561	ERS492461
112	SUR	TARA_N000001870	SAMEA2622536	ERS492426
113	SUR	TARA_N000001896	SAMEA2622605	ERS492516
122	DCM	TARA_N000001948	SAMEA2622699	ERS492708
122	SUR	TARA_N000001938	SAMEA2622661	ERS492651
123	SUR	TARA_N000001992	SAMEA2622717	ERS492740
124	SUR	TARA_N000002037	SAMEA2622770	ERS492825
125	SUR	TARA_N000002019	SAMEA2622826	ERS492897
128	DCM	TARA_N000002291	SAMEA2622936	ERS493111
128	SUR	TARA_N000002289	SAMEA2622914	ERS493057
129	DCM	TARA_N000002312	SAMEA2622983	ERS493175
129	SUR	TARA_N000002310	SAMEA2622957	ERS493132
130	SUR	TARA_N000002343	SAMEA2622997	ERS493199
131	DCM	TARA_N000002354	SAMEA2623037	ERS493264
131	SUR	TARA_N000002352	SAMEA2623018	ERS493224
132	DCM	TARA_N000002418	SAMEA2623092	ERS493353
132	SUR	TARA_N000002416	SAMEA2623072	ERS493313
135	DCM	TARA_N000002191	SAMEA2623228	ERS493557
135	SUR	TARA_N000002179	SAMEA2623206	ERS493519
136	SUR	TARA_N000002961	SAMEA2623266	ERS493612
137	DCM	TARA_N000002936	SAMEA2623304	ERS493679
137	SUR	TARA_N000002925	SAMEA2623284	ERS493645
138	DCM	TARA_N000003011	SAMEA2623379	ERS493797
138	SUR	TARA_N000003001	SAMEA2623359	ERS493761
139	SUR	TARA_N000003037	SAMEA2623419	ERS493853
142	DCM	TARA_N000003104	SAMEA2623504	ERS493997
142	SUR	TARA_N000003083	SAMEA2623479	ERS493954
143	DCM	TARA_N000003144	SAMEA2623564	ERS494087
143	SUR	TARA_N000003137	SAMEA2623543	ERS494036
145	SUR	TARA_N000003219	SAMEA2623641	ERS494184
146	SUR	TARA_N000003253	SAMEA2623685	ERS494248
147	SUR	TARA_N000002103	SAMEA2623723	ERS494304
148	SUR	TARA_N000002113	SAMEA2623743	ERS494341
149	SUR	TARA_N000002135	SAMEA2623783	ERS494403
150	DCM	TARA_N000002145	SAMEA2623839	ERS494501
150	SUR	TARA_N000002697	SAMEA2623817	ERS494454
151	DCM	TARA_N000002723	SAMEA2623879	ERS494570
151	SUR	TARA_N000002741	SAMEA2623861	ERS494529
152	SUR	TARA_N000002789	SAMEA2623901	ERS494594

Table S2. Post-hoc Tukey test pairwise pvalues for each of the 4 significant environmental parameters according to the Kruskal-wallis test.

Reference 1	Reference 2	Temperature p.Value	Oxygen p.Value	Chlorophyll a p.Value	NitrateNitrite p.Value
M.commoda	M.pusilla	2,94E-05	6,65E-04	5,58E-09	5,21E-03
M.commoda	O.lucimarinus	1,43E-08	0,00E+00	6,59E-14	9,89E-02
M.commoda	O.RCC809	9,98E-01	1,00E+00	1,00E+00	9,64E-01
M.commoda	B.prasinos	4,40E-09	1,27E-13	9,71E-05	1,21E-02
M.commoda	TOSAG39-1	1,00E+00	1,00E+00	1,00E+00	1,00E+00
M.pusilla	O.lucimarinus	1,00E+00	9,67E-01	1,00E+00	1,00E+00
M.pusilla	O.RCC809	4,13E-07	1,94E-07	7,74E-05	1,20E-01
M.pusilla	B.prasinos	1,00E+00	1,00E+00	8,43E-01	9,86E-01
M.pusilla	TOSAG39-1	1,86E-04	1,72E-04	6,20E-05	2,21E-01
O.lucimarinus	O.RCC809	1,23E-12	0,00E+00	7,49E-09	4,68E-01
O.lucimarinus	B.prasinos	9,68E-01	3,28E-01	5,39E-01	1,00E+00
O.lucimarinus	TOSAG39-1	3,49E-10	0,00E+00	7,10E-08	3,15E-01
O.RCC809	B.prasinos	3,11E-15	0,00E+00	1,82E-03	1,90E-01
O.RCC809	TOSAG39-1	9,83E-01	9,99E-01	1,00E+00	1,00E+00
B.prasinos	TOSAG39-1	1,05E-12	0,00E+00	1,76E-03	1,09E-01

Table S3. Relative metagenomic abundances for the six Mamiellales genomes in all samples.

	Bathycoccus RCC1105	Bathycoccus TOSAG39.1	Micromonas commoda	Micromonas pusilla	Ostreococcus RCC809	Ostreococcus lucimarinus
4DCM	0,3016	2,4123	0,0254	0,0030	0,5061	0,0473
4SUR	0,1049	0,1048	0,0026	0,0006	0,0154	0,0037
6DCM	0,3393	0,4141	0,0221	0,1893	0,1456	0,7040
6SUR	0,9600	0,1153	0,0728	0,8597	0,1221	2,1583
7DCM	0,5002	0,0482	0,0443	0,0281	0,0123	0,1772
9DCM	0,2778	1,0352	0,0260	0,0292	0,1265	0,0272
22SUR	0,2531	0,0071	0,0011	0,0005	0,0005	0,0001
32DCM	0,0083	0,7651	0,3369	0,0000	0,0606	0,0000
32SUR	0,0012	0,1122	0,0418	0,0000	0,0087	0,0000
34DCM	0,0060	0,4605	0,0567	0,0004	0,0785	0,0001
34SUR	0,0032	0,2674	0,4845	0,0004	0,5996	0,0004
36DCM	0,0012	0,0994	0,0369	0,0003	0,3392	0,0003
38DCM	0,0049	0,4244	0,1016	0,0003	0,4977	0,0003
39DCM	0,0043	0,3786	0,2086	0,0004	0,4089	0,0003
39SUR	0,0009	0,0840	0,2860	0,0003	0,1523	0,0001
41DCM	0,0090	0,8063	0,0847	0,0006	4,0650	0,0023
42DCM	0,0140	1,0935	0,0232	0,0004	1,1736	0,0009
43SUR	0,0013	0,1088	0,6990	0,0005	0,2310	0,0002
46SUR	0,0002	0,0134	1,7034	0,0006	1,3585	0,0007
51DCM	0,0078	0,6048	0,1596	0,0003	0,1579	0,0001
58DCM	0,0050	0,4017	0,2022	0,0005	0,0462	0,0001
58SUR	0,0009	0,0678	0,2027	0,0003	0,0288	0,0001
64DCM	0,0105	1,0781	1,0346	0,0004	0,1220	0,0001
64SUR	0,0140	1,3955	1,1471	0,0004	0,1253	0,0001
65DCM	0,0207	1,7449	1,2841	0,0004	0,6256	0,0003
66DCM	1,0049	0,1679	0,2130	0,0011	0,1488	1,0283
66SUR	0,7600	0,1264	0,1639	0,0008	0,1469	1,3684
67SUR	0,9094	0,0156	0,0468	0,4950	0,0003	0,3759
68DCM	0,2007	0,1053	0,0382	0,0005	0,0063	0,0025
68SUR	0,4034	0,0800	0,0362	0,0005	0,0032	0,0027
72DCM	0,0111	0,8385	0,0003	0,0005	0,0450	0,0001
76DCM	0,0032	0,2430	0,0150	0,0005	0,0224	0,0001
78DCM	0,1187	0,8901	0,2124	0,0120	0,0124	0,0003
78SUR	0,1017	0,0179	0,0423	0,0024	0,0004	0,0002
80DCM	0,3654	0,9161	0,1673	0,0288	0,0861	0,0104
80SUR	0,9215	0,3204	0,1197	0,0244	0,0290	0,0319
81DCM	2,5447	0,0362	0,0434	0,0028	0,0013	2,1993
81SUR	1,3416	0,0202	0,0190	0,0020	0,0006	1,0386
82SUR	0,2231	0,0067	0,0681	0,0007	0,0001	0,0014
83SUR	0,2110	0,0057	0,0215	0,0156	0,0001	0,0096
89SUR	0,2641	0,0065	0,0324	0,0006	0,0001	0,0002
92SUR	0,0287	0,0023	0,0030	0,0009	0,0013	2,3655
93DCM	0,5890	0,0277	0,0023	0,0010	0,0117	0,0396

93SUR	0,8711	0,0177	0,0025	0,0004	0,0010	0,0261
96DCM	0,0057	0,3918	0,0003	0,0004	0,0003	0,0002
102DCM	0,0061	0,4058	0,0004	0,0005	0,0086	0,0001
106DCM	0,0108	0,4008	0,0003	0,0005	0,0397	0,0076
110DCM	0,0097	0,3332	0,0003	0,0005	0,1319	0,0327
113SUR	0,0000	0,0033	0,0213	0,0004	0,1887	0,0001
123SUR	0,0001	0,0018	0,0005	0,0008	0,5463	0,0006
124SUR	0,0001	0,0019	0,0003	0,0004	0,1213	0,0002
132DCM	0,0069	0,2791	0,0002	0,0004	0,1110	0,0001
135DCM	2,3674	0,2333	0,0145	0,0011	0,0026	0,1016
135SUR	1,6224	0,0338	0,0070	0,0004	0,0001	0,0146
137DCM	0,0205	1,4474	0,0003	0,0004	0,4456	0,0004
138DCM	0,0055	0,3902	0,0004	0,0005	0,6681	0,0005
142DCM	0,0007	0,0581	0,5909	0,0004	0,0002	0,0001
142SUR	0,0005	0,0345	0,6418	0,0003	0,0001	0,0001
145SUR	1,3493	0,1263	0,1175	1,4651	0,0075	0,1732
146SUR	0,1010	1,8254	1,0767	0,0135	1,1766	0,0220
147SUR	0,0906	0,8466	1,2885	0,0109	0,5693	0,0049
148SUR	0,0569	0,4622	0,5862	0,0011	0,5447	0,0004
149SUR	0,4981	0,7095	0,9794	0,5541	0,7503	0,0759
150DCM	0,5359	0,8846	0,8000	0,0009	0,8947	0,0132
150SUR	0,3085	0,2685	0,2221	0,0006	0,3569	0,0272
151DCM	0,6281	0,4705	0,0248	0,0007	0,5691	0,0010
151SUR	0,1295	0,0797	0,0253	0,0008	0,2093	0,0003
152SUR	0,4794	0,0329	0,0162	0,0005	0,0021	0,0011

Annexe 3: Article "Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans"

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-02235-3 OPEN

Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans

Yoann Seeleuthner et al.#

Single-celled eukaryotes (protists) are critical players in global biogeochemical cycling of nutrients and energy in the oceans. While their roles as primary producers and grazers are well appreciated, other aspects of their life histories remain obscure due to challenges in culturing and sequencing their natural diversity. Here, we exploit single-cell genomics and metagenomics data from the circumglobal *Tara* Oceans expedition to analyze the genome content and apparent oceanic distribution of seven prevalent lineages of uncultured heterotrophic stramenopiles. Based on the available data, each sequenced genome or genotype appears to have a specific oceanic distribution, principally correlated with water temperature and depth. The genome content provides hypotheses for specialization in terms of cell motility, food spectra, and trophic stages, including the potential impact on their lifestyles of horizontal gene transfer from prokaryotes. Our results support the idea that prominent heterotrophic marine protists perform diverse functions in ocean ecology.

Correspondence and requests for materials should be addressed to M.S. (email: mike.sieracki@gmail.com) or to C.d.V. (email: vargas@sb-roscoff.fr) or to P.W. (email: pwincker@genoscope.cns.fr)
#A full list of authors and their affliations appears at the end of the paper

The microbial loop in planktonic ecosystems is the process by which suspended organic matter produced within food webs is channeled through heterotrophic prokaryotes and their tiny grazers and eventually transferred to higher trophic levels or remineralized[1]. Very small but numerous marine heterotrophic protists play key roles in these processes. Since most of them remain uncultured, their functions remain largely unknown[2]. A recent DNA metabarcoding survey based on *Tara* Oceans global plankton samples has revealed the existence of thousands of heterotrophic protist taxa in eukaryotic communities[3] that potentially participate in numerous species interaction networks in yet-to-be defined ways[4]. An extensive genome-level description of abundant marine heterotrophic protists could therefore be a key step toward understanding their ecological roles. Currently, the only way to obtain such information is through single-cell sequencing, although the technology is still in its infancy for eukaryotic cells[5–10], since generated assemblies are highly fragmented and rarely complete.

Here, we integrate single-cell genomics with metagenomic and metatranscriptomic sequence data for exploring the ecological and functional complexity of uncultured micro-eukaryotes, key players in the world's largest ecosystem. We selected for our study 40 single cells representative of three uncultured stramenopile clades that are known to be abundant in marine pico-nano plankton. Marine stramenopile group 4 (MAST-4) representatives are small, flagellated, bacterivorous cells that are abundant in temperate and tropical oceans[11,12]. A partial genome of a MAST-4 clade D was previously characterized using single-cell sequencing[8]. In this study, we present three distinct genomes from clades A, C, and E, clearly divergent from clade D. MAST-3[11] is a very diverse group of small flagellated organisms that includes a potential diatom epibiont and one cultured strain[13,14]. Heterotrophic chrysophytes from the Clade H additionally appear to be abundant in the ocean, according to environmental DNA surveys[15]. It has been postulated that all of these lineages originated from a presumably autotrophic stramenopile ancestor[16], although lack of genome information has hindered understanding of the evolution of heterotrophy vs. autotrophy within the stramenopiles. Assessment of the genes involved in the degradation of organic matter may thus be relevant for elucidating their roles in marine ecosystems and biogeochemical cycles[17].

Results

Assembly strategy. More than 900 single-cell amplified genomes (SAGs) were generated from small heterotrophic protists selected from eight *Tara* Oceans sampling stations representing contrasting environments in the Mediterranean Sea and Indian Ocean. SAGs belonging to the target lineages were identified by PCR and subsequent sequencing of their 18S rRNA gene. A total of 40 SAGs were sequenced[18]: 23 from three MAST-4 lineages (MAST-4A, MAST-C, and MAST-E), six from two lineages of MAST-3 (MAST-3A and MAST-F), and 11 from two lineages of chrysophytes (Chrysophytes H1 and H2). We also generated metagenomic and metatranscriptomic datasets from the 0.8 to 5 µm size fraction collected from 76 and 68 *Tara* Oceans sampling sites, respectively, to assist the removal of potential contaminants from nuclear sequences and to improve gene structures (see section "Methods"; Supplementary Fig. 1, and companion papers[18,19]). The characteristics of each composite genome are summarized in Table 1. The MAST-4A cells were co-assembled as two independent sets of sequences, for use as an internal control for subsequent analyses and because they originated from two different water masses; however, they were very similar in genome composition (Supplementary Fig. 2) and a single assembly would have been possible[20].

Functional repertoires. To assess variation in the functional repertoires of the sequenced uncultured stramenopiles and to provide further context, we predicted functional domains (Pfams) in each annotated protein from each of the lineages, and compared their diversity and abundance against each other and against other sequenced stramenopile genomes. We then calculated pairwise distances between genomes based on relative Pfam abundances. The resulting pattern (Fig. 1a) indicated that the uncultured heterotrophic stramenopiles contained a diversity of gene repertoires, comparable to those of the sequenced genomes of autotrophic stramenopiles. However, the composition of each genome clustered primarily according to the trophic mode of each organism, with groups corresponding to heterotrophs, single-celled autotrophs, multicellular autotrophs, and mixotrophs. Moreover, within the heterotrophs, the MAST lineages and the chrysophytes-clade H clustered into a single functional group despite their distant phylogenetic positions (Fig. 1b, Supplementary Fig. 3). They could also be clearly distinguished from the plant-parasitic and gut-commensal heterotrophic stramenopiles (Fig. 1a, groups 3, 5, and 6), suggesting ecosystem-specific functional diversification, which needs further investigation.

Within the marine SAG genomes, many gene families showed differential abundances, indicating that functional capacities are distinct (Supplementary Table 1). One extreme pattern was observed for genes encoding the axonemal dynein heavy chain (DHC), which is an essential flagellar component. Almost all SAG genomes contained a family of genes encoding DHCs, with the exception of MAST-3A, for which we could not detect a single full-length gene and observed a significant decrease in the number of DHC Pfam domains (Supplementary Fig. 4). A closer examination of the MAST-3A genome regions containing the DHC-associated Pfam domains showed evidence of advanced

Table 1 SAGs assembly and annotation summary

Name	Number of cells	Raw assembly size (Mbp)	Cross SAG sequences (Mbp)	Outlier sequences (Mbp)	Final assembly size (Mbp)	N50	BUSCO v2 complete genes (%)	Number of predicted genes
Chrysophyte H1	8	16.7	0.1	0.6	15.9	25,581	57	3050
Chrysophyte H2	3	14.3	1.1	0.3	10.6	10,194	27	1637
MAST-3A	4	20.0	0	1.0	18.9	6223	53	3289
MAST-3F	2	21.5	0	0.3	21.1	7132	37	2694
MAST-4A1	6	33.4	0	1.0	31.8	10,950	59	8018
MAST-4A2	4	37.1	3.0	1.1	32.8	11,577	64	8537
MAST-4C	4	31.2	0	0.9	30.0	8097	54	5478
MAST-4E	9	30.3	0.2	1.4	28.4	9788	61	4652

SAG single amplified genome, *N50* length of the shortest scaffold from the minimal set of scaffolds representing 50% of the assembly size, *BUSCO v2* number of complete genes found using the BUSCO program (Benchmarking Universal Single-Copy Orthologs)

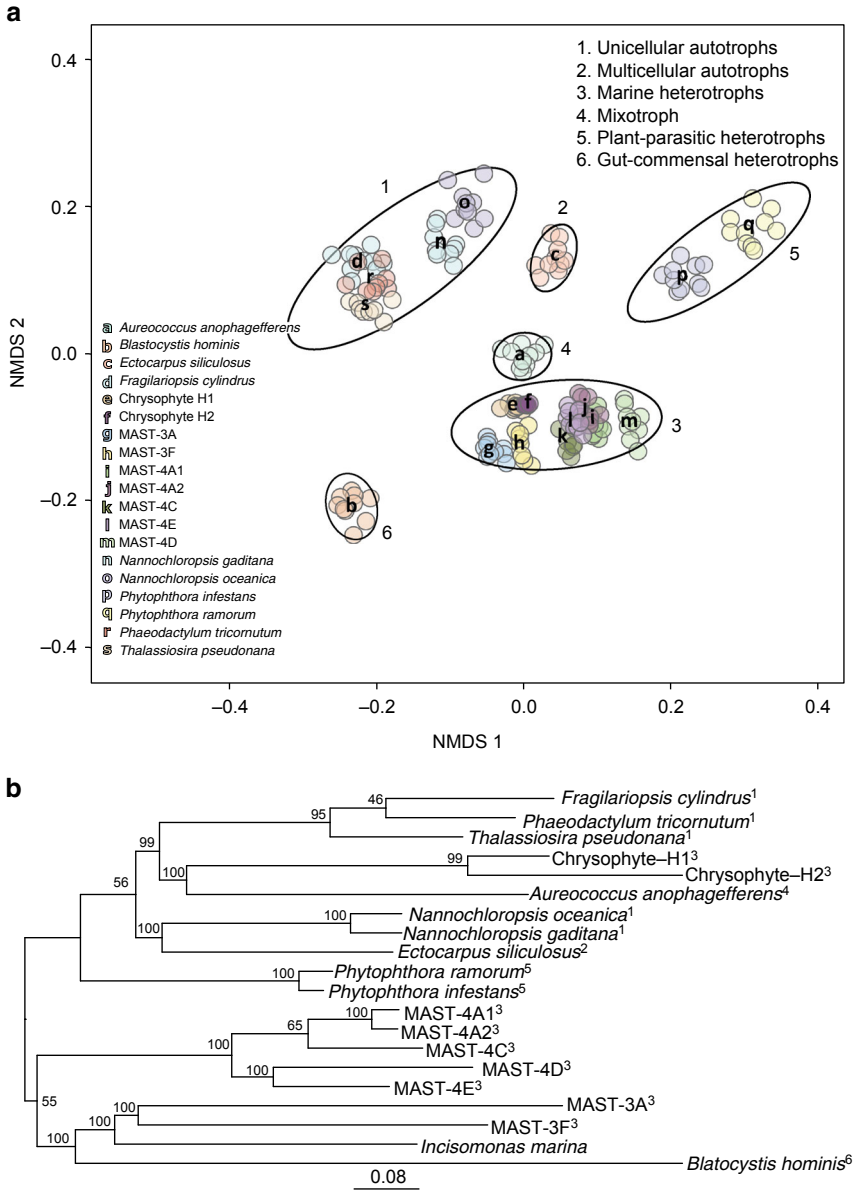


Fig. 1 Marine heterotrophic SAG lineages form a functional group distinct from autotrophs and other heterotrophs. **a** Non-metric multidimensional scaling (NMDS) projection of a Bray–Curtis distance matrix that shows Pfam motif occurrences in various stramenopile genomes. Because the genome sequences are incomplete, a rarefying procedure was applied to obtain 1400 Pfam motifs per genome. Ten independent rarefied samples were obtained and used for NMDS. Ellipses (at 95% confidence limit) were drawn by using the 'ordiellipse' function of the vegan package in R, with the group defined by life history mode (indicated by number in top right). Letters indicate the positions of the mean coordinates of the 10 rarefied Pfam counts per organism. The analysis was conducted on 19 stramenopile genomes, which included MAST-4D[8]. Marine heterotrophic stramenopiles from this study form a large but coherent group (Group 3), which is distinct from autotrophic species and heterotrophic species from other environments. **b** Phylogenetic tree from the analysis of a total of 160 conserved eukaryotic proteins using maximum likelihood. Protein sequences of *Incisomonas marina* from ref.[44] are included. Indices indicate life history mode as in panel **a**. Bootstrap values are represented on internal nodes. The branch length represents the mean number of substitutions per site

pseudogenization (Supplementary Fig. 4c and e–i), indicating that relatively recent gene loss events are responsible for the absence of DHC-encoding genes. Although we did not observe DHC reduction in the MAST-3F genome (Supplementary Table 1; Supplementary Fig. 4a), previous morphological analyses of other MAST-3 members had indicated reduced motility and the presence of only a single flagellum[12,13]. *Solenicola setigera* (MAST-3I clade) is found living epiphytically on diatoms, while the cultured *Incisomonas marina* (MAST-3J clade) seems to be a bad swimmer, with cells generally attaching to surfaces. Motility may therefore have been dispensed with on multiple occasions in these organisms, and may be congruent with the switch to epiphytic or parasitic lifestyles in several MAST-3 lineages.

We further observed the presence of rhodopsin coding genes exclusively in the MAST-4C lineage, suggesting again functional adaptation. Two rhodopsin classes with distinct functions are known: sensory rhodopsins act as light sensors for diverse signal transduction pathways, whereas proteorhodopsins are light-driven proton pumps that synthesize ATP independently of photosynthesis[21]. Phylogenetic analysis of these two rhodopsin genes revealed that they are related to previously described proteorhodopsins of diatoms, dinoflagellates and haptophytes,

and are evolutionarily distant from prokaryotic proteorhodopsins[22,23] (Supplementary Fig. 5). MAST-4C rhodopsins are thus eukaryotic proteorhodopsins, not derived from recent bacterial gene transfers. No proteorhodopsins were found in the other lineages, suggesting a specific genetic adaptation of MAST-4C to phototrophy. The MAST-4C proteorhodopsin genes appear to be highly expressed in surface samples, representing more than 3% of the total MAST-4C transcripts (Supplementary Fig. 5b). We further observed that MAST-4C cells were preferentially detected in samples from tropical surface waters (see below).

We then explored the gene families related to organic carbon acquisition in the various MAST lineages, and used Carbohydrate-active enzymes (CAZymes) as indicators of nutrient acquisition and more generally of organismal glycobiological potential[24]. The CAZyme-encoding gene profiles indicated a large repertoire of glycoside hydrolases (GHs) in almost all genomes, with many bearing secretion peptide signals (Supplementary Table 2). This is consistent with the bacterivorous lifestyle proposed for most of these organisms, which have the capacity to degrade bacterial carbohydrates and to target them for degradation in phagosomes. MAST-4 was found to be the most CAZyme-rich group, consistent with it including only bacterivorous lineages. On the other hand, MAST-3F appears to have a very limited CAZyme repertoire, almost none of which appear to be secreted. The MAST-3F genome also encodes fewer hydrolytic enzymes of other types, such as proteases (Supplementary Table 1), indicating that MAST-3F may not be bacterivorous. The other most CAZyme-poor genomes are those of chrysophytes, a group containing many photosynthetic organisms with mixotrophic behavior. This suggests complex evolutionary patterns in chrysophyte genomes, with intricate losses and/or gains of genes involved in photosynthesis and heterotrophy.

Putative substrates were predicted on all encoded CAZymes theoretically capable of cleaving complex carbohydrates (GHs and polysaccharide lyases) to reveal which enzymes are involved in bacterivory and possible carbohydrate acquisition from other sources (Fig. 2). Identification of lysozymes from the GH25 family in most co-assembled genomes could be indicative of peptidoglycan breakdown. Moreover, in all MAST-4 and MAST-3A genomes, suites of genes encoding enzymes able to hydrolyze all the components of green and brown algal cell walls were detected, including cellulose, xylan, pectin, and agarose (Fig. 2). Interestingly, examination of sequences that were considered as contaminants during genome reconstruction revealed large fragments of chloroplast, and sometimes even nuclear, DNA from photosynthetic eukaryotes in two of the MAST-4A and one of the MAST-4E cells, but not in any of the other lineages (Supplementary Table 2). MAST-4 was previously shown to have the capacity to ingest eukaryotic microalgae in an experimental setting in the presence of high algal concentrations[25]. Our observations provide further evidence for the role of MAST-4 and MAST-3A in algal consumption, which could have a significant impact on the transfer of organic material from primary producers to higher trophic levels. Further function predictions identified candidate secreted enzymes for the breakdown of starch, chitin, and beta-1,3-glucans (Fig. 2). The above observations imply that the examined organisms may have the capacity to degrade organic materials from bacteria and algae, as well as from chitin-containing organisms, such as fungi, diatoms, and crustaceans, emphasizing their global involvement and differentiated roles in the microbial loop.

For the MAST-4A, MAST-4C, MAST-4E, and MAST-3A genomes, the number of GH genes exceeded that of glycosyltransferases (GTs), with the GH/GT ratio ranging from 1.6 to 2, reflecting the heterotrophic nature of these organisms. However, the MAST-3F and chrysophytes H1 and H2 genomes displayed higher numbers of GTs than GHs, indicating that these organisms may be less dependent on carbohydrate degradation.

Horizontally transferred genes. Another fundamental question is whether heterotrophic protists are impacted by horizontal gene transfer (HGT) from the prey they ingest. We assessed the extent to which genes had probably been acquired by horizontal transfer from prokaryotes in each SAG lineage (see section "Methods"). The proportion of potential HGT events was different among the studied genomes (Supplementary Table 3). The lowest observed value was for MAST-3F, which was also the genome lacking elements suggestive of a bacterivorous lifestyle (see above). A link could therefore exist between bacterivory and prokaryotic gene acquisition in the other lineages. Furthermore, the functional classification of candidate HGTs based on Clusters of Orthologous Groups (COGs)[26] showed a bias towards metabolic activities (Supplementary Fig. 6a and 6b). Refining the metabolic COG categories revealed an even more pronounced bias towards activities linked to carbohydrate and protein degradation, defense/resistance against bacteria and nitrogen utilization (Supplementary Table 4). Overall, our data indicate that each MAST lineage may have a different functional profile in terms of organic matter processing, and that HGT may have contributed to enabling this metabolic specialization.

Geographical distributions. Finally, we used metagenomic fragment recruitment from the 0.8 to 5 µm size-fraction of the *Tara* Oceans metagenomics dataset to explore the global distribution of the studied lineages and of MAST-4 D (Fig. 3). In addition to quantifying lineage-specific abundances, metagenomics data was used to obtain indications of genetic diversification by using the similarity of nucleotide sequences to each reference genome as a measure of divergence (Supplementary Fig. 7). Widely differing geographic distributions were observed. First, the previously sequenced MAST-4 D genome is encountered in only one coastal sample from the South Atlantic Ocean, indicating that open ocean populations of MASTs can differ from coastal ones. In the studied lineages, only one organism with a well-conserved genotype, MAST-4A, appears to be cosmopolitan, although it was not detected in the Southern Ocean. Another group, MAST-4C, displays high genetic homogeneity worldwide but with a geographic range restricted mostly to tropical and subtropical waters, except in the sub-tropical Atlantic Ocean. In other cases, we observed the existence of genotype subsets divergent from the reference genomes, with preferential geographic patterns (MAST-4E, MAST-3A, and chrysophyte H1). Finally, chrysophyte H2 and MAST-3F are low-abundance species encountered in different regions as divergent genotypes.

Each of the distributions was compared to the environmental parameters recorded at each sampling site[27,28] (the four most significant parameters are highlighted in Supplementary Fig. 8). The most significant parameter that discriminate the distributions (Kruskal–Wallis test p-value = 2.2×10^{-16}) was water temperature (Fig. 4a), suggesting that some of these species likely have preferential temperature ranges in which they are maximally abundant. Divergent MAST-3A and MAST-4E genomes were found in water temperatures distinct from where organisms with genomes more similar to the reference SAG genome thrive (Wilcoxon test, p-value $<2\times10^{-2}$ and p-value $<3\times10^{-4}$, respectively; Fig. 4b and c). Finally, depth-dependent distributions were also frequent, with MAST-4C and MAST-3A being located preferentially in the subsurface, while MAST-4E and Chrysophyte H1 were found predominantly at the deep chlorophyll maximum (DCM), except in well-mixed water columns (Fig. 3).

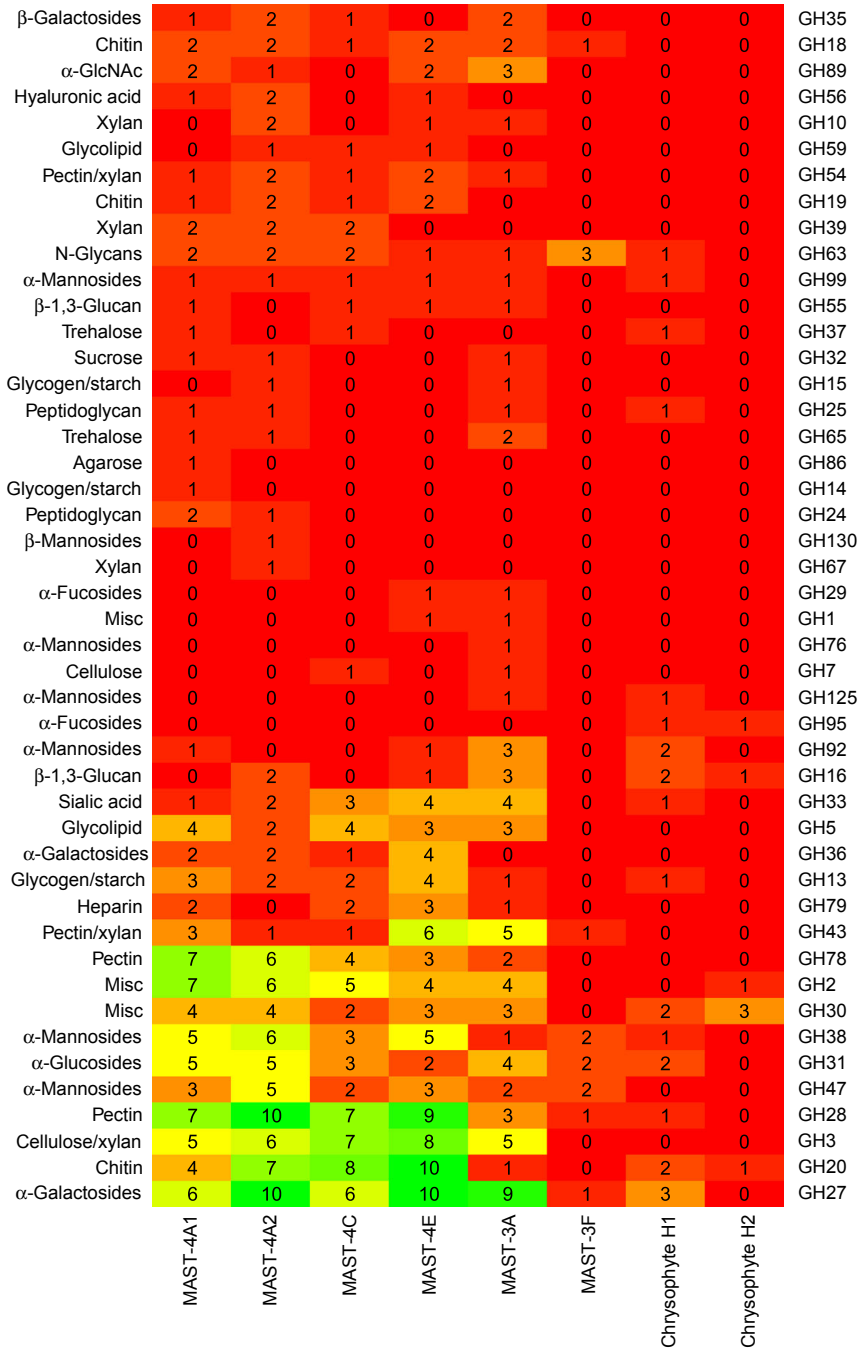


Fig. 2 SAG lineage glycoside hydrolases (GHs). GH families are numbered (right) according to the CAZyme database. Potential substrates are indicated on the left side. Internal numbers represent the number of genes in each genome predicted to belong to the GH category. Colors indicate the number of predicted GH genes per family, from low (red) to high (green)

Discussion

Our findings indicate that each of the examined taxa may have a specific spatial distribution that correlates with environmental parameters, principally ocean provinces, temperature, and depth. However, some limitations of the data set—mostly its single time point per location, the use of *Tara* Oceans metagenomes as the only resource, the relatively low resolution of sampling points per geographical area, and the absence of metagenomics replicates—may have under-estimated the true distribution of the organisms studied here. Notwithstanding, the *Tara* Oceans data set is by far the largest available today, and is the only extensive metagenomics effort tackling specifically the size fraction where these heterotrophic protists can be found (no additional location was revealed using the other available size fractions). The relatively low resolution of sampling locations is balanced by a careful choice of oceanographic situations in each sampled region. The depth of sequencing is also particularly significant compared to other studies (at about 25 Gb per sample), so the use of replicates will be of low utility for detecting the presence of the genomes under study here. The major limitation in our view is the absence of temporal information from each sampling location. Although *Tara* Oceans was a 3-year expedition that sampled plankton across all seasons, each location is currently described at a single time only and so it will be interesting to extend our results in future sampling campaigns by targeting sites of interest during different seasons.

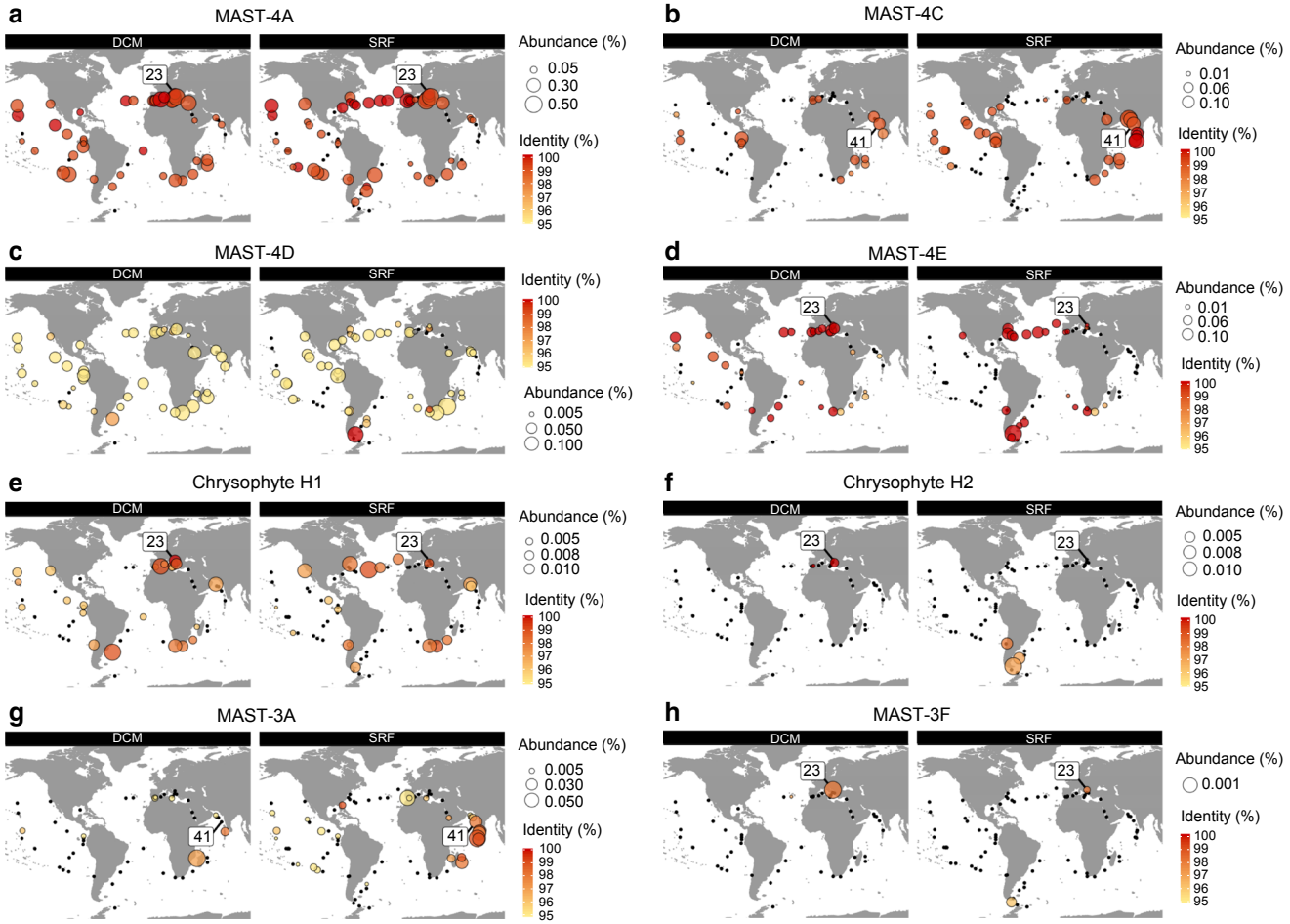


Fig. 3 Biogeographic distribution of the SAG lineages based on metagenome read recruitment with separation between deep chlorophyll maximum and subsurface. Global maps showing the presence of the SAG lineages based on metagenomics read mapping at each *Tara* Oceans station either as a black dot (no signal detected) or as a circle whose diameter indicates the species relative abundance. Abundance in samples from deep chlorophyll maximum (DCM, left panel) often differs from surface samples (SRF, right panel): only MAST-4A shows the same pattern in DCM and SRF samples (**a**). The color inside each circle provides the median percentage similarity of the reads to the reference. The station from where the SAG originates is indicated by its number. **a** MAST-4A; **b** MAST-4C; **c** MAST-4D; **d** MAST-4E; **e** Chrysophyte H1; **f** Chrysophyte H2; **g** MAST-3A; and **h** MAST-3F

Moreover, the differentiated gene content between taxa suggests specific distinctive functional capacities even within taxa. This indicates that, like prokaryotes and phytoplankton[29–31], heterotrophic protists are not interchangeable components of marine plankton ecosystems, but effectively participate from varied perspectives in the highly complex networks of interacting taxa[4,32].

Methods

Single-cell isolation and amplification. Aquatic samples were collected during the *Tara* Oceans expedition[23,33]. One-milliliter aliquots were amended with 6% (final concentration) glycine betaine and stored at −80 °C[34]. Flow-cytometric sorting, whole genome amplification, and sequencing of partial 18S rRNA genes of single cells were performed by the Bigelow Laboratory Single Cell Genomics Center (https://scgc.bigelow.org/), following previously described protocols[5,7] with a slight modification: 1x SYBR Green I (Life Technologies Corporation) was used instead of Lysotracker Green to stain the cells[18]. The 40 SAGs analyzed in this study came from the Mediterranean Sea (sampled in November 2009) and Indian Ocean (sampled in March 2010) (Table 1). Cell sorting was performed on cells lacking chlorophyll. Therefore all cells were considered heterotrophic.

Sequencing and assembly. The steps used for assembly, annotation, and contamination control are summarized in Supplementary Fig. 1a. Library preparation from single cells is described in Alberti et al.[18]. All cells were independently sequenced on a 1/8th Illumina HiSeq lane, which produced ~25 million 101-bp paired-end reads. Reads from SAGs with highly similar 18S were first co-assembled using the HyDA assembler[35]. Based on colored de Bruijn graphs, HyDA outputs the contribution of each library to each contig, which provides a criterion to determine which libraries can be co-assembled: only libraries that cover a large fraction of the longest contigs were pooled, which ensured that the genomes were close enough to be co-assembled. Libraries that were successfully co-assembled with HyDA were then re-assembled using SPAdes 2.4[36], which provided the best results in terms of assembly size, N50 and number of core eukaryotic genes recovered. Although SPAdes provides an integrated scaffolder, we re-scaffolded contigs with SSPACE v2[37] and filled gaps with GapCloser (SOAPdeNovo2 package [v 1.12-6][38]). Scaffolds shorter than 500 bp were discarded from the assembly. Accession numbers of generated assemblies can be found in Supplementary Table 5.

Removal of organelle sequences. Because we found nearly identical organellar DNA sequences in different SAG assemblies, we suspected a potential biological or technical contamination of these highly amplified sequences and decided to completely separate organellar sequences from the assemblies.

The presence of organellar scaffolds was searched using a combined approach. First a BLASTn analysis was done using scaffolds as queries against a database that contained all sequenced organelle genomes. Scaffolds similar to a known organelle genome (bit score >1000) were flagged. Then, a scaffold was considered to have an organelle origin if at least three predicted proteins from the scaffold showed similarities to proteins from the Curated Chloroplast Protein Clusters (CHL) or Curated Mitochondrial Protein Clusters (MTH) databases (http://www.ncbi.nlm.nih.gov/books/NBK3797/). Then, the two lists were merged. The scaffolds that were inferred to have come from organelles were retrieved from the SAG dataset for subsequent analysis and the corresponding proteins were removed from the nuclear protein dataset.

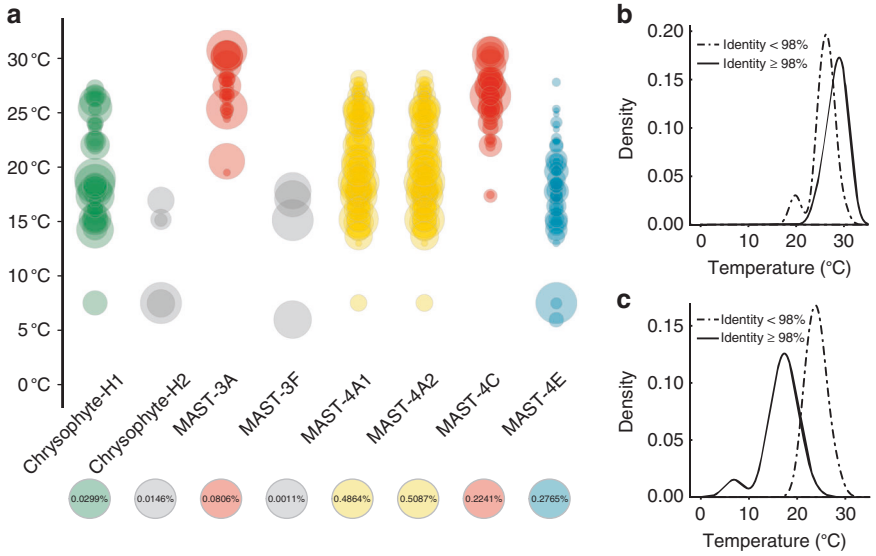


Fig. 4 Water temperature and distribution of the heterotrophic protists. **a** *x*-axis represents the lineage composite genomes, and *y*-axis represents surface temperatures in degrees Celsius at each sampling location. Relative abundances are represented by circle size (one per station/depth where the genome was detected). The scale for each column is indicated below the name of the lineage. **b** MAST-3A abundance distribution relative to temperature. A difference in the distributions is observed with a *p*-value $<2\times10^{-2}$ (Wilcoxon test). **c** MAST-4E abundance distribution relative to temperature. Means are statistically different with a *p*-value $<3\times10^{-4}$ (Wilcoxon test). In **b** and **c**, line type indicates median sequence similarity to the reference genome assembly

Cross-genera contamination removal. To detect identical scaffolds in the distantly related SAGs from this study, all scaffolds were cut into 1000 bp-long fragments along a 500-pb overlapping sliding window. We used entire sequences of scaffolds shorter than 1000 bp. We aligned these fragments on each target assembly with BLAT and kept alignments with ≥95% identity >80% length. For each assembly, we considered and discarded contigs with at least one selected match with a distant phylum as contaminants. We distinguished three taxa: chrysophytes, MAST-3, and MAST-4. Subsequently, we assessed assembly completion using the BUSCO v2 pipeline[39] with the eukaryotic set of genes.

Gene prediction. Protein-coding genes were predicted by combining alignments of proteins from a custom database built from Uniref100 and MMETSP, alignments of transcripts from the *Tara* Oceans collection and ab initio gene models. The combination step was performed using the GAZE framework.

The custom protein database was based on Uniref100, with the addition of curated translated CDS from MMETSP transcripts and in-house sequenced transcriptomes. The final dataset contained more than 26 million proteins that were aligned using a two-step strategy. Protein sequences were first aligned using the fast BLAT program and significant matches were then re-aligned using the more accurate Genewise v2.2.0 software.

Transcripts from the *Tara* Oceans metatranscriptomic dataset were mapped using BLAST + 2.2.28. Significant alignments were then refined using est2genome, in particular to properly define exon–intron boundaries. To select organism-specific transcripts and avoid false positives, we only retained transcripts with ≥95% identity and with ≥80% of their length aligned onto the assembly.

Ab initio models were predicted using SNAP (v2013-02-16) trained on complete protein matches. Because of the insufficient number of complete proteins matching the MAST-3 F assembly, SNAP was trained on MAST-3 A assembly before running on that of MAST-3 F (Supplementary Table 6).

GAZE framework was used to integrate these three types of resources, using different weights to reflect their reliability. The most reliable resources—transcript alignments—were weighted 6.0, whereas protein alignments were weighted 4.5 and ab initio models 1.0. The weight acts as a multiplier for the score of each resource to build the final gene structure. Gene predictions with a GAZE score ≥0 were selected.

Bacterial decontamination. Bacterial scaffolds were detected using the alien index (AI)[40] calculated on each predicted gene. The alien index was defined as log(best eukaryotic hit *e*-value + 10^{-200})–log(best non-eukaryotic hit *e*-value + 10^{-200}). Thus, purely eukaryotic genes have a negative value whereas prokaryotic genes have a positive value. Scaffolds with predicted genes having an AI > 45 exclusively were considered as bacterial scaffolds and discarded from the final assembly.

Metagenomic sequencing and mapping. We sequenced 122 samples (accession numbers and contextual data in Supplementary Data 1–3) from 76 stations from the 0.8 to 5 µm size fractions (the size fraction where the studied MAST lineages are most abundant), and obtained a total of 23.1×10^9 Illumina 101-bp paired-end reads. Reads from the 0.8 to 5 µm fraction size samples were mapped, in a three-step pipeline. In order to avoid the computation-intensive mapping of all reads, we first selected reads with at least one 25-mer in common with the target assembly. We then mapped the selected reads using bowtie2 2.1.0 aligner[41] with default parameters. Finally, we filtered alignments that correspond to low complexity regions using the DUST algorithm: alignments with <95% mean identity or <30% of high complexity bases were discarded.

Discarding contaminants through metagenomic signatures. The presence of unrelated sequences in the assembly was analyzed using a combination of approaches to obtain a list of scaffolds with atypical or suspect content. First, eukaryotic and prokaryotic signatures were determined for each scaffold. For this, a BLASTx analysis was conducted using the predicted gene as query against the nr-prot database (*e*-value threshold $<1\times10^{-0.5}$) followed by taxonomic assignment of each hit. A scaffold was determined to have a eukaryotic signature if it presented either at least one prediction assigned to one eukaryotic organism or none of the gene predictions had any similarities in the database. The scaffolds without these signatures were removed from the dataset. Second, we developed a new method to identify a population of scaffolds that co-vary in representation in the metagenomic data (see details below). This method identified outlier and inlier genes. The outlier dataset included genes with atypical behavior relative to the whole population of genes. Scaffolds that contained all genes that belonged to the outlier dataset were discarded. Supplementary Fig. 1b depicts an example of two different outlier scaffold groups (red), compared with the inlier scaffolds (blue). The three approaches were combined, which facilitated generation of a cleaned scaffold dataset and a corresponding cleaned gene dataset.

Gene functional analysis: comparison of Pfam domain content between stramenopile genomes. CDD search 3.11 was used for functional annotation of SAG genomes. Annotation was conducted on the cleaned gene dataset (see above) including outlier genes contained within single-gene scaffolds. We retrieved the Pfam motifs from CDD search output. Multiple occurrences of the same Pfam motif in one protein were counted as one. To perform a comparative analysis of the Pfam signature in the stramenopile taxa, we retrieved the protein dataset of representative available stramenopile genomes. To homogeneize these datasets from different projects, functional annotation of these gene datasets was performed. Proteins with similarities to CHL and MTH clusters were retrieved from the prior analysis. Because genome completeness was not similar between SAG lineages, random sorting of 1400 Pfam domains was independently performed 10 times for each genome. This threshold was selected because 1414 was the lowest number of Pfams, found per genome. A matrix with Pfam motif occurrence for all stramenopiles (10 random samplings per organism) was obtained. To visualize differences between Pfam content in stramenopile communities, we used non-metric multi-dimensional scaling (NMDS) based on Bray–Curtis dissimilarity distance. Bray–Curtis was used instead of Pearson correlation factor, because Bray–Curtis is unaffected by the addition or removal of Pfam motifs that are not

present in two gene repertoires. Moreover, it is unaffected by the addition of a new genome in the analysis. If Euclidean distance measures were used, the presence of double zeros in Pfam matrix abundance data may result in two genomes without any Pfam motifs in common being found to be more similar than other genome pairs with shared motifs. Bray–Curtis calculation and NMDS were created using the vegan package (v1.17-11) in R. Ellipses (95% confidence limit) were drawn in vegan using the ordiellipse function, with each group defined by common life history mode.

Phylogenomic analysis. The maximum likelihood phylogenetic tree of sequenced stramenopiles was reconstructed from conserved eukaryotic proteins detected using the BUSCO v2 pipeline. A total of 160 protein sequences present in at least four SAG assemblies were aligned using MUSCLE v3.8.31. Alignments were manually inspected to remove non-orthologous proteins (false positive detection with BUSCO). Subsequently, they were trimmed with Gblocks v0.91b using more relaxed parameters than default (*–b4*=5 *–b3*=4). Remaining trimmed sequences were concatenated. Because the selected 160 proteins were not present in all genomes, missing sequences were replaced by gaps ('-', character). Thus, the effective number of sequences used to infer phylogeny was often much lower than 160 (Chrysophyte H2: 51; MAST-4D: 72; MAST-3F: 73; MAST-3A: 88; MAST-4C: 90; MAST-4A1: 113; Chrysophyte H1: 113; MAST-4E: 115; MAST-4A2: 115). Phylogeny was inferred using RAxML v8.2.9 under the GAMMA model of heterogeneity in evolutionary rates among sites and using the JTT substitution model. Branch support was evaluated using 100 bootstrap pseudoreplicates.

CAZyme analysis. Using BLASTp[42], each encoded protein model was compared to the proteins listed in the CAZy database[24] (http://www.cazy.org/). Proteins with >50% identity over the entire domain length of an entry in CAZy were directly assigned to the same family, whereas proteins with 15–50% identity to a protein in CAZy were all manually inspected, aligned, and searched for conserved features, such as catalytic residues. Functional prediction was performed by BLASTp comparison of the candidate CAZymes against a library constructed with only the biochemically characterized CAZymes reported in the CAZy database under the 'characterized' tab of each family[43].

HGT detection. The presence of putative HGT events was determined using two methods. First, in the AI method[40], the 'inlier' gene dataset was used to query nr-prot (April 2014 version), and the BLASTx search output was used to calculate the AI. Additionally, a second step was also added to the AI method because the AI calculation is made using the first best hit from eukaryotes and prokaryotes: If a gene is wrongly assigned as prokaryotic, it would be erroneously considered an HGT event (false positive). Alternatively, if a closely related organism with a common HGT event is present in the database used for the BLAST search, a gene could be excluded from the putative HGT list (false negative). Consequently, the first 1000 hits were retrieved, taxonomically assigned, and classified in eukaryotic and prokaryotic classes. We considered genes with an AI > 45, predicted internally on a scaffold with more than five predicted genes as putative HGTs.

To validate these putative HGTs, we constructed a phylogenetic tree of the predicted protein and its 200 best BLASTp matches (Supplementary Data 4), but only allowing a maximum of three matches from the same genus to extend the sampled diversity. If less than 10 eukaryotic sequences were present in the 200 best BLAST matches, we included the 10 closest eukaryotic matches of all BLAST matches (8000 max). Sequences were aligned using MUSCLE 3.8.31 and non-conserved positions were discarded using GBlocks 0.91b with relaxed parameters (*–b3*=10 *–b4*=5 *–b5*=*h*). Phylogeny was inferred using RAxML 8.2.9 with JTT model and gamma model of rate heterogeneity (*–m* PROTGAMMAJTTX parameter). We considered the tree to support the horizontal transfer hypothesis if the investigated gene did not cluster with other eukaryotic sequences (bootstrap value >50). In the other case, the putative HGT was eliminated and considered as a False Positive of the alien index method.

Annotation of bacterial enzymatic activities in HGT. A functional classification of HGTs was obtained using Intepro and Pfam motifs, and functional categories were determined using COG. The HGT protein sequences were used for protein-versus-protein alignments, using the BL2 option (BLAST allowing gaps) and a BLOSUM62 score matrix against UniProtKB. Those that had >30% identity over at least 80% of the length of the smaller of two compared sequences were kept. The best hit for each HGT was then selected. For each best hit, Interpro and Pfam classification identifiers were retrieved using the UniProtKB interface. Each HGT protein was then manually assigned to one functional category (cellular process and signaling, information storage and processing, metabolism, or poorly characterized) using their best hit functional annotation and signatures.

Biogeography inlier/outlier detection. The measurement of an organism's relative abundance from short-read metagenomic information is very difficult, because some genes may be highly homologous to orthologous genes from others organisms and attract cross-mapping metagenomic reads. Here, we present a statistical approach to discriminate genes with atypical mapping behavior. This analysis relies on the assumption that the values of the metagenomic RPKM (number of mapped reads per gene (intron plus exon) per kb per million of mapped reads) per gene follow a normal distribution. The presence of genes with mapping values distant from the majority of genes could have numerous causes, such as (i) presence of a scaffold coming from another organism, (ii) cross mapping, or (iii) genes with a high copy number. Outlier presence was determined using the Grubb's test. The test was conducted for a station if at least 20% of the organism's genes were detected. A gene was considered detected if at least one read mapped with 95% identity on 100% of the read length. The outlier lists for each station were merged to provide the outlier gene list. This detection allowed clear discernment of genes usable for relative abundance measurement (the inlier dataset) from unusable genes with noisy or random signal (the outlier dataset). Organism abundance measurements across stations is highly dependent on this filter (Supplementary Fig. 9a, b, f, and g), necessary for this type of analysis. However, the abundance measured in one station resulted from the combination of inlier and outlier genes (as in station 89 and 85 at surface, Supplementary Fig. 9c). The high number of stations sampled during the *Tara* Oceans expedition allowed us to show that outlier genes were detected in a large number of stations, which is expected for non-specific signals (Supplementary Fig. 9d, e).

Biogeographic distributions. Genes detected as outliers were removed from the biogeographic analysis. The relative abundance of an organism was measured as the sum of the number of mapped reads per gene divided by the total number of reads sequenced per station. Because only genes and not intergenic regions were used, a correction factor was applied to the relative abundance values: corrected relative abundance = raw relative abundance × assembly size/(size of the mapped genome × genome completion). The abundance in a geographical area was calculated as the mean of the relative abundance of all stations in the corresponding geographical area (Atlantic Ocean, Mediterranean Sea, Indian Ocean, Southern Ocean, and Pacific Ocean). For the world maps (e.g., Fig. 3), and to compare the SAG lineage abundance and reveal common patterns of occurrence, the data were normalized by dividing the relative abundance by the maximal relative abundance per organism. The world maps were generated using the R packages maps_2.1-6, mapproj 1.1-8.3, gplots_2.8.0, and mapplots_1.4.

Correlations to environmental parameters. We tested whether the SAG lineage presence and/or abundance in *Tara* Oceans samples were correlated with local physico-chemical conditions. We used physico-chemical parameter values obtained from each sampling site during the expedition, which are available in the PANGAEA database[27]. For each parameter, we performed a Kruskal–Wallis one-way test and a post-hoc Tukey's test. We statistically delineated SAG lineage classes. Only stations for which we detected at least 20% of genes from each composite assembly lineage were considered. MAST-3F was not present at a sufficient number of stations and was therefore excluded from statistical analyses.

Code availability. Computer code used to perform comparative genomics, calculate relative abundances and represent biogeographies is available from the corresponding authors upon request.

Data availability. Sequencing data are archived at ENA under the accession number PRJEB6603 for the SAGs (see Supplementary Table 5 for details) and PRJEB4352 for the metagenomics data (see Supplementary Data 3). All other relevant data supporting the findings of the study are available in this article and its Supplementary Information files, or from the corresponding authors upon request.

Received: 10 May 2017 Accepted: 15 November 2017
Published online: 22 January 2018

References

1. Azam, F. & Malfatti, F. Microbial structuring of marine ecosystems. *Nat. Rev. Microbiol.* **5**, 782–791 (2007).
2. Worden, A. Z. et al. Environmental science. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594 (2015).
3. de Vargas, C. et al. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
4. Lima-Mendez, G. et al. Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
5. Heywood, J. L., Sieracki, M. E., Bellows, W., Poulton, N. J. & Stepanauskas, R. Capturing diversity of marine heterotrophic protists: one cell at a time. *ISME J.* **5**, 674–684 (2011).
6. Yoon, H. S. et al Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714–717 (2011).
7. Martinez-Garcia, M. et al Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *ISME J.* **6**, 703–707 (2012).

8. Roy, R. S. et al. Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci. Rep.* **4**, 4780 (2014).
9. Lasken, R. S. Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10**, 631–640 (2012).
10. Vannier, T. et al. Survey of the green picoalga Bathycoccus genomes in the global ocean. *Sci. Rep.* **6**, 37900 (2016).
11. Massana, R. et al Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl. Environ. Microbiol.* **70**, 3528–3534 (2004).
12. Massana, R., del Campo, J., Sieracki, M. E., Audic, S. & Logares, R. Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J.* **8**, 854–866 (2014).
13. Gomez, F., Moreira, D., Benzerara, K. & Lopez-Garcia, P. Solenicola setigera is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3. *Environ. Microbiol.* **13**, 193–202 (2011).
14. Cavalier-Smith, T. & Scoble, J. M. Phylogeny of Heterokonta: *Incisomonas marina*, a uniciliate gliding opalozoan related to Solenicola (Nanomonadea), and evidence that Actinophryida evolved from raphidophytes. *Eur. J. Protistol.* **49**, 328–353 (2013).
15. del Campo, J. & Massana, R. Emerging diversity within chrysophytes, choanoflagellates and bicosoecids based on molecular surveys. *Protist* **162**, 435–448 (2011).
16. Reyes-Prieto, A. & Bhattacharya, D. Phylogeny of nuclear-encoded plastid-targeted proteins supports an early divergence of glaucophytes within Plantae. *Mol. Biol. Evol.* **24**, 2358–2361 (2007).
17. Giering, S. L. et al Reconciliation of the carbon budget in the ocean's twilight zone. *Nature* **507**, 480–483 (2014).
18. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4**, 170093 (2017).
19. Carradec, Q. et al. A global ocean atlas of eukaryotic genes. *Nature Commun.* https://doi.org/10.1038/s41467-017-02342-1.
20. Mangot, J. F. et al. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci. Rep.* **7**, 41498 (2017).
21. Beja, O. et al Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 1902–1906 (2000).
22. Slamovits, C. H., Okamoto, N., Burri, L., James, E. R. & Keeling, P. J. A bacterial proteorhodopsin proton pump in marine eukaryotes. *Nat. Commun.* **2**, 183 (2011).
23. Marchetti, A. et al Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proc. Natl. Acad. Sci. USA* **109**, E317–E325 (2012).
24. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
25. Massana, R. et al Grazing rates and functional diversity of uncultured heterotrophic flagellates. *ISME J.* **3**, 588–596 (2009).
26. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–D269 (2015).
27. Pesant, S. et al. Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**, 150023 (2015).
28. Tara Oceans Consortium, C., Tara Oceans Expedition, Participants. *Methodological context of all samples from the Tara Oceans Expedition (2009–2013).* (2015).
29. Brown, M. V. et al. Global biogeography of SAR11 marine bacteria. *Mol. Syst. Biol.* **8**, 595 (2012).
30. Martiny, A. C., Tai, A. P., Veneziano, D., Primeau, F. & Chisholm, S. W. Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus. *Environ. Microbiol.* **11**, 823–832 (2009).
31. Swan, B. K. et al Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci. USA* **110**, 11463–11468 (2013).
32. Guidi, L. et al Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
33. Karsenti, E. et al. A holistic approach to marine eco-systems biology. *PLoS Biol.* **9**, e1001177 (2011).
34. Swan, B. K. et al Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**, 1296–1300 (2011).
35. Chitsaz, H. et al Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat. Biotechnol.* **29**, 915–921 (2011).
36. Bankevich, A. et al SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
37. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
38. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18 (2012).
39. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
40. Gladyshev, E. A., Meselson, M. & Arkhipova, I. R. Massive horizontal gene transfer in bdelloid rotifers. *Science* **320**, 1210–1213 (2008).
41. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
42. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
43. Cantarel, B. L. et al The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233–D238 (2009).
44. Derelle, R., Lopez-Garcia, P., Timpano, H. & Moreira, D. A phylogenomic framework to study the diversity and evolution of Stramenopiles (=Heterokonts). *Mol. Biol. Evol.* **33**, 2890–2898 (2016).

Acknowledgements

We thank the commitment of the following people and sponsors who made this singular expedition possible: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, the French Government 'Investissement d'Avenir' programs Oceanomics (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), Fund for Scientific Research—Flanders, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, ANR (projects 'PHYTBACK/ANR-2010-1709-01', POSEIDON/ANR-09-BLAN-0348, PROMETHEUS/ANR-09-PCS-GENM-217, TARA-GIRUS/ANR-09-PCS-GENM-218), EU FP7 (MicroB3/No. 287589, IHMS/HEALTH-F4-2010-261376), ERC Advanced Grant Award to CB (Diatomite: 294823), US NSF grant DEB-1031049 to M.E.S. and R.S., FWO, BIO5, Biosphere 2, agnès b., the Veolia Environment Foundation, Region Bretagne, World Courier, Illumina, Cap L'Orient, the EDF Foundation EDF Diversiterre, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, the *Tara* schooner and its captain and crew. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge C. Scarpelli for support in high-performance computing. This article is contribution number 63 of *Tara* Oceans.

Author contributions

R.M., O.J., M.Si., C.d.V., and P.W. designed the study. P.W. wrote the paper with substantial input from S.M., Y.S., Q.C., V.d.B., E.K., C.B., D.I., R.S., R.M., B.H., O.J., M.S., S. Su., C.d.V., P.H. and M.B.S. C.D., M.P., S.K.L., S.Se., and S.P. collected and managed *Tara* Oceans samples. J.P. and K.L. coordinated the genomic sequencing. N.P., R.S., and M.S. conducted SAG generation and identification. S.M., Y.S., Q.C., E.P., M.W., J.L., V.L., J.F. M., R.L., V.d.B., M.Sa., R.M., J.M.A., B.H., and O.J. analyzed the genomic data. D.I. analyzed oceanographic data. *Tara* Oceans Coordinators provided a creative environment and constructive criticism throughout the study. All authors discussed the results and commented on the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-017-02235-3.

Competing interests: The authors declare no competing financial interests.

Yoann Seeleuthner[1,2,3], Samuel Mondy[1,2,3], Vincent Lombard[4,5,6], Quentin Carradec[1,2,3], Eric Pelletier[1,2,3], Marc Wessner[1,2,3], Jade Leconte[1,2,3], Jean-François Mangot[7], Julie Poulain[1], Karine Labadie[1], Ramiro Logares[7], Shinichi Sunagawa[8,9], Véronique de Berardinis[1,2,3], Marcel Salanoubat[1,2,3], Céline Dimier[10,11,12], Stefanie Kandels-Lewis[8,13], Marc Picheral[14], Sarah Searson[15], Tara Oceans Coordinators, Stephane Pesant[16,17], Nicole Poulton[18], Ramunas Stepanauskas[18], Peer Bork[8], Chris Bowler[12], Pascal Hingamp[19], Matthew B. Sullivan[20], Daniele Iudicone[21], Ramon Massana[7], Jean-Marc Aury[1], Bernard Henrissat[4,5,6,22], Eric Karsenti[12,15,16], Olivier Jaillon[1,2,3], Mike Sieracki[23], Colomban de Vargas[10,11] & Patrick Wincker[1,2,3]

[1]CEA - Institut de biologie François Jacob, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [2]CNRS, UMR 8030, CP5706 Evry, France. [3]Université d'Evry, UMR 8030, CP5706 Evry, France. [4]Centre National de la Recherche Scientifique, UMR 7257, F-13288 Marseille, France. [5]Aix-Marseille Université, UMR 7257, F-13288 Marseille, France. [6]INRA, USC 1408 AFMB, F-13288 Marseille, France. [7]Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC), E-08003 Barcelona, Catalonia, Spain. [8]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany. [9]Institute of Microbiology, Department of Biology, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zürich, Switzerland. [10]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [11]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [12]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France. [13]Directors' Research European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany. [14]Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'océanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche sur Mer, France. [15]Department of Oceanography, University of Hawaii, 96815 Honolulu, Hawaii, USA. [16]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, 28359 Bremen, Germany. [17]MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany. [18]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA. [19]Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO UM 110, 13288 Marseille, France. [20]Departments of Microbiology and Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH 43210, USA. [21]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [22]Department of Biological Sciences, King Abdulaziz University, Jeddah, 21589, Saudi Arabia. [23]National Science Foundation, Arlington, VA 22230, USA. Yoann Seeleuthner and Samuel Mondy contributed equally to this work

Tara Oceans Coordinators

Silvia G. Acinas[7], Emmanuel Boss[24], Michael Follows[25], Gabriel Gorsky[16], Nigel Grimsley[26,27], Lee Karp-Boss[24], Uros Krzic[28], Fabrice Not[11], Hiroyuki Ogata[29], Jeroen Raes[30,31,32], Emmanuel G. Reynaud[33], Christian Sardet[16,34], Sabrina Speich[35,36], Lars Stemmann[16], Didier Velayoudon[37] & Jean Weissenbach[1,2,3]

[24]School of Marine Sciences, University of Maine, Orono, Maine, 04469, USA. [25]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [26]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [27]Sorbonne Universités, Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [28]Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [29]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-001, Japan. [30]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [31]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [32]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [33]Earth Institute, University College Dublin, Dublin 4, Ireland. [34]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France. [35]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [36]Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France. [37]DVIP Consulting, 92310 Sèvres, France

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Analyses de variations génomiques liées à la biogéographie des picoalgues Mamiellales.

Mots clés : Métagénomique, Mamiellales, *Tara* Oceans, Ecologie marine, Génomes eucaryotes

Résumé : Les Mamiellales sont un ordre d'algues vertes unicellulaires cosmopolites comprenant des espèces d'importance écologique telles que *Bathycoccus*, *Micromonas* ou encore *Ostreococcus*, des contributeurs majeurs à la production primaire. Cette thèse prend pour modèle d'étude ce groupe phytoplanctonique aux génomes de référence connus afin d'analyser au mieux l'impact de l'environnement sur le plancton grâce aux échantillons provenant de l'expédition *Tara* Oceans.
Pour cela, différentes analyses ont été menées afin de définir leur biogéographie et leurs préférences écologiques, d'abord dans les eaux tempérées puis dans les eaux froides et riches en nutriments de l'océan Arctique. Dans les deux cas, il a été montré que la température était le principal facteur distinguant l'environnement dans lequel les différentes espèces ont été trouvées. Nous avons ensuite réalisé une étude plus poussée en particulier sur *Bathycoccus prasinos*, une espèce abondante dans ces deux milieux distincts afin d'établir la structure de ses populations, qui s'avère séparer clairement trois groupes: les échantillons australx, arctiques et tempérés, montrant encore une fois un impact de la température mais pas uniquement au vu de la distance génomique entre les deux premiers bassins.
Finalement, notre étude a pu être étendue avec diverses collaborations, nous permettant d'observer également un groupe de protistes hétérotrophes, les straménopiles, et de réaliser des analyses à l'échelle beaucoup plus large des communautés. L'ensemble de ces résultats concluent encore une fois, entre autre, à un fort impact de la température, menant à un questionnement sur le contexte actuel de changements climatiques et son potentiel impact sur le plancton.

Title: Analysis of genomic variations related to biogeography of Mamiellales picoalgae.

Keywords: Metagenomics, Mamiellales, *Tara* Oceans, Marine ecology, Eukaryotic genomes

Abstract: Mamiellales are an order of unicellular cosmopolitan green algae with ecologically important species such as *Bathycoccus*, *Micromonas* or *Ostreococcus*, major contributors to the primary production. This thesis uses this phytoplankton group with known reference genomes as a study model in order to better analyze the impact of the environment on plankton using samples from the *Tara* Oceans expedition.
To do this, different analyses were carried out to define their biogeography and ecological preferences, first in temperate waters then in the cold, nutrient-rich waters of the Arctic Ocean. In both cases, temperature was shown to be the main factor distinguishing the environment in which the different genomes were found. We then carried out a more detailed study in particular on *Bathycoccus prasinos*, a species abundant in these two distinct environments, in order to establish its population structure, which proved to be clearly separated into three groups: southern , arctic and temperate samples, again showing an impact of temperature but not only in view of the genomic distance between the first two basins.
Finally, our study was extended with various collaborations, allowing us to observe a group of heterotrophic protists, the stramenopiles, and to perform analyses at the much larger scale of communities. All of these results conclude once again, among other things, on the strong impact of temperature, leading us to contribute to the question about the current context of climate change and its impact on plankton.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France