

Rôle des protistes hétérotrophes marins dans le cycle du carbone océanique par génomique en cellule unique.

NNT : 2018SACLE002

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université d'Évry Val d'Essonne

École doctorale n°577 Structure et dynamique des systèmes vivants

Spécialité de doctorat: Sciences de la Vie et de la Santé

Thèse présentée et soutenue à Évry, le 26/02/2018, par

Yoann Seeleuthner

Composition du Jury :

Cécile Fairhead Professeur à l'Université Paris-Saclay	Présidente
Didier Debroas Professeur à l'Université Clermont Auvergne (UMR 6023)	Rapporteur
Hervé Moreau Directeur de recherche CNRS, UPMC (UMR 7232)	Rapporteur
Claire Lemaitre Chargée de recherche, INRIA Rennes – Bretagne Atlantique	Examinatrice
Colomban De Vargas Directeur de recherche CNRS (UMR 7144)	Examineur
Patrick Wincker Directeur de recherche, CEA (UMR 8030)	Directeur de thèse
Jean-Marc Aury Ingénieur chercheur, CEA	Co-directeur de thèse

Remerciements

Je tiens en premier lieu à remercier mon directeur de thèse, Patrick Wincker, qui m'a permis de réaliser cette thèse et qui s'est rendu disponible pour m'encadrer et me diriger pendant ces trois années, toujours avec une extrême patience. Merci également à Jean-Marc Aury, pour son encadrement, ses conseils avisés et l'aide qu'il m'a apportée, en particulier au début de ma thèse.

Je souhaiterais également remercier les membres de mon jury de thèse, qui ont accepté d'évaluer ce travail. Merci aux deux rapporteurs, Didier Debroas et Hervé Moreau, qui ont accepté de lire ce manuscrit et de donner leur avis sur ce travail. Un grand merci également à Cécile Fairhead, Claire Lemaitre et Colomban De Vargas pour avoir pris de leur temps afin d'être examinateurs durant ma soutenance.

Un énorme merci à toutes les personnes qui sont ou ont été au Genoscope, pour m'avoir si bien accueilli et pour tous les moments passés ensemble, je pense ne jamais oublier les années passées là-bas.

En particulier, je tiens à remercier les personnes avec qui j'ai beaucoup échangé : Amine Madoui, Thomas Vannier, Jade Leconte, Quentin Carradec, Sarah Farhat, Kevin Sugier, Janaina Rigonato. Merci à vous pour votre bonne humeur et pour les discussions autour d'un verre ou d'un plat de nouilles ! Merci aussi à tous les autres membres du LAGE, Éric Pelletier, Olivier Jaillon, Betina Porcel, France Denoed et Julie Poulain pour vos conseils et votre gentillesse.

Je remercie également Marc Wessner, Léo D'Agata, Corinne Da Silva, Gaëlle Samson, Shahinaz Gas, Adriana Alberti, Karine Labadie et Franck Aniere pour m'avoir apporté leur aide dans un nombre varié de tâches.

Merci également à Catherine Sarlande, Nancy Delpeche et Catherine Contrepois pour m'avoir aidé dans tous types de démarches administratives avec une incroyable efficacité.

Un grand merci à mes collègues de bureau qui m'ont supporté tout ce temps : Benjamin Noel, Marion Dubarry, Artem Kourlaiev, Tsinda Rukwavu et Amos Kirilovsky. J'ai eu la chance de partager un bureau avec des gens travailleurs et plein d'humour !

Un merci particulier à Samuel Mondy, qui s'est également beaucoup impliqué sur le sujet et a contribué à ce travail.

Pour finir, je tiens à remercier mes proches pour m'avoir soutenu et supporté. Merci à mes parents, à mes frères, à toute ma famille ainsi qu'à ma conjointe, Émilie, pour m'avoir soutenu durant les bons moments comme dans les moments un peu plus difficiles.

Merci à tous.

Table des matières

A.	Synthèse bibliographique	1
A.1.	Protistes marins	1
A.1.1.	Diversité des protistes marins	1
A.1.2.	Les protistes marins dans le cycle du carbone océanique	4
A.2.	Un embranchement très diversifié d'eucaryotes : les straménopiles	8
A.2.1.	Caractéristiques et phylogénie des straménopiles	8
A.2.2.	Les straménopiles marins MAST	10
A.2.3.	Les chrysophytes	13
A.3.	La génomique pour l'étude d'organismes non cultivés	15
A.3.1.	La méta-omique	16
A.3.2.	Les codes-barres génétiques	17
A.3.3.	La génomique en cellule unique	19
A.4.	L'expédition <i>Tara Oceans</i> : l'étude du plancton de la surface des océans à l'échelle globale 22	
A.4.1.	Parcours de la goélette durant l'expédition <i>Tara Oceans</i>	23
A.4.2.	Stratégie d'échantillonnage	25
A.4.3.	Séquençage métagénomique	28
A.4.4.	Construction d'un catalogue de gènes eucaryotes	28
A.4.5.	Isolation de cellules par cytométrie en flux	29
	Objectifs de la thèse	31
B.	Chapitre 1 : Analyse de génomes à partir de cellules uniques	33
B.1.	Introduction	33
B.2.	Matériels et méthodes	34
B.2.1.	Reconstruction des génomes	34
B.2.2.	Suppression de la contamination entre assemblages	35
B.2.3.	Annotation syntaxique des génomes	36
B.2.4.	Décontamination basée sur la signature métagénomique	39
B.3.	Conclusions	40
C.	Chapitre 2 : Diversité fonctionnelle de straménopiles incultivés	42
C.1.	Introduction	42
C.2.	Article 1 : <i>Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans</i>	42
C.3.	Étude globale des rhodopsines de protistes marins.	54
C.3.1.	Présentation de la famille protéique des rhodopsines	54
C.3.2.	Matériels et méthodes	56
C.3.3.	Résultats	57
D.	Chapitre 3 : Instantané de l'état physiologique des populations de MAST-4 clade A dans	

l'environnement	61
D.1. Introduction	61
D.2. Article 2 : <i>Probing metabolic states of the uncultured marine protist MAST-4 A using environmental metatranscriptomics</i>	62
D.3. Conclusions	87
Conclusions et perspectives	88
Références	92
Annexes.....	98
Annexe 1. Informations supplémentaires de l'article <i>Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans</i>	98
Annexe 2. <i>Survey of the green picoalga Bathycoccus genomes in the global ocean</i>	117
Annexe 3. <i>A global ocean atlas of eukaryotic genes</i>	129

A.Synthèse bibliographique

A.1. Protistes marins

A.1.1. Diversité des protistes marins

Le terme « protiste » a été introduit par Haeckel en 1866 pour désigner un règne du vivant, aux côtés des plantes et des animaux, regroupant tous les organismes « inférieurs » tels que les eucaryotes unicellulaires, les bactéries et les champignons. Après différentes révisions du système de Haeckel, puis par la suite grâce aux analyses phylogénétiques, la classification moderne à 3 règnes est apparue (procaryotes, archées et eucaryotes) et le terme « protiste » a été utilisé pour désigner les eucaryotes unicellulaires. Même si la paraphylie de ce groupe ne fait aujourd'hui aucun doute, la multicellularité étant apparue plusieurs fois durant l'évolution (Schlegel and Hülsmann, 2007), le terme a été conservé par commodité.

Les protistes représentent la grande majorité de la diversité des eucaryotes (Figure 1). On estime que, dans les océans, plus de 85% de la diversité eucaryote est due aux protistes (De Vargas et al., 2015). Ils présentent une grande variété de tailles, de formes et de modes de vie. On retrouve à la fois des organismes capables de réaliser la photosynthèse – ils utilisent l'énergie lumineuse pour transformer le carbone inorganique (CO₂) en carbone organique – ainsi que des organismes hétérotrophes, qui obtiennent leur carbone par consommation de proies (phagotrophie), par relation symbiotique (de la symbiose au parasitisme) ou par récupération de la matière organique dissoute ou en décomposition (osmotrophie ou saprotrophie). Beaucoup d'organismes sont également mixotrophes, capables à la fois de réaliser la photosynthèse et d'ingérer des proies, souvent en fonction de la quantité de nutriments disponible dans leur milieu (Johnson, 2015). Mais il existe aussi une diversité parmi les mixotrophes, certaines algues étant majoritairement photosynthétiques mais supplémentent la photosynthèse avec la capture de proies, ou au contraire, sont surtout phagotrophes, mais utilisent la photosynthèse pour

survivre lorsque les proies manquent (Rottberger et al., 2013).

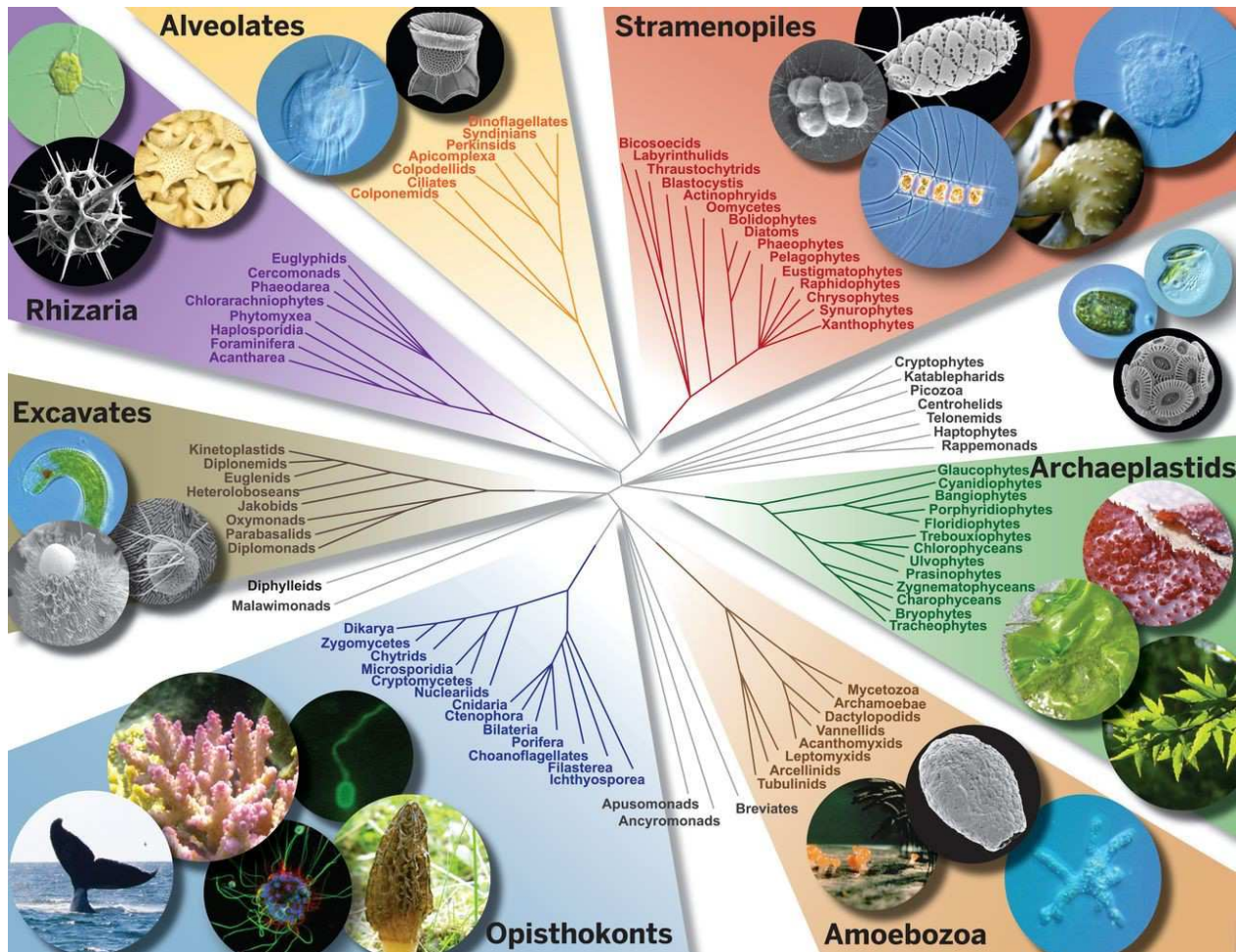


Figure 1 : Représentation schématique de l'arbre phylogénétique des eucaryotes. Les protistes sont présents dans quasiment toutes les branches à l'exception de la plupart des opistoconthes (métazoaires + *fungi*) et des plantes terrestres (*Tracheophytes* + *Bryophytes* + *Charophyceans* + *Zygnematophyceans*). Tiré de Worden et al. (2015)

On retrouve des protistes dans les cinq supergroupes d'eucaryotes :

- Amibozoaires (*Amoebozoa*) : ce sont communément des protistes vivant en forme libre, qui se déplacent par contraction cytoplasmique et phagocytent des bactéries. La plupart des amibozoaires font entre 10 et 20 µm.
- *Archaeplastida* (lignée verte) : les organismes de la lignée verte possèdent un ou plusieurs chloroplastes, descendants d'un ancêtre commun ayant réalisé la première endosymbiose avec une cyanobactérie. Ces organismes sont photosynthétiques, mais certaines espèces sont également capables d'ingérer des bactéries (*Micromonas pusilla* par exemple (McKie-

Krisberg and Sanders, 2014)). C'est dans ce groupe que l'on retrouve les plantes terrestres, mais également des protistes importants pour la production primaire comme les mamiellales.

- Excavés (*Excavata*) : ces organismes partagent une structure particulière, le cytostome, cavité par laquelle les particules sont phagocytées. Les excavés les plus étudiés sont des parasites humains : les trypanosomes *Trypanosoma brucei* et *Trypanosoma cruzi*, responsables respectivement de la maladie du sommeil et de la maladie de Chagas, *Trichomonas vaginalis* qui provoque une urétrite, ou encore *Leishmania* sp., responsable de la leishmaniose.
- Opisthoconthes (*Opisthokonta*) : ce supergroupe rapproche entre autres les animaux (métazoaires) et les champignons (*fungi*). Certains protistes proches des métazoaires sont particulièrement étudiés, comme les choanoflagellés qui sont des organismes unicellulaires ressemblant fortement aux choanocytes des éponges et peuvent vivre en colonies. Leur étude pourrait permettre de comprendre comment la multicellularité est apparue chez les animaux.
- SAR : Ce groupe contient les **Straménopiles**, **Alvéolés** et **Rhizaires** et est donc extrêmement diversifié.

Les straménopiles (*Stramenopila*) sont des flagellés possédant des modes de vie très divers. Comme ce sont des membres de ce groupe qui ont été étudiés au cours de cette thèse, les straménopiles seront présentés plus précisément au point A.2.

Les alvéolés (*Alveolata*) sont des protistes ayant pour caractéristique commune la présence de vésicules appelées « alvéoles » sous la membrane plasmique. Les principaux sous-groupes d'alvéolés sont les apicomplexes, les ciliés et les dinoflagellés. Les apicomplexes sont des endoparasites d'animaux, comme *Plasmodium falciparum*, responsable du paludisme, ou *Toxoplasma gondii*, l'agent de la toxoplasmose. Les ciliés (*Ciliata*) possèdent pour la plupart des cils vibratiles leur permettant de se mouvoir. Les ciliés marins sont essentiellement des consommateurs de pico- (0.2 à 2 µm) nano- (2 à 20 µm)

plancton (Rassoulzadegan et al., 1988; Sherr and Sherr, 2002). Les dinoflagellés peuvent également être des prédateurs importants du pico- nanoplancton (JEONG, 1999), comme *Oxyrrhis marina*, mais d'autres sont photosynthétiques (les zooxanthelles associées aux coraux par exemple), mixotrophes ou encore parasites, comme les *amoebophrya* qui ont un rôle important dans la régulation des populations de microalgues générant des blooms toxiques (Park et al., 2004; Chambouvet et al., 2008). Les dinoflagellés sont particulièrement abondants dans les données de *metabarcoding* (Massana and Pedros-Alio, 2008; Not et al., 2009; De Vargas et al., 2015), reflétant un rôle écologique majeur dans les océans.

Les rhizaires (*Rhizaria*) ont une importance écologique longtemps sous-estimée, en raison de la difficulté à les prélever sans destruction (Caron, 2016). Cela a contribué à sous-estimer les effectifs de rhizaires dans les comptages. Certains membres de ce groupe comme les foraminifères ou les radiolaires sont étudiés pour leur squelette siliceux, qui peut être détecté dans les registres fossiles. Les rhizaires sont souvent trouvés en association avec des microalgues, ce qui leur permet de survivre dans les régions oligotrophes (Caron, 2016).

Du fait de leur diversité, les protistes sont présents dans tous les milieux et sont des contributeurs majeurs aux cycles biogéochimiques de la planète, en particulier dans le cycle du carbone océanique.

A.1.2. Les protistes marins dans le cycle du carbone océanique

Le cycle du carbone est l'un des cycles biogéochimiques majeurs sur Terre, du fait de l'utilisation intensive du carbone par la vie. Ce cycle implique des échanges complexes entre l'atmosphère, l'océan et les écosystèmes terrestres. Nous allons ici

nous concentrer sur la partie océanique du cycle du carbone, qui est responsable de la séquestration d'une grande partie du carbone de la planète : on estime que l'océan contient approximativement 39 000 Giga tonnes (Gt) de carbone, majoritairement sous forme d'ions dissous, contre 800 Gt dans l'atmosphère et 2 000 Gt dans la biosphère terrestre (Siegenthaler and Sarmiento, 1993). Il est responsable de la séquestration en profondeur et dans les sédiments de 2 Gt de carbone atmosphérique (7 Gt de CO₂) par an. À titre de comparaison, les activités humaines produisent entre 5 et 10 Gt de carbone (18-36 Gt de CO₂) par an (Canadell et al., 2007; Houghton, 2007).

Les protistes interviennent à différents niveaux dans le cycle du carbone océanique (Figure 2). Tout d'abord, les protistes photosynthétiques ont un rôle de producteurs primaires. Ils utilisent le dioxyde de carbone naturellement dissous à l'interface avec l'atmosphère pour produire du carbone organique *via* la photosynthèse. Même si leur nombre est souvent moins important que celui des cyanobactéries, la biomasse totale des picoeucaryotes photosynthétique peut être aussi importante que celle des cyanobactéries dans certaines régions (Blanchot and Rodier, 1996; Zubkov et al., 1998; Buitenhuis et al., 2012). Les facteurs induisant ou limitant leur croissance sont maintenant relativement bien compris. Dans la plupart des régions, la croissance du phytoplancton est limitée par les nutriments disponibles, en particulier l'azote et le phosphore (Howarth, 1988). Dans les régions HNLC (*High Nutrients Low Chlorophyll*), riches en nutriments mais pauvres en organismes photosynthétiques, on note que le facteur limitant de la croissance du phytoplancton est le fer (Kolber et al., 1994).

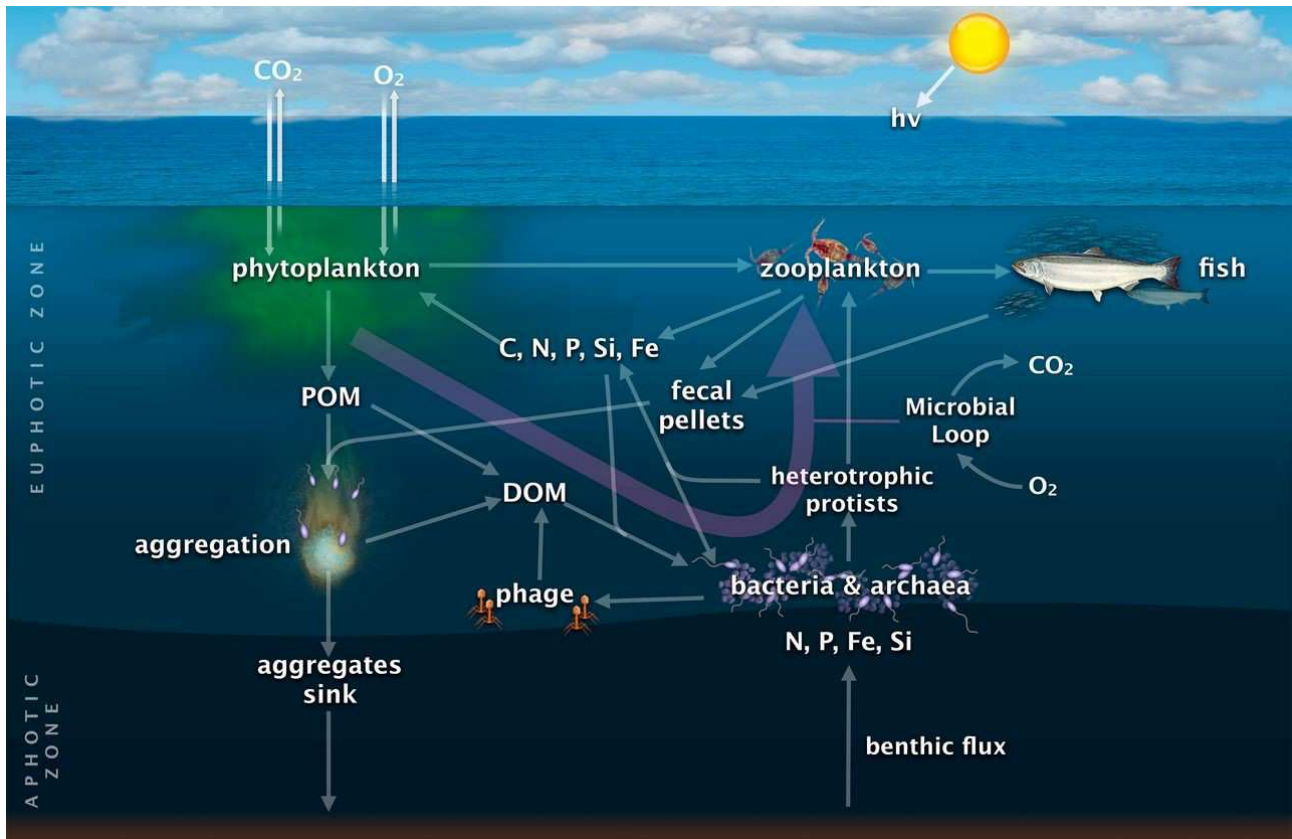


Figure 2: Schéma des réseaux trophiques dans la zone euphotique de l'océan. Les flèches représentent les liens trophiques entre les compartiments du plancton. Tiré de Worden et al. (2015)

À la mort des organismes phytoplanctoniques, les débris cellulaires s'accumulent sous forme de particules colloïdales de matière organique et coulent vers le fond où ils peuvent être stockés dans les sédiments. Cependant, une grande partie est décomposée en chemin par les bactéries hétérotrophes, qui vont reminéraliser une partie de ce carbone organique en CO₂ 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 (Stone and Weisburd, 1992; Fenchel and Finlay, 2008). Les picoeucaryotes hétérotrophes, en consommant des bactéries, vont participer à la régulation de cette boucle microbienne. En étant ensuite eux même les proies d'organismes plus grands, qui peuvent difficilement se nourrir de bactéries directement, les picoeucaryotes bactérivores vont permettre de faire passer le carbone organique aux niveaux trophiques supérieurs. Leur respiration

va également jouer un rôle important dans la reminéralisation du carbone organique (Burns and Galbraith, 2007; Das and Pandey, 2015).

A.2. Un embranchement très diversifié d'eucaryotes : les straménopiles

A.2.1. Caractéristiques et phylogénie des straménopiles

Les straménopiles (ou hétérokontes) sont un groupe très varié d'eucaryotes, appartenant au supergroupe SAR (Straménopiles, Alvéolés, Rhizaires). Ce groupe est essentiellement basé sur les données moléculaires (Van de Peer and De Wachter, 1997), mais une caractéristique morphologique partagée est la présence de deux flagelles de longueurs différentes au cours d'au moins une partie de leur vie : le flagelle antérieur est long et couvert d'une ou plusieurs rangées de « poils » appelés mastigonèmes tripartites (en trois parties : une base aplatie, une partie tubulaire et un ou des filaments terminaux), tandis que le flagelle postérieur est court et lisse.

Les flagelles des straménopiles servent essentiellement à se mouvoir, mais également à attraper leurs proies, comme chez le chrysophyte *Epipyxis pulchra*, où le long flagelle antérieur va battre rapidement en présence d'une proie pour créer un courant fort qui amènera la bactérie au contact, avant de la saisir par ses deux flagelles simultanément (Wetherbee and Andersen, 1992).

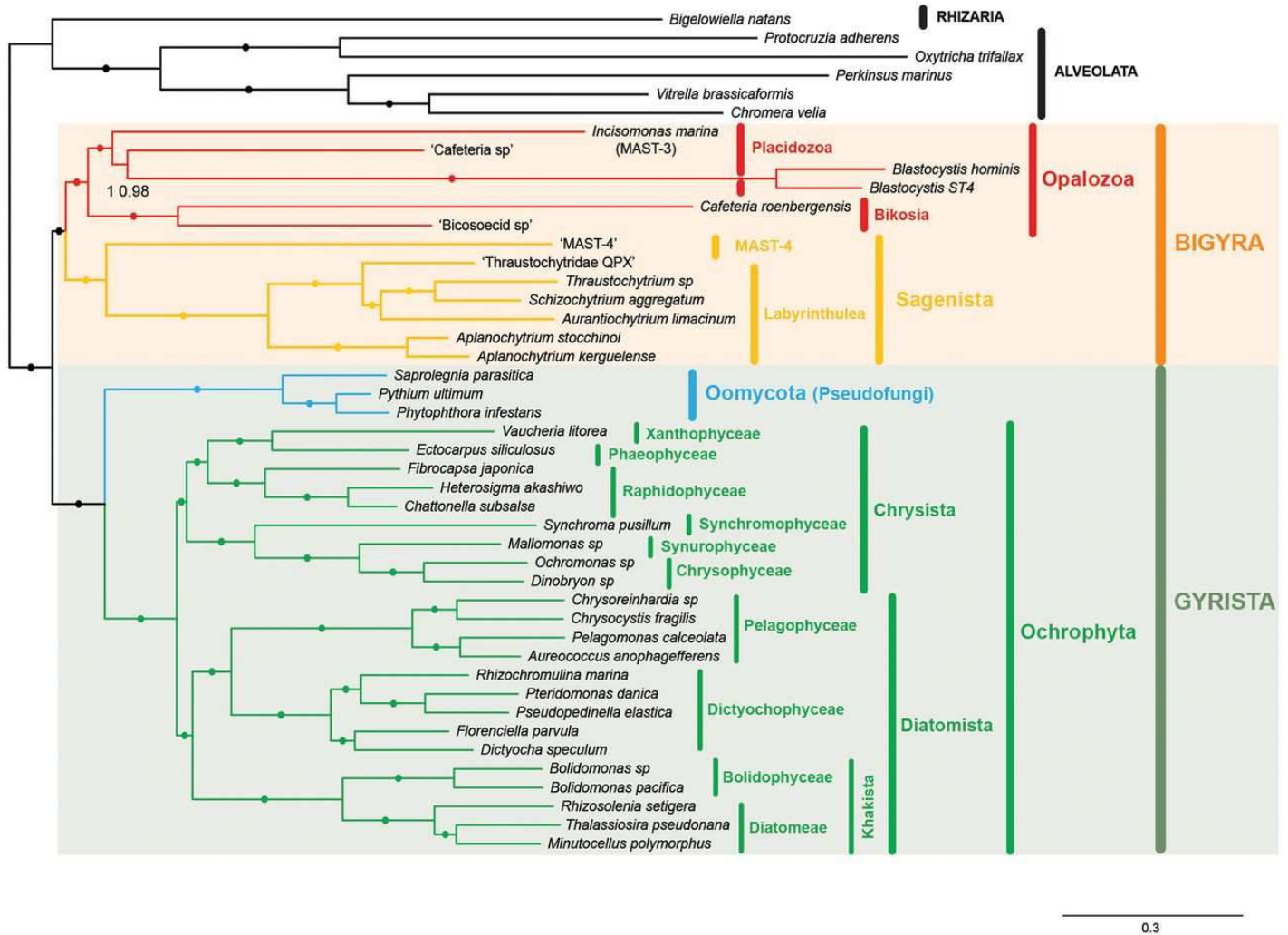


Figure 3 : Arbre phylogénomique des straménopiles, basé sur 339 alignements protéiques. Tiré de Derelle et al. (2016)

D'après les études les plus récentes (Derelle et al., 2016), les straménopiles sont divisés en deux groupes : les *Gyrista*, qui regroupe les straménopiles majoritairement photosynthétiques – les diatomées (*Bacillariophyta*), les bolidophytes, les *dictyochophyceae*, les *pelagophyceae* et les chrysophytes – et les oomycètes, qui ont longtemps été considérés comme des champignons du fait de leur morphologie filamenteuse, qui peuvent être saprophytes ou bien parasites, en particulier parasites de plantes (le genre *Phytophthora* par exemple) ou de poissons (*Saprolegnia parasitica*). Le groupe frère des *Gyrista* est le groupe des *Bigyra* qui ne contient que des organismes hétérotrophes, qu'il s'agisse de bactérivores (*Cafeteria*, certains MAST) ou de parasites (*Blastocystis*, les *labyrinthulea*).

A.2.2. Les straménopiles marins MAST

En 2002, une étude à large échelle des ADN ribosomiques 18S a mis en évidence une forte abondance de straménopiles inconnus dans les océans ouverts (Massana et al., 2002). Ces straménopiles de petite taille (entre 1 et 5 μm) ont également été détectés comme abondants dans des échantillons côtiers dans différents environnements marins tels que l'Atlantique Nord, la mer Méditerranée, la Manche ou la mer du Nord (Massana et al., 2004). Ces straménopiles ont été nommés MAST pour *Marine Stramenopiles* et ont été groupés sur la base de la similarité de séquence de leur ADNr 18S pour former 12 ribogroupes appelés MAST-1 à MAST-12. L'origine de la plupart de ces groupes est proche de la racine de l'arbre des straménopiles. Ceux-ci sont donc éloignés des straménopiles pour lesquels nous possédons des données génomiques. L'absence de pigmentation des organismes de ces 12 lignées laisse supposer que ces organismes sont hétérotrophes, et vraisemblablement bactérivores pour certains.

Par la suite, d'autres groupes de MAST ont été découverts, mais la plupart de ces groupes se sont révélés être des erreurs d'identification de diatomées, de *bicosoecida* ou de labyrinthulomycètes (Massana et al., 2014). De même, le groupe MAST-5 a été supprimé, ce ribogroupe était en fait construit à partir de séquences chimériques (Massana et al., 2014). Finalement, les groupes de MAST confirmés à ce jour sont MAST-1 à MAST-4, MAST-6 à MAST-12, MAST-16, MAST-23, MAST-24 et MAST-25. Les groupes les plus abondants dans les eaux de surface des océans sont les MAST-1, les MAST-3, les MAST-7 et les MAST-4 (Massana et al., 2014). D'autres groupes, comme les MAST-6 sont principalement retrouvés dans les sédiments (Figure 4).

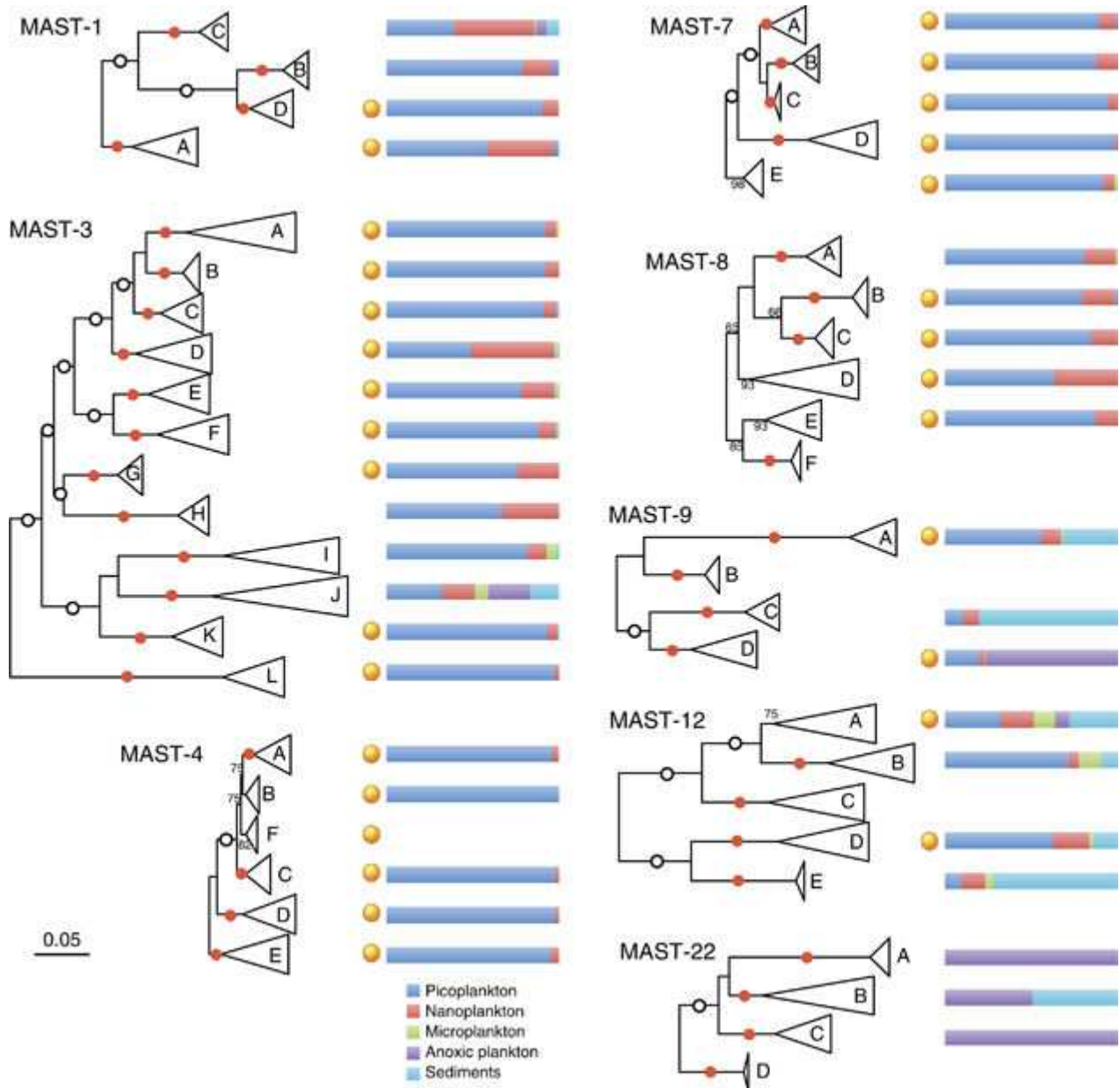


Figure 4 : Arbres phylogénétiques des ribogroupes MAST et distribution selon la taille et les préférences environnementales. Tiré de Massana et al. (2014)

Le comportement trophique de certains MAST a pu être testé en laboratoire. Des individus appartenant aux groupes MAST 1 et 4 ont ainsi pu être conservés quelques jours en aquarium et plusieurs expériences d'hybridation in situ en fluorescence (FISH) ont démontré l'ingestion de bactéries MED₁₃₄ (*Roseobacter*) et MED₄₇₉ (α -*proteobacteria*), mais aussi de picoalgues comme *Micromonas pusilla* et *Ostreococcus sp.*, lorsque la concentration de ces dernières était suffisamment importante (Massana et al., 2009). Des individus appartenant au groupe MAST-4 ont

également été retrouvés en association avec *Pelagibacter ubique* (Martinez-Garcia et al., 2012). L'éventail des proies possibles pour les espèces appartenant au groupe MAST-4 semble donc très étendu.

Concernant le groupe MAST-3, deux espèces avaient pu être mises en culture auparavant. La première, *Solenicola setigera* a été observé très tôt (1904 d'après Gomez et al. (2011), 1908 d'après Gomez (2007)), en raison du fait que les individus de cette espèce vivent en colonies de quelques dizaines d'individus fixés sur le frustule de la diatomée *Leptocylindrus mediterraneus* (Gomez et al., 2011). Une matrice extracellulaire muqueuse recouvre le frustule de la diatomée colonisée et peut héberger d'autres micro-organismes, en particulier une cyanobactérie du genre *Synechococcus* qui est fréquemment rencontrée (Buck and Bentham, 1998). Cependant, l'association potentielle entre *Solenicola setigera* et la cyanobactérie du genre *Synechococcus* n'est pas encore claire, les auteurs mettant en avant une symbiose potentielle, où la cyanobactérie trouverait une niche appauvrie en oxygène propice à sa croissance et multiplication, et où *S. setigera* pourrait directement récolter la cyanobactérie pour se nourrir. Cette relation est encore à éclaircir et reste facultative puisque la cyanobactérie n'est pas observée dans toutes les colonies de MAST-3.

Solenicola setigera en lui-même fait entre 4 et 7 μm de long, est incolore (absence de chlorophylle) et les observations en microscopie ne permettent de distinguer tout au plus qu'un seul flagelle de longueur variable (jusqu'à 24 μm). L'association diatomée – *S. setigera* est ubiquitaire, retrouvée à la fois dans les régions tempérées, tropicales et polaires (Gomez, 2007).

La seconde espèce en culture, *Incisomonas marina*, vit en forme libre, les individus mesurent 3-4 μm et sont phagotrophes. Les cellules sont capables de glisser sur des surfaces, ou de nager difficilement en utilisant leur unique flagelle postérieur (Cavalier-Smith and Scoble, 2013).

La caractéristique commune à *Solenicola setigera* et *Incisomonas marina* étant

une perte secondaire du flagelle antérieur, le nom d'*Uniciliatida* a été proposé pour nommer tous les descendants de leur ancêtre commun (Cavalier-Smith and Scoble, 2013).

La diversité des MAST-3 semble très importante, tant en nombre de taxons qu'en modes de vie. Des organismes appartenant au groupe MAST-3 ont ainsi été retrouvés dans des environnements anoxiques, sulfureux (Gomez et al., 2011) et hypersalins (Stock et al., 2012).

A.2.3. Les chrysophytes

Les chrysophytes (*chrysophyceae*) forment un large groupe chez les straménopiles, avec pour le moment plus de 1 200 espèces décrites. Les chrysophytes ont un large éventail de stratégies trophiques, de la phototrophie à l'hétérotrophie, en passant par la mixotrophie. Les chrysophytes phototrophes sont communément appelés « algues dorées », en raison de la couleur que leur confère les chlorophylles a et c (vert) et leur pigment accessoire, la fucoxanthine (brunâtre). Beaucoup d'algues dorées sont également mixotrophes, comme les genres *Ochromonas* ou *Dinobryon*.

Cependant, certains chrysophytes ne possèdent pas de pigmentation, ceux-ci étant probablement phagotrophes ou saprophytes.

Les chrysophytes ont surtout été étudiés en eaux douces, où ils peuvent être largement dominants parmi les straménopiles (Kammerlander et al., 2015), mais les chrysophytes marins sont encore peu connus. Des analyses environnementales ont cependant essayé d'évaluer la diversité des chrysophytes dans les milieux marins et de créer des clades monophylétiques à partir des séquences ribosomiques 18S (del Campo and Massana, 2011). Cette étude a abouti à la création de 12 clades (A, B₁, B₂, C, D, E, F₁, F₂, G, H, I, J), dont plus de la moitié d'entre eux (7) comprennent à la fois

des séquences issues d'organismes d'eaux douces et de milieux marins. À l'intérieur du clade H en particulier, les séquences d'origine marine forment un sous-groupe monophylétique, indiquant vraisemblablement une origine en eau douce pour ce clade, puis une adaptation aux eaux salées de l'un des descendants.

A.3. La génomique pour l'étude d'organismes non cultivés

Les organismes incultivés représentent la majeure partie de la biodiversité océanique. En effet, la plupart des microorganismes observés par microscopie dans les écosystèmes océaniques (mais aussi terrestres) ne poussent pas dans les milieux nutritifs utilisés pour la mise en culture. Ce problème, observé initialement pour les bactéries que l'on tentait de cultiver sur boîtes de Petri, a été appelé la 'grande anomalie du comptage sur boîte' (Staley and Konopka, 1985). Il a été estimé que moins de 0.1% des bactéries marines peuvent être mises en culture avec les méthodes usuelles (Kogure et al., 1979). Un phénomène analogue existe sans doute pour les protistes, puisque les collections sont largement biaisées en faveur des organismes phototrophes et des parasites (del Campo et al., 2014). Principalement en raison de la difficulté à déterminer précisément leurs besoins alimentaires (Kiy, 1998), la plupart des protistes hétérotrophes ne sont pas encore cultivés, malgré leur abondance et leur importance écologique certaine.

Parallèlement aux efforts développés pour la mise en culture de ces organismes (Massana et al., 2006), il est nécessaire de pouvoir utiliser des méthodes qui s'affranchissent de cette étape limitante dans l'étude des microorganismes marins.

Le nombre de solutions pour l'étude moléculaire d'organismes non cultivés est aujourd'hui relativement limité. La quantité d'ADN disponible n'étant souvent pas suffisante pour un séquençage *de novo* ciblé sur un organisme, il faut alors séquencer toute la communauté présente dans un échantillon (méta-omique), séquencer un gène marqueur (code-barre génétique) ou bien utiliser des techniques d'amplification (génomique en cellule unique). Ces différentes méthodes ont été mise en œuvre dans le projet *Tara Oceans*, et nous allons détailler ici les avantages et inconvénients de chacune.

A.3.1. La méta-omique

Le terme méta-omique regroupe l'ensemble des méthodes faisant appel au séquençage global de communautés complexes. Les deux principales méthodes employées sont la métagénomique, qui consiste à séquencer les génomes de tous les individus présents dans un échantillon, et la métatranscriptomique, qui séquence tous les ARN messagers présents dans un échantillon environnemental.

La métagénomique est utilisée avec succès pour l'étude des communautés bactériennes : il est possible d'assembler les génomes d'organismes procaryotes non cultivés (Tyson et al., 2004), d'analyser les fonctions présentes (Sunagawa et al., 2015) ou encore d'évaluer la variabilité génique d'une population (Schloissnig et al., 2013). Cependant, des difficultés se posent lorsque l'on s'intéresse aux communautés eucaryotes. En effet, la taille des génomes eucaryotes est extrêmement variable (de 12.56 Mbp pour la picoalgue *Ostreococcus tauri* (Derelle et al., 2006) à environ 200 Gbp en taille de génome estimée pour le dinoflagellé *Prorocentrum micans* (Hou and Lin, 2009). À titre de comparaison, le génome classique d'une bactérie est compris entre 1 et 10 Mbp). Les grands génomes nécessiteront une profondeur de séquençage plus importante pour être assemblés, mais du fait du nombre important de répétitions et d'éléments transposables dans ces génomes, l'assemblage à partir de courtes lectures sera très difficile et probablement fragmenté. Cette différence entre tailles de génomes est également un biais important dans la représentation des génomes dans les lectures issues du séquençage. Le recrutement des lectures peut en effet être utilisé pour mesurer l'abondance de génomes de références, mais cette abondance est relative puisqu'elle dépend à la fois de l'abondance de l'organisme dans l'échantillon, mais également de la taille des génomes présents dans la communauté.

Pour pallier à ces problèmes, la métatranscriptomique peut être employée (Bailly et al., 2007). Au lieu de séquencer tout le génome des organismes présents dans l'échantillon, la métatranscriptomique séquence tous les transcrits présents. La taille des génomes n'entre plus en compte étant donné que le nombre de gènes est relativement constant entre les organismes (facteur 10 au maximum, contre un

facteur 10 000 entre les tailles de génomes). De plus, la métatranscriptomique permet de connaître les gènes qui sont réellement exprimés dans un environnement donné et ainsi d'étudier les fonctions réalisées par une communauté. Les gènes non transcrits au moment de l'échantillonnage ne seront en revanche pas séquencés.

A.3.2. Les codes-barres génétiques

Le séquençage de codes-barres génétiques est une méthode permettant d'identifier des espèces 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 (Valentini et al., 2009).

En général, les gènes codant les sous-unités du ribosome sont utilisés, puisqu'ils ont l'avantage d'être suffisamment conservés pour concevoir des amorces universelles, mais sont suffisamment divergents entre taxons pour réaliser une identification relativement précise par comparaison de séquences aux bases de données. En particulier, c'est souvent le gène codant pour la petite sous unité du ribosome qui est utilisé (16S chez les procaryotes, 18S chez les eucaryotes), voire uniquement une sous partie de ce gène pour les eucaryotes (régions hypervariables V₄ ou V₉ (Figure 5)). Mais son pouvoir de résolution est parfois insuffisant. Il a par exemple été démontré que deux écotypes de la picoalgue *Bathycoccus prasinus* ayant la même séquence ADNr 18S ont en fait des génomes très divergents (Vannier et al., 2016). Pour augmenter ce pouvoir de résolution, certaines études utilisent d'autres gènes marqueurs, comme les ITS (*Internal Transcribed Spacers*) (Chase and Fay, 2009) (Figure 5) ou encore le gène mitochondrial de la sous unité 1 de la cytochrome c oxydase (CO₁) pour les animaux (Hebert et al., 2003).

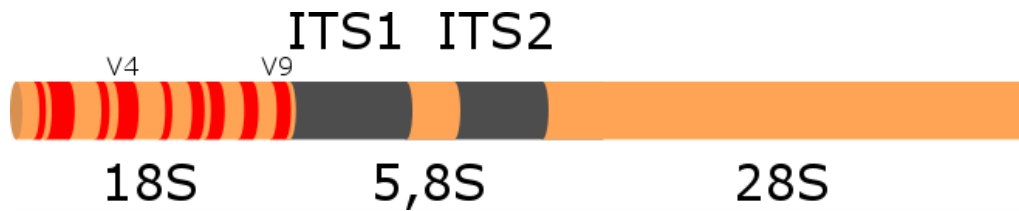


Figure 5 : Structure d'une copie d'un gène ribosomique chez les eucaryotes. ITS : *Internal Transcribed Spacer*, région transcrite mais non traduite en protéine ; 18S ; 5,8S ; 28S : gènes codant les sous-unités du ribosome, nommés d'après leur coefficient de sédimentation en Svedberg; Rouge : régions hypervariables de l'ADNr 18S, nommées V1 à V9.

Les codes-barres génétiques peuvent être utilisés pour identifier les espèces présentes dans un échantillon environnemental. On parle alors de *metabarcoding*, une méthode très employée pour comparer les compositions taxonomiques de différents échantillons complexes. Cette technique est en effet moins coûteuse et plus facile à mettre en œuvre que la métagénomique puisque l'amplification est ciblée grâce aux amorces et seul le gène marqueur est séquencé.

Cependant, l'identification correcte des taxons repose fortement sur les bases de données déjà préétablies. Même si de nouveaux groupes peuvent être créés 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'organismes dont la séquence du gène marqueur partage moins de 80% d'identité avec la séquence de référence la plus proche.

Un autre problème potentiel est que le nombre de copies d'ADNr peut varier fortement selon les organismes, ce qui compromet l'analyse quantitative du *metabarcoding*. Cependant, ce problème peut en partie être corrigé puisqu'il a été montré pour le plancton que le nombre de copies d'ADNr est corrélé à la taille des cellules (Zhu et al., 2005) et au biovolume (Godhe et al., 2008). Mais cette correction reste imprécise, et chez les dinoflagellés par exemple, le nombre de copies de l'ADN ribosomique est bien supérieur à celui estimé d'après la taille des cellules (Zhu et al., 2005).

A.3.3. La génomique en cellule unique

La génomique en cellule unique consiste à amplifier l'ADN présent dans une seule cellule (une seule molécule) à des niveaux suffisant pour le séquençage (quelques nanogrammes). Cette technique permet donc de séquencer un génome *de novo* sans étape de mise en culture, à partir d'une cellule prélevée dans son environnement. Elle est également utilisée dans le domaine médical pour génotyper des lignées cellulaires, comme des cellules cancéreuses (Saadatpour et al., 2015), ou pour du diagnostic prénatal (Peng et al., 2007).

Cette méthode s'est développée avec l'apparition de l'amplification par déplacements multiples (MDA, pour *Multiple Displacement Amplification*)(Dean et al., 2002) qui permet d'obtenir quelques microgrammes à partir d'une seule molécule d'ADN, avec un taux d'erreur réduit et des fragments beaucoup plus grands comparativement à la DOP-PCR (*Degenerate Oligonucleotide Primed-Polymerase Chain Reaction*, une PCR se basant sur des amorces dégénérées (Carter et al., 1992)) parfois utilisée auparavant pour du génotypage (Cheung and Nelson, 1996). Brièvement, l'amplification MDA utilise l'amorçage aléatoire (*random priming*) pour cibler le génome entier. L'ADN polymérase $\phi 29$ (provenant du bactériophage $\phi 29$ de *Bacillus subtilis*) est utilisée pour l'élongation en raison de sa fiabilité (peu d'erreurs de copie), de sa processivité (nombre de bases parcourues avant de se détacher du brin matrice) et de son activité de déplacement de brin. En effet, celle-ci peut continuer l'élongation du brin matrice en déplaçant le brin complémentaire à son passage, ce qui permet de copier à plusieurs reprises la même région en déplaçant les brins complémentaires néo-synthétisés (Figure 6).

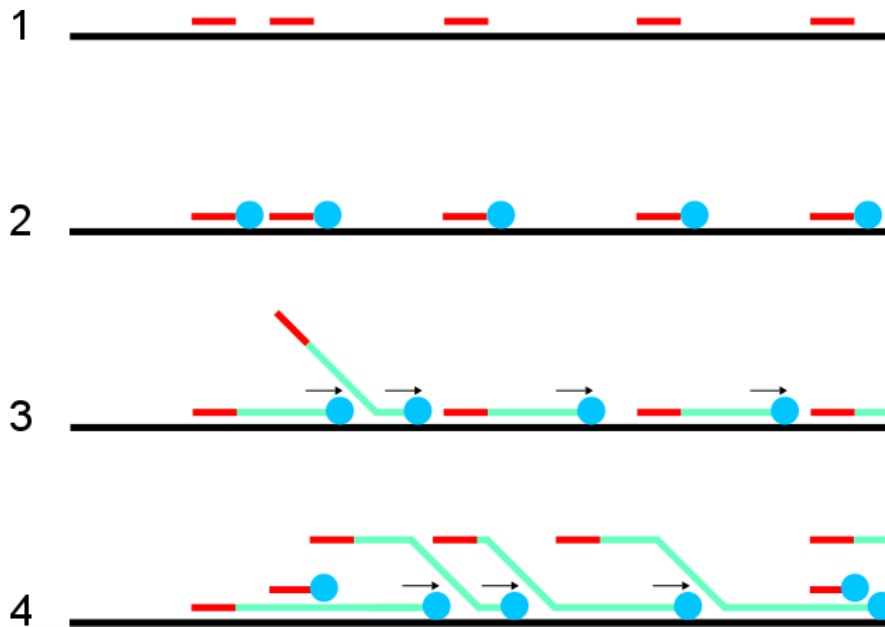


Figure 6 : Représentation schématique de l’amplification par déplacements multiples. 1. Hybridation des amorces aléatoires (rouge) sur le brin matrice (noir). 2. Fixation de l’ADN polymérase $\phi 29$ (en bleu). 3. Début de polymérisation et exemple de déplacement d’un brin néo-synthétisé. 4. Déplacement des brins néo-synthétisés et hybridation avec de nouvelles amorces.

Un avantage de cette méthode est qu’elle est isotherme, toute la réaction pouvant se dérouler à 30°C, ce qui facilite sa mise en œuvre.

Malheureusement, l’amplification MDA présente plusieurs problèmes. Le premier est la production de séquences chimériques, fragments formés à partir de deux régions distantes du génome amplifié. La polymérase peut en effet « sauter » lors du déplacement de brin, s’amorcer sur un brin voisin et continuer l’élongation (Lasken and Stockwell, 2007). Selon les auteurs, la proportion de fragments chimériques varie entre 6% et 50% (Hutchison and Venter, 2006). Le second problème est que l’amplification étant exponentielle, certains fragments sont extrêmement amplifiés tandis que d’autres ne le sont quasiment pas, ce qui entraîne des biais de représentation des fragments. D’autres méthodes ont été développées ensuite dans le but d’atténuer ce biais. Par exemple, l’amplification MALBAC (*Multiple Annealing and Looping Based Amplification Cycles*) apparue en 2012 (Zong et al., 2012) est une amplification quasi-linéaire qui permet d’obtenir une couverture plus uniforme du génome cible. Pour cela, les amorces sont en partie dégénérées (8 nucléotides aléatoires), mais contiennent une base commune de 27 nucléotides qui

serviront à faire boucler les amplicons complets par complémentarité et ainsi éviter qu'ils ne soient eux même réamplifiés. Pour finir, les amplicons complets, sont amplifiés par PCR en utilisant la séquence des 27 nucléotides connus comme amorces. En raison de cette étape de PCR, l'amplification MALBAC génère plus d'erreurs de séquences que la MDA (Chen et al., 2014) et peut être plus difficile à utiliser pour le génotypage (Zong et al., 2012).

A.4. L'expédition *Tara Oceans* : l'étude du plancton de la surface des océans à l'échelle globale

À partir du XIX^{ème} siècle, des expéditions maritimes scientifiques ont été lancées dans le but d'étudier la diversité planctonique et de comprendre le fonctionnement des écosystèmes marins à l'échelle de la planète. L'expédition du *Challenger* (1872-1876) fut la première expédition océanographique, lancée par la *Royal Society* dans le but d'étudier la distribution des formes de vies le long de la colonne d'eau ainsi que sur le plancher océanique, et de mesurer les conditions physico-chimiques dans différents environnements à différentes profondeurs. Cette expédition a permis de décrire et répertorier environ 4 000 espèces inconnues à cette époque.

À l'avènement du séquençage haut débit, Craig Venter lança l'expédition *Global Ocean Sampling* du *Sorcerer II* (2003-2006). Cette fois-ci, la génomique allait remplacer la microscopie pour l'évaluation de la diversité microbienne dans 44 échantillons prélevés dans les eaux de surface de 41 sites allant de l'Atlantique Nord (Nouvelle Écosse) à l'océan Pacifique sud (Polynésie française) en passant par la mer des Caraïbes. Cette expédition a permis de constater le manque d'informations taxonomiques précises sur les espèces bactériennes abondantes dans l'océan (Rusch et al., 2007; Yooseph et al., 2007). Cette approche de séquençage global a permis de détecter de nouvelles familles de protéines mais reste encore loin d'avoir exploré toute la diversité microbienne et virale (Yooseph et al., 2007). De plus, le fractionnement par taille a ciblé les organismes entre 0.1 et 0.8 µm, ciblant ainsi principalement les procaryotes (Rusch et al., 2007).

L'expédition *Tara Oceans* a débuté en 2009 et s'est fixé 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. L'approche holistique de ce projet permet d'étudier les organismes, leurs interactions dans leur milieu, ainsi que les fonctions exprimées par les communautés planctoniques.



Figure 7 : Goélette *Tara*, navire ayant servi à la collecte des échantillons lors de l'expédition *Tara Oceans*.

Des équipes scientifiques se sont relayées sur la goélette *Tara* (Figure 7) afin de prélever des échantillons d'eau de la zone photique des océans, dont les organismes seront analysés au niveau moléculaire afin d'étudier à la fois les espèces présentes (*metabarcoding*) dans les écosystèmes, leurs génomes (métagénomique), ainsi que les gènes exprimés par ces communautés (métatranscriptomique). Les paramètres environnementaux comme la température, la salinité ou la concentration en nutriments ont également été mesurés à chaque point d'échantillonnage afin de pouvoir contextualiser les résultats obtenus.

A.4.1. Parcours de la goélette durant l'expédition *Tara Oceans*

La goélette *Tara* a parcouru durant 2 ans et demi plus de 115 000 km autour du globe pour prélever des échantillons sur plus de 130 sites (Figure 8). Ces sites, appelés

« stations » par la suite, ont été choisis afin d'explorer une grande quantité d'environnements contrastés et de phénomènes océaniques remarquables, comme les *upwellings*, les régions anoxiques ou oligotrophes.

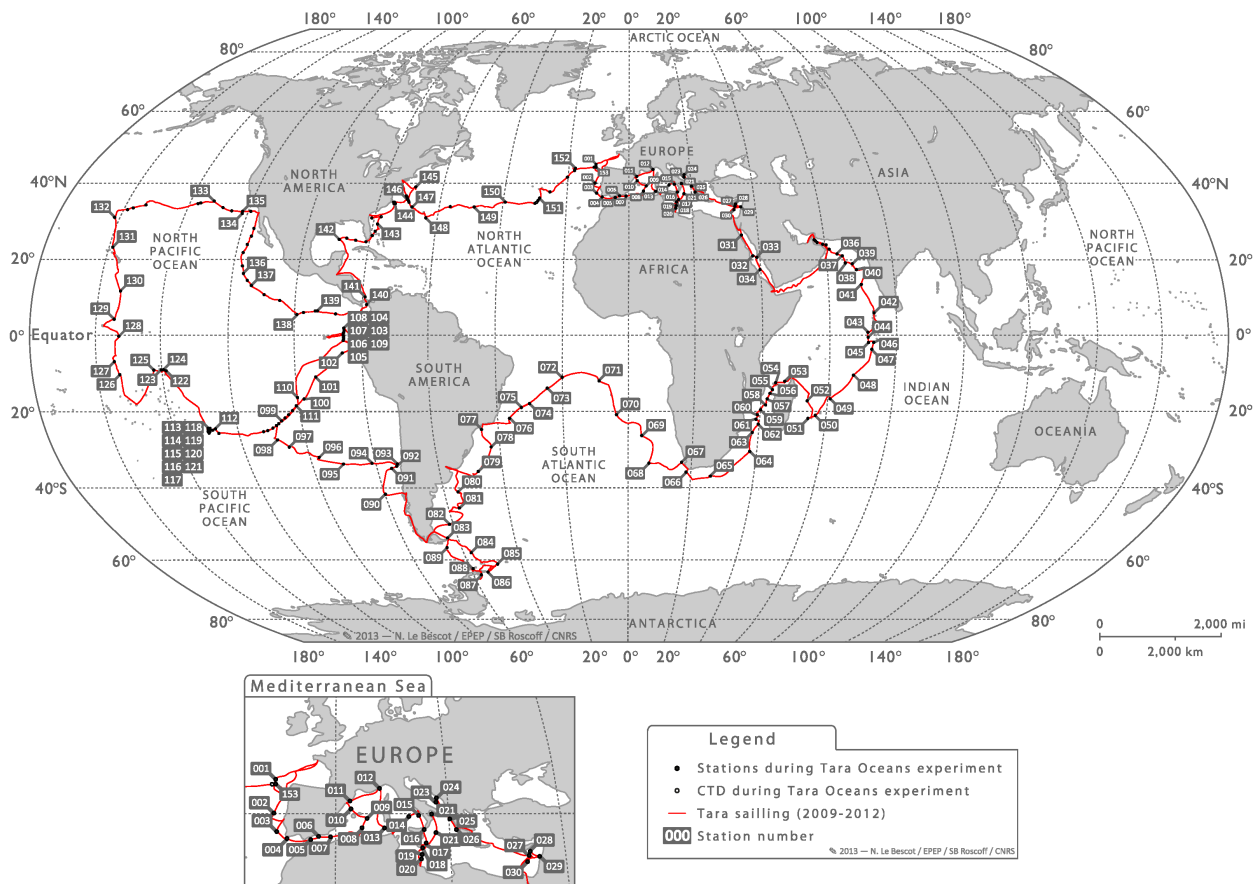


Figure 8: Parcours de la goélette durant l'expédition *Tara Oceans* (2009-2012). En rouge, le parcours du bateau, annoté d'une étiquette grise à chaque point de prélèvement.

La goélette a débuté son trajet depuis le port de Lorient en septembre 2009, a passé le cap de Gibraltar pour échantillonner les principaux courants de la mer Méditerranée puis a traversé la mer Rouge pour atteindre l'océan Indien six mois plus tard. Le navire est ensuite passé par le canal du Mozambique puis le cap de Bonne-Espérance en suivant le courant des aiguilles. Ce courant marque la délimitation entre l'océan Indien et l'océan Atlantique sud. Au contact des eaux de l'Atlantique, plus froides, le courant des aiguilles génère des tourbillons de plusieurs centaines de kilomètres de diamètres, appelés anneaux des aiguilles. La goélette a suivi ces anneaux le long du gyre de l'atlantique sud avant d'effectuer quelques prélèvements dans l'océan austral. Après avoir passé le cap Horn, l'équipage a longé les côtes

chiliennes pour prélever des échantillons dans l'*upwelling* du Chili. Un grand nombre d'échantillons ont été prélevés dans l'océan Pacifique, afin de rendre compte de la diversité des environnements présents dans cet océan. En effet, l'océan Pacifique est globalement divisé en deux gyres : le gyre subtropical du Pacifique sud, une région oligotrophe (pauvre en nutriments), considérée comme un désert océanique et le gyre subtropical du Pacifique plus riche en nutriments. La goélette *Tara* a ensuite passé le canal de Panama, a effectué quelques prélèvements dans le golfe du Mexique avant de remonter suivant le *gulf stream* et de finir par une traversée de l'Atlantique nord d'ouest en est pour regagner son port d'attache en mars 2012.

A.4.2. Stratégie d'échantillonnage

La position exacte des lieux de prélèvements a été décidée en temps réel, en utilisant les données des instruments de bord ainsi que les données satellitaires, afin d'échantillonner au mieux les régions d'intérêt (Pesant et al., 2015).

Globalement, l'échantillonnage s'est déroulé à trois profondeurs différentes : les échantillons de surface, prélevés entre 3 et 9 mètres de profondeur ; les échantillons DCM (*Deep Chlorophyll Maximum*), prélevés à la profondeur où la concentration en chlorophylle est maximale, généralement entre 20 et 100 mètres sous la surface, suffisamment proche de la surface pour que la photosynthèse soit efficace, mais suffisamment profond pour que les organismes photosynthétiques soient protégés des rayons ultraviolets; et enfin quelques échantillons mésopélagiques ont été prélevés entre 300 et 400 mètres de profondeur, où la lumière est complètement absente.

La stratégie d'échantillonnage comportait également un fractionnement par taille des organismes cibles. En effet, il existe globalement une corrélation négative entre la taille des organismes et leur abondance dans les océans. Pour pouvoir étudier le plancton de la manière la plus exhaustive possible, les prélèvements d'eau ont été

filtrés en utilisant différentes tailles de mailles. Les volumes d'eau filtrés ont été plus importants lorsque les organismes étaient moins abondants, afin d'obtenir une quantité suffisante de matériel pour les analyses moléculaires (Figure 9).

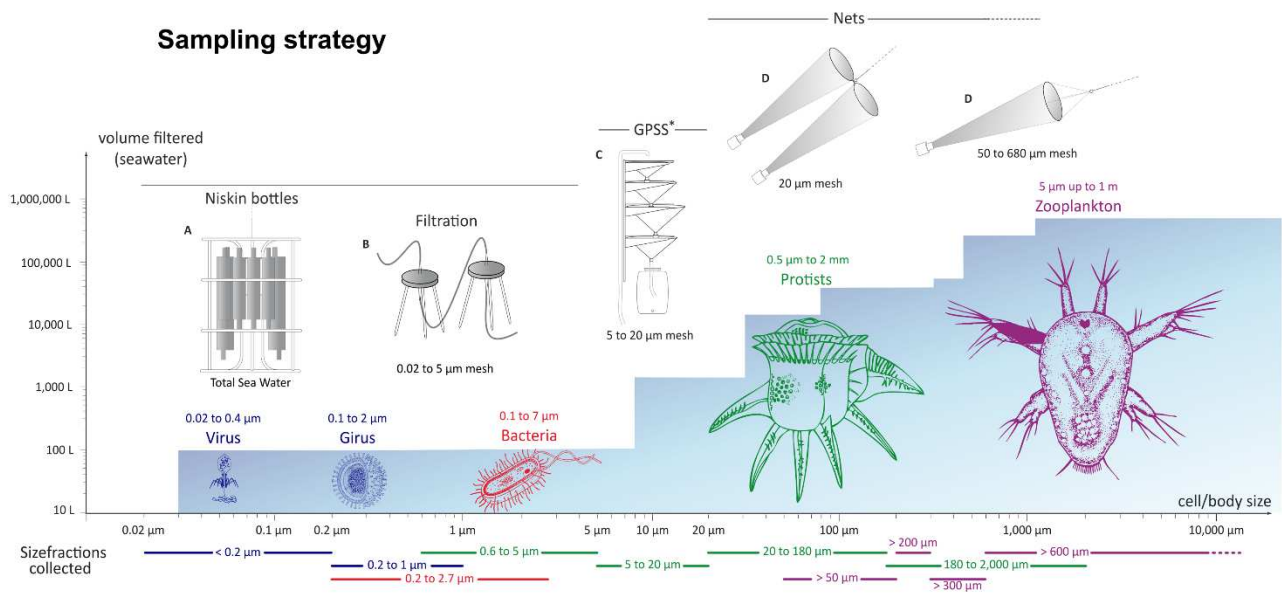


Figure 9 : Méthode de prélèvement en fonction de la taille des organismes cibles. Plus la taille des organismes est importante (axe des abscisses), plus le volume d'eau filtré (bleu) est important. La méthode de filtration (pompe péristaltique ou filet) varie également selon la taille des organismes. Tiré de Karsenti et al. (2011)

Les principales fractions de tailles et les organismes ciblés par chacune d'entre elles sont indiqués dans le Tableau 1.

Size fraction (μm)	Targeted organisms
< 0.2	Viruses
0.2 – 0.8	Giruses Bacteria
0.8 – 5	Bacteria Nanoeukaryotes
5 – 20	Eukaryotes
20 – 180	Eukaryotes
180 – 2,000	Zooplankton

Tableau 1: Principales fractions réalisées durant l'échantillonnage *Tara Oceans*. Cette liste n'est pas exhaustive, mais représente les échantillons les plus étudiés.

L'expédition *Tara Oceans* permet ainsi d'obtenir un très grand jeu de données homogène, autorisant des comparaisons à l'échelle globale des échantillons prélevés. Cette homogénéité supprime une partie des biais résultant de la comparaison d'échantillons issus de différentes études.

A.4.3. Séquençage métagénomique

Dans chacun des échantillons d'eau, le matériel biologique a été récupéré sur une membrane filtrante correspondant au maillage le plus fin (0.8 μm pour la fraction 0.8-5 μm par exemple). L'ADN a ensuite été extrait à partir d'un protocole spécialement étudié pour effectuer une lyse complète des cellules et des noyaux de protistes et de métazoaires (Alberti et al., 2017). Les fragments d'ADN ont ensuite été fractionnés en séquences de 300 bp en moyenne et préparés pour du séquençage pairé (*paired-end*).

Le séquençage des échantillons s'est majoritairement déroulé sur des plateformes Illumina HiSeq2000 et HiSeq2500, produisant en moyenne 160 millions de lectures de 2×101 bp par échantillon.

La qualité des données générées a été contrôlée à la fin du séquençage, écartant les échantillons trop peu complexes (nombre de duplicats PCR élevé) et les échantillons fortement contaminés par des champignons ou des bactéries, pour les échantillons des plus grandes fractions de taille. Les lectures ont également été nettoyées : les séquences correspondant aux adaptateurs Illumina ont été supprimées et les lectures avec une partie de séquence de faible qualité (score Phred) ont été raccourcies.

A.4.4. Construction d'un catalogue de gènes eucaryotes

Afin d'identifier les différents gènes exprimés par le plancton de la zone euphotique des océans, un catalogue de gènes a été réalisé à partir des données métatranscriptomiques du projet *Tara Oceans*. Les échantillons provenant de 68 stations de prélèvement, à deux profondeurs différentes (en surface et à la profondeur correspondant à la concentration maximale en chlorophylle), pour 4 fractions de taille (0.8-5 μm , 5-20 μm , 20-180 μm , 180-2000 μm). Pour chacun de ces échantillons, les ARN messagers ont été récupérés à partir de leur queue polyadénylée, afin de sélectionner uniquement les ARN messagers eucaryotes non

ribosomiques.

Une synthèse des ADN complémentaires a ensuite été réalisée à partir des ARN messagers sélectionnés, qui ont servi à la préparation des banques *paired-end* Illumina. Le séquençage de chacune des banques a été effectué sur 1 piste de séquenceur Illumina HiSeq2000, produisant en moyenne 160 millions de lectures de 2×101 bp.

Les lots de séquences ont ensuite été assemblés indépendamment en utilisant l'outil *Velvet* avec une taille de k-mer relativement importante (63 bp) afin d'éviter la création de chimères durant l'assemblage. Les contigs plus petits que 150 bp ont été supprimés. Les contigs restants provenant de tous les échantillons ont ensuite été regroupés et le logiciel CD-HIT a été utilisé afin d'éliminer la redondance et d'obtenir un seul représentant pour des séquences similaires. Ainsi, les gènes retrouvés dans différents échantillons ne sont présents qu'une fois dans le catalogue final. Les gènes correspondant à des séquences ribosomiques, chloroplastiques ou mitochondriales ont été écartés du catalogue final.

Ce catalogue de gènes contient un peu plus de 116 millions de séquences non redondantes, appelées « unigènes ». Ces unigènes ont été par la suite utilisés pour l'annotation de génomes séquencés en cellule unique (Chapitre 1).

A.4.5. Isolation de cellules par cytométrie en flux

En plus des échantillons obtenus par fractionnement, des cellules ont été isolées par cytométrie en flux afin de pouvoir être analysées par génomique en cellule unique. Les cellules ont été triées sur la base de leur taille et de la présence d'ADN, détecté grâce au fluorochrome SYBR Green I. Les cellules ne contenant pas d'ADN ont ainsi été écartées. La présence de chlorophylle a également été détectée afin de séparer les organismes phototrophes des organismes hétérotrophes. Ce tri des cellules a été réalisé au *Bigelow Laboratory for Ocean Sciences* (États-Unis).

Cette thèse se base en grande partie sur ces cellules isolées durant l'expédition

Tara Oceans.

Objectifs de la thèse

Le principal objectif de cette thèse est d'étudier à l'échelle génomique le compartiment hétérotrophe du pico- nanoplancton, élément important de la chaîne trophique, peu étudié en raison du faible nombre d'organismes cultivés à ce jour. En particulier, l'importance écologique des straménopiles marins incultivés (MAST, Chrysophytes) a été récemment mise en évidence grâce aux données de *metabarcoding*. Des cellules de ces organismes abondants ont été prélevées durant l'expédition *Tara Oceans* en vue d'un séquençage en cellule unique pour analyser les spécificités de leurs répertoires de gènes et émettre des hypothèses quant à leurs modes trophiques.

Nous décrivons dans le premier chapitre la mise en place d'une méthode d'assemblage et d'annotation des génomes séquencés, qui diffèrent des méthodes utilisées en génomique classique. Cette méthode a été utilisée pour reconstruire partiellement le génome de sept lignées de straménopiles marins relativement abondants dans les océans. Elle a également été utilisée dans d'autres études (Vannier et al., 2016; Mangot et al., 2017).

L'analyse génomique de ces organismes a permis de mettre en évidence des spécificités propres à certains d'entre eux, en termes de trophisme et de motilité. L'utilisation des données métagénomiques de *Tara Oceans* a également permis de décrire la biogéographie de ces organismes et de mettre en évidence que la température discrimine le mieux ces distributions géographiques. Ce travail a fait l'objet d'un article accepté dans *Nature Communications*.

L'analyse biogéographique d'un organisme en particulier, MAST-4 clade A, a montré que celui-ci était très abondant dans tous les bassins océaniques excepté dans les régions polaires, avec une diversité génétique étonnamment faible. Ces caractéristiques ont permis d'étudier l'expression relative des gènes de cet organisme dans différentes conditions environnementales en utilisant les données métatranscriptomiques de *Tara Oceans*. La comparaison de l'expression des grandes fonctions biologiques a permis de déterminer que les fonctions liées à la

phagotrophie semblent être les fonctions dont l'expression est la plus flexible dans les populations naturelles de MAST-4 A. Les résultats ont été soumis sous forme d'article à la revue *Environmental Microbiology*.

B. Chapitre 1 : Analyse de génomes à partir de cellules uniques

B.1. Introduction

Durant l'expédition *Tara Oceans*, des centaines de cellules de protistes ont été isolées par cytométrie en flux pour amplification et séquençage. L'objectif est de pouvoir analyser les génomes d'organismes abondants dans les océans ouverts, mais qu'il n'est aujourd'hui pas possible de mettre en culture. Cela contribue également à la création de génomes de référence d'organismes non cultivés afin d'aider à l'analyse des données métagénomiques ou métatranscriptomiques, pour lesquels l'assignation taxonomique repose sur les génomes disponibles dans les bases de données.

La génomique en cellule unique est cependant une technique relativement récente, qui pose encore plusieurs problèmes par rapport à la génomique « classique » (Hutchison and Venter, 2006; Walker and Parkhill, 2008). En effet, en partant d'une seule molécule d'ADN nucléaire, la moindre dégradation durant la préparation avant amplification ne peut pas être récupérée et se traduit par une impossibilité de séquencer une partie du génome d'intérêt. De plus, l'amplification utilisée est non linéaire, ce qui génère des biais de couverture verticale après séquençage. La plupart des méthodes d'assemblages se basant sur la couverture verticale pour éviter les erreurs d'assemblage et générer des contigs plus longs, de nouveaux outils ont dû émerger pour reconstruire les génomes *de novo* à partir de données issues de séquençage en cellule unique.

L'annotation de ces génomes est également difficile, puisque, pour les génomes eucaryotes, celle-ci se base généralement sur des transcriptomes obtenus à partir de cultures dans différentes conditions ou par transfert d'annotation réalisées sur des génomes d'espèces proches. Or la transcriptomique en cellule unique était encore à ses balbutiements durant l'expédition *Tara Oceans*, les transcriptomes de ces organismes ne sont donc pas disponibles, et les organismes ciblés sont bien souvent très éloignés des espèces dont les génomes ont déjà été séquencés et annotés.

Ces problèmes ont été résolus pour aboutir à la création d'une méthode

d'assemblage et d'annotation de génomes séquencés en cellule unique, utilisant également les données métagénomiques pour la décontamination et les données métatranscriptomiques pour l'annotation syntaxique des génomes. Cette méthode, présentée dans ce chapitre, a été utilisée sur différentes lignées de protistes marins isolés durant l'expédition *Tara Oceans*, dont les premiers résultats ont été utilisés durant ce projet de thèse.

B.2. Matériels et méthodes

B.2.1. Reconstruction des génomes

Les cellules ont été amplifiées au *Bigelow Laboratory for Ocean Sciences* suivant la méthode d'amplification par déplacements multiples décrite au point A.3.3. Le résultat de l'amplification de chaque isolat a ensuite été séquencé indépendamment sur séquenceur Illumina HiSeq 2000 ou HiSeq 2500 en 2×101 paires de bases (séquençage « *paired-end* ») sur $1/8^{\text{ème}}$ de piste, produisant approximativement 25 millions de lectures par lot de séquences.

Afin de maximiser la proportion du génome reconstruite, nous avons assemblé ensemble les lots de séquences provenant d'individus avec des génomes quasi identiques. Cette méthode demande ainsi d'identifier *a priori* les génomes qui sont suffisamment proches pour être co-assemblés.

Pour ce faire, nous réalisons un premier assemblage avec le logiciel *HyDA* (Movahedi et al., 2016) qui utilise les graphes de De Bruijn colorés pour garder la trace de chaque lot de séquence après assemblage. Le logiciel permet alors de générer des contigs en utilisant des lectures de plusieurs lots de séquences (correspondant ici à différents individus), et de reporter, pour chaque contig, la couverture en lectures provenant de chacun des lots de séquences. Une distance est alors calculée entre paires de lots de séquences i et j en utilisant la formule suivante :

$$d = \frac{\text{Nombre de bases provenant de contigs couverts uniquement par } i}{\text{Nombre de bases provenant de contigs couverts par } i \text{ et } j}$$

Les contigs sont considérés comme couverts lorsque la moyenne de la couverture verticale est supérieure à un seuil $\varepsilon = 1$.

Ce premier assemblage permet de détecter les lots de séquences qu'il est possible de co-assembler afin d'obtenir un génome partiel représentatif de plusieurs individus génétiquement très proches (plus de 99% d'identité). Grâce à cela, les co-assemblages sont plus grands et moins fragmentés que les assemblages individuels, comme reporté dans d'autres études (Movahedi et al., 2012).

Les lots de séquences sont ensuite assemblés une seconde fois, en utilisant un assembleur plus performant que *HyDA*. En effet, d'autres outils dédiés à l'assemblage de génomes séquencés en cellule unique produisent de meilleurs résultats, tant au niveau de la longueur des assemblages qu'au niveau de la continuité. Le programme d'assemblage retenu est le logiciel *SPAdes* (Bankevich et al., 2012) qui produit de bons résultats et est maintenant largement utilisé tant pour l'assemblage de génomes bactériens, pour lesquels il a été conçu, que pour l'assemblage de génomes eucaryotes (De Bourcy et al., 2014; Roy et al., 2014; Mangot et al., 2017)

Un scaffolding des contigs est ensuite réalisé par le logiciel *SSPACE* (Boetzer et al., 2010), ce qui permet d'utiliser l'information du séquençage paillé pour lier des contigs entre eux et créer des scaffolds, plus grands.

B.2.2. Suppression de la contamination entre assemblages

Aux premières comparaisons effectuées entre les assemblages de génomes séquencés en cellule unique, nous avons détecté des contigs partagés entre assemblages de génomes provenant d'organismes appartenant à des phyla très différents. Ceci s'explique par une contamination entre les différents assemblages, provenant potentiellement du séquençage. En effet, le multiplexage de plusieurs

échantillons sur une même piste peut mener à l'identification incorrecte de l'échantillon auquel appartient une lecture (phénomène d'*index switching*, aussi appelé *index hopping*). Ce phénomène a été rapporté comme étant particulièrement important sur la plateforme Illumina HiSeq4000 (Sinha et al., 2017), alors que nous avons utilisés des séquenceurs HiSeq2000 et HiSeq2500, mais il semblerait que les banques construites à partir d'amplification MDA soient particulièrement concernées par ce phénomène.

Cette contamination a été éliminée après assemblage, en comparant les contigs d'assemblages d'organismes appartenant à des phyla génétiquement éloignés. Pour ce faire, les *scaffolds* de chaque assemblage ont été à tour de rôle fractionnés en séquences de 1 000 bp et alignées contre les autres assemblages. Les *scaffolds* où s'alignent au moins une séquence avec plus de 95% d'identité sur 80% de la séquence sont considérés comme de la contamination et sont écartés.

Cette approche est conservative puisque l'on écarte des assemblages des séquences qui appartiennent bien au génome de l'un des organismes séquencés. Mais ne pouvant connaître leur origine avec certitude, les séquences ne sont pas intégrées à l'assemblage final.

B.2.3. Annotation syntaxique des génomes

L'annotation syntaxique des génomes consiste à retrouver la position et la structure des gènes. Chez les eucaryotes, l'annotation syntaxique se fait principalement à partir de données transcriptomiques de l'organisme cible. Le séquençage des transcrits va en effet permettre d'identifier les régions codantes en alignant les transcrits sur le génome. Cependant, nous ne possédons pas de telles données pour les génomes amplifiés en cellule unique étudiés dans cette thèse. Nous avons donc utilisé le jeu de données métatranscriptomiques de l'expédition *Tara Oceans* pour recruter des lectures issues de transcrits de l'organisme d'intérêt. Dans

le détail, les lectures métatranscriptomiques de chacun de échantillons sont comparées au génome (recherche d'un k-mer commun de 31 bp) puis alignées sur le génome en utilisant *STAR* (Dobin et al., 2013), un aligneur capable de gérer l'alignement épissé. En effet, à la différence des procaryotes, les eucaryotes peuvent avoir des introns dans leurs gènes, qui seront épissés durant l'étape de maturation des ARN messagers. Lors du séquençage des messagers, certaines lectures vont correspondre à deux exons différents, qui seront éloignés sur le génome. Les programmes classiques d'alignement de courtes séquences tels que BWA ou Bowtie ne peuvent pas gérer ces cas. Au contraire, les programmes comme BLAST ou Blat, permettant des alignements locaux, ne sont pas capables utiliser les courtes lectures du séquençage Illumina. Après alignement, la structure exon/intron des gènes est dessinée à partir de la couverture verticale en lectures metatranscriptomiques en utilisant le programme G-Mo.R-Se (<http://www.genoscope.cns.fr/externe/gmorse/>). Dans une précédente version de la méthode d'annotation (voir l'article du chapitre 2), ce n'était pas les lectures qui étaient utilisées, mais les unigènes issus du catalogue de gènes eucaryotes de *Tara Oceans* qui étaient directement alignés en utilisant *est2genome* (Rice et al., 2000). Cependant, sur les premiers génomes annotés, l'information obtenue grâce aux lectures était un peu plus complète que celle obtenue avec l'alignement des unigènes.

Cette information est combinée à deux autres ressources pour produire l'annotation finale (Figure 10): des alignements protéiques issus d'organismes déjà séquencés et annotés présents dans les bases de données (alignements protéiques), ainsi que des prédictions *ab initio* réalisées à partir d'un échantillon de gènes prédits à partir alignements complets des deux ressources précédentes. L'outil chargé de la combinaison de ces 3 ressources est *GMOVE*, un outil réalisé au Genoscope qui permet de retenir le meilleur de chaque ressource pour proposer des modèles de gènes. Dans la précédente version de la méthode d'annotation syntaxique, l'outil utilisé pour la combinaison des ressources était *GAZE* (Howe et al., 2002).

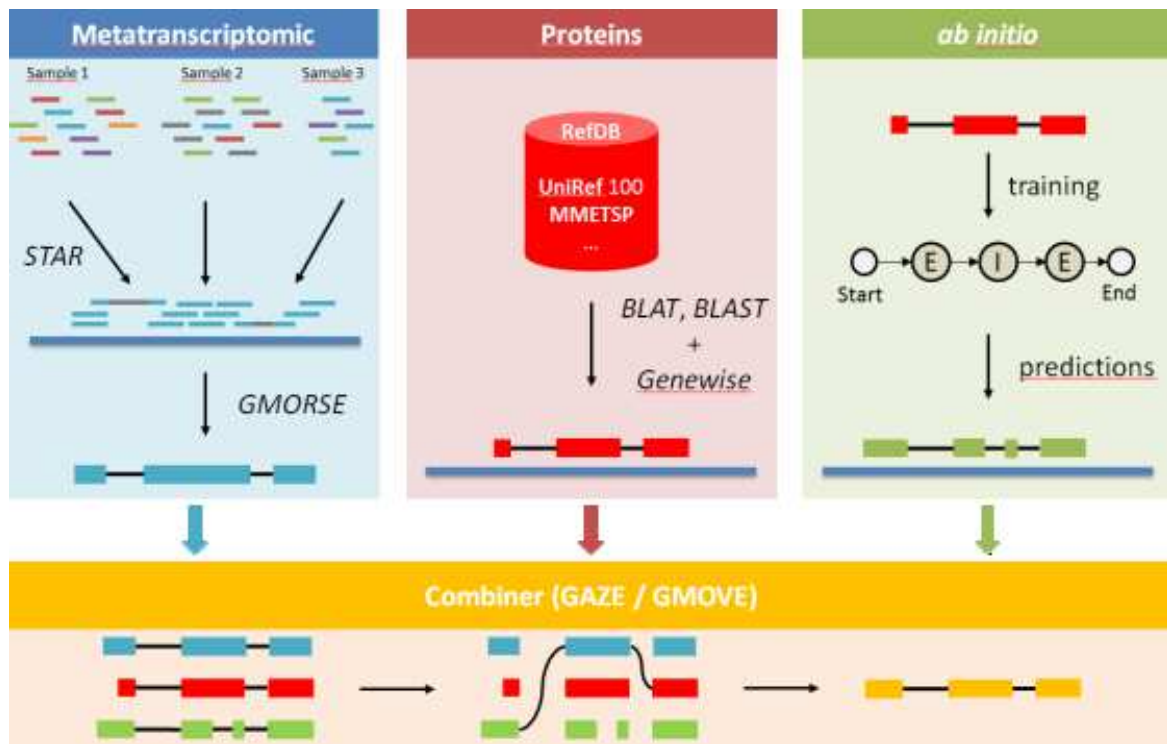


Figure 10: Méthode d'annotation syntaxique utilisée actuellement pour les génomes issus de séquençage en cellule unique. Trois ressources principales sont utilisées : des données d'expression, récupérées à partir de données métatranscriptomiques (bleu), les protéines disponibles dans les bases de données publiques (rouge), ainsi que des prédictions *ab initio* réalisées à partir d'un échantillon de gènes de l'organisme (vert) obtenu à partir des ressources complètes précédentes. Les résultats sont ensuite combinés et filtrés pour aboutir aux prédictions de gènes finales (jaune).

B.2.4. Décontamination basée sur la signature métagénomique

Malgré la première décontamination entre assemblages, il reste à écarter la contamination biologique, due aux potentielles proies ingérées, symbiotes, parasites ou virus. Cette contamination est difficile à détecter puisqu'il peut s'agir de séquences virales, procaryotes ou eucaryotes éloignées des génomes déjà séquencés et disponibles dans les bases de données. Nous avons donc opté pour une approche se basant sur les profils d'abondance dans les échantillons métagénomique de chacun des gènes. Le génome de chaque organisme possède en effet une distribution géographique quasiment unique dans les centaines d'échantillons métagénomiques de *Tara Oceans*, selon les stations où celui-ci est présent, et son abondance relative. Sur ce principe, l'abondance relative de chaque gène dans les échantillons métagénomiques est calculée et les gènes dont la signature s'éloigne significativement de la signature moyenne sont considérés comme aberrants. Si tous les gènes d'une même *scaffold* sont considérés comme aberrants, le *scaffold* appartient vraisemblablement au génome d'un autre organisme et est écarté. La figure 11 donne un exemple de différentes signatures métagénomiques retrouvées dans un même assemblage.

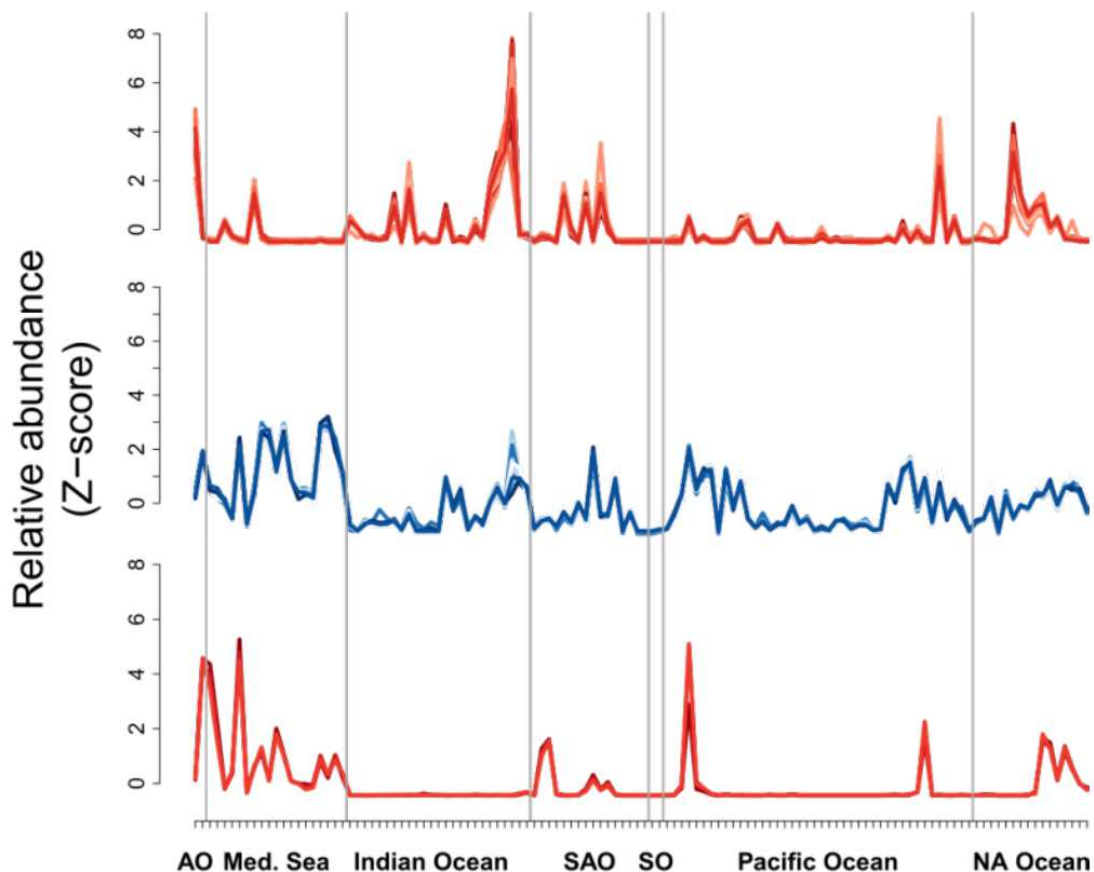


Figure 11 : Signatures d'abondance métagénomique de différentes séquences retrouvées dans un même assemblage. Dans l'assemblage de MAST-4 A, les séquences non contaminantes (bleu) ont une signature très différente des séquences contaminantes (rouge). En haut, l'assignation taxonomique nous indique que ces séquences proviennent d'un génome de *Bathycoccus prasinos*, une microalgue potentiellement ingérée par MAST-4 A. En bas, la recherche de similarité de séquence indique que ces séquences proviennent du génome d'une souche de la cyanobactérie *Prochlorococcus*, que MAST-4 A est capable d'ingérer (Massana et al., 2009).

B.3. Conclusions

Nous avons mis en place une méthode d'assemblage et d'annotation des génomes séquencés en cellule unique, en tenant compte des problèmes relatifs à la génomique en cellule unique (biais de couverture, génomes incomplets), à la distance phylogénétique des organismes étudiés avec les organismes dont le génome a déjà été séquencé et annoté ainsi qu'aux contaminations potentielles, inévitables pour des cellules directement prélevées dans leur milieu naturel.

La méthode présentée a déjà été utilisée pour assembler et annoter le génome de sept lignées de straménopiles marins présentées au chapitre 2, ainsi que le

génomique d'un autre écotype de *Bathycoccus prasinos* (Vannier et al., 2016). Cependant, près d'un millier de cellules ont été isolées durant l'expédition *Tara Oceans*, et la mise en place de cette méthode au Genoscope pourra servir à reconstruire le génome d'autres organismes non cultivés.

C. Chapitre 2 : Diversité fonctionnelle de straménopiles incultivés

C.1. Introduction

Les nanoflagellés hétérotrophes sont considérés dans les modèles écologiques comme des brouteurs de bactéries, sans spécificité autre que leur taux d'ingestion de proies. Cependant, ces organismes sont assez peu étudiés, en raison du faible nombre de cultures disponibles et le mode de vie des espèces écologiquement importantes n'est pas connu exactement. Afin d'en apprendre plus sur les organismes abondants dans les océans ouverts, une grande quantité de cellules ont été prélevées durant l'expédition *Tara Oceans*. Nous présentons dans ce chapitre les résultats obtenus concernant 7 lignées de straménopiles marins *a priori* relativement abondants dans les océans : trois lignées de MAST-4 (clades A, C et E), 2 lignées de MAST-3 (clades A et F) relativement éloignés de *Solenicola setigera* et *Incisomonas marina* et 2 lignées de chrysophytes clade H.

C.2. Article 1 : *Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans*

Résumé en français :

Les eucaryotes unicellulaires (protistes) sont des acteurs clés dans les cycles biogéochimiques globaux des nutriments et de l'énergie dans les océans. Bien que leurs rôles en tant que producteurs primaires et brouteurs soient bien compris, d'autres aspects de leur mode de vie demeurent obscurs en raison des défis que posent la mise en culture et le séquençage de leur diversité naturelle. Nous exploitons ici des données de génomique en cellule unique et de métagénomique issues de l'expédition circumglobale *Tara Oceans* pour analyser le contenu du

génomique et la distribution océanique de sept lignées prédominantes de straménopiles hétérotrophes non cultivées. D'après les données disponibles, chaque génome séquencé ou génotype semble avoir une distribution océanique spécifique, principalement corrélée à la température et à la profondeur. Le contenu du génome fournit des hypothèses de spécialisation en termes de motilité cellulaire, de spectres alimentaires et niveaux trophiques, y compris l'impact potentiel sur leur mode de vie du transfert horizontal de gènes à partir de procaryotes. Nos résultats appuient l'idée que des protistes marins hétérotrophes de premier plan remplissent diverses fonctions dans l'écologie océanique.

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 affiliations 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 remineralized¹. Very small but numerous marine heterotrophic protists play key roles in these processes. Since most of them remain uncultured, their functions remain largely unknown². A recent DNA metabarcoding survey based on *Tara* Oceans global plankton samples has revealed the existence of thousands of heterotrophic protist taxa in eukaryotic communities³ that potentially participate in numerous species interaction networks in yet-to-be defined ways⁴. 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 sequencing⁸. In this study, we present three distinct genomes from clades A, C, and E, clearly divergent from clade D. MAST-3¹¹ 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 surveys¹⁵. It has been postulated that all of these lineages originated from a presumably autotrophic stramenopile ancestor¹⁶, 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 cycles¹⁷.

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 sequenced¹⁸: 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 possible²⁰.

Functional repertoires. To assess variation in the functional repertoires of the sequenced uncultured stramenopiles and to provide further context, we predicted functional domains (Pfam) 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)

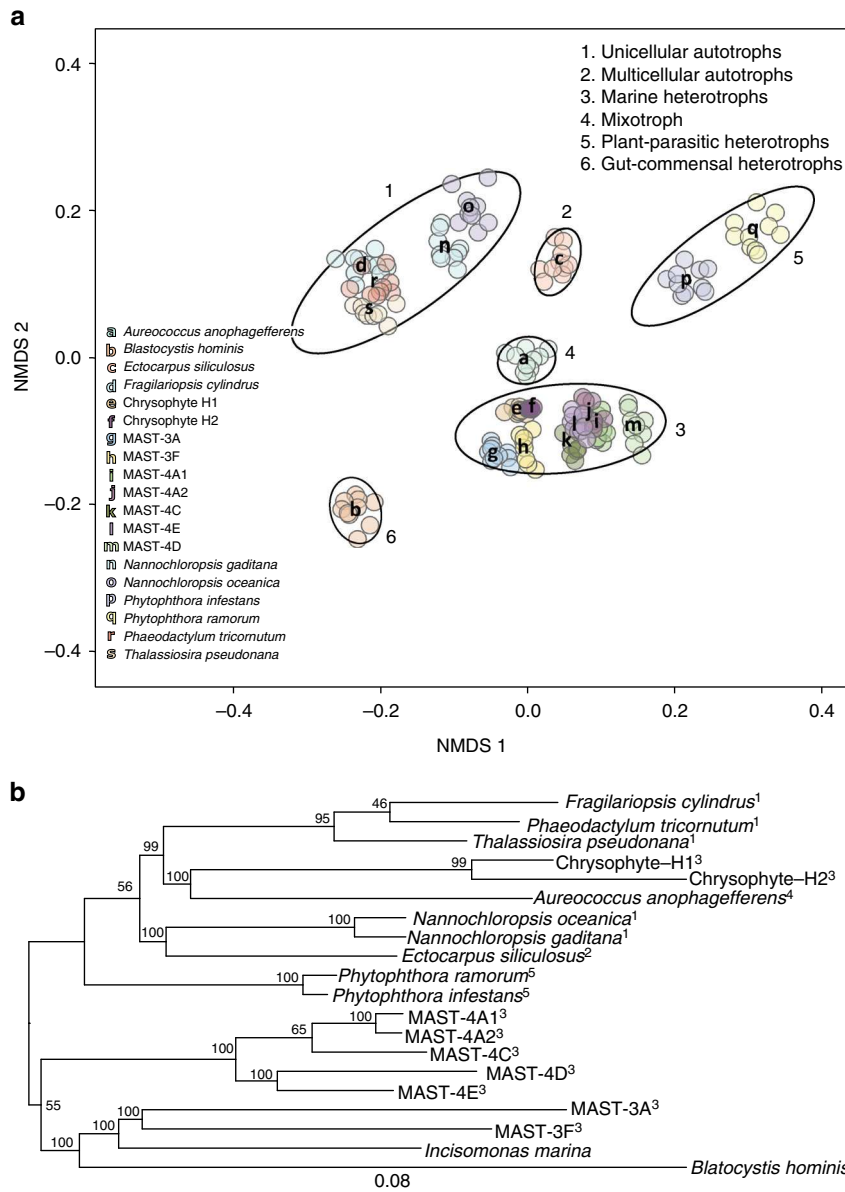


Fig. 1 Marine heterotrophic SAG lineages form a functional group distinct from autotrophs and other heterotrophs. **a** Non-metric multidimensional scaling (NMDS) projection of a Bray-Curtis distance matrix that shows Pfam motif occurrences in various stramenopile genomes. Because the genome sequences 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-4D⁸. 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 photosynthesis²¹. 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 glyco-biological potential²⁴. 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 concentrations²⁵. 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 glycosyl-transferases (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 sub-tropical 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 discriminates 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 $< 2 \times 10^{-2}$ and p -value $< 3 \times 10^{-4}$, 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).

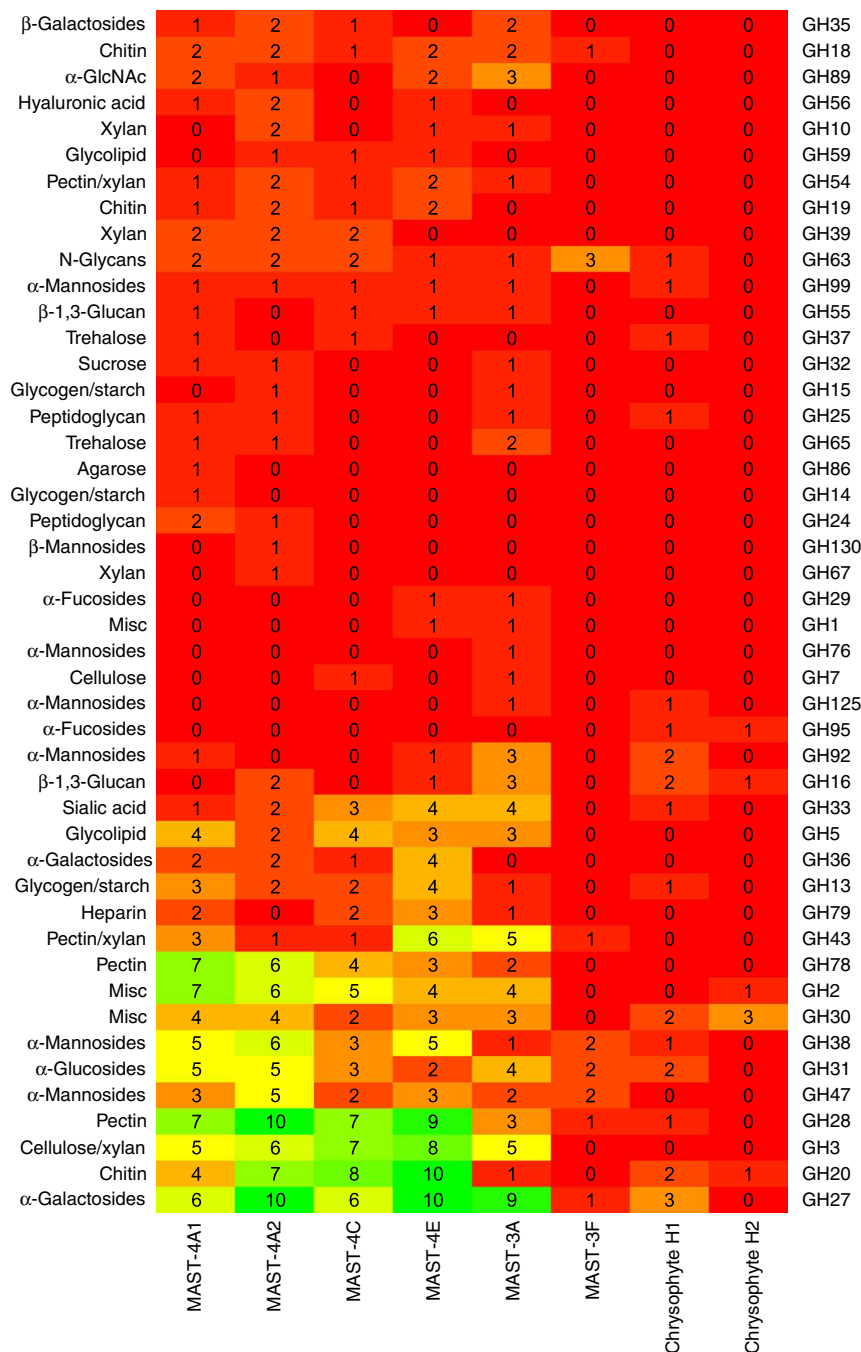


Fig. 2 SAG lineage glycoside hydrolases (GHs). GH families are numbered (right) according to the CAZyme database. Potential substrates are indicated on the left side. Internal numbers represent the number of genes in each genome predicted to belong to the GH category. Colors indicate the number of predicted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

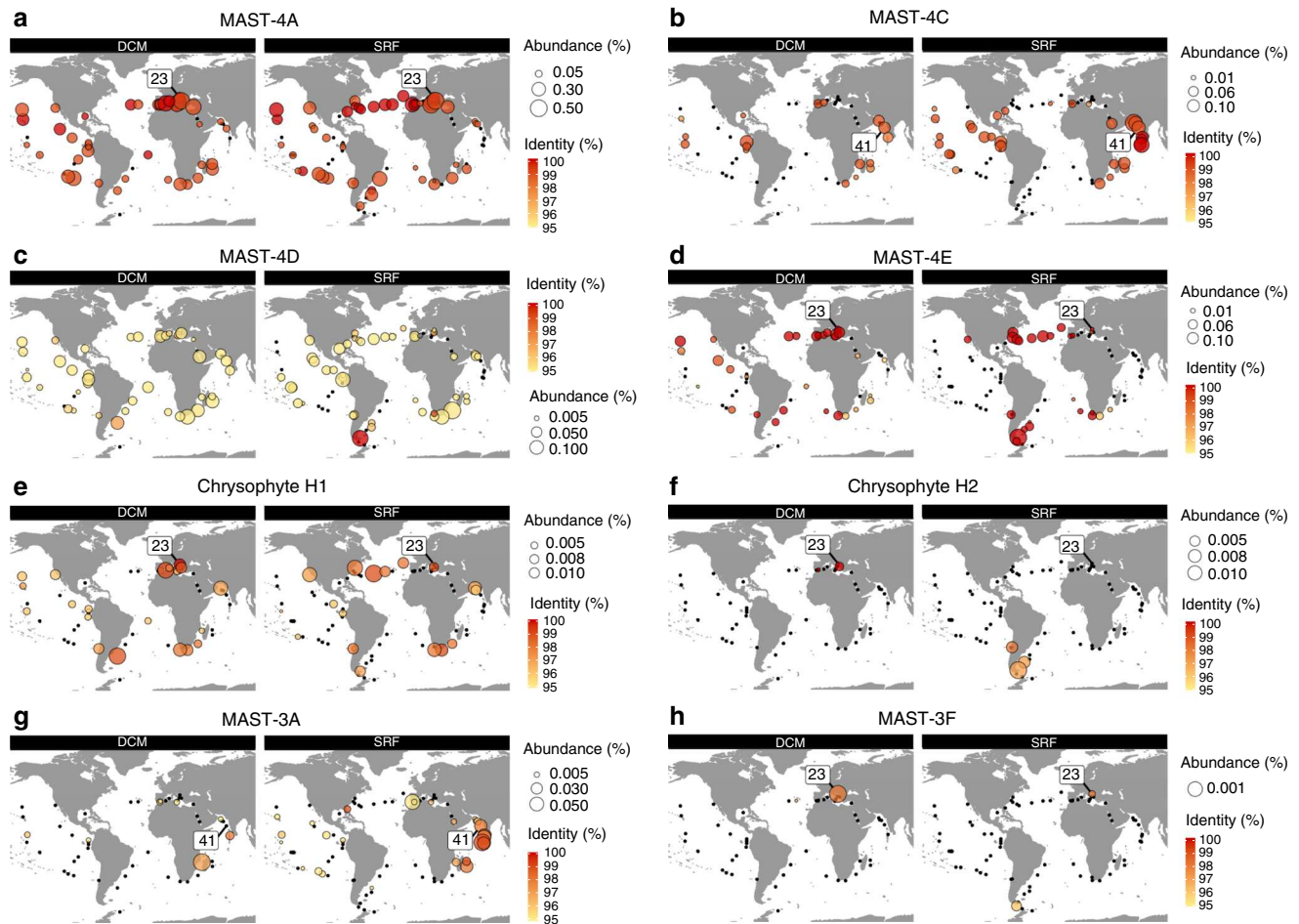


Fig. 3 Biogeographic distribution of the SAG lineages based on metagenome read recruitment with separation between deep chlorophyll maximum and subsurface. Global maps showing the presence of the SAG lineages based on metagenomics read mapping at each *Tara* Oceans station either as a black dot (no signal 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 cells¹⁸. 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 18th 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 assembler³⁵. 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 2.4³⁶, 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 v2³⁷ 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 organelle DNA sequences in different SAG assemblies, we suspected a potential biological or technical contamination of these highly amplified sequences and decided to completely separate organelle sequences from the assemblies.

The presence of organelle 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.

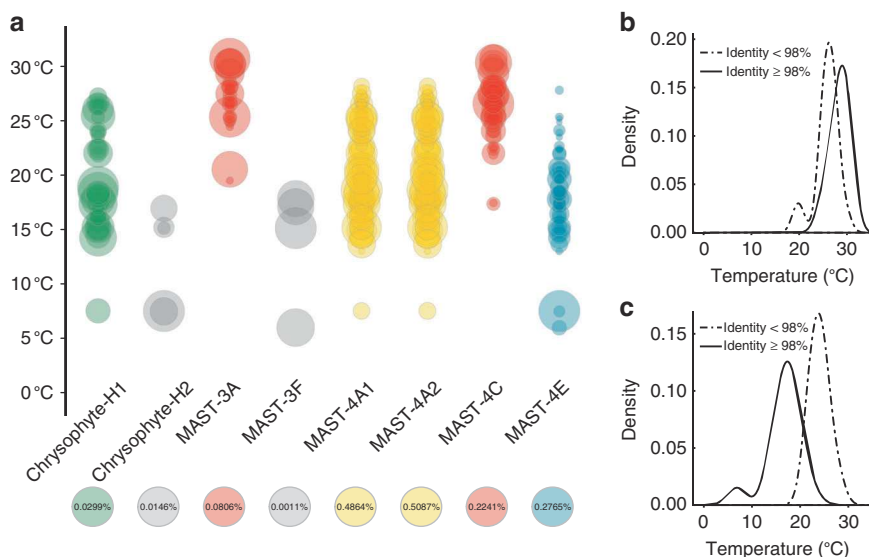


Fig. 4 Water temperature and distribution of the heterotrophic protists. **a** x-axis represents the lineage composite genomes, and y-axis represents surface temperatures in degrees Celsius at each sampling location. Relative abundances are represented by circle size (one per station/depth where the 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 $< 2 \times 10^{-2}$ (Wilcoxon test). **c** MAST-4E abundance distribution relative to temperature. Means are statistically different with a p -value $< 3 \times 10^{-4}$ (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-bp 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 $\geq 95\%$ 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 pipeline³⁹ 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 $\geq 95\%$ identity and with $\geq 80\%$ 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 $\log(\text{best eukaryotic hit } e\text{-value} + 10^{-200}) - \log(\text{best non-eukaryotic hit } e\text{-value} + 10^{-200})$. 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 aligner⁴¹ 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 $< 1 \times 10^{-0.5}$) 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 homogenize 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 (Chrysoophyte H2: 51; MAST-4D: 72; MAST-3F: 73; MAST-3A: 88; MAST-4C: 90; MAST-4A1: 113; Chrysoophyte 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 BLASTp⁴², each encoded protein model was compared to the proteins listed in the CAZy database²⁴ (<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 family⁴³.

HGT detection. The presence of putative HGT events was determined using two methods. First, in the AI method⁴⁰, 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 ($-m$ PROTAMMAJTTX 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 Interpro 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 other 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 \times assembly size / (size of the mapped genome \times 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 PAN-GAEA database²⁷. 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. Martínez-García, 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 opalozoon 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. & Arkipova, 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.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

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 Mangot⁷, Julie Poulain¹, Karine Labadie¹, Ramiro Logares⁷, 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 Picheral¹⁴, Sarah Searson¹⁵, Tara Oceans Coordinators, Stephane Pesant^{16,17}, Nicole Poulton¹⁸, Ramunas Stepanauskas¹⁸, Peer Bork⁸, Chris Bowler¹², Pascal Hingamp¹⁹, Matthew B. Sullivan²⁰, Daniele Iudicone²¹, Ramon Massana⁷, Jean-Marc Aury¹, Bernard Henrissat^{4,5,6,22}, Eric Karsenti^{12,15,16}, Olivier Jaillon^{1,2,3}, Mike Sieracki²³, Colombar de Vargas^{10,11} & Patrick Wincker^{1,2,3}

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

Tara Oceans Coordinators

Silvia G. Acinas⁷, Emmanuel Boss²⁴, Michael Follows²⁵, Gabriel Gorsky¹⁶, Nigel Grimsley^{26,27}, Lee Karp-Boss²⁴, Uros Krzic²⁸, Fabrice Not¹¹, Hiroyuki Ogata²⁹, Jeroen Raes^{30,31,32}, Emmanuel G. Reynaud³³, Christian Sardet^{16,34}, Sabrina Speich^{35,36}, Lars Stemmann¹⁶, Didier Velayoudon³⁷ & Jean Weissenbach^{1,2,3}

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

C.3. Étude globale des rhodopsines de protistes marins.

C.3.1. Présentation de la famille protéique des rhodopsines

Les rhodopsines sont des protéines photosensibles liées au rétinal, un chromophore. Ces protéines partagent une structure en 7 domaines transmembranaires, mais sont classées en deux catégories distinctes (Figure 12) :

- Les rhodopsines de type I, appelées rhodopsines microbiennes, sont présentes dans tous les domaines du vivant (bactéries, archées, eucaryotes). Celles-ci peuvent avoir différents rôles : pompes à protons (protéorhodopsines), pompes à ions chlorure (halorhodopsines), pompes à sodium ou encore rhodopsines sensorielles qui vont médier la phototaxie chez les procaryotes.
- Les rhodopsines de type II sont des récepteurs couplés aux protéines G (GPCR) qui enclenchent une cascade d'activations à la réception d'un photon. Les rhodopsines de type II permettent la vision chez les animaux.

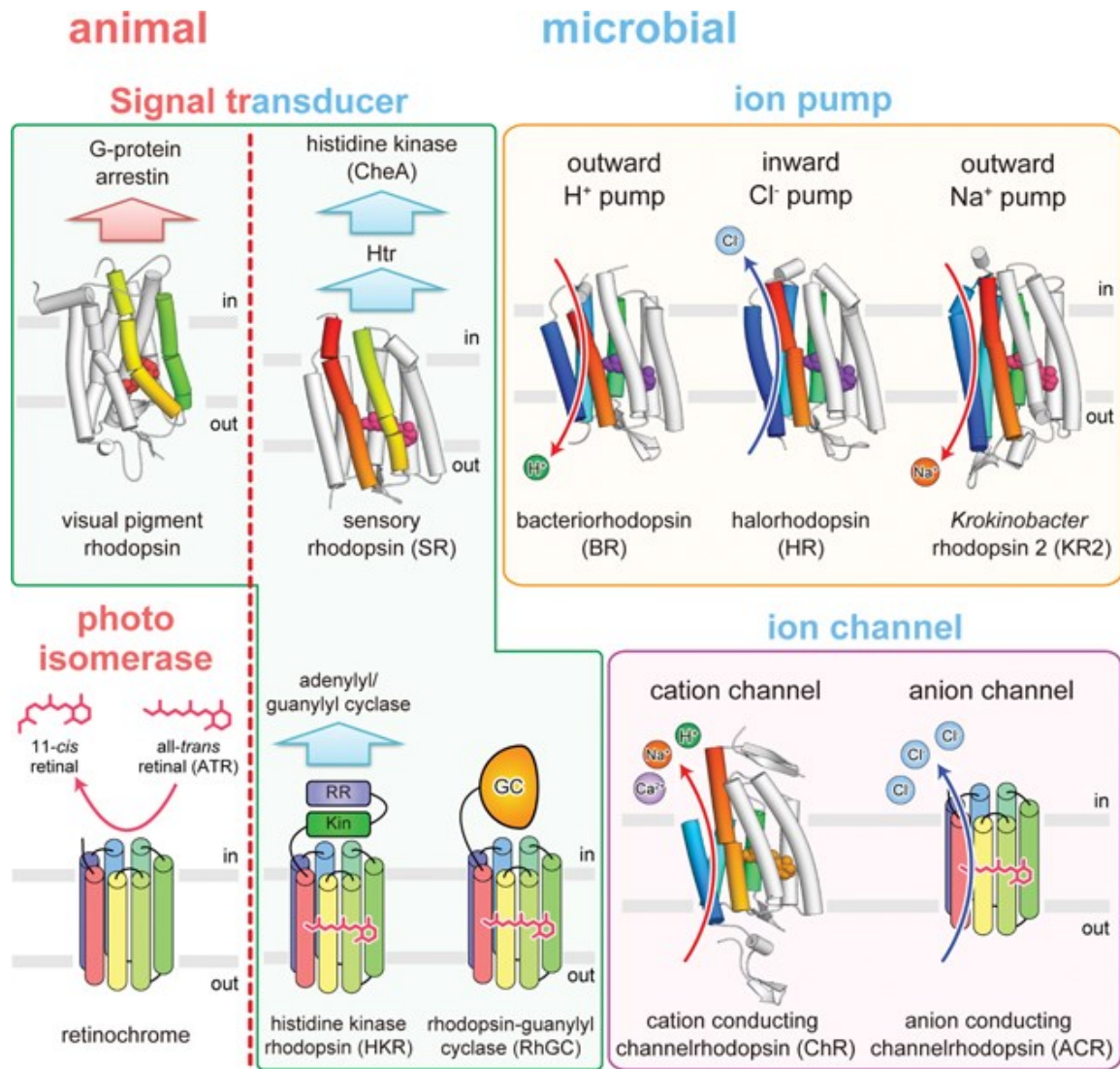


Figure 12 : Différentes catégories fonctionnelles de rhodopsines. Bleu : rhodopsines de type I, Rouge : rhodopsines de type II.

Nous avons vu dans le point précédent que certains organismes hétérotrophes comme MAST-4 C possèdent des rhodopsines du type I, homologues aux protéorhodopsines. Les protéorhodopsines ont été découvertes relativement récemment chez les bactéries marines (Beja et al., 2000) et encore un peu plus tard chez les protistes (Slamovits et al., 2011). Couplées avec une H⁺-ATPase, les protéorhodopsines permettent la production d'ATP à partir du gradient de protons généré (Beja et al., 2001). L'avantage de cette phototrophie sans chlorophylle n'est pas encore tout à fait élucidé, même s'il semblerait que la phototrophie utilisant les protéorhodopsines permette la survie et la croissance des bactéries même dans des environnements très pauvres (Gomez-Consarnau et al., 2007; Gomez-Consarnau et

al., 2010). Le rôle des protéorhodopsines dans le cycle du carbone océanique est sans doute majeur et il apparaît nécessaire d'étudier plus en détail les eucaryotes possédant ces protéines.

Nous nous sommes donc intéressés à la diversité des rhodopsines de type I chez les eucaryotes marins à travers le catalogue de gènes de *Tara Oceans* afin de dresser un catalogue exhaustif des différents types de rhodopsines présentes dans l'océan et de connaître les taxa utilisant majoritairement cette phototrophie non chlorophyllienne.

C.3.2. Matériels et méthodes

Pour évaluer la diversité des protéorhodopsines dans l'océan, nous avons recherché le motif PFAM (protein family) PF01036 'bac_rhodopsin', spécifique des rhodopsines de type I, dans le catalogue d'unigènes eucaryotes de *Tara Oceans* à l'aide de hmmscan (issu du package HMMER v3.1b2). Nous avons également constitué une banque des séquences de référence recherchant le motif PFAM sur tous les transcrits de la MMETSP (Marine Microbial Eukaryotic Transcriptome Sequencing Project (Keeling et al., 2014)). Nous y avons également inclus toutes les séquences des bases de données publiques qui possèdent le motif.

Un total de 71 576 unigènes et 4 679 séquences de référence ont été comparées au niveau protéique en utilisant BLAST+ 2.6.0. Les séquences ont ensuite été groupées en utilisant l'algorithme MCL (Enright et al., 2002). La valeur $-\log(\textit{evaluate})$ a été utilisée comme poids pour les arêtes, et le paramètre d'inflation a été fixé à 1.4 (paramètre par défaut conseillé). Au final, 78 groupes de plus de 10 séquences ont été créés, les deux plus gros groupes agrégeant plus de 70% des séquences.

Pour chacun des trois plus grands groupes, les séquences protéiques ont été alignées avec le programme MAFFT 7.310 et les positions avec moins de 50% de trous ont été conservées. Les séquences consensus ont été réalisées avec weblogo 3 et les segments transmembranaires ont été détectés en utilisant TMHMM Server 2.0 sur les

séquences consensus.

C.3.3. Résultats

Plus de 77% des séquences portant le motif PFAM PF01036 font partie des 3 plus grands groupes MCL. La plupart des unigènes eucaryotes (73%) sont assignés aux alvéolés, montrant la grande importance de ce groupe au niveau de la photosensibilité médiée par les rhodopsines. Le groupe 1, le plus important en nombre de séquences, contient des protéines similaires aux xanthorhodopsines, dont la séquence contient les résidus essentiels à la fonction de pompe à protons (Figure 13). À la différence des protéorhodopsines, les xanthorhodopsines connues utilisent en plus du rétinol une antenne de caroténoïdes comme pigment (Balashov et al., 2005). Presque toutes les séquences du groupe 1 sont assignées aux alvéolés, aux straménopiles et aux haptophytes, ce qui dans l'hypothèse d'un transfert horizontal de gène (Slamovits et al., 2011), ferait remonter le transfert à l'ancêtre commun du groupe SAR et des *haptista*.

Le deuxième groupe contient des séquences similaires aux séquences de protéorhodopsines connues, mais le groupe agrège très peu de séquences de référence eucaryotes (37). La plupart des unigènes sont assignés aux alvéolés, incluant les syndiniales, un groupe d'alvéolés parasites. L'utilisation de la phototrophie chez les parasites est aujourd'hui inconnue et pose des questions quant au rôle exact de ces protéines chez les syndiniales.

Il est intéressant de noter que l'acide aminé responsable de l'ajustement de la longueur d'onde absorbée (Man-Aharonovich et al., 2004) est différent entre les séquences consensus du groupe 1 et du groupe 2. La séquence consensus du groupe 1 possède une leucine en position 105 (notation de Man-Aharonovich et al. (2004)), correspondant à une absorption dans le vert, tandis que la séquence consensus du groupe 2 possède une glutamine à cette position, correspondant à une absorption dans le bleu.

Enfin, le troisième groupe est bien couvert par les séquences de référence et contient presque toutes les séquences de rhodopsines sensorielles connues. De plus,

l'acide aminé accepteur de proton (E76 sur la séquence consensus) essentiel à la fonction de pompe à proton n'est pas conservé dans ce groupe (Figure 13). Il semblerait donc que le groupe 3 soit constitué en partie au moins de rhodopsines sensorielles, médiateur de la phototaxie. La distribution phylogénétique de ce groupe est plus variée que dans les groupes 1 et 2. On trouve un grand nombre de dinoflagellés et de straménopiles, mais également des séquences assignées aux champignons. Cependant, plus de 50% des séquences n'ont pas d'assignation plus précise que 'Eucaryote', mettant en lumière le manque évident d'organismes de référence utilisant les rhodopsines.

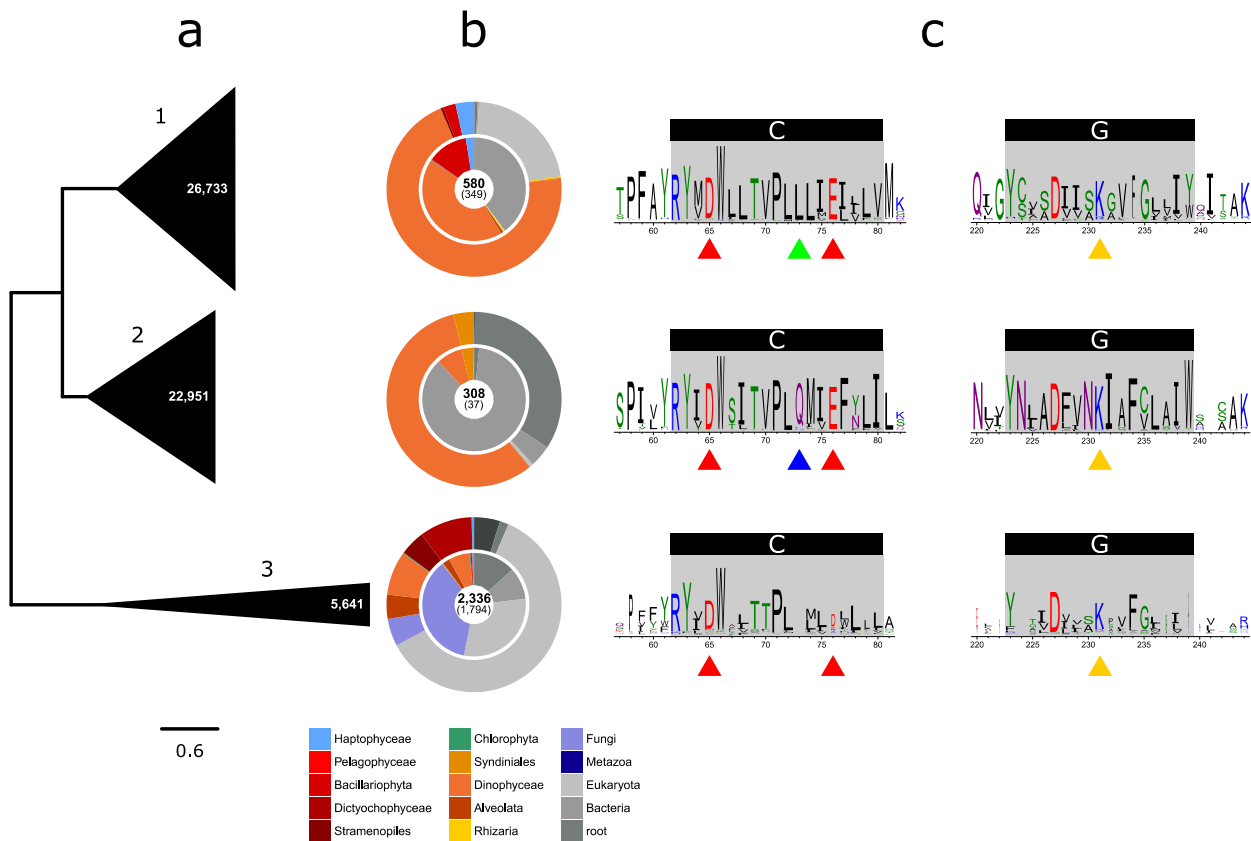


Figure 13 : Caractérisation des 3 sous-familles de rhodopsines de type I les plus exprimées dans le catalogue de gènes *Tara Oceans*. **a.** Arbre phylogénétique des 3 plus grands groupes MCL obtenus à partir de l'échantillonnage de 300 séquences dans chaque groupe. L'inférence a été réalisée grâce à FastTree (approximation du maximum de vraisemblance). La hauteur des triangles représente le nombre de séquences dans chaque groupe MCL (aussi indiquée explicitement en blanc) et la largeur représente la longueur maximum de branche dans chaque groupe après élimination des 5% de branches les plus longues. **b.** Assignment taxonomique des séquences de référence (diagramme interne) et des unigènes (diagramme externe) de chaque groupe MCL. Le nombre de séquences de référence est indiqué en gras, avec le nombre de références eucaryotes entre parenthèses. **c.** Logo des séquences consensus de chaque groupe après alignement global. Deux régions d'intérêt (hélices C et G et acides aminés voisins) contenant des résidus fonctionnels et des résidus conservés sont représentées. Certains acides aminés fonctionnels sont indiqués par des flèches. Rouge : donneur de protons (D^{65}) et accepteur de protons (E^{76}). Vert : résidu spécifique des protéorhodopsines sensibles à la lumière verte. Bleu : résidu spécifique des protéorhodopsines sensibles à la lumière bleue. Jaune : lysine liée au rétinol.

Globalement, les rhodopsines les plus abondantes et les plus exprimées sont des pompes à protons (xanthorhodopsines et protéorhodopsines), permettant potentiellement d'utiliser la phototrophie par couplage à une ATP-synthase. Dans l'océan, ces rhodopsines sont surtout exprimées par des alvéolés. On observe en

particulier la présence de protéorhodopsines chez des dinoflagellés parasites, les syndiniales. L'utilisation de la phototrophie basée sur les rhodopsines par des organismes parasites n'a pour le moment jamais été étudiée et pourrait être un sujet extrêmement intéressant à l'avenir. De plus, une grande partie des séquences de protéorhodopsines n'ont pas pu être assignées à un taxon, indiquant qu'un grand nombre d'espèces utilisent sans doute la phototrophie non chlorophyllienne de manière insoupçonnée.

D. Chapitre 3 : Instantané de l'état physiologique des populations de MAST-4 clade A dans l'environnement

D.1. Introduction

Dans le chapitre 2, nous avons décrit la distribution géographique de sept lignées de straménopiles marins incultivés, et mis en évidence des distributions géographiques très différentes. En particulier, un génome de MAST-4 clade A a pu être détecté abondant dans beaucoup d'échantillons d'eaux tempérées et tropicales, avec une faible diversité génétique des populations majoritairement présentes. Ces caractéristiques en font un modèle pour l'étude de la variation de l'expression génique de populations naturelles. Nous avons ainsi utilisé les données métatranscriptomiques de l'expédition *Tara Oceans* pour avoir un aperçu des fonctions exprimées par le nanoflagellé MAST-4 A dans différents bassins océaniques. Ce travail a fait l'objet d'un article soumis au journal *Environmental Microbiology*.

D.2. Article 2 : *Probing metabolic states of the uncultured marine protist MAST-4 A using environmental metatranscriptomics*

Résumé :

Les eucaryotes unicellulaires marins sont une composante majeure des cycles biogéochimiques océaniques. Malgré leur abondance et l'importance des rôles écologiques de ces organismes, la plupart des protistes hétérotrophes sont encore peu étudiés, en raison de la difficulté et du temps nécessaire pour les mettre en culture. En particulier, les MAST-4 (*Marine Stramenopiles 4*) sont des eucaryotes bactérivores d'une taille de 2-3 μm , qui malgré leur forte abondance dans les eaux tempérées et tropicales sont encore aujourd'hui non cultivés en laboratoire.

Dans cette étude, nous utilisons la génomique en cellule unique et les données metatranscriptomiques de l'expédition *Tara Oceans* pour avoir un aperçu de la physiologie des populations de nanoflagellés MAST-4 clade A. À partir de l'expression relative des gènes dans un nombre important d'échantillons, nous avons pu trouver des différences de profils de transcription reflétant probablement l'état physiologique des populations de MAST-4 A dans son environnement. Nous avons examiné plus en détail le profil d'expression des gènes codant pour les protéines ribosomiques et mettons en évidence une corrélation négative entre l'expression de ces gènes et la température de l'eau. Nous avons également élucidé la fonction de l'un des gènes les plus exprimés chez MAST-4 A et nous avons trouvé que ce gène code pour une protéine impliquée dans la structure des mastigonèmes. Nous avons également trouvé d'autres gènes avec une structure et un profil d'expression similaires, et émettons l'hypothèse que ces gènes sont également impliqués dans le fonctionnement du flagelle.

En utilisant la génomique en cellule unique et des données environnementales, nous avons observé des différences dans le profil d'expression des populations naturelles de MAST-4 A, laissant supposer que les populations sont au moins partiellement synchronisées. Nous notons également que les différences majeures entre les profils d'expression ont trait à la phagotrophie, laissant supposer que la disponibilité des proies est le principal facteur déclencheur de la réponse

transcriptionnelle de MAST-4 A dans son environnement.

1 **Probing metabolic states of the uncultured marine protist MAST-4 A using environmental**
2 **metatranscriptomics**

3

4 **Running title:** MAST-4 A transcription profiles in the environment

5

6 **Authors:** Yoann Seeleuthner^{1,2,3}, Jade Leconte^{1,2,3}, Quentin Carradec^{1,2,3}, Patrick Wincker^{1,2,3*}

7 ¹ CEA – Institut de biologie François Jacob, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry France.

8 ² CNRS, UMR 8030, CP5706, Evry France.

9 ³ Université d'Evry, UMR 8030, CP5706, Evry France

10

11 *** Correspondence:** Patrick Wincker (pwincker@genoscope.cns.fr).

12

13 **Keywords:** marine stramenopile, MAST, uncultured microbes, single-cell genomics,
14 **metatranscriptomics, expression profiles**

15

16 **Abstract**

17 Marine unicellular eukaryotes (protists) are a major component of ocean biogeochemical cycles.

18 Despite their important abundance and ecological roles, most heterotrophic protists are still

19 understudied due to the difficulty and time requirements to grow them in culture. In particular,

20 marine stramenopiles 4 (MAST-4) are bacterivorous eukaryotes that are relatively abundant in

21 tropical and temperate epipelagic waters, but cannot be cultured yet.

22 In this study, we used single-cell genomics and metatranscriptomic data from the *Tara* Oceans

23 expedition to get an insight into the physiology of the heterotrophic nanoflagellate MAST-4 clade A.

24 Based on relative gene expression across multiple samples, we found different transcriptional
25 signatures enlightening MAST-4 A physiological states in the environment, at different time points
26 and locations. We further examined the ribosomal protein genes expression pattern and pointed out a
27 negative correlation with water temperature. Additionally, we computationally investigated the
28 function of one of the most expressed gene of MAST-4 A and found that it encodes a protein
29 involved in the structure of stramenopile mastigonemes. We found other genes with a similar
30 structure and expression pattern, and make the hypothesis that these genes are also involved in the
31 structure of the flagellum. Overall, based on environmental data and single-cell genomics, we
32 observe that the differences in transcriptional profiles of natural populations of MAST-4 A are
33 mainly due to phagotrophy-related processes.

34

35 **Introduction**

36 Marine protists are unicellular eukaryotes with a phenomenal diversity (Burki, 2014; De Vargas et
37 al., 2015) that have key roles in the planktonic food web. Photosynthetic protists are primary
38 producers responsible for the fixation of dissolved CO₂ into organic matter. Heterotrophic protists
39 transfer the organic matter to higher trophic levels and remineralize a part of it *via* respiration.
40 Whereas photosynthetic protists such as diatoms, green algae and some dinoflagellates have been
41 extensively studied, heterotrophic marine protists lifestyles are still obscure. This may be mainly due
42 to the lack of culture representatives and the difficulty to precisely identify them by microscopy, in
43 particular for organisms $\leq 5 \mu\text{m}$ (Lim et al., 2001) such as nanoflagellates. To fill this gap in
44 knowledge, single-cell amplified genomes from unexplored branches of stramenopiles have been
45 sequenced (Roy et al., 2014; Mangot et al., 2017; Seeleuthner et al., 2018).

46 One of these stramenopiles, MAST-4 is a bacterivorous nanoflagellate of 2-3 μm in size that has
47 been detected highly abundant in all oceans except in polar regions (Logares et al., 2012; Rodriguez-

48 Martinez et al., 2012). In particular, MAST-4 clade A is broadly dispersed among different marine
49 environments, but surprisingly, the genetic diversity of the dominant populations is relatively low
50 (Seeleuthner et al., 2018), making it a good model for studying gene expression in different natural
51 environments. The cosmopolitan distribution of MAST-4 A is partly explained by its very large food
52 spectrum, being able to predate different bacteria such as the α -proteobacteria MED479 (Massana et
53 al., 2009), the bacteroidetes MED134 (Massana et al., 2009) or *Pelagibacter ubiquus* (Martinez-
54 Garcia et al., 2012). The genome analysis of MAST-4 A also revealed a high number of glycoside
55 hydrolases, enzymes that degrade carbohydrates, with some of them that are predicted to target algae
56 cell wall components (Seeleuthner et al., 2018).

57 Until recently, it would not have been possible to study uncultured microbes for their transcriptional
58 plasticity. In the present work, we use *Tara* Oceans metatranscriptomic data to study the
59 transcriptional states of MAST-4 A in its environment. We make hypotheses about the different
60 physiological states regarding the expression profiles and link the ribosomal genes expression pattern
61 to temperature, a result already known from cold stress experiments (Sahara et al., 2002; Kim et al.,
62 2004), but shown here on a marine heterotrophic protist in environmental conditions. We also
63 examined the most frequently differentially expressed genes and found an enrichment in phagotrophy
64 related functions. Finally, we inspected the expression pattern of the second most expressed gene of
65 MAST-4 A, a structural protein part of the mastigoneme, and propose other candidate genes to this
66 function, based on their expression patterns.

67

68 **Results and discussion**

69 **Environmental functional signatures of MAST-4 A**

70 We selected all 39 samples (Fig. 1) from the 0.8-5 μm size fraction of the *Tara* Oceans
71 metatranscriptomic dataset where at least 50% of MAST-4 A's genes were detected as transcribed.

72 For each sample, we compared the expression level of genes involved in KEGG pathways to the
73 median expression level of the same genes across the 39 selected samples and computed the log fold
74 change. We used paired Wilcoxon signed-rank tests to find KEGG pathways significantly diverging
75 from the median expression profile (Fig. 2). Surprisingly, we observed significant differences in
76 global expression profiles between environmental samples (Methods). This observation means that
77 environmental populations of MAST-4 A are sufficiently synchronized in most of samples to observe
78 an emerging profile. Moreover, despite MAST-4 A is usually considered as a cosmopolitan bacterial
79 grazer with no differentiated states, we can observe different physiological states using
80 metatranscriptomics.

81 After clustering samples based on correlation distance, two major different expression patterns can be
82 segregated based on KEGG expression profiles. First, a group composed of samples 135DCM,
83 9DCM, 52DCM, 51DCM, 97DCM, 18DCM, 25DCM, 18SRF, 80DCM, 93SRF, 132DCM, 9SRF
84 and 143SRF (see Methods for naming convention) is globally characterized by the under-expression
85 of genes involved in respiration (oxidative phosphorylation), citrate cycle, and proteasome, ribosome,
86 and lysosome activities. On the contrary, genes involved in the ribosome biogenesis and in the
87 spliceosome are slightly overexpressed compared to the median expression across samples. Globally,
88 genes involved in transcription/translation and in the cell cycle are apparently more expressed than in
89 most samples.

90 The second cluster contains samples where genes involved in respiration, citrate cycle, ribosome
91 and phagosome/lysosome are significantly overexpressed (samples 22SRF, 83SRF, 152MXL, 81SRF,
92 152SRF, 65SRF, 150DCM, 150SRF, 151SRF, 134SRF, 135SRF, 23SRF, 95SRF). Except for the ribosome
93 structure, these pathways are involved in phagotrophy, carbohydrate and energy metabolism, and
94 could have a direct link with the phagotrophy capabilities of MAST-4 A. The overexpression of these
95 processes is at the cost of different other processes such as DNA replication (samples 150DCM,

96 95SRF, 102DCM, 131DCM, 4SRF, 7DCM), transcription and translation (samples 151SRF, 135SRF,
97 23SRF, 95SRF, 102DCM, 4SRF, 7DCM) or in the peroxisome and the endocytosis (samples 145SRF,
98 83SRF, 81SRF, 152SRF).

99 It has been shown in the mixotrophic microalga *Prymnesium parvum* that the TCA cycle is
100 stimulated by the availability of preys (Liu et al., 2015). By analogy, we make the hypothesis that the
101 population of MAST-4 A is feeding in samples of cluster 2, and not (or not sufficiently) in samples of
102 cluster 1. However, we note that fatty acid metabolism ('Fatty acid metabolism' and 'Fatty acid
103 degradation' KEGG pathways) is not overexpressed in the 'feeding' situation, although this is one of
104 the major pathway overexpressed in *P. parvum* when preys are available (Liu et al., 2015).

105 Interestingly, in samples 4SRF and 7DCM, genes involved in 'protein processing in endoplasmic
106 reticulum' (particularly chaperonins), in proteolysis and in the proteasome structure are significantly
107 overexpressed compared to other samples. We interpret this result as a strong stress response of
108 MAST-4 A populations in these two samples, with a high expression level of chaperonins at the cost
109 of expression of genes involved in replication. We observe, to a lesser extent, a similar trend in the
110 102DCM sample, where numerous chaperonins are highly expressed. We are not able to explain this
111 stress response with environmental measures (temperature and/or oxygen concentration for example).

112 One possible explanation may be that biotic parameters might be responsible for the observed stress-
113 response, for example a high concentration of predators of MAST-4 A.

114 Despite a lack of biological replicates, we used a relatively high number of samples to make the
115 median expression profile more robust. We are relatively confident in the significance of our results
116 for several reasons. First, a pair of samples could be considered as replicates: 152SRF and 152MXL
117 samples have both been sampled at the same location in a mixed layer (homogeneous euphotic zone
118 because of turbulence). As expected, we find that their expression profiles are very similar (Pearson's
119 correlation coefficient = 0.9). Moreover, the clustering result of KEGG pathways makes sense. For

120 example, pathways related to cell division ('DNA replication', 'Purine/Pyrimidine metabolism', 'Cell
121 cycle', 'Nucleotide excision repair') group together. This result cannot be explained by the number of
122 shared genes between these pathways, because this association is rarely observed when randomly
123 swapping gene expression profiles (Supplementary Data 1).

124

125 **Relative expression of cytoplasmic ribosomal protein genes is negatively correlated with** 126 **temperature**

127 We investigated further the expression of ribosomal protein genes (RP genes) because the p-values
128 associated to the Wilcoxon tests were extremely low ($< 1.10^{-5}$) in multiple samples for these genes.
129 The highly significant results are explained by the similarity between RP genes expression patterns.
130 This similarity is even more striking when excluding mitochondrial ribosomal protein genes and RP
131 genes with low expression (Fig. 3 A). This suggests that MAST-4 A's cytoplasmic RP genes are co-
132 regulated, probably in the same way that previously elucidated in yeast (Reja et al., 2015). Because
133 we were able to retrieve similar expression profiles for genes with divergent sequences that are
134 known to be coregulated, this result is a solid validation of the environmental metatranscriptomic
135 mapping approach.

136 To investigate the role of environment on RP genes expression, we correlated the main expression
137 pattern of these RP genes to 11 environmental parameters (Fig. 3B) such as nutrient concentration,
138 salinity, temperature, depth and chlorophyll concentration (see materials and methods). We found
139 that water sample temperature is slightly but significantly negatively correlated with RP genes
140 relative expression (Fig. 3C): RP genes are more expressed in colder samples. This result is
141 consistent with cold-stress experiments on yeast (Sahara et al., 2002), on plant (Kim et al., 2004) and
142 during environmental studies of phytoplankton (Toseland et al., 2013; Pearson et al., 2015), but with
143 a relatively low number of comparisons. Here, this result is validated at the global scale for a

144 cosmopolitan heterotrophic marine organism. The generally accepted explanation is that
145 overexpression of RP genes is necessary under low temperatures to maintain protein production
146 while translation efficiency decreases (Sahara et al., 2002; Toseland et al., 2013). Overexpression of
147 RP genes seems to be a widespread acclimation strategy to lower temperatures and can be measured
148 *in situ*, not only in controlled experiments.

149

150 **Most frequently differentially expressed genes**

151 We performed pairwise comparisons between samples to detect differentially expressed genes (see
152 methods). We looked for Gene Ontology (GO) enrichment in the most frequently differentially
153 expressed genes among pairwise comparisons. After multiple testing corrections, we observed a
154 significant enrichment of 11 GO terms in frequently differentially expressed genes (Table 1). Genes
155 of these categories include subunits of the vacuolar ATPase (GO:0015991 , GO:0033179 ,
156 GO:0016820, GO:0033178) – proton pumps that acidify the phagosomes and lysosomes in the cell –,
157 ABC transporters (GO:0005215, GO:0016887, GO:0006810) and structural constituents of
158 cytoskeleton (GO: 0005200). These processes have a direct link to phagotrophy: phagosome
159 formation requires cytoskeleton rearrangements (May and Machesky, 2001) and digestion then
160 occurs in low pH lysosomes. Transporters could be involved in the excretion of unused compounds.
161 This result tends to show that genes involved in phagotrophy are regulated with more flexibility than
162 other genes in environmental conditions. MAST-4 A is not in a constant state and food availability
163 seems to have an impact on its gene transcription.

164

165 **Expression of genes encoding EGF-2 domain-containing proteins**

166 Epidermal growth factor like (EGF-like) domains are cysteine rich domains that forms disulfide
167 bonds. In particular, EGF-2 domains are EGF-like domains found in a variety of extracellular or
168 transmembrane proteins such as integrins, peptidases, or tenascins in animals.

169 Genes encoding EGF-2 domain-containing proteins are highly expressed in the sunlit ocean, mainly
170 from stramenopiles (Carradec et al.), but their exact functions are not known for now. In the MAST-
171 4 A partial genome, 20 genes encode EGF-2 domain-containing proteins. But one gene represents
172 approximately 65% of the total genes expression; on average, it is one of the most transcribed
173 predicted gene. A protein BLAST search indicated that this protein is similar to the *mas3* protein
174 (49% identity with the *sig3* protein of *Thalassiosira weissflogii*) – named *sig3* in *Thalassiosira*
175 *weissflogii* (Armbrust, 1999) – involved in the structure of the tripartite mastigonemes (hairs onto the
176 longest flagellum of stramenopiles) of the stramenopile *P. nicotianae* (Blackman et al., 2011). The
177 expression patterns of *mas* homologs are nearly identical (Fig. 4A), a feature observed with genes
178 encoding structural proteins. Interestingly, other genes with EGF-2 domains exhibit the same
179 expression profile (GSMAS4A2.ASY1.ANO1.7266.1, GSMAS4A1.ASY1.ANO1.5313.1 and
180 GSMAS4A2.ASY1.ANO1.7438.1) but are not detected as part of the *mas* gene family based on
181 sequence similarity (Fig. 4B). However, the encoded proteins are conserved among other
182 stramenopiles and their exact function remains unknown. An interesting hypothesis for future
183 research on those orphan genes may be that they could encode other proteins involved in the
184 mastigoneme structure directly or in the mastigoneme assembly or regulatory process.

185

186 **Conclusions**

187 Using single-cell genomics and metatranscriptomic data, we were able to highlight different
188 transcriptional states in the uncultured widely distributed picoeukaryote MAST-4 clade A. This result
189 demonstrates that MAST-4 A populations are responding in a relatively synchronized manner in

190 environmental samples. This is generally observed for autotrophic species, where individual cells
191 are reacting collectively in the same way to abiotic factors, but our results indicate that
192 heterotrophs may behave in a similar coordinated fashion, although the nature of the triggers of
193 the transcriptomic response are unknown. Because many of the differences in transcriptional
194 pattern are related to phagotrophy, we hypothesize that MAST-4 A transcriptional flexibility in the
195 environment may be mainly related to its feeding state, and that entire populations are involved in
196 this transcriptional state.

197 Additionally, we observed a consistent expression profile of ribosomal protein genes across
198 samples, an argument in favor of a coexpression of ribosomal protein genes, and correlated their
199 expression pattern to water temperature at the sampling location.

200 Moreover, we investigated further the expression profile of the *mas* genes, coding proteins involved
201 in the structure of the mastigonemes and propose that other proteins with a similar structure (same
202 protein domains) and expression pattern may play a role in the mastigoneme structure or assembly
203 process.

204

205

206

207 **Materials and methods**

208 ***Tara* Oceans samples naming convention**

209 *Tara* Oceans samples are named using the *Tara* station number followed with the sampling depth.

210 Stations are numbered sequentially from the beginning to the end of the expedition (Fig. 1). Depth is

211 encoded on three letters as follow: SRF for surface, DCM for deep chlorophyll maximum and MXL

212 for mixed layer. For example, 151SRF designates the sample collected at the surface of station
213 TARA_151 (Pesant et al., 2015).

214

215 **Mapping of *Tara* Oceans metatranscriptomic readsets**

216 *Tara* Oceans metatranscriptomic readsets (Alberti et al., 2017) from the 0.8-5 μm size-fraction
217 (MAST-4 A's size is 2-3 μm) were mapped using bowtie2 2.2.6 (Langmead et al., 2009) (default
218 parameters) onto MAST-4 A1's and MAST-4 A2's predicted genes, separately. To dismiss
219 nonspecific mapping, reads with $\geq 70\%$ of low-complexity bases were filtered out using a custom
220 script based on DUST. PCR duplicates were removed with 'samtools rmdup' command after
221 alignment. Reads with a mapping quality score (MAPQ) ≤ 5 and with $\leq 95\%$ identity to the reference
222 were also removed. Genes that exhibited an outlier pattern in at least one *Tara* Oceans
223 metagenomic sample were not considered further because their relative expression could result
224 from cross-mapping with closely related species reads (see Seeleuthner *et. al* for details). Briefly, we
225 considered that gene read counts should follow a normal distribution for a given species in a given
226 metagenomic sample. Genes with a read count significantly outside the distribution in at least one
227 sample were not further investigated in expression analyses. With this method, a total of 663 (8%)
228 and 655 (8%) genes were removed from the gene sets of MAST-4 A1 and MAST-4 A2 respectively to
229 create 'inlier' genesets.

230

231 **Merging of MAST-4 A 1's and MAST-4 A 2's predicted genesets**

232 Because orthologous genes of MAST-4 A 1 and MAST-4 A 2 are nearly identical (98.5% mean
233 identity at the nucleotide level), 'inlier' genesets of both genomes were merged to create a more
234 exhaustive dataset. After the union, redundant proteins were detected by performing reciprocal

235 best hits using BLAST+ (Camacho et al., 2009) (e-value < 10^{-10}). For each couple of redundant
236 proteins, the MAST-4 A 1's copy was removed from the dataset. In the end, 4,007 proteins (34%)
237 originate from MAST-4 A1's inlier geneset and 7,797 proteins (66%) from MAST-4 A 2's inlier
238 geneset. Read counts from independent mappings were used so that mapping reads are not
239 distributed on two orthologs if ortholog detection failed.

240

241 **Functional annotation**

242 Protein domains were identified using hmmscan from package HMMER 3.1b1 (<http://hmmer.org/>)
243 against the Pfam database version 28 (Finn et al., 2014). The search was performed using the
244 default parameters and the gathering threshold (--cut_ga option).

245 Amino-acid sequences were compared to the Kyoto Encyclopedia of Genes and Genomes using
246 BlastKOALA (Kanehisa et al., 2016) ('protists' taxonomy group; 'family_eukaryotes' database). Using
247 the pathway file of the KEGG website (http://www.kegg.jp/kegg-bin/get_htext?br08901.keg), each
248 gene was associated to a level-C group. KEGG groups involved in 'human diseases' and 'organismal
249 systems' categories of the KEGG pathway maps were considered as irrelevant for a marine single-
250 celled organism and were not considered further.

251

252 **Statistical analysis of KEGG pathways expression**

253 Read counts of genes from the merging of MAST-4 A 1 and MAST-4 A 2 transcriptomes with a KEGG
254 annotation (1326 genes) were quantile-normalized using the 'normalize.quantiles' function of the
255 'preprocessCore' R package (v1.32.0) over samples where at least 50% of MAST-4 A's genes were
256 detected as transcribed (at least 1 mapping read). Multiple Wilcoxon signed-rank tests were then
257 performed to compare the expression of each gene in a considered KEGG pathway to their average

258 expression across all samples. Resulting p-values were adjusted using Benjamini-Hochberg's false
259 discovery rates procedure (Benjamini and Hochberg, 1995).
260 Log fold change between the median expression in sample and median expression across samples
261 were represented using 'heatmap.2' function of the 'gplots' R package (v2.17.0) (Warnes et al.,
262 2015). Samples and pathways were clustered using the correlation distance (1 - Pearson's
263 correlation coefficient).
264 Negative controls were performed by randomly swapping read counts between genes and doing the
265 same analysis as above (Supplementary Data 1).

266

267 **Detection of frequently differentially expressed genes**

268 We performed all vs. all sample comparisons and compared the read count of each gene in the two
269 compared samples after quantile normalization. To detect differentially expressed genes, we
270 considered the two pairs of samples 152SRF/152MXL and 22SRF/23SRF as pseudo replicates. We
271 modelled a threshold from the MA-plots in such a way that no more than 1% of genes are considered
272 as differentially expressed in pseudo replicates.

273 Thus, a gene is considered as differentially expressed between samples i and j if and only if:

$$\log_2\left(\frac{rc_i}{rc_j}\right) \geq \frac{35}{\log_2(rc_i + rc_j)}$$

$$rc_i \geq 20 \text{ and } rc_j \geq 20$$

274 with rc_i being the read count in sample i and rc_j being the gene read count in sample j .

275 Genes that were detected as differentially expressed in at least 10 comparisons were considered as
276 frequently differentially expressed.

277

278 **Correlation of ribosomal protein genes expression to environmental parameters**

279 Ribosomal protein genes were detected using interproscan 5.17.56.0. We excluded mitochondrial RP
280 genes from the analysis because their expression pattern was not consistent. Additionally, we
281 excluded genes with a read count < 10, to have a robust representative pattern. We z-transformed the
282 expression of each RP gene across samples and correlated the mean z-score of each pattern to
283 different environmental parameters measured at sampling locations: depth, salinity, temperature,
284 nitrite concentration, nitrite+nitrate concentration, phosphorus concentration, chlorophyll
285 concentration, oxygen concentration, net primary production and angular scattering coefficient at
286 470nm (suspended particles). We also correlated the mean expression of RP genes to iron
287 concentration, obtained using the MIT Darwin Project (<http://darwinproject.mit.edu/>). After
288 adjustment for multiple testing using the Bonferroni correction, RP gene expression has been found
289 to be significantly correlated to temperature ($R^2 = 0.317$, adjusted p-value < 0.0012).

290

291 **Canonical correspondence analysis**

292 The canonical correspondence analysis was performed using the 'VEGAN' 2.4-1 R package. Gene
293 normalized read counts were used as species and were constrained according to environmental
294 data (see previous paragraph). Resulting plot was species-centered and thus samples are not to
295 scale.

296

297 **Detection of *mas* homologs and reconstruction of *mas3* phylogenetic tree**

298 Mas orthologs were detected using the MCL algorithm (Li et al., 2003) on the reference sequences
299 of *mas* genes (from *Phytophthora infestans*, *Phytophthora nicotianae*, *Phytophthora sojae*,
300 *Phytophthora capsici*, *Phytophthora ramorum*, *Ochromonas danica*, *Aureococcus anophagefferens*,
301 *Thalassiosira pseudonana* and *Thalassiosira weissflogii*), MAST-4 A predicted genes and all the
302 eukaryotic unigenes of the Tara Ocean catalog. Several inflation thresholds were tested (1.4, 3.0,

303 4.0, 5.0): no threshold can effectively distinguish *mas1* subgroup from *mas2* subgroup family
304 without splitting reference genes of the *mas2* subgroups in different MCL clusters. Inflation
305 threshold 1.4 was unable to distinguish reference *mas1* genes from reference *mas2* genes and
306 inflation threshold 5.0 was too stringeant and clusterized reference *mas2* genes in two different
307 MCL clusters. We utilized inflation thresholds 3.0 and 4.0, with identical results.
308 Phylogenetic tree was reconstructed using the phylogeny.fr pipeline (Dereeper et al., 2008).
309 MUSCLE 3.5 was used to perform alignments, Gblocks 0.91b for curation (Min. seq. for flank pos.:
310 85%, Max. contig. conserved pos.: 8, Min. block length: 10) and PhyML+aLRT to find the maximum-
311 likelihood phylogenetic tree.

312

313 **Acknowledgements**

314 We thank the commitment of the people and sponsors who made this singular expedition possible.
315 Funding is provided through the French Gouvernement 'Investissement d'Avenir' programs
316 Oceanomics (ANR-11-BTBR-0008) and FRANCE GENOMIQUE (ANR-10-INBS-09. We thank the
317 *Tara* schooner and its captain and crew. *Tara* Oceans would not exist without continuous support
318 from 23 institutes (<http://oceans.taraexpeditions.org>). We also acknowledge C. Scarpelli for support
319 in high-performance computing. This article is contribution number XX of *Tara* Oceans.

320

321 **Conflict of Interest**

322 The authors declare no conflict of interest.

323 **References**

324 Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I. et al. (2017) Viral to metazoan
325 marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci Data* 4: 170093.

- 326 Armbrust, E.V. (1999) Identification of a new gene family expressed during the onset of sexual
327 reproduction in the centric diatom *Thalassiosira weissflogii*. *Appl Environ Microbiol* 65: 3121-3128.
- 328 Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful
329 approach to multiple testing. *Journal of the royal statistical society Series B (Methodological)*: 289-
330 300.
- 331 Blackman, L.M., Arikawa, M., Yamada, S., Suzuki, T., and Hardham, A.R. (2011) Identification of a
332 mastigoneme protein from *Phytophthora nicotianae*. *Protist* 162: 100-114.
- 333 Burki, F. (2014) The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb*
334 *Perspect Biol* 6: a016147.
- 335 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L.
336 (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.
- 337 Carradec, Q., Pelletier, E., Silva, C.D., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R. et al. A global
338 ocean atlas of eukaryotic genes. *Nature Communications* (in press).
- 339 De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R. et al. (2015) Ocean plankton.
340 Eukaryotic plankton diversity in the sunlit ocean. *Science* 348: 1261605.
- 341 Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F. et al. (2008) Phylogeny.fr:
342 robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* 36: W465-469.
- 343 Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R. et al. (2014) Pfam: the
344 protein families database. *Nucleic Acids Res* 42: D222-230.
- 345 Kanehisa, M., Sato, Y., and Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG Tools for
346 Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* 428: 726-731.
- 347 Kim, K.Y., Park, S.W., Chung, Y.S., Chung, C.H., Kim, J.I., and Lee, J.H. (2004) Molecular cloning of
348 low-temperature-inducible ribosomal proteins from soybean. *J Exp Bot* 55: 1153-1155.
- 349 Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and memory-efficient
350 alignment of short DNA sequences to the human genome. *Genome Biol* 10: R25.
- 351 Lim, E.L., Dennett, M.R., and Caron, D.A. (2001) Identification of heterotrophic nanoflagellates by
352 restriction fragment length polymorphism analysis of small subunit ribosomal DNA. *J Eukaryot*
353 *Microbiol* 48: 247-257.
- 354 Liu, Z., Jones, A.C., Campbell, V., Hambright, K.D., Heidelberg, K.B., and Caron, D.A. (2015) Gene
355 expression in the mixotrophic prymnesiophyte, *Prymnesium parvum*, responds to prey availability.
356 *Front Microbiol* 6: 319.
- 357 Logares, R., Audic, S., Santini, S., Pernice, M.C., de Vargas, C., and Massana, R. (2012) Diversity
358 patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing.
359 *ISME J* 6: 1823-1833.

360 Mangot, J.F., Logares, R., Sanchez, P., Latorre, F., Seeleuthner, Y., Mondy, S. et al. (2017) Accessing
361 the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells.
362 Sci Rep 7: 41498.

363 Martinez-Garcia, M., Brazel, D., Poulton, N.J., Swan, B.K., Gomez, M.L., Masland, D. et al. (2012)
364 Unveiling in situ interactions between marine protists and bacteria through single cell sequencing.
365 ISME J 6: 703-707.

366 Massana, R., Unrein, F., Rodriguez-Martinez, R., Forn, I., Lefort, T., Pinhassi, J., and Not, F. (2009)
367 Grazing rates and functional diversity of uncultured heterotrophic flagellates. ISME J 3: 588-596.

368 May, R.C., and Machesky, L.M. (2001) Phagocytosis and the actin cytoskeleton. J Cell Sci 114: 1061-
369 1077.

370 Pearson, G.A., Lago-Leston, A., Canovas, F., Cox, C.J., Verret, F., Lasternas, S. et al. (2015)
371 Metatranscriptomes reveal functional variation in diatom communities from the Antarctic
372 Peninsula. ISME J 9: 2275-2289.

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

375 Reja, R., Vinayachandran, V., Ghosh, S., and Pugh, B.F. (2015) Molecular mechanisms of ribosomal
376 protein gene coregulation. Genes Dev 29: 1942-1954.

377 Rodriguez-Martinez, R., Rocap, G., Logares, R., Romac, S., and Massana, R. (2012) Low evolutionary
378 diversification in a widespread and abundant uncultured protist (MAST-4). Mol Biol Evol 29: 1393-
379 1406.

380 Roy, R.S., Price, D.C., Schliep, A., Cai, G., Korobeynikov, A., Yoon, H.S. et al. (2014) Single cell
381 genome analysis of an uncultured heterotrophic stramenopile. Sci Rep 4: 4780.

382 Sahara, T., Goda, T., and Ohgiya, S. (2002) Comprehensive expression analysis of time-dependent
383 genetic responses in yeast cells to low temperature. J Biol Chem 277: 50015-50021.

384 Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., Wessner, M. et al. (2018) Single-
385 cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity
386 across oceans. Nature Communications.

387 Toseland, A., Daines, S.J., Clark, J.R., Kirkham, A., Strauss, J., Uhlig, C. et al. (2013) The impact of
388 temperature on marine phytoplankton resource allocation and metabolism. Nature Climate Change
389 3: 979-984.

390 Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T. et al. (2015) gplots:
391 various R programming tools for plotting data. R package version 2.17. 0. Computer software]
392 Available online at: <http://CRAN.R-project.org/package=gplots>.

393

395 **Tables**

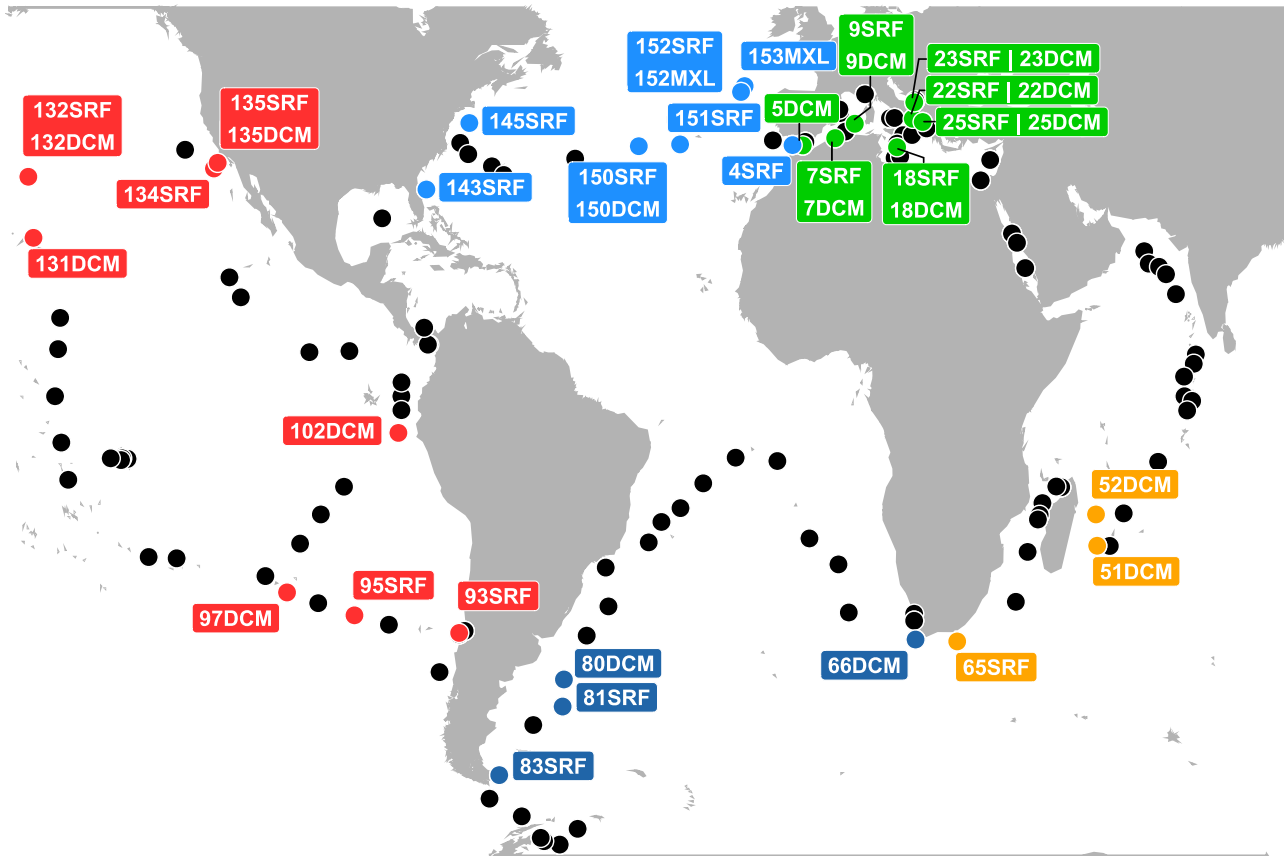
GO	Name	p-value	n	N
GO:0015991	ATP hydrolysis coupled proton transport	8.26.10 ⁻⁶	7	15
GO:0005215	transporter activity	0.00067	10	48
GO:0033179	proton-transporting V-type ATPase, V0 domain	0.00108	4	8
GO:0016887	ATPase activity	0.00209	14	96
GO:0006810	Transport	0.00218	16	120
GO:0004601	peroxidase activity	0.00706	3	6
GO:0016820	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	0.00706	3	6
GO:0033178	proton-transporting two-sector ATPase complex, catalytic domain	0.00706	3	6
GO:0015078	hydrogen ion transmembrane transporter activity	0.01329	4	12
GO:0005200	structural constituent of cytoskeleton	0.02083	4	13
GO:0020037	heme binding	0.03370	6	30

396 **Table 1.** Enriched GO terms for most frequently differentially expressed genes of MAST-4 A. n:
397 number of genes in the GO category that are frequently differentially expressed; N: total number of
398 genes in the GO category.

399

400

401 **Figures**

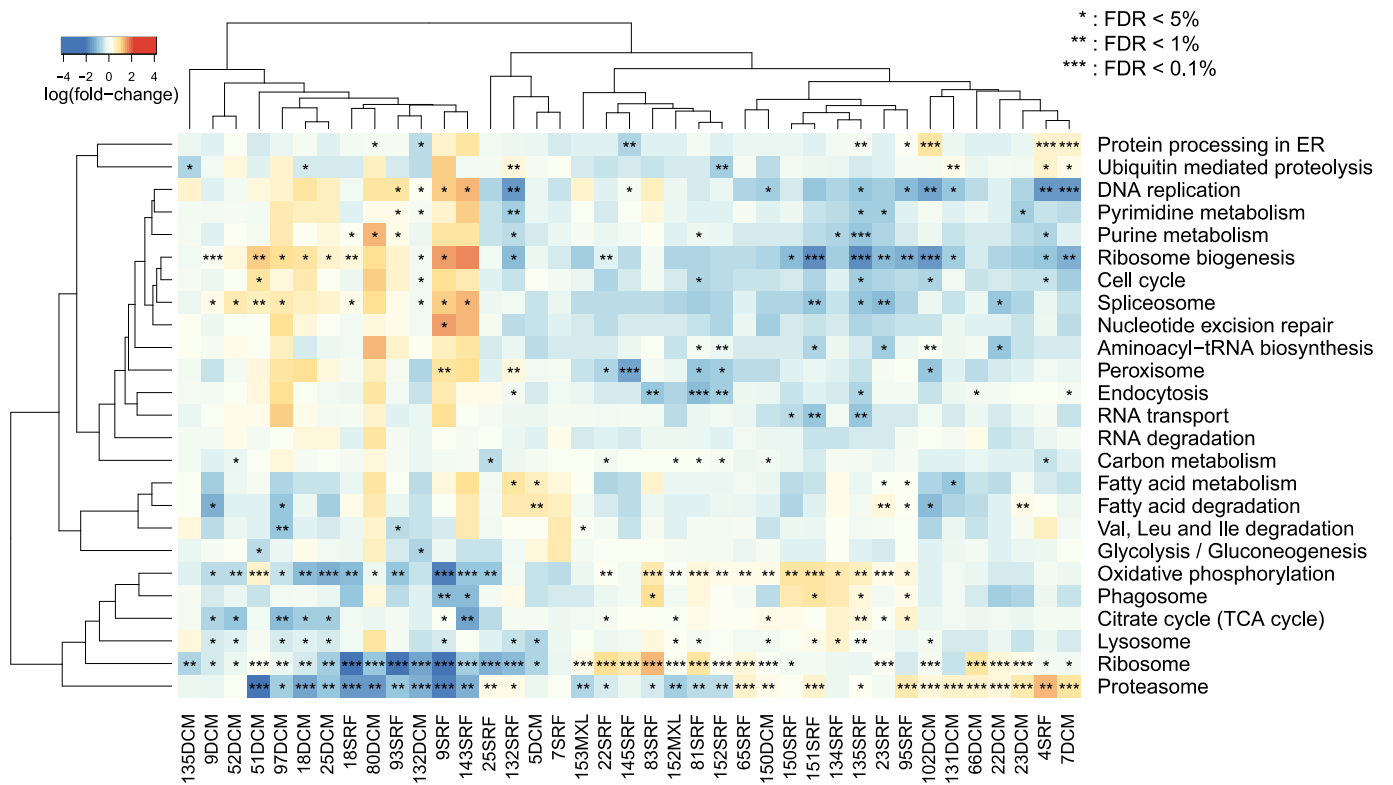


402

403 **Figure 1. Sampling locations of *Tara Oceans* metatranscriptomic samples used in this study.** Each
404 environmental sample used in this study is labelled with the number of the *Tara Oceans* station
405 followed with the sampling depth (SRF for surface, DCM for deep chlorophyll maximum and MXL for
406 mixed layer). Colors represent the oceanic basin: green: Mediterranean Sea; yellow: Indian Ocean;
407 deep blue: South Atlantic Ocean; red: Pacific Ocean; light-blue: North Atlantic Ocean. Black dots
408 represent other sampling sites of the *Tara Oceans* expedition.

409

410



411

412 **Figure 2. Relative expression profiles of MAST-4 A's pathways in environmental**

413 **metatranscriptomics samples.** Log fold change between normalized expression of genes and their

414 median expression level across samples, grouped by KEGG pathways. Samples and pathways were

415 clustered using correlation distance (1 – Pearson's correlation coefficient). Warm colors (yellow to

416 red) indicate that gene expression in the sample is higher than their median expression, whereas

417 cold colors (blue-gray to dark blue) indicate the contrary. Wilcoxon pairwise test have been

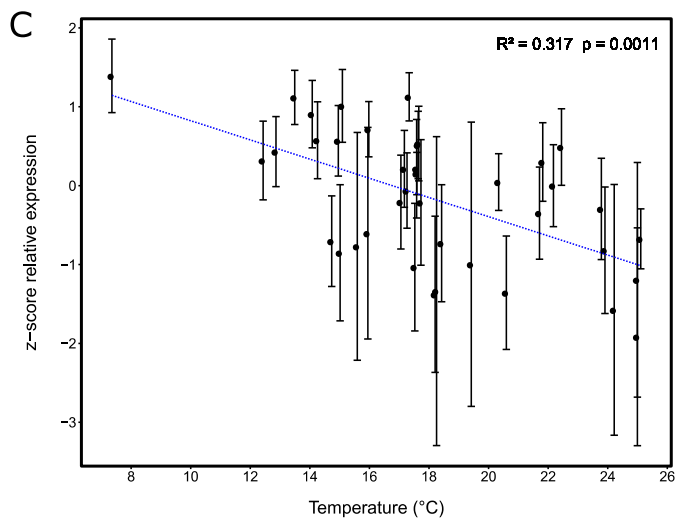
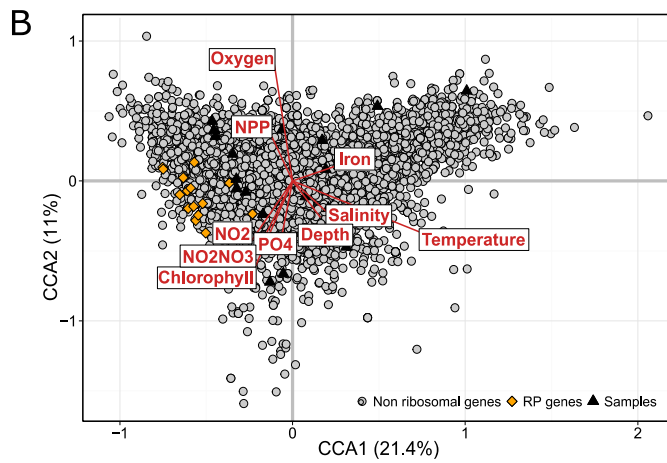
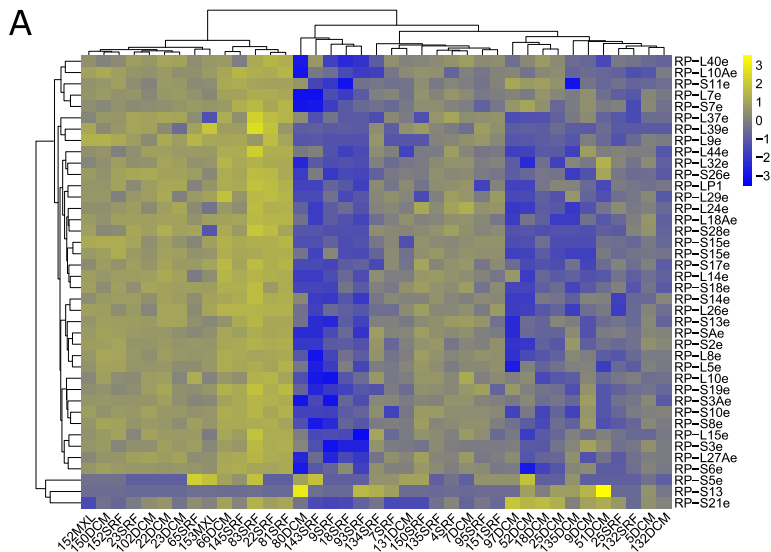
418 performed and p-values were adjusted for multiple testing using Benjamini-Hochberg's false

419 discovery rate (FDR) method and significant values are annotated with 1, 2 or 3 stars, according to

420 the FDR threshold (0.05, 0.01 and 0.001 respectively).

421

422



423

424 **Figure 3. A. Expression patterns of genes encoding ribosomal proteins in MAST-4** A. Genes with a

425 low read count (< 10) and genes encoding mitochondrial ribosomal proteins were discarded.

426 Samples with less than 20% of MAST-4 A's genes detected were discarded. Ribosomal protein genes
427 exhibit a very similar expression pattern (median Pearson's correlation coefficient = 0.71).

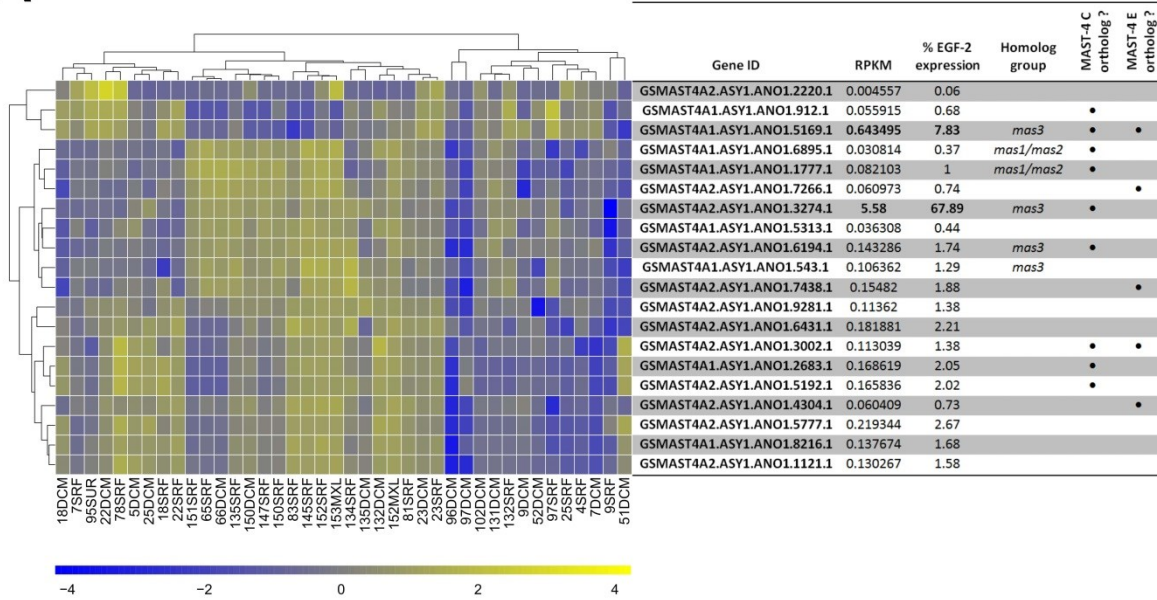
428 **B. Canonical correspondence analysis between MAST-4 A gene expression levels and**
429 **environmental parameters.** Percentages on axes represent the proportion of inertia of each axis.

430 Ribosomal protein genes (orange diamonds) are on the left of the first axis, of which temperature is
431 the most important component, meaning that their expression could be mainly explained by
432 temperature. NPP: net primary production; PO₄: dissolved phosphate concentration; NO₂:
433 dissolved nitrite concentration; NO₂NO₃: dissolved nitrite + nitrate concentration.

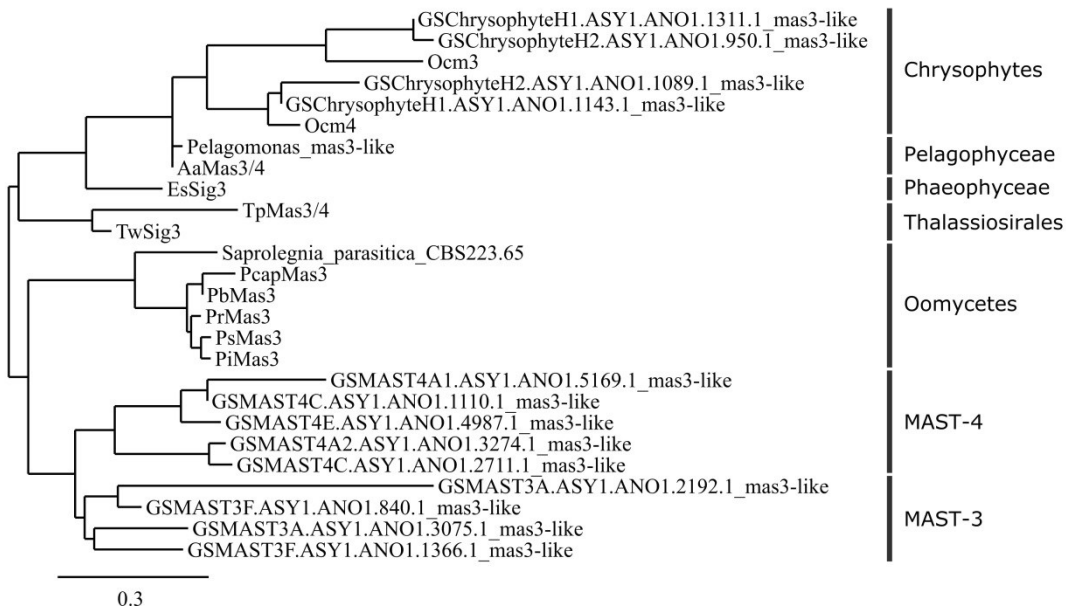
434 **C. Correlation between median z-score of genes encoding ribosomal proteins and temperature in**
435 **the sample.** Error bars represent the interquartile range of z-scores in a sample and the black line
436 represents the linear regression, showing a negative correlation between ribosomal gene
437 expression and sample temperature. The correlation is still significant (adjusted p-value < 0.003)
438 when removing the leftmost point at 8°C.

439

A



B



440

441 **Figure 4.** A. Genes predicted to encode EGF-2 domain-containing proteins in MAST-4 A genome and
 442 their mean normalized expression (RPKM). Some of these genes are similar to the *mas* gene family
 443 which encodes proteins constitutive of the stramenopile mastigonemes. Gene
 444 GSMAS4A2.ASY1.ANO1.3274.1 is orthologous to *mas3* and represents more than 67% of the
 445 expression of genes encoding EGF-2 domain containing proteins.

446 B. Unrooted phylogenetic tree of the *mas3*-like proteins of stramenopiles using Phylogeny.fr
447 (Dereeper et al., 2008). Data from Blackman *et al.* (2011), with the addition of proteins from single
448 amplified genomes of MASTs and marine chrysophytes.

449

D.3. Conclusions

Grâce à la génomique en cellule unique et à la métatranscriptomique environnementale, nous avons pu en apprendre plus sur les états transcriptionnels des populations naturelles de MAST-4 A. Du fait de l'émergence de différents profils transcriptionnels, nous pensons que les populations de MAST-4 A sont en grande partie synchronisées dans l'environnement. La plupart des différences entre les profils de transcription sont liées à la phagotrophie, montrant que le comportement trophique de MAST-4 A est sans doute responsable d'une grande part de sa plasticité transcriptionnelle.

Nous avons également mis en évidence que l'un des gènes les plus exprimés chez MAST-4 A code pour une protéine impliquée dans la structure des mastigonèmes. Des protéines avec des domaines similaires (domaines EGF-2, qui permettent la création de ponts disulfures) et un profil d'expression très similaire ont été retrouvées. Nous émettons l'hypothèse que ces protéines sont également impliquées dans la structure des mastigonèmes, ou dans le contrôle de leur assemblage.

Conclusions et perspectives

Durant ce projet de thèse, nous avons utilisé la génomique en cellule unique pour en apprendre plus sur le mode de vie de plusieurs lignées de nanoflagellés hétérotrophes marins encore non cultivés à ce jour. Souvent considérés comme des brouteurs de bactéries où seul le taux de capture est pris en compte dans les modélisations écologiques, nous avons mis en évidence une réalité plus complexe.

Pour ce faire, nous avons tout d'abord dû mettre en place une méthode d'assemblage, de décontamination et d'annotation de ces génomes séquencés en cellule unique qui a été utilisée pour l'étude de sept lignées de straménopiles. La principale originalité de cette méthode est qu'elle utilise fortement les données métagénomiques et métatranscriptomiques, qui apportent des informations biologiques importantes pour la décontamination et l'annotation syntaxique des génomes.

La génomique comparative de sept lignées de straménopiles marins non cultivés appartenant aux groupes MAST-3, MAST-4 et aux chrysophytes a mis en évidence une diversité fonctionnelle importante chez ces organismes hétérotrophes, mais cependant discernable des fonctions retrouvées chez les straménopiles photoautotrophes. Mais malgré des répertoires de gènes relativement proches au niveau global, des spécificités ont été trouvées dans chaque génome. Le génome de MAST-4 clade C par exemple possède deux gènes codant pour des protéorhodopsines, indiquant que cet organisme est potentiellement capable d'utiliser la lumière comme source d'énergie (Béjà et al., 2001). Ces protéorhodopsines auraient été acquises par transfert horizontal (Slamovits et al., 2011). Si cette hypothèse est exacte, et dans le cas d'un unique transfert depuis un procaryote, l'origine de ce dernier remonterait à l'ancêtre des alvéolés et des straménopiles. Cependant, il peut également s'agir d'un transfert eucaryote - eucaryote, beaucoup plus difficile à mettre en évidence.

Globalement, l'analyse fonctionnelle des gènes potentiellement d'origine eucaryote et acquis par transferts horizontaux montrent un enrichissement en

fonctions liées au métabolisme, en particulier dans la dégradation des carbohydrates et des protéines, dans l'utilisation de l'azote et dans la défense/résistance contre les bactéries. Les transferts sélectionnés ont donc probablement contribué à la spécialisation métabolique de ces organismes.

Nous avons également observé l'absence de gènes fonctionnels codant pour la chaîne lourde de la dynéine chez MAST-3 A. Ces protéines sont impliquées dans la structure du flagelle, indiquant que MAST-3 A est probablement non motile et vit peut-être fixé sur un substrat, à l'instar de *Solenicola setigera*. Cette réduction du nombre de gènes codant pour la chaîne lourde de la dynéine n'est pas observée chez MAST-3 F. Cependant, ce dernier possède un nombre relativement réduit d'enzymes de dégradation des carbohydrates, ainsi qu'une faible proportion de transferts horizontaux d'origine bactérienne, nous amenant à émettre l'hypothèse que cet organisme ne serait peut-être pas bactériophage.

Les deux génomes de chrysophytes H possèdent également relativement peu de glycosides hydrolases, ce qui peut s'expliquer par l'histoire évolutive de ce groupe phylogénétique où beaucoup d'espèces sont phototrophes et mixotrophes. Cela suggère des gains et pertes fréquentes de gènes impliqués dans la phagotrophie et la photosynthèse.

L'utilisation des données métagénomiques nous a également permis de connaître la distribution géographique de ces organismes, plus précisément qu'en utilisant les codes-barres génétiques. Le paramètre environnemental le plus déterminant pour la distribution de ces organismes semble être la température, à l'instar de ce qu'il se passe pour les procaryotes (Sunagawa et al., 2015). Seul un organisme, MAST-4 A, est retrouvé dans tous les bassins océaniques, excepté aux régions polaires où la température diffère énormément. Ce succès écologique pourrait s'expliquer par le nombre de proies que MAST-4 A est capable d'ingérer. Il possède en effet un équipement enzymatique capable de dégrader un nombre important de carbohydrates, impliqués notamment dans la composition des parois bactériennes, mais certaines enzymes sont également capables de dégrader les composants de la paroi des microalgues.

Le plus étonnant est que MAST-4 A, malgré sa distribution globale, possède une faible variabilité génétique. Nous avons donc pu nous intéresser à la transcription relative des gènes de MAST-4 A dans les nombreux échantillons où celui-ci est abondant, et mis en évidence des différences d'expression de fonctions liées à la phagotrophie. La réponse transcriptionnelle de MAST-4 A semble indiquer que les populations environnementales répondent principalement à la présence ou l'absence de proies.

Nous avons également pu établir une corrélation négative entre l'expression des gènes ribosomiques et la température. Ce résultat était déjà observé chez le phytoplancton (Toseland et al., 2013), mais il est ici mis en évidence chez un organisme hétérotrophe et vérifié sur un grand nombre d'échantillons. L'explication communément avancée est que les organismes compensent la baisse d'efficacité de la traduction lorsque la température baisse par l'expression plus forte des gènes codant les protéines ribosomiques.

En complément, il pourrait être intéressant de s'intéresser à la structure des populations de MAST-4 A en utilisant les variations présentes dans les données métagénomiques, et de déterminer les facteurs sous tendant cette structure. Par exemple, il a été démontré que les courants étaient un facteur important dans la structure des populations méditerranéennes du copépode *Oithona nana* (Madoui et al., 2017).

La détermination des modes trophiques de ces organismes à partir de leur génome séquencé en cellule unique et de la flexibilité transcriptionnelle d'un organisme bactériophage fortement abondant dans les océans questionnent leur rôle monolithique dans les modélisations écologiques d'un compartiment essentiel du cycle du carbone dans les océans.

Une intégration de la complexité des protistes hétérotrophes dans ces modèles semble nécessaire à une meilleure représentation des chaînes trophiques du plancton. La plasticité transcriptionnelle observée chez MAST-4 A laisse également ouverte la question des interactions avec les autres compartiments du plancton, champ d'étude encore inexploré pour ces organismes.

Références

- Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I. et al. (2017) Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific data* **4**: 170093.
- Bailly, J., Fraissinet-Tachet, L., Verner, M.-C., Debaud, J.-C., Lemaire, M., Wésolowski-Louvel, M., and Marmeisse, R. (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. *ISME J* **1**: 632-642.
- Balashov, S.P., Imasheva, E.S., Boichenko, V.A., Anton, J., Wang, J.M., and Lanyi, J.K. (2005) Xanthorhodopsin: a proton pump with a light-harvesting carotenoid antenna. *Science* **309**: 2061-2064.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S. et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology* **19**: 455-477.
- Beja, O., Spudich, E.N., Spudich, J.L., Leclerc, M., and DeLong, E.F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* **411**: 786-789.
- Beja, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P. et al. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902-1906.
- Béjà, O., Spudich, E.N., Spudich, J.L., Leclerc, M., and DeLong, E.F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* **411**: 786-789.
- Blanchot, J., and Rodier, M. (1996) Picophytoplankton abundance and biomass in the western tropical Pacific Ocean during the 1992 El Niño year: results from flow cytometry. *Deep Sea Research Part I: Oceanographic Research Papers* **43**: 877-895.
- Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2010) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**: 578-579.
- Buck, K., and Bentham, W. (1998) A novel symbiosis between a cyanobacterium. *Synechococcus sp.*: 349-355.
- Buitenhuis, E.T., Li, W.K., Vaultot, D., Lomas, M.W., Landry, M., Partensky, F. et al. (2012) Picophytoplankton biomass distribution in the global ocean. *Earth System Science Data* **4**: 37-46.
- Burns, C.W., and Galbraith, L.M. (2007) Relating planktonic microbial food web structure in lentic freshwater ecosystems to water quality and land use. *Journal of plankton research* **29**: 127-139.
- Canadell, J.G., Le Quere, C., Raupach, M.R., Field, C.B., Buitenhuis, E.T., Ciais, P. et al. (2007) Contributions to accelerating atmospheric CO₂ growth from economic activity, carbon intensity, and efficiency of natural sinks. *Proc Natl Acad Sci U S A* **104**: 18866-18870.
- Caron, D.A. (2016) Ocean science: The rise of Rhizaria. *Nature* **532**: 444-445.
- Carter, N.P., Bebb, C.E., Nordenskjöld, M., Ponder, B.A., and Tunnacliffe, A. (1992) Degenerate

oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**: 718-725.

Cavalier-Smith, T., and Scoble, J.M. (2013) 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.

Chambouvet, A., Morin, P., Marie, D., and Guillou, L. (2008) Control of toxic marine dinoflagellate blooms by serial parasitic killers. *Science* **322**: 1254-1257.

Chase, M.W., and Fay, M.F. (2009) Ecology. Barcoding of plants and fungi. *Science* **325**: 682-683.

Chen, M., Song, P., Zou, D., Hu, X., Zhao, S., Gao, S., and Ling, F. (2014) Comparison of multiple displacement amplification (MDA) and multiple annealing and looping-based amplification cycles (MALBAC) in single-cell sequencing. *PLoS ONE* **9**: e114520.

Cheung, V.G., and Nelson, S.F. (1996) Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proceedings of the National Academy of Sciences* **93**: 14676-14679.

Das, N., and Pandey, A. (2015) Role of Nanoplanktons in Marine food-webs. *International Letters of Natural Sciences* **43**.

De Bourcy, C.F., De Vlaminck, I., Kanbar, J.N., Wang, J., Gawad, C., and Quake, S.R. (2014) A quantitative comparison of single-cell whole genome amplification methods. *PloS one* **9**: e105585.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R. et al. (2015) Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P. et al. (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences* **99**: 5261-5266.

del Campo, J., and Massana, R. (2011) Emerging diversity within chrysophytes, choanoflagellates and bicosoecids based on molecular surveys. *Protist* **162**: 435-448.

del Campo, J., Sieracki, M.E., Molestina, R., Keeling, P., Massana, R., and Ruiz-Trillo, I. (2014) The others: our biased perspective of eukaryotic genomes. *Trends in ecology & evolution* **29**: 252-259.

Derelle, E., Ferraz, C., Rombauts, S., Rouze, P., Worden, A.Z., Robbens, S. et al. (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci U S A* **103**: 11647-11652.

Derelle, R., Lopez-Garcia, P., Timpano, H., and Moreira, D. (2016) A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). *Mol Biol Evol* **33**: 2890-2898.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S. et al. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**: 1575-1584.

Fenchel, T., and Finlay, B. (2008) Oxygen and the spatial structure of microbial communities. *Biol*

Rev Camb Philos Soc **83**: 553-569.

Godhe, A., Asplund, M.E., Harnstrom, K., Saravanan, V., Tyagi, A., and Karunasagar, I. (2008) Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl Environ Microbiol* **74**: 7174-7182.

Gomez, F. (2007) The consortium of the protozoan *Solenicola setigera* and the diatom *Leptocylindrus mediterraneus* in the Pacific Ocean. *Acta Protozoologica* **46**: 15.

Gomez, F., Moreira, D., Benzerara, K., and Lopez-Garcia, P. (2011) *Solenicola setigera* is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3. *Environ Microbiol* **13**: 193-202.

Hebert, P.D., Ratnasingham, S., and deWaard, J.R. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci* **270 Suppl 1**: S96-99.

Hou, Y., and Lin, S. (2009) Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. *PLoS ONE* **4**: e6978.

Houghton, R.A. (2007) Balancing the Global Carbon Budget. *Annu Rev Earth Planet Sci* **35**: 313-347.

Howe, K.L., Chothia, T., and Durbin, R. (2002) GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome research* **12**: 1418-1427.

Hutchison, C.A., and Venter, J.C. (2006) Single-cell genomics. *Nature biotechnology* **24**: 657-658.

JEONG, H.J. (1999) The ecological roles of heterotrophic dinoflagellates in marine planktonic community. *Journal of Eukaryotic Microbiology* **46**: 390-396.

Johnson, M.D. (2015) Inducible Mixotrophy in the Dinoflagellate *Prorocentrum minimum*. *J Eukaryot Microbiol* **62**: 431-443.

Kammerlander, B., Breiner, H.W., Filker, S., Sommaruga, R., Sonntag, B., and Stoeck, T. (2015) High diversity of protistan plankton communities in remote high mountain lakes in the European Alps and the Himalayan mountains. *FEMS Microbiol Ecol* **91**.

Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J. et al. (2011) A holistic approach to marine eco-systems biology. *PLoS biology* **9**: e1001177.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A. et al. (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* **12**: e1001889.

Kiy, T. (1998) Heterotrophic Protists—A New Challenge in Biotechnology? *Protist* **149**: 17-21.

Kogure, K., Simidu, U., and Taga, N. (1979) A tentative direct microscopic method for counting living marine bacteria. *Can J Microbiol* **25**: 415-420.

Kolber, Z.S., Barber, R.T., Coale, K.H., Fitzwateri, S.E., Greene, R.M., Johnson, K.S. et al. (1994) Iron limitation of phytoplankton photosynthesis in the equatorial Pacific Ocean. *Nature* **371**: 145-149.

Lasken, R.S., and Stockwell, T.B. (2007) Mechanism of chimera formation during the Multiple

Displacement Amplification reaction. *BMC biotechnology* **7**: 19.

Madoui, M.A., Poulain, J., Sugier, K., Wessner, M., Noel, B., Berline, L. et al. (2017) New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod *Oithona*. *Molecular ecology*.

Man-Aharonovich, D., Sabehi, G., Sineshchekov, O.A., Spudich, E.N., Spudich, J.L., and Beja, O. (2004) Characterization of RS29, a blue-green proteorhodopsin variant from the Red Sea. *Photochem Photobiol Sci* **3**: 459-462.

Mangot, J.F., Logares, R., Sanchez, P., Latorre, F., Seeleuthner, Y., Mondy, S. et al. (2017) Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci Rep* **7**: 41498.

Martinez-Garcia, M., Brazel, D., Poulton, N.J., Swan, B.K., Gomez, M.L., Masland, D. et al. (2012) Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *ISME J* **6**: 703-707.

Massana, R., and Pedros-Alio, C. (2008) Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* **11**: 213-218.

Massana, R., Guillou, L., Díez, B., and Pedrós-Alió, C. (2002) Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean. *Appl Environ Microbiol* **68**: 4554-4558.

Massana, R., Guillou, L., Terrado, R., Forn, I., and Pedrós-Alió, C. (2006) Growth of uncultured heterotrophic flagellates in unamended seawater incubations. *Aquatic microbial ecology* **45**: 171-180.

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

Massana, R., Unrein, F., Rodríguez-Martínez, R., Forn, I., Lefort, T., Pinhassi, J., and Not, F. (2009) Grazing rates and functional diversity of uncultured heterotrophic flagellates. *ISME J* **3**: 588-596.

Massana, R., Castresana, J., Balagué, V., Guillou, L., Romari, K., Groisillier, A. et al. (2004) Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl Environ Microbiol* **70**: 3528-3534.

McKie-Krisberg, Z.M., and Sanders, R.W. (2014) Phagotrophy by the picoeukaryotic green alga *Micromonas*: implications for Arctic Oceans. *ISME J* **8**: 1953-1961.

Movahedi, N.S., Forouzmand, E., and Chitsaz, H. (2012) De novo co-assembly of bacterial genomes from multiple single cells. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*: IEEE, pp. 1-5.

Movahedi, N.S., Embree, M., Nagarajan, H., Zengler, K., and Chitsaz, H. (2016) Efficient Synergistic Single-Cell Genome Assembly. *Front Bioeng Biotechnol* **4**: 42.

Not, F., del Campo, J., Balagué, V., de Vargas, C., and Massana, R. (2009) New insights into the diversity of marine picoeukaryotes. *PLoS ONE* **4**: e7143.

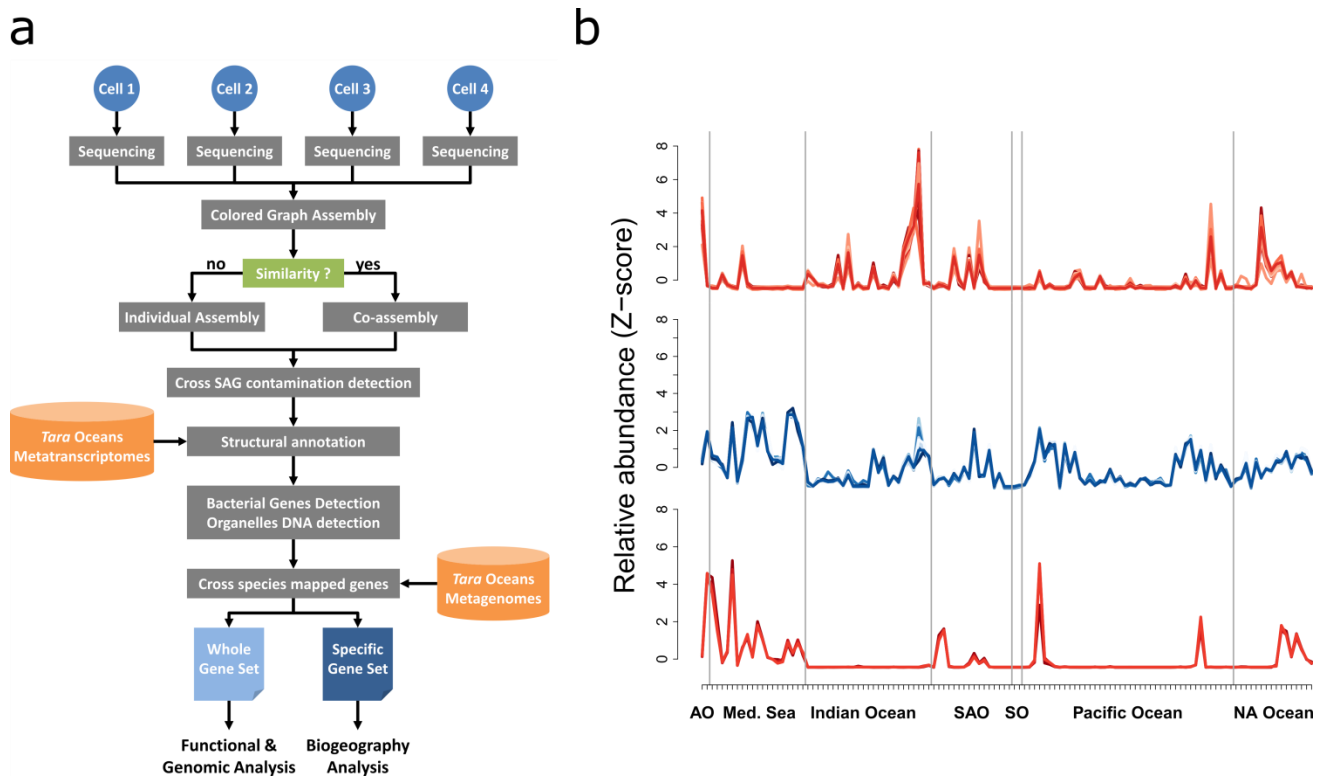
- Park, M.G., Yih, W., and Coats, D.W. (2004) Parasites and phytoplankton, with special emphasis on dinoflagellate infections. *J Eukaryot Microbiol* **51**: 145-155.
- Peng, W., Takabayashi, H., and Ikawa, K. (2007) Whole genome amplification from single cells in preimplantation genetic diagnosis and prenatal diagnosis. *Eur J Obstet Gynecol Reprod Biol* **131**: 13-20.
- Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G. et al. (2015) Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* **2**: 150023.
- Rassoulzadegan, F., Laval-Peuto, M., and Sheldon, R. (1988) Partitioning of the food ration of marine ciliates between pico-and nanoplankton. *Hydrobiologia* **159**: 75-88.
- Rice, P., Longden, I., and Bleasby, A. (2000) EMBOS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276-277.
- Rottberger, J., Gruber, A., Boenigk, J., and Kroth, P.G. (2013) Influence of nutrients and light on autotrophic, mixotrophic and heterotrophic freshwater chrysophytes. *Aquatic microbial ecology* **71**: 179-191.
- Roy, R.S., Price, D.C., Schliep, A., Cai, G., Korobeynikov, A., Yoon, H.S. et al. (2014) Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci Rep* **4**: 4780.
- Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S. et al. (2007) The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology* **5**: e77.
- Saadatpour, A., Lai, S., Guo, G., and Yuan, G.C. (2015) Single-Cell Analysis in Cancer Genomics. *Trends Genet* **31**: 576-586.
- Schlegel, M., and Hülsmann, N. (2007) Protists – A textbook example for a paraphyletic taxon. *Organisms Diversity & Evolution* **7**: 166-172.
- Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A. et al. (2013) Genomic variation landscape of the human gut microbiome. *Nature* **493**: 45-50.
- Sherr, E.B., and Sherr, B.F. (2002) Significance of predation by protists in aquatic microbial food webs. *Antonie Van Leeuwenhoek* **81**: 293-308.
- Siegenthaler, U., and Sarmiento, J. (1993) Atmospheric carbon dioxide and the ocean. *Nature* **365**: 119-125.
- Sinha, R., Stanley, G., Gulati, G.S., Ezran, C., Travaglini, K.J., Wei, E. et al. (2017) Index switching causes “spreading-of-signal” among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv*: 125724.
- Slamovits, C.H., Okamoto, N., Burri, L., James, E.R., and Keeling, P.J. (2011) A bacterial proteorhodopsin proton pump in marine eukaryotes. *Nat Commun* **2**: 183.
- Staley, J.T., and Konopka, A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology* **39**: 321-346.
- Stock, A., Breiner, H.-W., Pachiadaki, M., Edgcomb, V., Filker, S., La Cono, V. et al. (2012) Microbial eukaryote life in the new hypersaline deep-sea basin Thetis. *Extremophiles* **16**: 21-34.

- Stone, L., and Weisburd, R.S. (1992) Positive feedback in aquatic ecosystems. *Trends in ecology & evolution* **7**: 263-267.
- Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G. et al. (2015) Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**: 1261359.
- Toseland, A., Daines, S.J., Clark, J.R., Kirkham, A., Strauss, J., Uhlig, C. et al. (2013) The impact of temperature on marine phytoplankton resource allocation and metabolism. *Nature Climate Change* **3**: 979-984.
- Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M. et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37-43.
- Valentini, A., Pompanon, F., and Taberlet, P. (2009) DNA barcoding for ecologists. *Trends Ecol Evol* **24**: 110-117.
- Van de Peer, Y., and De Wachter, R. (1997) Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA. *Journal of molecular evolution* **45**: 619-630.
- Vannier, T., Leconte, J., Seeleuthner, Y., Mondy, S., Pelletier, E., Aury, J.M. et al. (2016) Survey of the green picoalga *Bathycoccus* genomes in the global ocean. *Sci Rep* **6**: 37900.
- Walker, A., and Parkhill, J. (2008) Single-cell genomics. *Nature Reviews Microbiology* **6**: 176-177.
- Wetherbee, R., and Andersen, R. (1992) Flagella of a chrysophycean alga play an active role in prey capture and selection. *Protoplasma* **166**: 1-7.
- Worden, A.Z., Follows, M.J., Giovannoni, S.J., Wilken, S., Zimmerman, A.E., and Keeling, P.J. (2015) Environmental science. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347**: 1257594.
- Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K. et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS biology* **5**: e16.
- Zhu, F., Massana, R., Not, F., Marie, D., and Vaultot, D. (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52**: 79-92.
- Zong, C., Lu, S., Chapman, A.R., and Xie, X.S. (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**: 1622-1626.
- Zubkov, M.V., Sleigh, M.A., Tarran, G.A., Burkill, P.H., and Leakey, R.J. (1998) Picoplanktonic community structure on an Atlantic transect from 50 N to 50 S. *Deep Sea Research Part I: Oceanographic Research Papers* **45**: 1339-1355.

Annexes

Annexe 1. Informations supplémentaires de l'article *Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans*

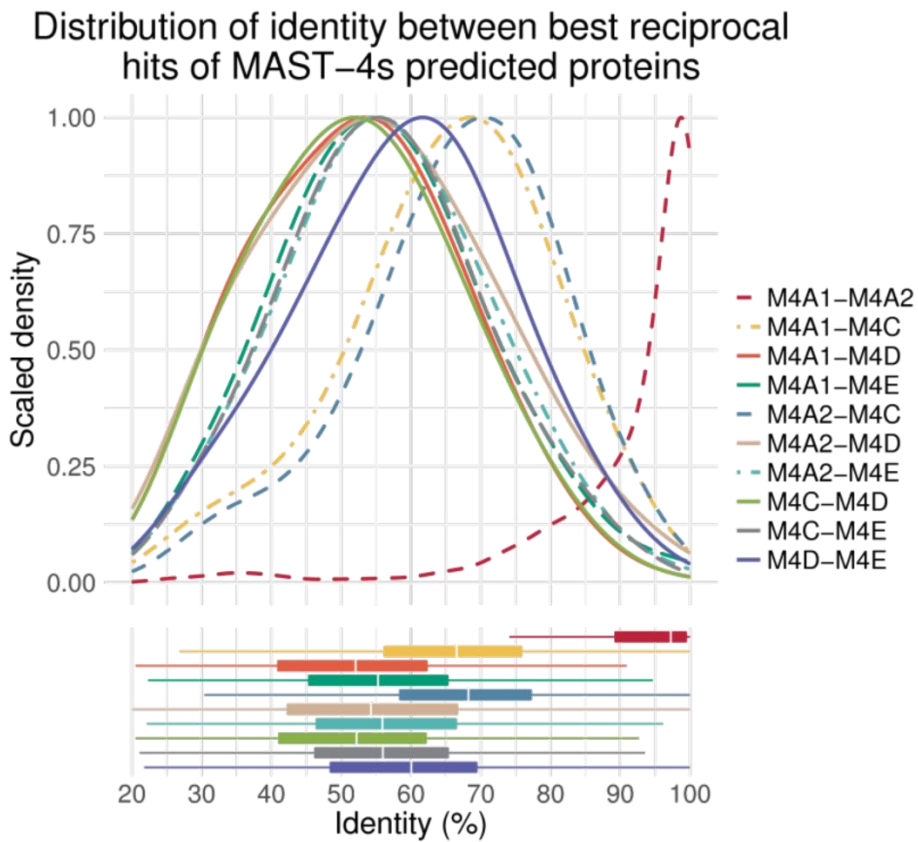
Supplementary Information



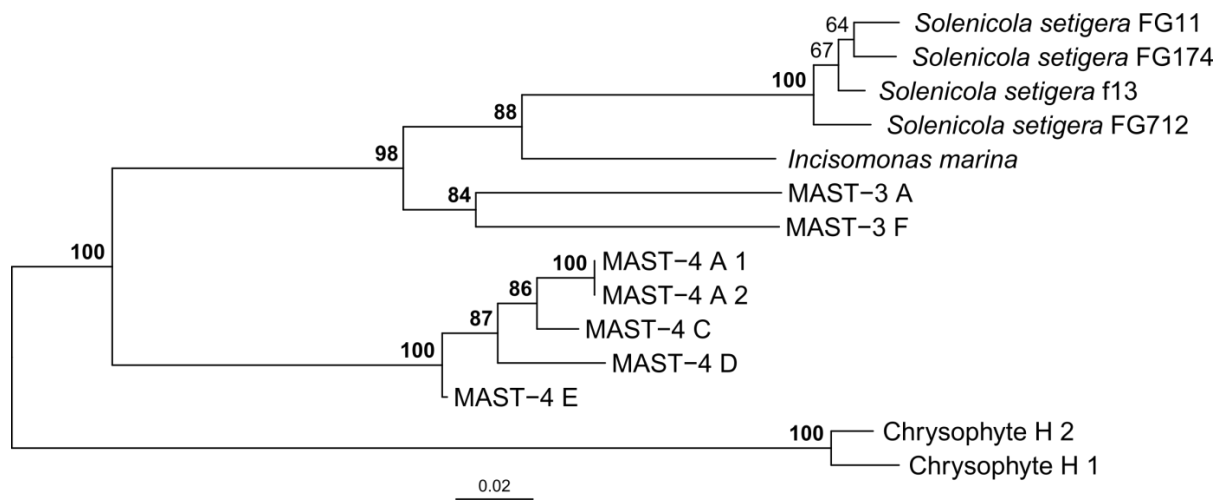
Supplementary Figure 1. Schematic pipeline for single-cell genome assembly, cleaning

and annotation. a. The assembly process was optimized to account for multiple cells putatively originating from the same species. We used metatranscriptomes from *Tara* Oceans to improve gene detection accuracy. We exploited metagenomic fragment recruitment results to filter out assembled genomic regions that likely correspond to other species and highly conserved genes that can be mapped by reads from other species. **b.** Detection of foreign contigs in the MAST4-A1 assembly by fragment recruitment analysis. The x-axis represents *Tara* Oceans metagenomes from the 0.8-5 μm size fraction (AO: Atlantic Ocean, Med. Sea: Mediterranean Sea, SAO: South Atlantic Ocean, SO: Southern Ocean, NA Ocean: North Atlantic Ocean). The y-axis represents the z-score of the abundance of mapped metagenomic reads for each metagenome. The central (blue) graph shows the typical pattern obtained for the 20 longest contigs of MAST-4A1 as an example. The red graphs show two subsets of

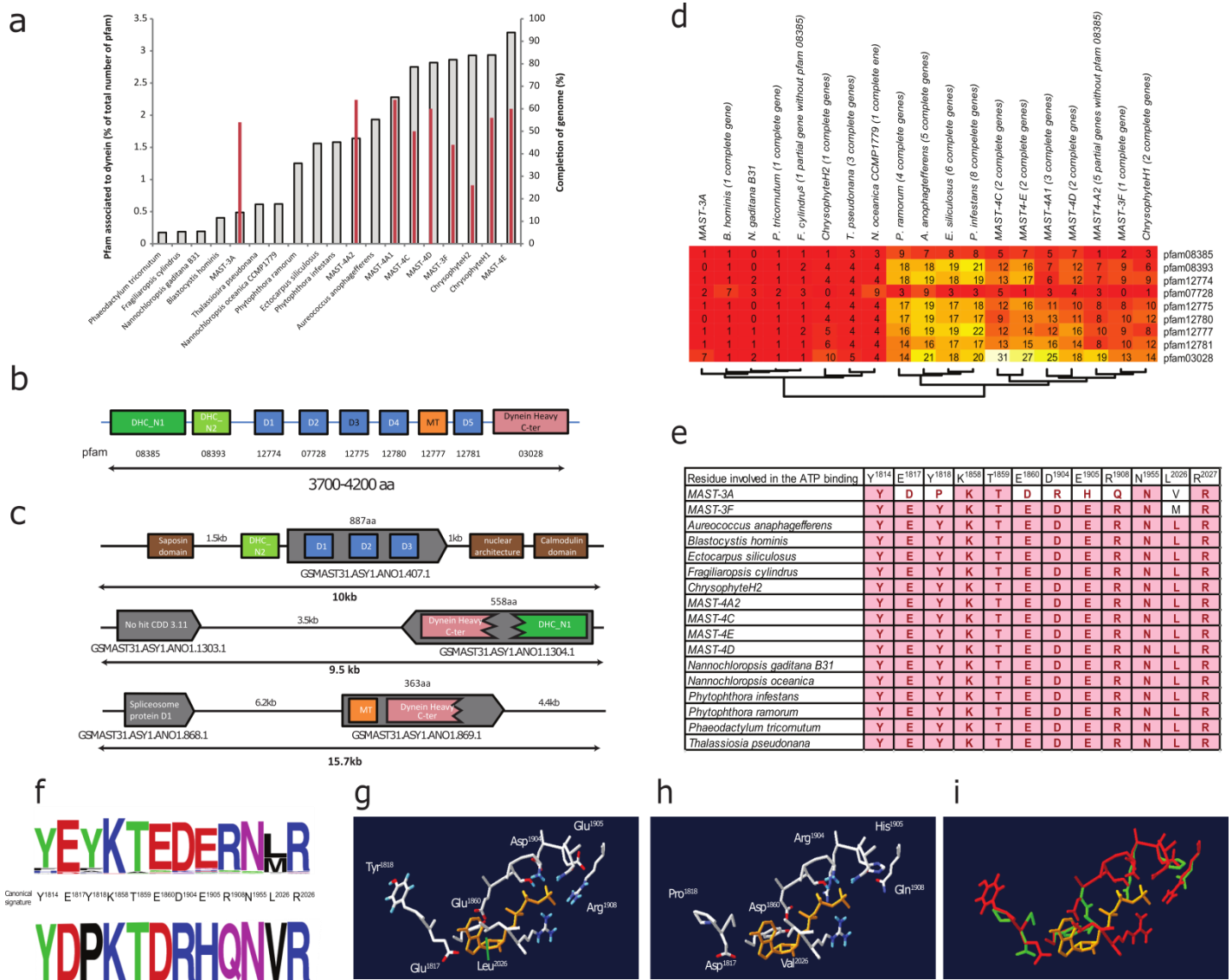
contigs from the MAST-4A1 assembly that were rejected by the filter because of a signature statistically deviating from the MAST4-A1 signature. Two different patterns are shown: the upper red plot is a subset of contigs taxonomically assigned to *Bathycoccus prasinus*, whereas the lower corresponds to contigs assigned to *Prochlorococcus* MED4.



Supplementary Figure 2: Distribution of identity between best reciprocal hits of MAST-4 predicted proteins. All-versus-all comparisons of the distribution of identity between MAST-4 predicted proteomes. Except the 2 assemblies of MAST-4 A that came from nearly identical genomes, the median identity between MAST-4 orthologs ranges from 53% to 68%. Median identity with the previously sequenced MAST-4 D ranges from 53% to 60%.



Supplementary Figure 3. Maximum-likelihood phylogenetic tree inferred from the 18S rDNA sequences of the SAGs and close taxa. Bootstrap values are indicated for each node and the scale bar represents the expected number of substitutions per site.

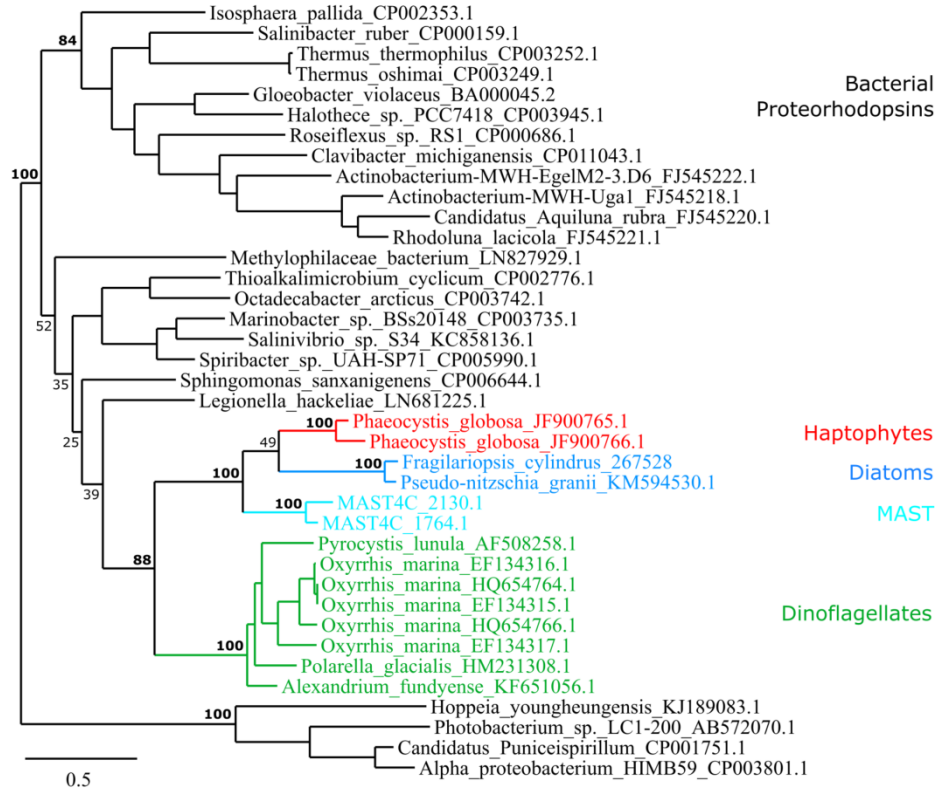


Supplementary Figure 4. Dynein heavy chains encoded in the co-assembled genomes.

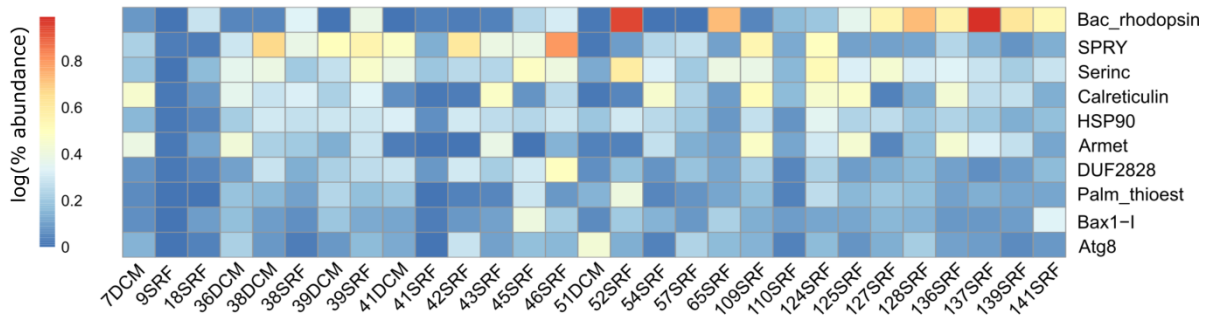
a. Number of Pfam domains associated with dynein heavy chain (DHC) protein genes per genome (grey bars) for the heterotrophic lineages and in genomes from a range of reference organisms including from marine environments. The estimated genome completion of each co-assembled genome calculated from BUSCO output is indicated by red bars. The MAST-3A genome has particularly few DHC domains. **b.** Canonical DHC gene structure. Nine Pfam domains were detected for DHCs. The ATP binding and hydrolysis occur in the D1 module (associated with pfam12774). **c.** Genomic structure in MAST-3A of genes encoding proteins

with similarities to the DHC-associated Pfam. Grey boxes indicate genes predicted by GAZE. The colored boxes indicate the domain found by CDD search analysis after a six-frame scaffold sequence translation. Broken boxes indicate truncated domains. All regions contain incomplete genes, and remnant sequences are indicative of pseudogenization. **d.** Heatmap of Pfam domains associated with dynein heavy chains in stramenopiles. The number of genes per organism with all Pfam domains is indicated next to the name of the organism. Hierarchical clustering was performed between organisms. **e.** Alignment of residue involved in ATP binding. For each organism, the protein with the highest conservation of the canonical residues was selected and used for alignment. Red indicates the residue conservation in the corresponding protein. **f.** Conserved canonical signature of the DHC D1 module (conserved residues are implied in ATP binding) in all proteins with similarities to the Pfam12774 motif in SAG organisms (upper panel). Sequence of the same residues in the only protein from MAST-3A with similarities to the Pfam12774 domain (lower panel). Nature and position of the residues shown in the structural models (**g-i**) are indicated in the middle. **g.** Three-dimensional structure of residues contacting ATP (indicated in orange) with a residue that corresponds to the canonical signature. **h.** Three-dimensional structure of the residue found in a MAST-3A protein with Pfam12774 signature. **i.** Comparison between the canonical residue (red) and MAST-3A residue (green). Modified residues are more distant from ATP, which indicates decreased affinity.

a



b



c

```

MAST-4C_1764  MMASTIFYWMMVSNVKPRYSALTITGLVTFIAAYHYFRIFNSWVEAYRYPVPGGSSKTTIGNPELTGKPFNDAYRYMDWLLTVPLLLIEIIFVMDLKPE
MAST-4C_2130  MMASTIFYWMMVSNVKPKFRSALTITGLVTFIAAYHYFRIFNSWVEAYRYPVPGGSSKTTIGNPELTGKPFNDAYRYMDWLLTVPLLLIEIVFMELKPE
*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****

MAST-4C_1764  ETSKAWQLGASAALMIIILGYPGELILEADKLSRWVYWALAMIPFLFVVYTLVGLAGATRNETPEVASAIRYAQMWTVLSWCTYPIVYIIPMFGAKGS
MAST-4C_2130  ETSKAWQLGASAALMIIILGYPGELILEADKLSRWVYWALAMIPFLFVVYTLVGLAGALRDESPEIASSIRTAQMWTVISWCTYPIVYIIPMFGAKGA
*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****

MAST-4C_1764  NAVVGIQVGYCIADVISKCGVGFVIYINITARKSAQSSDKDGYNPIQN
MAST-4C_2130  NAVVGIQLGYCIADVISKCGVGFIIYINITAKKSAL-TDKDGYRAVQ-
*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****.*****

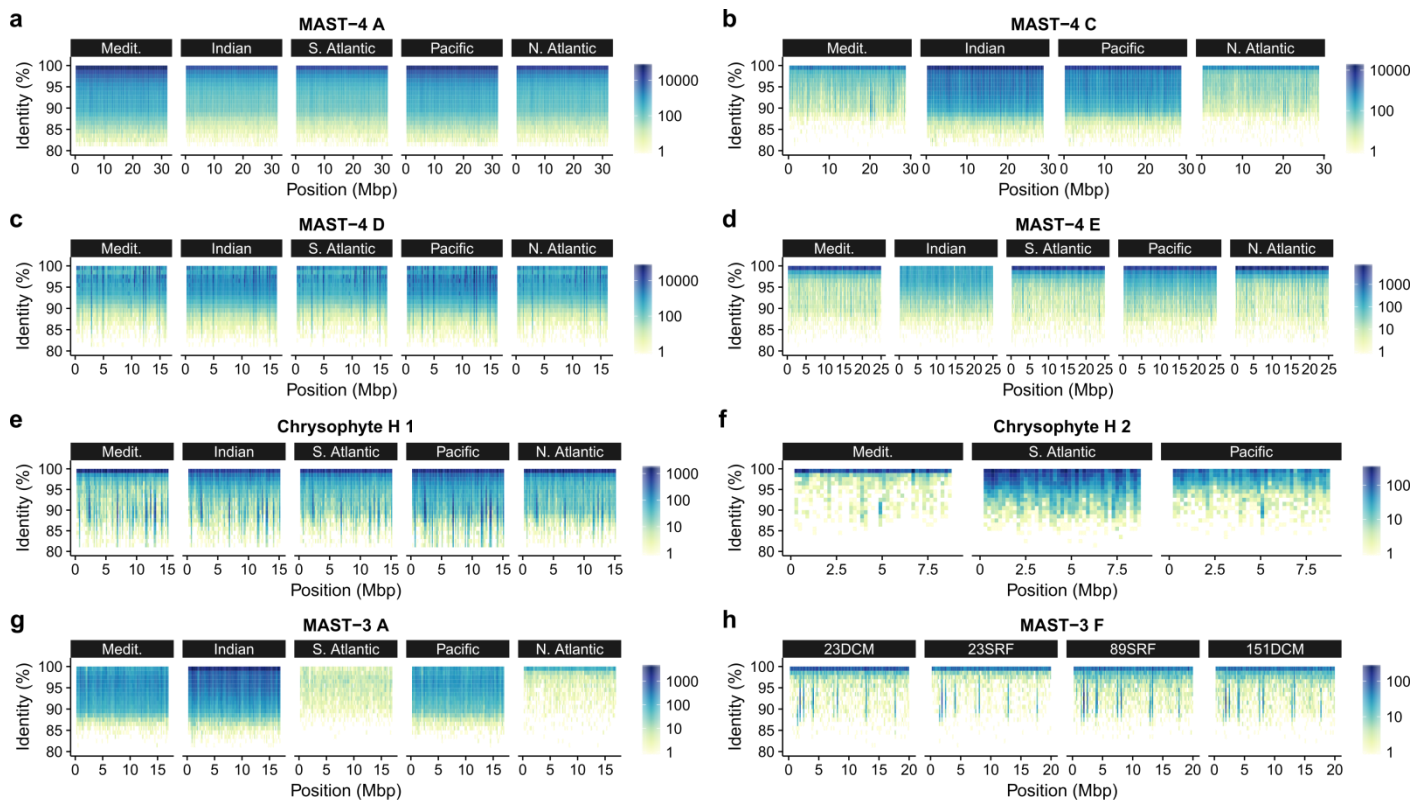
```

Supplementary Figure 5. The proteorhodopsin gene candidates in MAST-4C are typical of eukaryotic sequences and represent some of the most expressed transcripts in different environments.

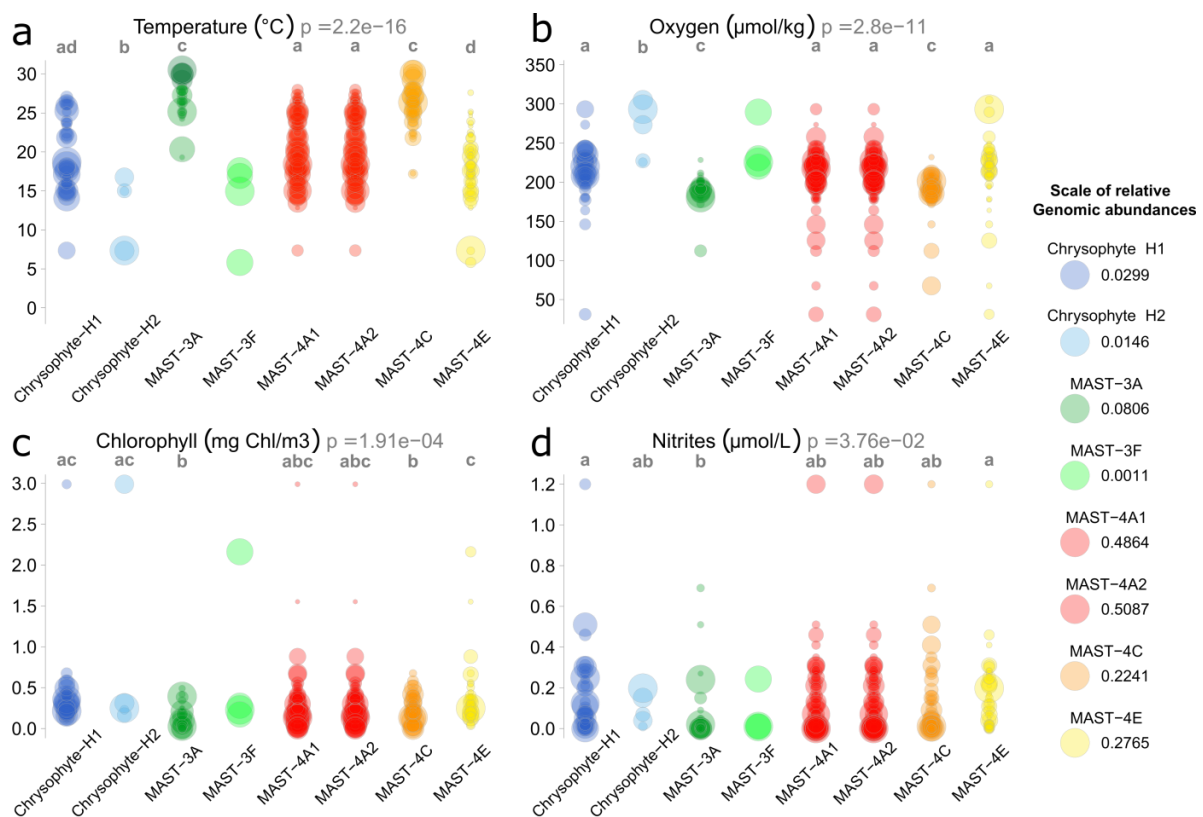
a. Phylogenetic tree of proteorhodopsins based on the tBLASTN best hits of the MAST4C_1764.1 gene against the nr-nuc database. The two MAST-4C and *Phaeocystis globosa* proteorhodopsins were added manually. Proteorhodopsin sequences were aligned with MUSCLE 3.7 and the maximum likelihood tree was constructed with 100 bootstraps. Bootstrap values of important nodes are reported and bootstrap values > 80 are in bold. **b.** Heatmap of the log relative metatranscriptomic RPKM values (in log percentages) for the 10 most expressed – on average – Pfam domains in the *Tara* Oceans samples of the 0.8-5 μ m size fraction where more than 50% of MAST-4 C genes are expressed. The proteorhodopsin candidates (two genes) constitute the most expressed category in these samples. **c.** MUSCLE alignment of the two MAST-4 C proteorhodopsins. Residues implicated in the proton-pump function are colored in red.



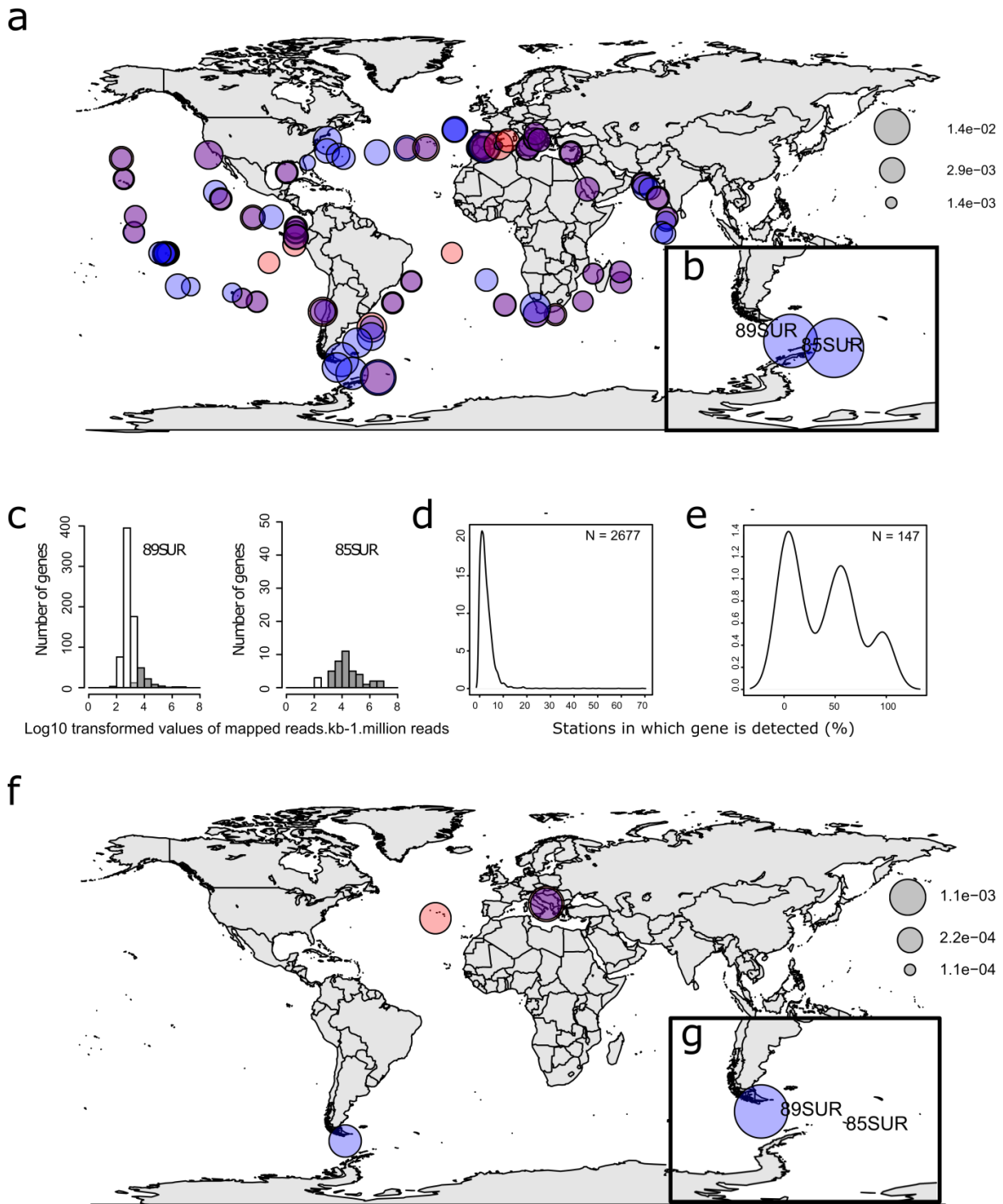
Supplementary Figure 6. Relative proportions and functions of putative Horizontally Transferred Genes. **a.** Fraction of four reference genomes annotated according to different COG categories (blue: metabolism; red: cellular processes and signaling; green: information storage and processing; purple: poorly characterized). **b.** Horizontally transferred genes annotated according to different COG categories. COG categories linked to metabolism (blue) that are significantly over-represented (one sided Chi-squared test p -value < 0.05) in the MAST and chrysoyhte genomes compared to bacterial genomes are annotated with a star (*).



Supplementary Figure 7. Fragment recruitment plots with *Tara* Oceans metagenomic samples grouped by ocean regions. The x-axis corresponds to the matching positions of metagenomics reads on arbitrarily concatenated scaffolds, and the y-axis shows the identity level of alignments. The 2D space is binned (200 kbp on the x-axis and 1% identity on the y-axis) to improve readability. Colour scale represents the density of reads recruited by bins. Bins that contained highly conserved genes were removed from this representation. For the same SAG lineage, the centre of identity distributions may vary among regions (for MAST-3A, MAST-4E, and Chrysophyte H2), or be relatively constant (for MAST-4C, Chrysophyte H1, MAST-4A, and MAST-3F). The mean identity value reflects the genetic distance of the assemblies from the genetically closest abundant genome in the considered region; this is more precisely reported for each metagenomic sample on a world map in Figure 2.



Supplementary Figure 8. Abundance plots based on several contextual parameters measured at each sampling site. Each parameter is indicated above the corresponding graph. The x-axis corresponds to each lineage, and the y-axis shows the value of the parameter (unit indicated above each figure). Circle size corresponds to relative co-assembled genome abundance at a particular station/depth measured by metagenomic read mapping. This is normalized on the figure, and each correspondence between circle size and relative abundance is indicated on the right. Kruskal-Wallis probabilities are indicated in grey close to parameter names. Lineage statistical classes were computed on parameters (when Kruskal-Wallis probabilities $< 10^{-2}$). Temperature is the highest discriminant parameter; groupings included MAST-4A1 and MAST-4A2 (class a), and MAST-3A and MAST-4C (class c), but MAST-4E1 was an independent class (class d). Chrysophyte H1 (class ad) is present in a range of temperatures that corresponds to both classes a and d of MAST-4A and MAST-4E. Chrysophyte H2 is independently classified (class b).



Supplementary Figure 9. Importance of detecting specifically-matching genes to study heterotrophic protist biogeography. **a.** Relative abundance of MAST-3F calculated from the initial whole gene dataset (cleaned for organelle scaffolds). Circle size is proportional to the relative organismal abundance (scale indicated by the grey circles). Blue circles indicate surface stations and red circles indicate DCM stations. This organism was apparently ubiquitously distributed. These results were inconsistent with distributions observed by V9

sequencing and may be evidence of cross-mapping. **b.** Close-up of relative MAST-3F abundances from stations 85SUR and 89SUR. **c.** Histogram plot of metagenomic RPKM values of inlier (white bar) and outlier (grey bar) gene dataset (x-axis : log₁₀-transformed values of metagenomic RPKM, y-axis: number of genes). **d.** Inlier and **e.** Outlier datasets of occurrence of genes per station was calculated for each SAG lineage. The plot represents the density of genes detected (metagenomic RPKM >0) in a specific percentage of stations (x-axis: percentage of station positive for the gene). In this example, the MAST-3F genes were detected in 3 to 5% of the *Tara* Oceans stations (**d**). However, genes from the outlier data set showed a different occurrence pattern; some were detected in all stations or about 50% of the stations. The MAST-3F distribution shown with the inlier dataset is more discrete, with two major locations, in Mediterranean Sea (where MAST-3F was sampled) and in South Atlantic Ocean (**f**). **g.** The use of the inlier dataset showed that the entire signal in station 85 was due to non-specific matches.

pfam00069	2.99	3.01	2.97	3.77	3.54	3.24	2.99	3.5	3.42	3.06	Protein kinase domain
pfam12796	3.98	1.29	1.49	0.63	1.42	3.41	3.18	1.3	1.65	4.06	Ankyrin repeats (3 copies)
pfam00225	0.98	1.56	1.13	0.77	1.87	1.31	1.21	1.18	1.49	0.95	Kinesin motor domain
pfam00271	0.75	1.21	1.34	1.13	1.46	0.65	0.83	0.74	0.78	0.47	Helicase conserved C-terminal domain
pfam13499	0.4	0.63	0.5	0.17	0.42	0.84	0.72	0.79	0.89	1.55	EF-hand domain pair
pfam00071	0.96	0.55	0.85	0.77	0.83	0.44	0.5	0.66	0.84	0.53	Ras family
pfam00270	0.53	0.94	0.78	0.9	1.29	0.44	0.53	0.49	0.63	0.45	DEAD/DEAH box helicase
pfam00226	0.47	0.51	0.35	0.57	0.54	0.58	0.5	0.39	0.57	0.68	DnaJ domain
pfam03028	0.19	0.55	0.71	0.17	0.5	0.54	0.42	0.71	0.71	0.47	Dynein heavy chain and region D6 of dynein motor
pfam00443	0.37	0.39	0.64	0.47	0.46	0.3	0.4	0.47	0.5	0.53	Ubiquitin carboxy-terminal hydrolase
pfam00415	0.31	0.23	0.25	0.2	0.42	0.42	0.5	0.39	0.5	0.81	Regulator of chromosome condensation (RCC1) repeat
pfam00076	0.53	0.43	0.14	0.47	0.5	0.33	0.31	0.37	0.37	0.47	RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)
pfam00632	0.2	0.27	0.28	0.37	0.63	0.42	0.42	0.42	0.42	0.24	HECT-domain (ubiquitin-transferase)
pfam00112	0.09	0.27	0.59	0.43	0.08	0.35	0.42	0.54	0.47	0.26	Papain family cysteine protease
pfam00894	0.1	0.2	0.25	0.04	0.49	0.7	0.42	0.47	0.32	0.32	Sulfatase
pfam12781	0.18	0.43	0.42	0.42	0.35	0.18	0.3	0.39	0.37	0.37	ATP-binding dynein motor region D5
pfam00027	0.21	0.43	0.35	0.13	0.29	0.35	0.31	0.52	0.6	0.92	Cyclic nucleotide-binding domain
pfam00085	0.43	0.31	0.21	0.33	0.12	0.4	0.35	0.27	0.26	0.45	Thioredoxin
pfam00378	0.13	0.31	0.14	0.4	0.08	0.65	0.44	0.44	0.26	0.32	Enoyl-CoA hydratase/isomerase family
pfam12774	0.19	0.31	0.28	0.03	0.37	0.14	0.15	0.32	0.44	0.32	Hydrolytic ATP binding site of dynein motor region D1
pfam12777	0.2	0.31	0.35	0.03	0.37	0.26	0.22	0.27	0.37	0.42	Microtubule-binding stalk of dynein motor
pfam00063	0.26	0.39	0.28	0.57	0.29	0.3	0.2	0.42	0.21	0.11	Myosin head (motor domain)
pfam12780	0.18	0.43	0.28	0.42	0.28	0.18	0.2	0.34	0.29	0.29	P-loop containing dynein motor region D4
pfam12775	0.18	0.35	0.28	0.03	0.33	0.23	0.18	0.27	0.42	0.26	P-loop containing dynein motor region D3
pfam03133	0.21	0.31	0.21	0.17	0.42	0.21	0.31	0.49	0.42	0.21	Tubulin-tyrosine ligase family
pfam00171	0.08	0.23	0.28	0.37	0.29	0.26	0.22	0.27	0.26	0.11	Aldehyde dehydrogenase family
pfam08393	0.2	0.23	0.28	0.37	0.16	0.13	0.3	0.42	0.32	0.32	Dynein heavy chain, N-terminal region 2
pfam00240	0.19	0.2	0.35	0.1	0.12	0.56	0.35	0.05	0.29	0.37	Ubiquitin family
pfam07714	0.57	0.12	0.21	0.23	0.08	0.3	0.33	0.1	0.13	0.29	Protein tyrosine kinase
pfam13561	0.14	0.04	0.3	0.08	0.42	0.4	0.47	0.31	0.11	0.11	Enoyl-(Acyl carrier protein) reductase
pfam00520	0.13	0.27	0.35	0.07	0.08	0.21	0.2	0.22	0.44	0.44	Ion transport protein
pfam00648	0.09	0.27	0.28	0.17	0.12	0.16	0.22	0.32	0.16	0.26	Calpain family cysteine protease
pfam13424	0.78	0.23	0.14	0.37	0.33	0.3	0.1	0.08	0.45	0.45	Tetratricopeptide repeat
pfam06602	0.04	0.16	0.21	0.17	0.21	0.23	0.18	0.22	0.16	0.05	Myotubularin-like phosphatase domain
pfam00107	0.05	0.04	0.14	0.13	0.04	0.33	0.35	0.2	0.21	0.18	Zinc-binding dehydrogenase
pfam01926	0.22	0.35	0.28	0.17	0.5	0.14	0.15	0.1	0.16	0.08	50S ribosome-binding GTPase
pfam01363	0.22	0.12	0.21	0.23	0.04	0.16	0.13	0.15	0.13	0.32	FYVE zinc finger
pfam14580	0.15	0.12	0.07	0.13	0.12	0.16	0.13	0.27	0.21	0.26	Leucine-rich repeat
pfam00628	0.24	0.04	0.14	0.07	0.17	0.12	0.13	0.12	0.24	0.18	PHD-finger
pfam13855	0.8	0.27	0.21	0.07	0.12	0.14	0.13	0.07	0.13	0.32	Leucine rich repeat
pfam00135	0.03	0.27	0.21	0.13	0.04	0.19	0.15	0.12	0.08	0.08	Carboxylesterase family
pfam00933	0.01	0.17	0.17	0.12	0.13	0.17	0.21	0.16	0.16	0.16	Glycosyl hydrolase family 3 N terminal domain
pfam08016	0.1	0.14	0.15	0.14	0.15	0.17	0.18	0.15	0.18	0.5	Polycystin cation channel
pfam00118	0.13	0.16	0.42	0.13	0.25	0.12	0.02	0.02	0.1	0.08	TCP-1/cpn60 chaperonin family
pfam00026	0.06	0.08	0.21	0.13	0.04	0.07	0.11	0.12	0.18	0.21	Eukaryotic aspartyl protease
pfam00295	0.04	0.1	0.04	0.16	0.2	0.17	0.24	0.13	0.24	0.13	Glycosyl hydrolases family 28
pfam00169	0.19	0.16	0.1	0.04	0.09	0.04	0.12	0.21	0.21	0.32	PH domain
pfam13637	0.13	0.04	0.21	0.35	0.29	0.1	0.11	0.11	0.11	0.11	Ankyrin repeats (many copies)
pfam03060	0.01	0.04	0.07	0.13	0.19	0.31	0.25	0.16	0.03	0.03	Nitronate monooxygenase
pfam00083	0.18	0.2	0.35	0.07	0.12	0.05	0.04	0.02	0.08	0.21	Sugar (and other) transporter
pfam08385	0.08	0.12	0.07	0.03	0.08	0.12	0.02	0.12	0.18	0.18	Dynein heavy chain, N-terminal region 1
pfam01408	0.09	0.08	0.1	0.04	0.14	0.13	0.12	0.13	0.03	0.03	Oxidoreductase family, NAD-binding Rossmann fold
pfam13540	0.07	0.04	0.12	0.15	0.12	0.18	0.12	0.18	0.16	0.16	Regulator of chromosome condensation (RCC1) repeat
pfam13833	0.09	0.07	0.03	0.04	0.09	0.18	0.15	0.16	0.16	0.16	EF-hand domain pair
pfam03016	0.12	0.12	0.07	0.2	0.05	0.07	0.07	0.13	0.11	0.11	Exostosin family
pfam01915	0.01	0.17	0.12	0.13	0.07	0.21	0.07	0.21	0.11	0.11	Glycosyl hydrolase family 3 C-terminal domain
pfam00211	0.05	0.1	0.12	0.37	0.35	0.1	0.08	0.08	0.08	0.08	Adenylyate and Guanylyate cyclase catalytic domain
pfam01485	0.02	0.16	0.07	0.03	0.08	0.09	0.09	0.07	0.1	0.11	IBR domain
pfam02515	0.04	0.17	0.04	0.44	0.37	0.15	0.08	0.08	0.08	0.08	CoA-transferase family III
pfam12237	0.02	0.07	0.07	0.04	0.09	0.11	0.12	0.1	0.1	0.08	Phosphorylated CTD interacting factor 1 WW domain
pfam00728	0.08	0.07	0.03	0.07	0.15	0.17	0.24	0.08	0.08	0.08	Glycosyl hydrolase family 20, catalytic domain
pfam02065	0.08	0.04	0.07	0.1	0.07	0.09	0.07	0.16	0.13	0.13	Methylase
pfam00583	0.01	0.03	0.19	0.15	0.2	0.21	0.11	0.2	0.21	0.11	Acetyltransferase (GNAT) family
pfam09286	0.01	0.1	0.3	0.15	0.07	0.16	0.05	0.16	0.05	0.05	Pro-kumamolisin, activation domain
pfam00144	0.01	0.04	0.04	0.04	0.26	0.15	0.07	0.1	0.18	0.18	Beta-lactamase
pfam01391	0.03	0.04	0.07	0.07	0.07	0.05	0.08	0.18	0.18	0.18	Collagen triple helix repeat (20 copies)
pfam11527	0.17	0.23	0.21	0.03	0.02	0.07	0.05	0.03	0.05	0.03	The ARF-like 2 binding protein BART
pfam03151	0.05	0.08	0.14	0.07	0.04	0.05	0.04	0.02	0.05	0.03	Triose-phosphate transporter family
pfam01434	0.05	0.04	0.14	0.07	0.29	0.05	0.05	0.05	0.08	0.03	Peptidase family M41
pfam11028	0.02	0.04	0.04	0.04	0.51	0.46	0.25	0.18	0.05	0.05	Protein of unknown function (DUF2723)
pfam00066	0.02	0.04	0.03	0.12	0.13	0.07	0.13	0.05	0.05	0.05	LNR domain
pfam01074	0.01	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	Glycosyl hydrolases family 38 N-terminal domain
pfam03382	0.08	0.16	0.21	0.57	0.08	0.05	0.04	0.02	0.03	0.03	Mycoplasma protein of unknown function, DUF285
pfam00488	0.08	0.08	0.07	0.03	0.12	0.04	0.02	0.03	0.03	0.03	MutS domain V
pfam00620	0.03	0.03	0.03	0.04	0.12	0.07	0.02	0.13	0.11	0.11	RhoGAP domain
pfam01764	0.22	0.04	0.07	0.03	0.07	0.07	0.02	0.03	0.03	0.03	Lipase (class 3)
pfam13088	0.04	0.04	0.13	0.02	0.02	0.05	0.08	0.08	0.08	0.08	BNR repeat-like domain
pfam01833	0.02	0.04	0.12	0.2	0.2	0.13	0.18	0.13	0.18	0.18	IP/TIG domain
pfam05592	0.07	0.08	0.03	0.14	0.11	0.11	0.05	0.03	0.03	0.03	Bacterial alpha-L-rhamnosidase
pfam12848	0.02	0.08	0.03	0.07	0.07	0.07	0.05	0.03	0.03	0.03	ABC transporter

Supplementary Table 1. Most abundant Pfam domains annotated in the SAG co-assembled genomes and in the *Ectocarpus siliculosus* genome (as an example of photosynthetic stramenopile) and MAST-4D. Red and orange indicate the most abundant Pfam domains, and grey indicates complete absence of the domain in the annotated proteins. The values represent percentages of total Pfam domains.

	GH (number)	GH (% of gene models)	Potentially secreted	Putative algal cell wall degrading enzymes	Associated algal DNA (number of cells positive / total number of cells)
MAST-4A1	91	1.1%	39	8	Micromonas (1/4)
MAST-4A2	101	1.1%	44	10	Micromonas, Bathycoccus (1/5)
MAST-4C	73	1.3%	38	7	none
MAST-4E	98	2.1%	67	9	Pelagomonas (1/9)
MAST-3A	76	2.3%	33	3	none
MAST-3F	13	0.5%	2	1	none
ChrysophyteH1	23	0.8%	12	1	none
ChrysophyteH2	7	0.4%	3	0	none

Supplementary Table 2. Distribution of glycoside hydrolases (GHs) and algal DNA in the SAG lineages.

	Category 1	Category 2	Category 3	False Positives	Ambiguous	Total
MAST-3A	6	1	10	3	2	22
MAST-3F	1	1	1	1	0	4
MAST-4 A1	32	15	27	17	7	98
MAST-4 A2	25	22	23	25	10	105
MAST-4C	4	16	4	11	3	38
MAST-4E	2	10	12	7	0	31
Chrysophyte-H1	5	2	15	13	2	37
Chrysophyte-H2	2	2	5	1	0	10
Total	77 (22.3%)	69 (20.0%)	97 (28.1%)	78 (18.6%)	24 (6.4%)	345 (100%)

Supplementary Table 3. Summary of the validation of putative horizontal gene transfers using a tree-based method. Putative HGTs were classified in three categories, from the most recent events (category 1) to the less recent ones (category 3). Candidate HGTs that were branching with other eukaryotic proteins were considered as False Positives. Ambiguous cases result from poor alignments or insufficient matches.

	Antibiotic and stress resistance	Carbohydrate metabolism	Lipid metabolism	Nitrogen containing molecules	Proteolytic enzymes (peptidase/protease)	Transport
MAST-4A1	11	17	3	10	7	2
MAST-4A2	8	11	0	6	5	3
MAST-4C	6	7	0	3	0	2
MAST-4E	2	9	0	1	0	1
MAST-3A	1	2	4	2	1	0
MAST-3F	0	1	0	0	0	0
ChrysophyteH1	3	2	0	2	0	1
ChrysophyteH2	1	0	1	0	1	0

Supplementary Table 4. Main categories of HGT enzymes involved in metabolism.

Name	Scientific Name	Taxon ID	Accession numbers
MAST-4 A 1	Stramenopiles sp. TOSAG23-1	1735742	ERR1198936 ERR1198938 ERR1198948 ERR1198949 ERR1198925 ERR1198954
MAST-4 A 2	Stramenopiles sp. TOSAG23-2	1735743	ERR1138643 ERR1138644 ERR1138645 ERR1138646
MAST-4 C	Stramenopiles sp. TOSAG41-1	1735744	ERR1198926 ERR1198940 ERR1198945 ERR1198955
MAST-4 E	Stramenopiles sp. TOSAG23-3	1735745	ERR1189844 ERR1189846 ERR1189847 ERR1189854 ERR1198927 ERR1198928 ERR1198941 ERR1198946 ERR1198950
MAST-3 A	Stramenopiles sp. TOSAG41-2	1735746	ERR1198931 ERR1198953 ERR1198930 ERR1198943
MAST-3 F	Stramenopiles sp. TOSAG23-6	1735747	ERR1189848 ERR1189852
Chrysophyte H 1	Chrysophyceae sp. TOSAG23-4	1735748	ERR1189849 ERR1189855 ERR1198924 ERR1198933 ERR1198934 ERR1198937 ERR1198951 ERR1198956
Chrysophyte H 2	Chrysophyceae sp. TOSAG23-5	1735749	ERR1198929 ERR1198935 ERR1198944

Supplementary Table 5. Scientific name, taxon ID and accession numbers at the European


Nucleotide Archive of each Single-cell Amplified Genome.

Genome	Number of models	Calibrated on
ChrysophyteH1	428	Chrysophyte-H1
ChrysophyteH2	243	Chrysophyte-H2
MAST-3A	111	MAST-3A
MAST-3F	111	MAST-3A
MAST-4A1	370	MAST-4A1
MAST-4A2	393	MAST-4A2
MAST-4C	226	MAST-4C
MAST-4E	188	MAST-4E

Supplementary Table 6. Number of complete models used to train SNAP and source of calibration.

Annexe 2. Survey of the green picoalga Bathycoccus genomes in the global ocean

SCIENTIFIC REPORTS



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 Aury¹, Colomban de Vargas⁴, Michael Sieracki⁵, Daniele Iudicone⁶, Daniel Vaulot⁴, 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 production¹. Picocyanobacteria of the genera *Prochlorococcus* and *Synechococcus* dominate the prokaryotic component². However, small eukaryotes (picoeukaryotes; <2 μm) 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 groups⁵. Among them, three genera of green algae belonging to the order Mamiellales (class Mamiellophyceae⁶), *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 μm¹⁵ can be differentiated into at least four clades. Two *Ostreococcus* species have been formerly described: *O. tauri* and *O. mediterraneus*^{15,16}. Among these *Ostreococcus* clades, different strains seem to be adapted to different light ranges¹⁷. However, the ecological preferences of *Ostreococcus* strains are probably more complex, implying other environmental parameters such as nutrients and temperature⁹.

The genus *Bathycoccus* was initially isolated at 100 m from the deep chlorophyll maximum (DCM) in the Mediterranean Sea¹⁸ and cells with the same morphology (body scales) had been reported previously from the Atlantic Ocean¹⁹. *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 available²². Metagenomic data have suggested the existence of two *Bathycoccus* ecotypes^{10,11,23}, recently named B1 and B2¹¹. These two ecotypes have

¹CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. ²CNRS, UMR 8030, CP5706 Evry, France. ³Université d'Evry, UMR 8030, CP5706 Evry, France. ⁴Sorbonne Universités, UPMC Université Paris 06, CNRS, UMR7144, Station Biologique de Roscoff, 29680 Roscoff, France. ⁵National Science Foundation, 4201 Wilson Blvd., Arlington, VA 22230, USA. ⁶Stazione 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. ¹The 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. ²A + B + C + D corresponds to a non-redundant merging of contigs from individual assemblies. ³Combined 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 genes¹⁰. 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 metabarcodes²⁷. 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 bacteria³⁰ 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 set³³. 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 regions²⁷. During this expedition a large number of metagenomic data and single-cell amplified genomes (SAGs³⁷) 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 RCC1105²² 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 genes³⁹. 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

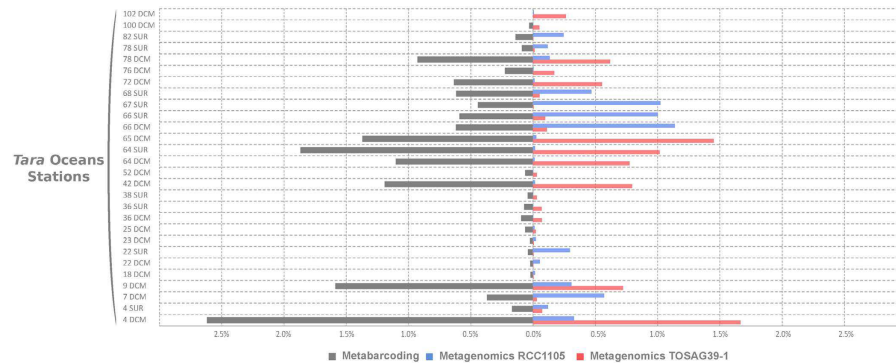


Figure 1. Comparisons of relative abundances of *Bathycoccus* in the 0.8–5 μm size fraction samples from Tara Oceans stations. Left: relative 18S rRNA V9 amplicons abundance (percent of reads). Right: relative metagenomic abundances (percent of metagenomic reads) from direct mapping of metagenomic reads 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 DCM⁴⁰. We also looked at the plastid 16S marker gene⁴¹ and to the PRP8 intein gene that has been proposed as markers for *Bathycoccus*¹⁰. The plastid 16S 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 database⁴² (Supplementary Figures S5). Metagenomes T142 and T149 from the South East Pacific²³ and transcriptome MMETSP1399 (strain CCMP1898, which is the type strain for *Bathycoccus prasinos*) correspond, or are closely related to RCC1105. The tropical Atlantic Ocean metagenome⁴⁰ and transcriptome MMETSP1460 (strain RCC716 from the Indian Ocean) correspond, or are closely related to TOSAG39-1. Direct amino acid BLAST⁴³ 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 gene²⁷ 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.

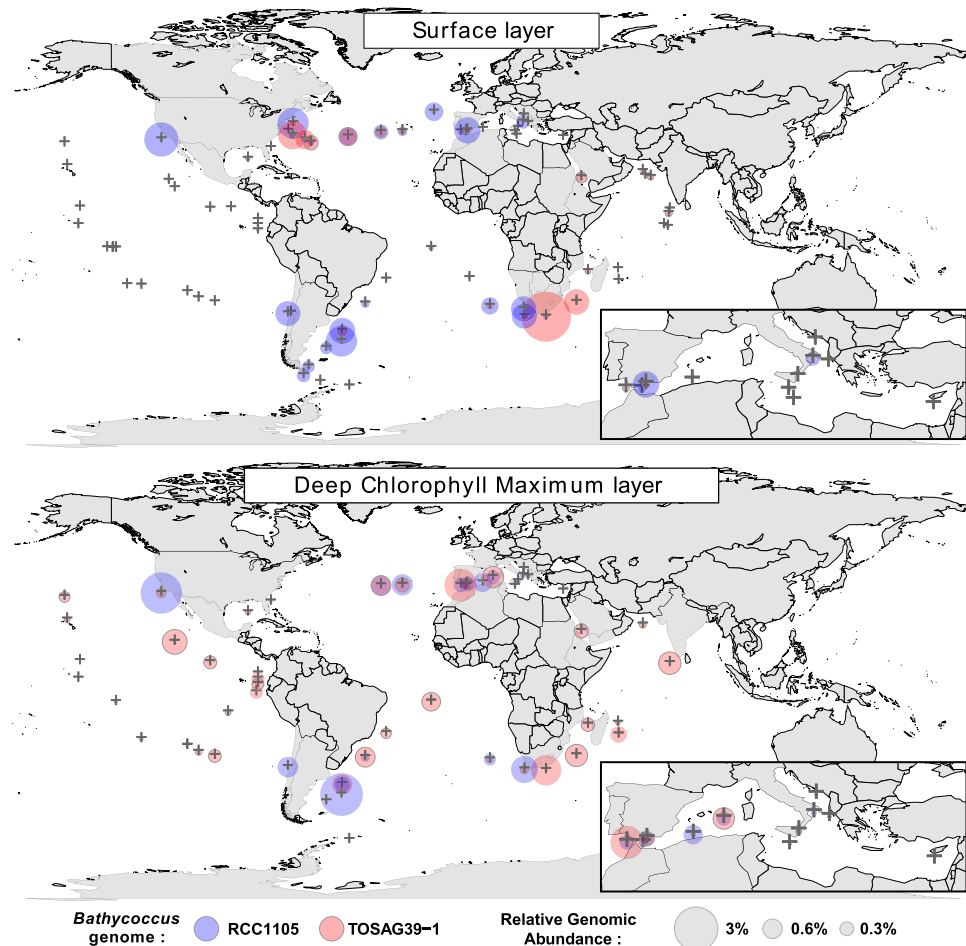


Figure 2. Geographical distribution of two *Bathycoccus* genomes, RCC1105 and TOSAG39-1, along *Tara* Oceans expedition stations from recruitments of metagenomic reads. Top and bottom maps correspond to the surface and deep chlorophyll maximum (DCM) samples respectively. Gray crosses indicate *Tara* 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 iron⁴⁴, *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 current⁴⁵.

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).

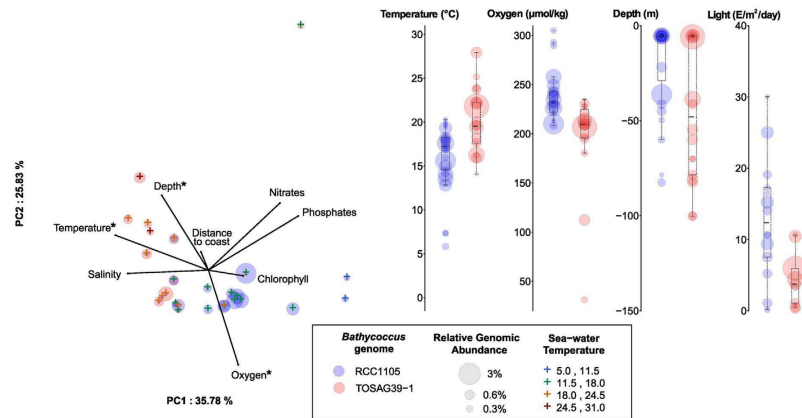


Figure 3. Relationships between environmental parameters and *Bathycoccus* genome abundance.

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 O₂ was low (31.2 µM, Fig. 3, Supplementary Figures S9 and S10), though no samples originated from anoxic waters⁴⁶.

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 concentrations⁴⁷ (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 populations⁵⁰. 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 19²². 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 of 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

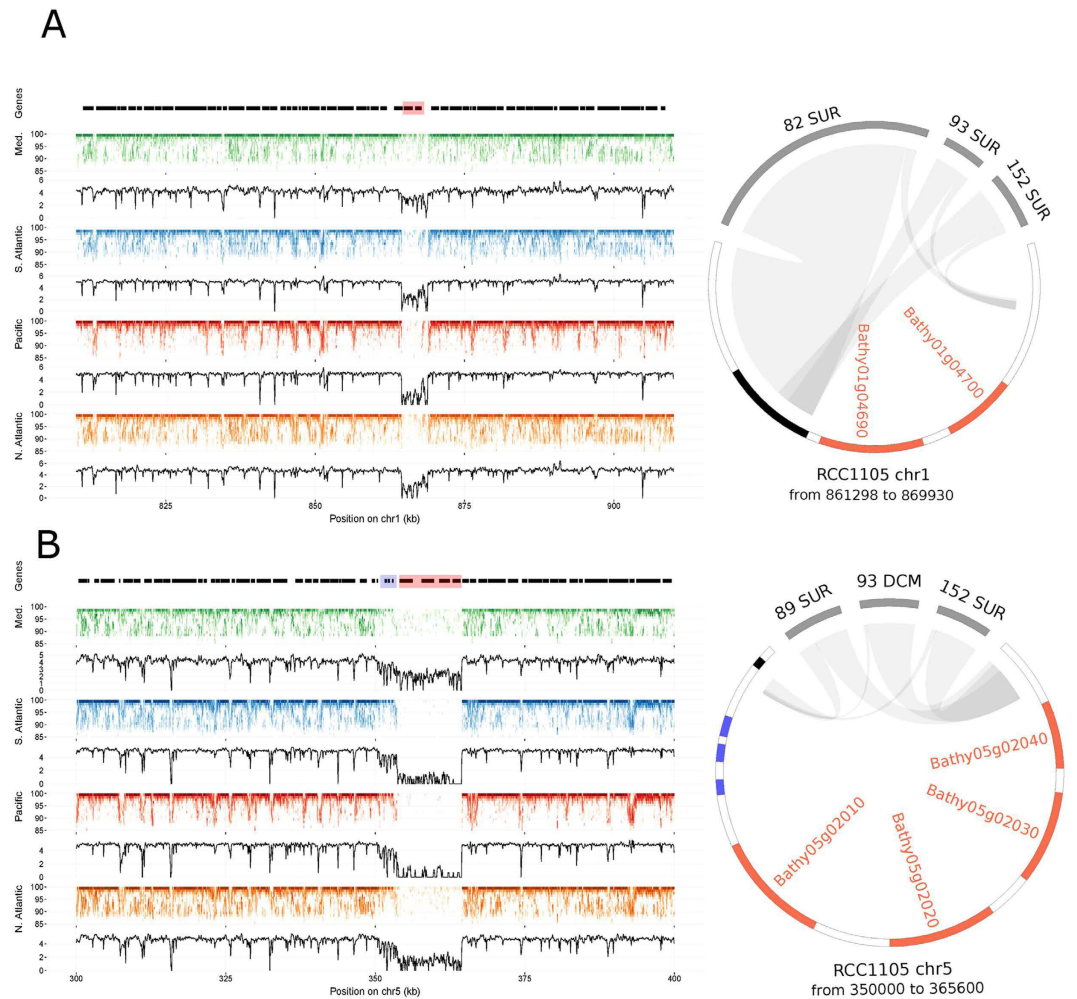


Figure 4. Evidence for cassettes of dispensable genes in *B. prasinos* RCC1105. Left and right sides of the figures represent fragment recruitment and genomic alignments of dispensable gene cassettes, respectively. Fragment recruitments plots are displayed by marine zones (left legend). Each dot corresponds 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 graphs⁷⁹ 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-occurrence 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 assemblies⁵¹. 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 recently¹¹. 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 clades⁵².

A previous study¹¹ 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 identity⁵³) 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 intercrossing⁵⁶ 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* clade⁵⁷, or the haptophyte species *Emiliania huxleyi* and *Gephyrocapsa oceanica*, which also have identical 18S rRNA gene sequences, but quite different morphologies⁵⁸.

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 genome⁶¹ and has been proposed, in some cases, to result from horizontal transfer. In *Prochlorococcus*, genomic islands are thought to be related to niche adaptation⁶³ because they host ecologically important genes⁶⁰. A comparison of two *Prochlorococcus* ecotypes revealed that differences in gene content were related to high-light vs. low-light adaptation⁶⁴. Such adaptations have been hypothesized in species closely related to *Bathycoccus*, like *Ostreococcus*¹⁷, but are still a matter of debate⁹. 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 O₂ concentrations at Costa Rica Dome station 138, an area of high biological production in the East equatorial Pacific⁶⁶ where picoplankton can be very abundant⁶⁷. 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 O₂.

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 lines⁶⁸. 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 properties⁷¹. 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 reference³⁶). 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 sequencing³⁷. 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 method⁷² 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 genome²² 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 databank⁷⁶ and the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) collection of marine protist transcriptomes⁴². In addition, we performed direct *ab initio* prediction by calibrating and running the Markov model implemented in snap⁷⁷. Integrating and combining all this evidence provided a final set of genes, using a process based on Gmorse software rationale⁷⁸. 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 reads²⁴. 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

- 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).
- Scanlan, D. J. *et al.* Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.* **73**, 249–299 (2009).
- 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).
- Wilkins, D. *et al.* Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics. *Environ. Microbiol.* **15**, 1318–1333 (2013).
- Vaulot, D., Eikrem, W., Viprey, M. & Moreau, H. The diversity of small eukaryotic phytoplankton ($\leq 3 \mu\text{m}$) in marine ecosystems. *FEMS Microbiol. Rev.* **32**, 795–820 (2008).
- 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).
- 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).
- Lovejoy, C. *et al.* Distribution, phylogeny, and growth of cold-adapted picoprasinophytes in arctic seas. *J. Phycol.* **43**, 78–89 (2007).
- Demir-Hilton, E. *et al.* Global distribution patterns of distinct clades of the photosynthetic picoeukaryote *Ostreococcus*. *ISME J.* **5**, 1095–1107 (2011).
- 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. & Thronsen, 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-Fabriz, S., Vault, 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. Vault, 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. Martínez-García, 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., Massignani, 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. Makalowski, 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. Ahlgrén, 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 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, 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.pangea.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/DBBL/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).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016


Annexe 3. A global ocean atlas of eukaryotic genes

ARTICLE

DOI: 10.1038/s41467-017-02342-1

OPEN

A global ocean atlas of eukaryotic genes

Quentin Carradec et al.[#] 

While our knowledge about the roles of microbes and viruses in the ocean has increased tremendously due to recent advances in genomics and metagenomics, research on marine microbial eukaryotes and zooplankton has benefited much less from these new technologies because of their larger genomes, their enormous diversity, and largely unexplored physiologies. Here, we use a metatranscriptomics approach to capture expressed genes in open ocean *Tara* Oceans stations across four organismal size fractions. The individual sequence reads cluster into 116 million unigenes representing the largest reference collection of eukaryotic transcripts from any single biome. The catalog is used to unveil functions expressed by eukaryotic marine plankton, and to assess their functional biogeography. Almost half of the sequences have no similarity with known proteins, and a great number belong to new gene families with a restricted distribution in the ocean. Overall, the resource provides the foundations for exploring the roles of marine eukaryotes in ocean ecology and biogeochemistry.

Correspondence and requests for materials should be addressed to E.P. (email: eric.pelletier@genoscope.cns.fr) or to C.B. (email: cbowler@biologie.ens.fr) or to P.W. (email: pwincker@genoscope.cns.fr). [#]A full list of authors and their affiliations appears at the end of the paper.

Single-celled microeukaryotes and small multicellular zooplankton account for most of the planktonic biomass in the world's ocean^{1,2}. They are involved in various processes that shape the biogeochemical cycles of the planet, from primary production, recycling of organic matter by predation and parasitism, sequestration of carbon to a depth, and the transfer of organic material to higher trophic levels in the food webs³. Yet, their analysis is confounded because they are represented by hundreds of thousands of different taxa belonging to almost all phylogenetic groups of eukaryotes⁴, and the vast majority of them cannot be cultured. Their highly variable genome sizes, spanning at least four orders of magnitude⁵, and the predominance of noncoding sequences are additional challenges that have impeded their genomic exploration. Consequently, their study has been limited principally to morphological description of diversity, as well as taxonomic and biogeographic characterizations using individual barcode genes^{6,7}. By contrast, global surveys of the functional potential of marine microbiota ($\leq 3 \mu\text{m}$) and double-stranded DNA viruses are advancing rapidly because of the availability of comprehensive gene catalogs^{8–12}, as has been performed for the human gut¹³. To help assess gene function in marine eukaryotes, transcriptome data sets from hundreds of cultured marine eukaryotes¹⁴ have been generated, as well as from some species of zooplankton¹⁵, which is helping to analyze features of the global eukaryotic proteome and to interpret the transcriptional responses of some components of eukaryotic communities to localized stimuli^{16,17}.

Herein, we use a metatranscriptomics approach using samples collected from the global ocean during the *Tara Oceans* expedition¹⁸ to generate a global ocean reference catalog of genes from planktonic eukaryotes and to explore their expression patterns with respect to biogeography and environmental conditions.

Results

The *Tara Oceans* catalog of expressed eukaryotic genes. To identify and characterize the transcriptionally active genes from the most abundant eukaryotic plankton in the global ocean, we selected samples collected during the *Tara Oceans* expedition at two main depths in the euphotic zone (subsurface (SRF) and deep chlorophyll maximum (DCM)), at 68 different geographic locations across all the major oceanic provinces except the Arctic¹⁹ (Fig. 1a). Four main organismal size fractions were sampled independently²⁰ to optimize the recovery of comprehensive metatranscriptomes from piconanoplanktonic, nanoplanktonic, microplanktonic, and mesoplanktonic communities, covering protists to zooplankton and fish larvae. High-coverage polyA-based (to avoid ribosomal, organellar, and bacterial RNA) RNA-Seq was performed on a total of 441 size-fractionated plankton communities (Fig. 1a), resulting in 16.5 terabases of raw data from which residual ribosomal RNA sequences were removed. The cDNA reads were individually assembled for each sample and then clustered together at 95% sequence identity to create a single, largely nonredundant resource of 116.8 million transcribed sequences of at least 150 bases in length, hereafter termed unigenes, with a N50 length of 635 bases. Rarefaction analysis revealed that, despite its magnitude, the sampling effort did not result in near saturation of the eukaryotic gene space, contrasting with the results obtained from the smallest prokaryote-enriched size fractions, analyzed by metagenomics from 243 *Tara Oceans* samples⁹ (Fig. 1b). We estimate that the unigene curve would reach saturation at 166–190 million sequences, if all ocean regions would be taxonomically homogeneous (Supplementary Data 1).

Annotation of the >116 million unigenes (Methods and Supplementary Fig. 1a) revealed that we could assign a taxonomy level (from “cellular organism” to species name) to only 48.3% of

the unigenes (Fig. 2a and Supplementary Fig. 1b). By mapping the unigenes onto known gene annotations from marine genomes, we found a mean value of 2.20 (s.d. = 0.47) unigenes per gene (Methods and Supplementary Data 2). We then estimated the number of distinct transcriptomes (originating from different species) that were present in the catalog by counting the mean number of copies of conserved ribosomal protein genes, which indicated that the catalog contains genes from 8823 (s.d. = 1532) different organisms (Supplementary Data 3). These values indicate that the unigenes are derived from around 53 (44–68) million genes, with a mean of 6014 (4226–9223) genes per sampled organism (Supplementary Data 4). All sequencing reads from the 441 samples, as well as the reads from a parallel metagenomics sequencing program, were mapped onto the unigenes to provide relative expression and abundance for each gene in every sample (Methods and Supplementary Fig. 1a).

With an equivalent sequencing effort, the complexity of the metatranscriptomes decreased from the smallest piconanoplanktonic communities to the largest, mesoplanktonic assemblages (Fig. 1c), matching the pattern observed in extensive rDNA metabarcoding data sets⁶. Rarefaction curves calculated individually per size fraction revealed the higher complexity of the piconano and nanoplankton communities (Fig. 1b), and we found that the 5–20 μm size fraction was the most gene rich, due to intersample dissimilarity and the presence of more gene-rich transcriptomes (Fig. 1b, c). All size fractions contained a significant number of genes not found in the others (8.7–29%; Supplementary Fig. 1c), indicating the importance of size fractionation to describe the global eukaryote gene content of the ocean. With the limitation that we are considering the most expressed genes in our samples rather than the total gene content, we observed that a breakdown of the rarefaction curve by oceanic provinces shows consistent richness and undersaturation of the gene space, with the notable exception of the Southern Ocean, and to a lesser extent of the Mediterranean Sea (Fig. 1b). A high-taxonomic level breakdown of the assignable unigenes across *Tara Oceans* stations and organismal size fractions shows a higher relative abundance of genes from photosynthetic protists in the piconano plankton, and their progressive replacement by metazoan transcripts in larger size fractions (Fig. 2b), confirming the efficiency of the fractionation-based approach. We observed 1.13% of unigenes that are affiliated to prokaryotes. These were not removed from the catalog, as they can be true nonpolyadenylated transcripts from this group, or alternatively to the low level of eukaryotic annotations with respect to prokaryotes in reference databases, or to horizontal gene transfers.

Our metatranscriptomic data also captured transcripts (or RNA genomes) of viruses actively infecting their eukaryotic hosts. Their activities were found to be pervasive across the geographic and organismal size ranges examined in this study. Of the taxonomically assignable unigenes, 33,870 (0.06%) were predicted to be of eukaryotic virus origin, the vast majority of which (86%) originated from nucleocytoplasmic large dsDNA viruses (NCLDVs)²¹ (Fig. 2c) likely due to the large number of genes encoded in these viruses. Eukaryotic viral unigenes were expressed (or present in the case of RNA viruses) in all 441 samples at a relative abundance ranging from 0.0006 to 0.4% (0.02% on average). NCLDV transcripts dominated the piconano-planktonic communities, while RNA virus sequences became dominant with increasing organism size (Supplementary Fig. 2).

Factors discriminating the most expressed functional classes. To investigate the functional structuring within eukaryotic plankton communities, we defined the main parameters

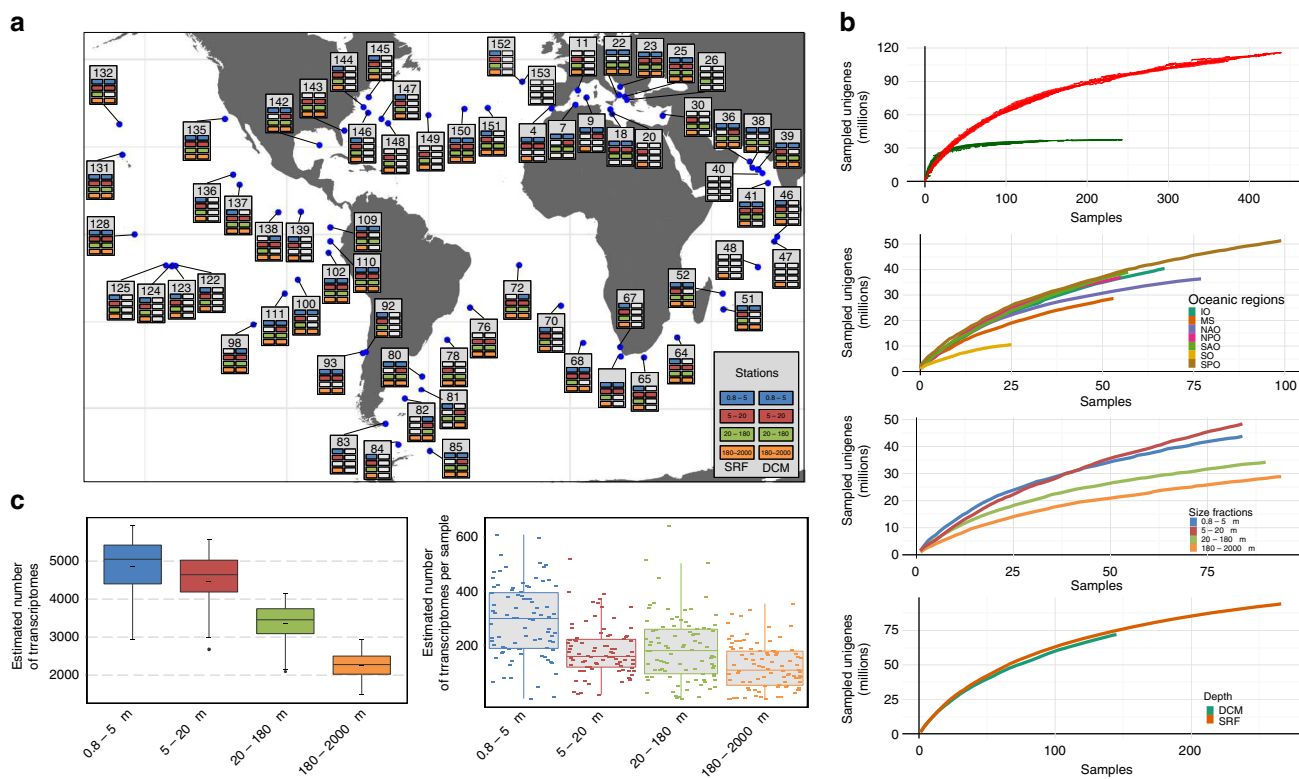


Fig. 1 The Tara Oceans eukaryote gene catalog. **a** Sampling map. Geographic distribution of 68 sampling stations at which seawater from the surface (SRF) and/or the deep chlorophyll maximum (DCM) was collected and size fractionated into four main groups: 0.8–5 μm (blue), 5–20 μm (red), 20–180 μm (green), and 180–2000 μm (orange). Availability of sequence data sets is indicated by the colored boxes at each sampling station. Two stations (TARA_40 and TARA_153) containing only atypical size fractions are shown on this map with empty boxes. **b** Rarefaction curves of detected genes. Top panel: rarefaction curves of 441 eukaryotic samples (red curve) compared to 139 prokaryotic samples (green curve) derived from Sunagawa et al.⁹. Other panels: rarefaction curve of eukaryotic samples by oceanic region (IO, Indian Ocean; MS, Mediterranean Sea; NAO, North Atlantic Ocean; NPO, North Pacific Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean), size fraction, and depth (SRF or DCM). For each curve, sampling order has been 10-fold permuted. **c** Estimated number of transcriptomes in eukaryotic samples. Left panel: distribution of the total number of transcriptomes estimated for each size fraction computed from the number of unigenes similar to a catalog of 24 single-copy ribosomal proteins. Right panel: distribution of the number of transcriptomes in each sample (small dashes) grouped by size fraction

discriminating the Pfam domain profiles using principal component analysis (PCA). The first two axes of the PCA are shown in Supplementary Fig. 3a. The main parameter explaining variance corresponded to differentiation between small-size and large-size fractions (horizontal axis), and the second major component of variance (vertical axis) separated the Southern Ocean (SO) samples from all the others. A few Gene Ontology (GO) terms show consistent patterns across all size fractions, highlighting major functional and taxonomical differences between SO regions and temperate or tropical oceans (Supplementary Fig. 3b), that can be either due to geographic segregation or to specific parameters of SO, e.g., low iron bioavailability. Samples from this region also tend to be more enriched in diatoms than at the other stations (mean 13%, s.d. = 3.8 in austral stations vs. 3%, s.d. = 2.2, in other samples) (Fig. 2b).

When looking at the most enriched gene categories between size classes, we observed small fractions being enriched in light-based energetic processes (photosynthesis and proteorhodopsins), transport of nutrients, carbohydrate metabolism, and flagellar movement, whereas large size fractions were associated with functions related to multicellularity, cell–cell contact, chitin metabolism, and muscular movement (Fig. 3a and Supplementary Fig. 4). This result demonstrates that the metatranscriptomics data capture not only the taxonomic differences observed previously⁶ but also the functional repertoires in each size fraction. We also observed that the relative expression of

photosynthesis genes (seen through chlorophyll-binding proteins) vs. proteorhodopsins (Bac_rhodopsin Pfam domain corresponding to type-I rhodopsins^{22,23}) showed a strong preference for photosynthesis in groups dominated by autotrophs, supporting that rhodopsin is not a major way of using light energy in these groups in natural conditions (Supplementary Fig. 5a). To further investigate the distribution of the expression of the rhodopsins present in the catalog, we isolated all the unigenes bearing a Bac_rhodopsin Pfam domain. We added to the dataset 2112 proteins—mainly from fungi (40%), bacteria (35%), and archaea (18%)—from public databases and 2538 eukaryotic protein sequences from MMETSP¹⁴. Protein sequences from the 71,576 unigenes carrying the Bac_rhodopsin Pfam domain were aligned and clustered with reference sequences to study their diversity (Methods section). We found that a large majority of annotated eukaryotic unigenes (82% of unigenes with the Bac_rhodopsin motif) were assigned to alveolates (73%), and contain conserved residues for proton-pumping activity, indicating that this group is the main contributor to proteorhodopsin-based light transduction in the open ocean. The three main clusters contain 55,325 unigenes (77%), and correspond to the three main groups observed based on references only²⁴ (Fig. 3b). Cluster 1 contains xanthorhodopsin-like proteins with conserved residues implicated in proton pumping (Fig. 3b, c and Supplementary Fig. 5b). The 26,733 unigenes of this cluster are almost exclusively derived

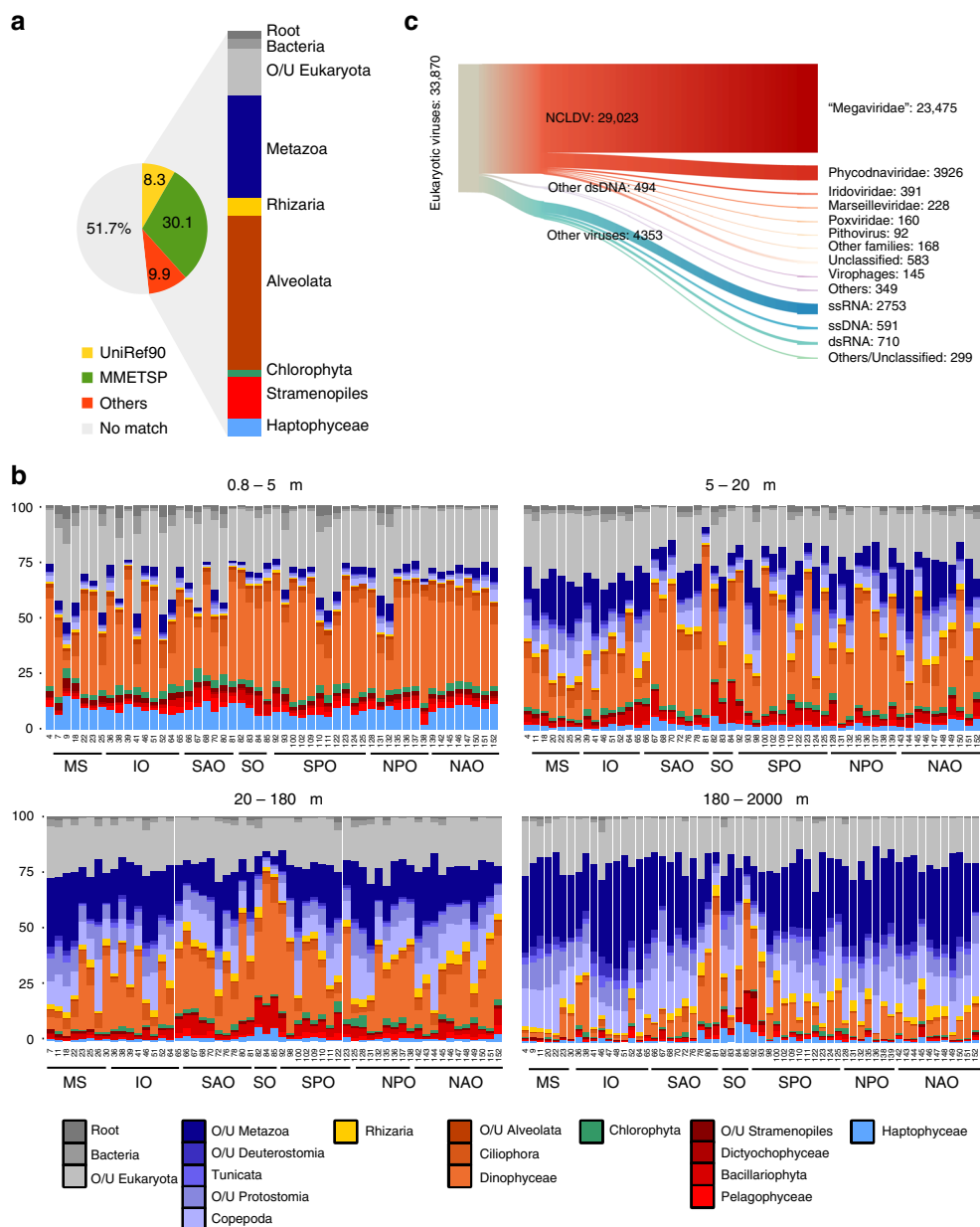


Fig. 2 Taxonomic composition of the gene catalog. **a** Origin of the best similarity sequence match as a fraction of the total in the circular diagram (MMETSP¹⁴: release of August 2014, with manual curation; UniRef90⁴²: release of September 2014; “Others”: are other reference transcriptomes that were added as reference to offset the lack of knowledge about organisms in large size fractions, in particular copepods and rhizaria; Methods section). Unigenes without significant matches (i.e., those with an e -value $>10^{-5}$ for their best similarity match) are tagged as “No match”. The proportion of unigenes affiliated to each major taxonomic group is indicated in the right column. O/U, other or unassigned. **b** Proportion of each major taxonomic group across *Tara* Oceans stations based on the mean number of unigenes classified as one of 24 different single-copy ribosomal proteins detected in each sample (IO, Indian Ocean; MS, Mediterranean Sea; NAO, North Atlantic Ocean; NPO, North Pacific Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean). **c** Eukaryotic viral unigenes. NCLDV unigenes are classified at the family level

from stramenopiles, alveolates, and haptophytes. This taxonomic distribution is consistent with the proposed single horizontal transfer from a bacterium to the common ancestor of the SAR group (Stramenopiles, Alveolates, and Rhizaria) and Haptista²⁴. The third cluster contains a large number of eukaryote references and most known sensory rhodopsins, but only 5641 unigenes with diverse taxonomies. Moreover, the proton acceptor residue E76, involved in the proton-pumping function, is not conserved, indicating that Cluster 3 proteins are likely to represent principally sensory rhodopsins (Fig. 3b, c and Supplementary Fig. 5c). Surprisingly, Cluster 2 contains only a few eukaryotic

references but is the second largest with 22,951 sequences, and displays the consensus sequence consistent with a proton-pumping function (Fig. 3b, c and Supplementary Fig. 5d). Most of these appear to be derived from alveolates, including the syndiniales parasites. This indicates that one of the most important categories of proteorhodopsins in the ocean is currently underestimated, possibly because of the lack of cultivated organisms bearing it, and that it may link photoheterotrophy with parasitism, a currently unexplored topic. Based on the hypothesis of a single lateral gene transfer event²⁴, the restricted taxonomic distribution of unigenes in Cluster 2 suggests

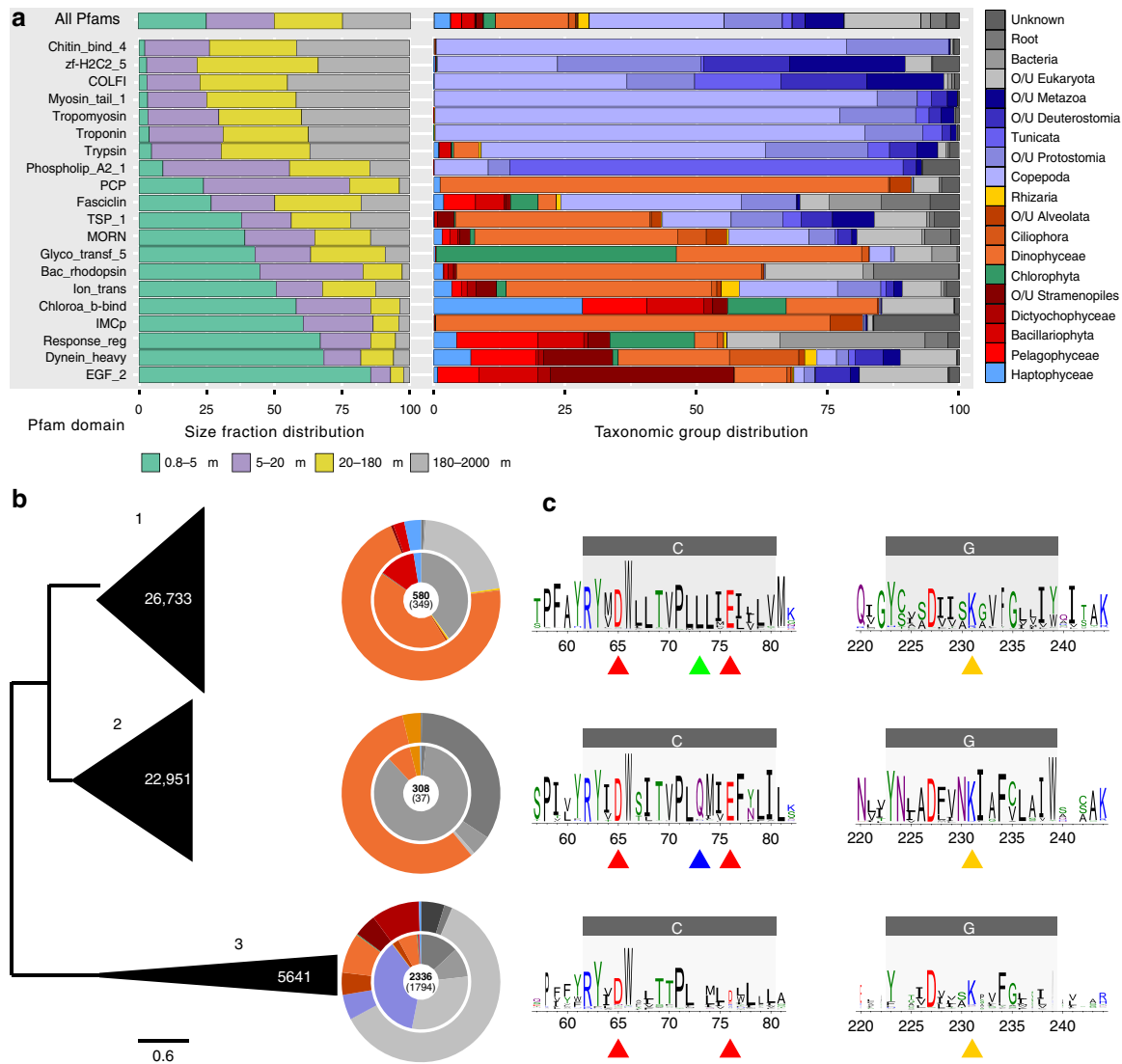


Fig. 3 Characterization of highly expressed gene families. **a** Major Pfam domains present in different size fractions and in different taxonomic groups. Among the highly expressed Pfam domains (Supplementary Fig. 4), those with specific patterns are shown. The relative expression of Pfam domains in the four filter sizes (left panel) and the contribution of each taxonomic group to the total expression of the Pfam domain (right panel) are shown as an average of all *Tara* Oceans SRF and DCM samples. O/U, other or unassigned. **b** Unrooted phylogenetic tree of type-I rhodopsin subfamilies (PF01036) obtained using sampling of 300 sequences of the three largest MCL clusters (see details in Supplementary Fig. 5b). The vertical size of the triangles represents the number of unigenes in each cluster (explicitly indicated in white) and their width represents the maximum branch length of 95% of sequences in the cluster. Taxonomic assignments of reference sequences (inner ring) and unigenes (outer ring) are indicated for each cluster with the color code of **a**. The number of reference sequences in each cluster is indicated in the center in bold, with the number of eukaryotic sequences in parentheses. **c** Logo consensus sequences, based on the global alignment of each cluster. Two regions of interest (helices C and G and their neighborhoods) containing functional and conserved residues are represented²⁵. Specific functional residues are indicated with arrows. Red: proton donor (D65) and acceptor (E76); green: residue specific to green light-sensitive proteorhodopsins; blue: amino acid specific to blue light-sensitive proteorhodopsins; yellow: lysine residue linked to retinal. Predicted transmembrane helices are represented as gray boxes

a more recent acquisition, which probably occurred before or during the radiation of the alveolate lineage. Interestingly, the consensus spectral tuning residue is different between Cluster 1 and Cluster 2: Cluster 1 protein sequences exhibit a leucine at position 105²⁵, indicating a maximal absorption of green light, whereas Cluster 2 sequences bear a glutamic acid at this position, indicating a peak absorption of blue wavelengths (Fig. 3c).

Gene novelty. The majority (51.2%) of unigenes currently have no matches in public sequence databases, which limits the insights that can be derived from the gene catalog. Some

sequences may be derived from non-coding genes or non-coding portions of coding genes, very short open reading frames, parts of genes where only another region is functionally known, or completely new open reading frames. To distinguish between these possibilities and better classify the catalog, we clustered all the unigenes according to a nucleic acid similarity threshold of >70% (Methods; Supplementary Fig. 6a). Despite its size, the gene catalog is not saturated, and accordingly we observed that 59.6% of unknown unigenes (UU) and 39.8% of known unigenes are represented by singletons (Fig. 4a, b). The clusters may thus be considered as being representative of gene family (GF) content of the catalog, with most singletons likely being derived from

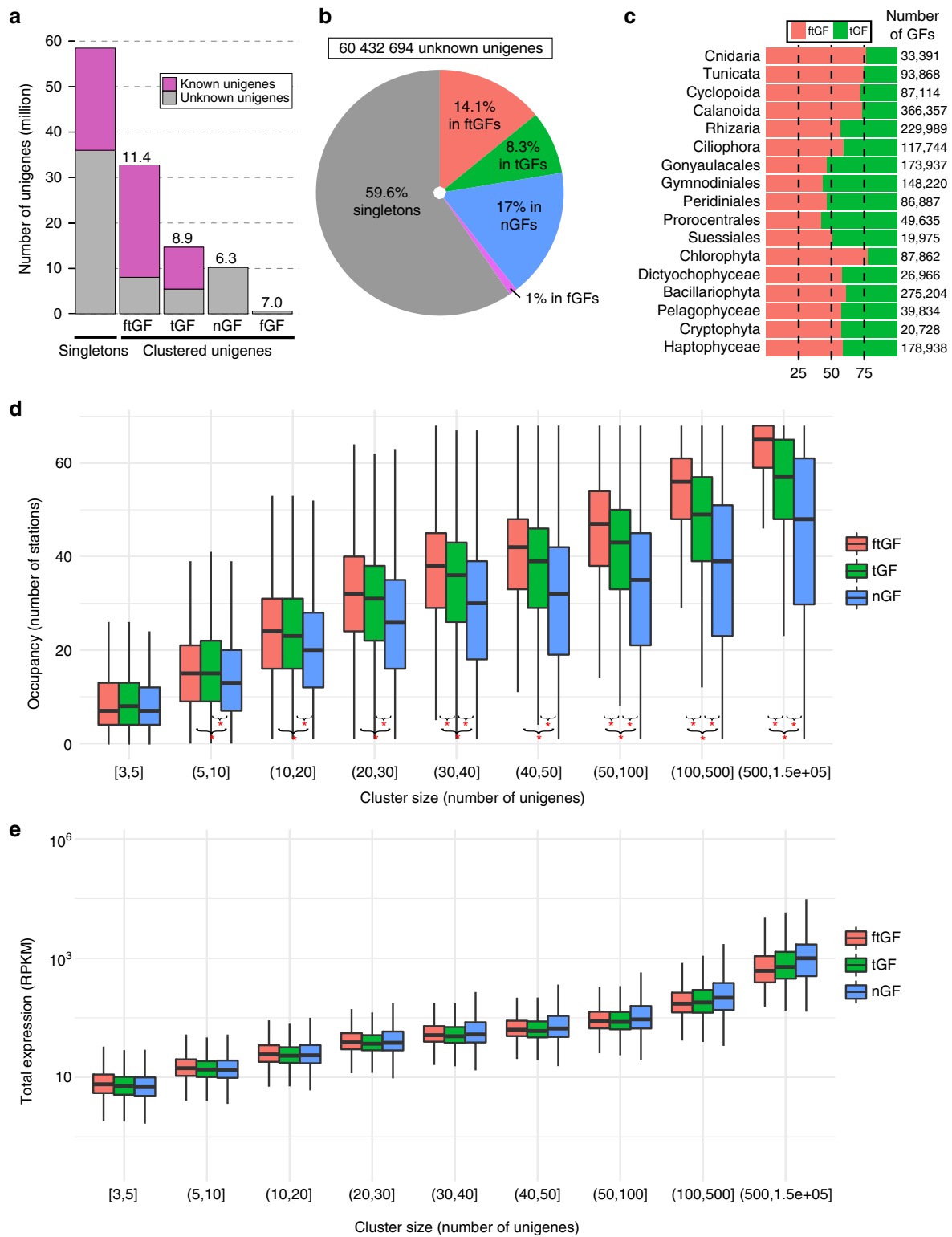


Fig. 4 Eukaryote gene catalog clustering and characterization of novel genes. **a** Global repartition of unigenes based on the gene catalog clustering. Unigenes were considered as singletons if they are in clusters of less than three units. Gene families are novel (nGF), taxonomically assigned (tGF), functionally assigned (fGF), or both (ftGF) (Methods). Numbers above each bar indicate the numbers of unigenes per cluster. **b** Distribution of unknown unigenes in the different categories described in **a**. **c** Ratio of tGFs vs. ftGFs in the main taxonomic groups. The total number of GFs assigned to each taxonomic group is indicated on the right. **d** Distribution of GF occupancy for the three main GF categories. GFs are classified according to their size (x-axis) and the y-axis indicates the number of stations where the GF family is expressed (at least one unigene detected with a coverage of more than 80% of the unigene length). Kolmogorov-Smirnov tests with $p < 10^{-5}$ between occupancy distributions are indicated with red stars. **e** Distribution of mean expression levels of the three different categories of GFs among all samples. GFs are classified according to their size (x-axis). The expression of a GF in a sample was determined by the sum of the expression of its unigenes in RPKM

smaller GFs that will grow with more sequencing effort. The 6.2 million GFs, encompassing 58.4 million unigenes, were subsequently subdivided into four classes based on taxonomic affiliation and functional annotation (see Methods; Fig. 4a–c): those with both functional and taxonomic assignments (ftGF), those with taxonomy-only assignments (tGF), those with function-only assignments (fGF), and those representing new GF (nGF). The fGF category was not considered further because it contains too few clusters (1.43%).

We searched for fundamental differences between these three types of GFs by observing in how many stations they were detected (Methods section). Regardless of GF size, nGFs were present in less stations than ftGFs, whereas tGFs showed intermediate occupancies (Fig. 4d and Supplementary Fig. 7a). This pattern was not due to higher mean expression levels of ftGFs or tGFs that would render them more detectable than nGFs (Fig. 4e). We conclude that the gene novelty detected corresponds to families that are present in fewer environments, yet are not less expressed than known families. Moreover, nGFs generally represent smaller GFs (6.3 unigenes per cluster) than fGFs (8.9) and ftGFs (11.4), suggesting that nGFs are conserved in a smaller range of species than characterized GFs (Fig. 4a and Supplementary Fig. 6b), or that they are present in less abundant taxonomic groups. It has been previously suggested that newly discovered genes are either biased taxonomically (which restrains their presence in databases), or that they correspond to genes that are necessary only in some conditions, potentially related to the adaptation of organisms to specific environments²⁶. We found evidence for both cases, as nGFs are more restricted in occupancy, whereas tGFs are more abundant in less-characterized phyla (Fig. 4c–e).

We further questioned whether the intermediate occupancies observed with tGFs can be due to an intrinsic property or to them being distributed between two types of families, looking either more like ftGFs or more like nGFs. The distribution of occupancies in tGFs indeed appears to be bimodal, with a group containing fewer UUs resembling the ftGF distribution, and another group containing a high proportion of UUs resembling the nGF distribution (Supplementary Fig. 7b,c). We conclude that some of the tGFs likely represent widely occurring genes that have no predicted functions, most likely because of their limited taxonomic distribution in the global tree of eukaryotes. The others may represent GFs with characteristics of nGFs that have few members matching with references, generally reflecting efforts to gain information on environmentally-important organisms such as the MMETSP effort¹⁴.

Although our metatranscriptomics sequencing effort is based on polyadenylated RNA and relatively shallow coverage per individual organism, and thus may not be able to capture non-coding RNAs significantly, we then consider the nGF category, asking if these new families can be coding. For this, we selected the central unigene of each cluster of more than 10 unigenes as a reference of the GF, then we looked for protein homologies between references (see Methods and Supplementary Fig. 6a). This created 75,175 protein groups of GFs, among which 11,431 link 30,558 nGFs only, and 22,072 link 130,501 tGFs only. Examples of nGFs are shown in Fig. 5 (protein group number 14079 for nGFs with restricted expression) and Supplementary Fig. 8a–d (protein group number 1540 for more broadly distributed nGFs). We were able to align ORFs from these clusters and found that they contain highly conserved amino acids that can provide clues about their structure (Fig. 5d, Supplementary Fig. 8d). Another example from a highly conserved tGF restricted to dinoflagellates and close relatives is shown in Supplementary Fig. 8e–h. Taken together, these data show that 3.26 million GFs with or without taxonomic

information are present as highly expressed families in the global ocean and do not correspond to defined domains. We suggest that these may be important targets for future definition of new protein domains to more faithfully encompass the functional diversity present in eukaryotes. The current database of protein domains such as Pfam²⁷ contains 16,712 different domains of known and unknown functions, whereas we detected 11,431 protein groups of nGFs, and 22,072 groups of tGFs based only on Clustering of the largest families, indicating the high discovery rate of new conserved domains that could be used to derive a more exhaustive list of conserved domains within eukaryotes.

In summary, we have found that UUs can be part of known GFs but that a large proportion are predicted to be novel protein-coding genes. As they are distributed less globally than known functions, their extent remains to be evaluated, although we have shown here that they represent a highly significant portion of the gene repertoire of eukaryotic plankton.

The environmental footprint of gene expression in phytoplankton. To highlight how the annotated gene catalog can be useful for studying environmental gene expression, we examined the five principal photosynthetic groups (Fig. 2c), namely diatoms (Bacillariophyta), chlorophytes, dinoflagellates (Dinophyceae), haptophytes, and pelagophytes, for some of their most highly expressed functions and their variations according to two environmental parameters, specifically iron and net primary production (NPP). Obligate autotrophs, such as diatoms and chlorophytes, showed a higher correlation to NPP for genes involved in photosynthesis and carbon fixation than the other groups that also contain mixotrophic representatives. Additionally, we observed an apparent lack of correlation between expression of genes important for photosynthesis and carbon fixation in dinoflagellates in conditions of high NPP (Supplementary Fig. 9). Although this could be explained by low reliance on transcriptional regulation in this group⁵, we observed an increased correlation of expression of genes encoding cell lytic components, such as proteases and lipases. Such changes in ecosystem function may be a consequence of alterations in the dominant dinoflagellates in the community or to switches in trophic strategy in mixotrophic species, and have significant implications for the functioning of marine food chains in different environmental conditions.

Differences in expression patterns of unigenes between two sampling stations can be linked to either (or both) changes in population composition and changes in expressed functions related to the environment. Comparison of metagenomes and metatranscriptomes allows assessment of the expression of genes from the catalog normalized to underlying gene abundances. To highlight this, we examined genes whose expression and/or copy number have been shown to be responsive to nutrient availability, specifically iron, an important yet often limiting nutrient in the ocean.

Phytoplankton are good models to study iron homeostasis as they have significant high demands of this metal due to its requirement for photosynthesis²⁸. One low iron response that occurs in the photosynthetic electron transport chain involves the replacement of the iron-sulfur containing electron carrier ferredoxin with flavodoxin, a less efficient protein that does not require iron^{29,30}. In addition to the canonical photosynthetic versions, there are a number of flavodoxins and ferredoxins involved in different metabolisms, or constituting functional domains of complex multidomain redox proteins²⁹. To study whether the flavodoxin/ferredoxin switch can be detected using our dataset, we carried out an analysis of the ferredoxin and flavodoxin families using the Pfam domains PF00111 and

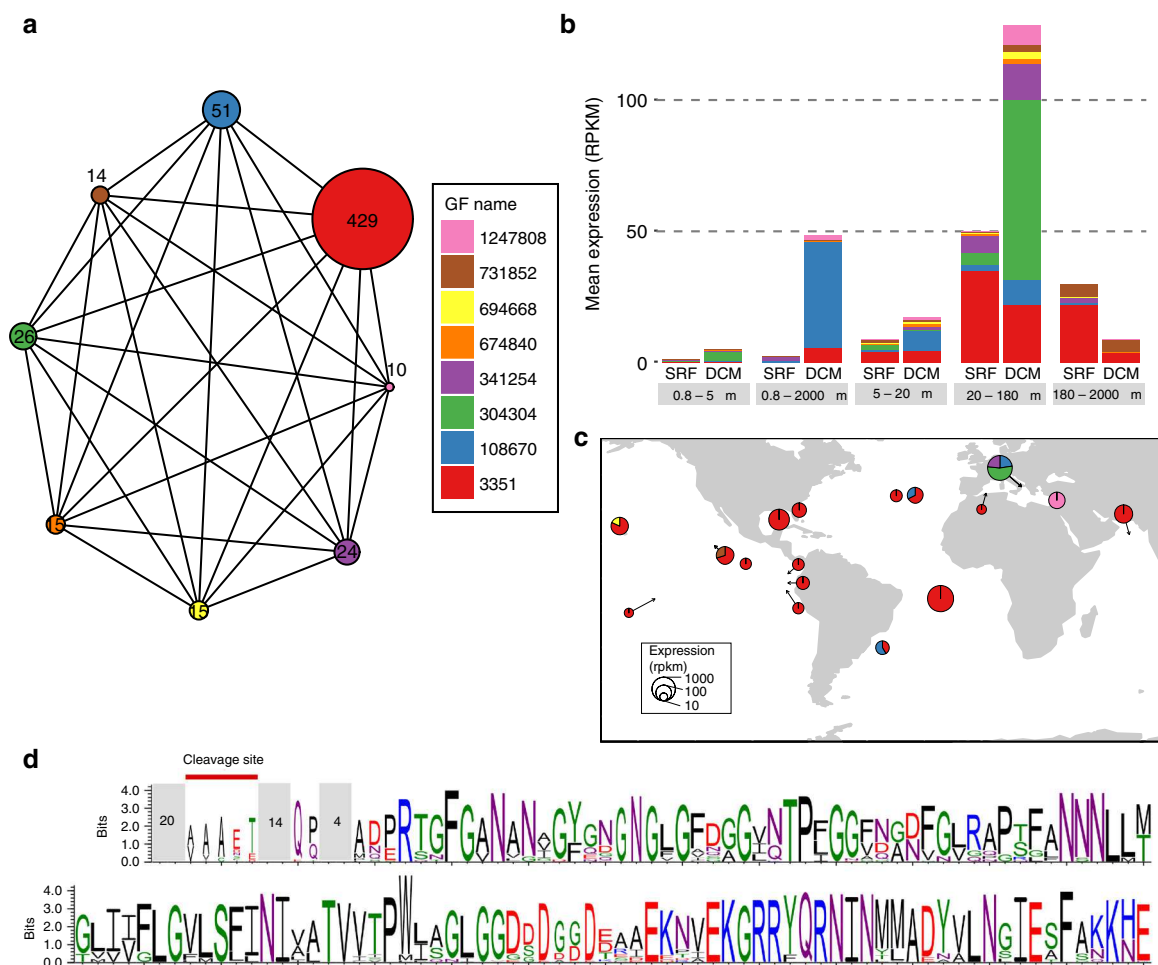


Fig. 5 New gene families expressed in 20–180 μm size fraction. **a** Graph representation of the protein group number 14079. Each GF of the protein group is represented by a node with a diameter proportional to the number of unigenes in the GF. Protein matches between GFs are represented by an edge. **b** Mean expression of GFs in different size-fractions and depths. Each color corresponds to a GF of protein group 14079. **c** World map representation of protein group 14079 expression in the 20–180 μm size fraction. SRF and DCM samples have been pooled. Circle diameters represent the relative expression of the protein group in RPKM. The contribution to expression of each GF is represented by the different colors. **d** Sequence logo of the multiple alignments of the protein group 14079. 45 ORFs (153 amino acids in average) of protein group 14079 were aligned and positions with more than 50% of gaps were removed. Mean numbers of amino acids on unaligned regions of the protein are indicated in gray boxes. A signal peptide cleavage site, indicated on the left part of the sequence logo was predicted on 21 sequences

PF00258. These families not only include the photosynthetic versions but also other isoforms and domains, and there is an overlap of redox properties between different members of these two families, being potential isofunctional proteins in many reactions²⁹. Thus, we studied the relative levels of the two families of genes in the five major phytoplankton groups by calculating the ratio of their relative abundances and expression (Fig. 6). With the exception of diatoms, gene abundances show little variations and only weak correlations with iron concentrations (Fig. 6a; “Metagenome” column and Supplementary Data 5). On the other hand, the ratios of relative expression show strong variations, particularly for chlorophytes, haptophytes and pelagophytes (Fig. 6; “Metatranscriptome” column), indicating that these three groups modulate the relative levels of ferredoxin and flavodoxin principally by regulation of mRNA levels. By contrast, diatoms tend to express flavodoxin genes more than ferredoxin genes, although a few mainly coastal stations showed a strong up-regulation of the latter. In this group, the metagenomics data indicate that diatom genomes display far more heterogeneity in ferredoxin/flavodoxin content than the other groups studied, suggesting that individual diatom species may be permanently

adapted to specific iron regimes in the ocean rather than maintaining transcriptional flexibility, as observed in haptophytes, chlorophytes and pelagophytes. Unlike any other groups, dinoflagellates appear to rely only weakly on gene abundance or expression variations (Fig. 6), which may again be related to their low transcriptional flexibility. These results suggest that nutrient limitations are dealt with in different ways among these main photosynthetic taxa, either by a genotypic commitment to a specific regime, or by the maintenance of transcriptional flexibility, and that the *Tara* Oceans eukaryote gene catalog may be a useful resource to distinguish the strategies of any plankton group to adapt to these limitations when transcript regulation or gene copy number is implicated.

Discussion

The global ocean transcript catalog reported here represents a first resource to study extensively and uniformly the gene content of eukaryotes and the dynamics of their expression in the environment, and notably adds to previous DNA-based resources that describe the viral and prokaryotic components of the

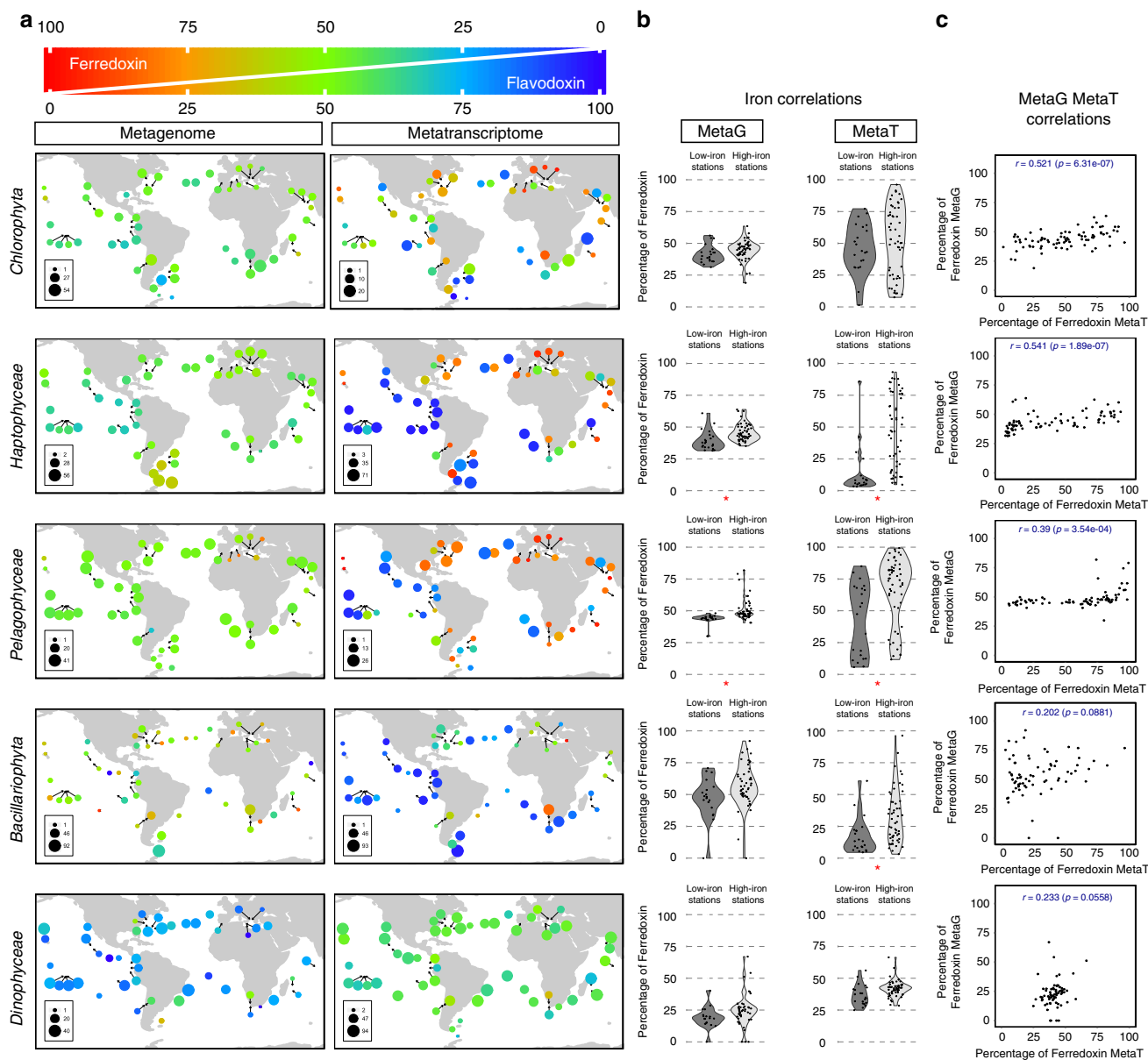


Fig. 6 Ratios of differential gene abundance and relative expression of ferredoxin vs. flavodoxin in the five major photosynthetic groups. **a** Representation of the relative abundance (left) and expression (right) of the two genes identified in surface samples for *Chlorophyta*, *Pelagophyceae*, *Haptophyceae* (from 0.8 to 5 μm filters), *Bacillariophyta* and *Dinophyceae* (from the 5 to 20 μm filters). The circle colors, from red to blue, represent the relative expression of one gene compared to the other, with the color code given in the top diagram. The sum of the expression levels of the two genes affiliated to each taxonomic group is represented by the circle diameter as a percentage of the total expression of these genes. **b** Distribution of the relative abundance (left) or expression (right) of ferredoxin in low iron stations ($<0.02 \mu\text{mol m}^{-3}$, 15 stations, dark gray) or iron rich stations ($>0.2 \mu\text{mol m}^{-3}$, 31 stations, light gray) according to a model of iron concentration in the oceans (Supplementary Data 5). Significant differences of expression between low and rich iron stations are indicated with red stars (non-parametric wilcoxon rank-sum test, $p < 10^{-3}$). **c** Correlations between the relative metagenome (MetaG) abundance and metatranscriptome (MetaT) expression of ferredoxin in SRF and DCM samples, expressed as a percentage of the total value of ferredoxin + flavodoxin. Pearson correlation coefficients (r) and their statistical significance (p) are indicated in each graph. Ferredoxins and flavodoxins were identified using the Pfams PF00111 and PF00258, respectively

ocean^{9–11}. The gene repertoire of planktonic eukaryotes is massive and diverse, much more so than the prokaryotic gene space⁹. The impressive number of genes without functionally-characterized homologs in databases points to the large numbers of understudied yet widely distributed genera inhabiting marine ecosystems, for which even widely conserved GFs have yet to be investigated. The restricted distribution of totally new GFs highlights the need to develop methods for revealing their roles without the support of homology-based hypotheses. Because

representatives of almost all of the eukaryote groups⁴ are abundant in oceanic plankton, they can likely inform us in new ways about the evolutionary trajectories of different eukaryotes, in particular those with parasitic and symbiotic lifestyles that have remained largely recalcitrant to study until now, although being a large part of the interacting species network within plankton ecosystems^{31,32}. The resource is also likely to be of great utility for exploring organisms within the zooplankton, including metazoans, that have to date been largely unexplored by genomics³³. As

we have shown for the principal groups of phytoplankton, it is possible to obtain insights between adaptive and acclimatory processes underlying organismal responses to their environment using as proxies the contrasts between metagenomics and metatranscriptomics, paving the way for similar studies in other organisms.

Methods

Sampling of eukaryotic plankton communities. The biological samples were collected during the *Tara* Oceans expedition from 68 sampling sites. Typically, two depths were sampled in the photic zone: subsurface (SRF) and deep-chlorophyll maximum (DCM). A detailed description of all *Tara* Oceans field sampling strategy and protocols is available in Pesant et al.²⁰. In short, planktonic eukaryotic communities were collected in the 0.8–2000 µm range and size-fractionated in four fractions (0.8–5 µm, 5–20 µm, 20–180 µm, and 180–2000 µm). A low-shear and non-intrusive industrial peristaltic pump was used for the 0.8–5 µm fraction and plankton nets for the others. The volumes of filtered seawater were scaled according to known organismal concentrations within each size fraction, from 0.1 m³ for the most concentrated pico-plankton to 148 ± 136 m³ for the most-dilute meso-plankton, in order to get near-exhaustive recovery of total eukaryotic biodiversity in each sample. Water was filtered immediately after sampling. Whole-plankton communities were subsequently filtered on polycarbonate membranes, rapidly flash-frozen and preserved in liquid nitrogen on board *Tara*.

Physicochemical parameters measured during the expedition are available in the Pangaea database (<https://www.pangaea.de/> and Supplementary Data 5) and described in Pesant et al.²⁰. Due to the sparse availability of direct observations of iron in the surface ocean, concentrations were derived from a global ocean simulation using the MITgcm ocean model configured with 18 km horizontal resolution and a biogeochemical simulation which resolves the cycles of nitrogen, phosphorus, iron and silicon³⁴. The biogeochemical parameterizations, including iron, are detailed in Follows et al.³⁵. Atmospheric deposition of iron was imposed using monthly fluxes from the model of Mahowald et al.³⁶. NPP values were derived from satellite measurements from 8-day composites of the vertically generalized production model³⁷. Physicochemical parameters of each station analyzed in this article are indicated in Supplementary Data 5.

Nucleic acid extraction, library construction, and sequencing. DNA and RNA were extracted simultaneously by cryogenic grinding of cryopreserved membrane filters using a 6770 Freezer/Mill or 6870 Freezer/Mill instrument (SPEX Sample-Prep, Metuchen, NJ) followed by nucleic acid extraction with NucleoSpin RNA Midi kits (Macherey-Nagel, Düren, Germany) combined with DNA Elution buffer kit (Macherey-Nagel). DNA and RNA were quantified by a fluorometric method using Qubit 2.0 Fluorometer (ThermoFisher Scientific, Waltham, MA). DNase treatments were applied to all RNA extractions. Metagenomic libraries were prepared manually or in a semi-automatic manner according to available DNA quantity. Genomic DNA was first sheared to a mean target size of 300 bp using a Covaris E210 instrument (Covaris, Woburn, MA). DNA inputs in fragmentation step were 30–100 ng in the case of a downstream manual preparation or 250 ng for semi-automated protocol. End repair, A-tailing and Illumina adapter ligation were then performed manually using NEBNext Sample Reagent Set (New England Biolabs) or with the SPRIWorks Library Preparation System and SPRI TE instrument (Beckmann Coulter Genomics), according to the manufacturers protocol. Ligation products were PCR-amplified using Illumina adapter-specific primers and Platinum Pfx DNA polymerase (Invitrogen). Amplified library fragments were size selected at around 300 bp on 3% agarose gels. For RNA samples, a poly (A)⁺ RNA selection strategy was used to limit rRNA quantity. Different cDNA synthesis protocols were applied according to the quantity of RNA. When at least 2 µg total RNA were available, cDNA synthesis was carried out using the TruSeq mRNA Sample preparation kit (Illumina, San Diego, CA). Samples with less than 2 µg of RNA were processed using the SMARTer Ultra Low RNA Kit (Clontech, Mountain View, CA). In these cases, fifty nanograms or less of total RNA were used for cDNA synthesis, followed by 12 cycles of PCR preamplification of cDNA and Covaris shearing to a 150–600 bp size range. cDNAs were then used for Illumina library preparation following the manual protocol described for metagenomic libraries, except that the size selection step on agarose gel was omitted. A detailed description of nucleic acid extractions and library construction protocols is available in Alberti et al.³⁸. After library profile analysis by Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and qPCR quantification (MxPro, Agilent Technologies, USA), libraries were sequenced on HiSeq2000 instruments (Illumina) with a read length of 101 bp in a paired-end mode. In average, 160 million reads per sample were obtained.

Reads, assembly and gene catalog construction. An Illumina filter was applied to remove the least reliable data from the analysis. The raw data were filtered to remove any clusters with too much intensity corresponding to bases other than the called base. Adapters and primers were removed on the whole read and low quality nucleotides were trimmed from both ends (while quality value is lower than 20). Sequences between the second unknown nucleotide (N) and the end of the read

were also removed, as were reads with a resulting length smaller than 30 bp, as well as their mates mapped onto run quality control sequences (PhiX genome). After cleaning, all single reads (fragment with one discarded read) were eliminated from further analyses. Ribosomal RNA-like reads were excluded using *sornterRNA*³⁹. Resulting reads from each metatranscriptomic sample were assembled using velvet v.1.2.07⁴⁰ with a kmer size of 63. Isoform detection was performed using oases 0.2.08⁴¹. Contigs smaller than 150 bp were removed from further analysis.

Assembly results and descriptive statistics for each sample are shown in Supplementary Data 6. Similar sequences from more than one sample were removed using Cdhit-est v 4.6.1, with the following parameters: -id 95 -aS 90 (95% of nucleic identity over 90% of the length of the smallest sequence). For each cluster of contigs, the longest sequence was kept as reference for the gene catalog. Ribosomal, chloroplastic, and mitochondrial sequences were removed from the resource after blast comparisons and Pfam domains identification. Prokaryote 16S-like unigenes were mega-BLAST scanned for removal. Mitochondrial or chloroplastic sequences were removed based either on the basis of a positive BLAST hit against dedicated reference databases manually curated, and having matches with at least 70% identity over at least 80% of the unigene length or at least 300 bp long, or based on the presence of specific protein domains identified by CDD search. Domains COX1, COX2, COX3, COX2_TM, Cytochrom_B_N_2, Cytochrom_B_C, Cytochrom_B_N, Oxidored_q1, Oxidored_q2, Oxidored_q3, Oxidored_q4, Oxidored_q5_N, Oxidored_q1_N, NADHdh, NDH_I_M, NDH_I_L, and ATP_synt_6_or_A were used as signature for mitochondrial based genes, domains and Photo_RC, PsaA_PsaB, PSII, RuBisCO_large, and RuBisCO_large_N for the chloroplastic ones, while unigenes also bearing domains Peptidase_M41, Gp_dh_N, or Gp_dh_C, GAPDH-I were kept in the resource, being considered as nuclear genes. In summary a unigene as defined here is a complete or partial transcript assemble from metatranscriptomic reads of at least one *Tara* Oceans station. The gene catalog is accessible at <http://www.genoscope.cns.fr/tara/>.

Taxonomic assignment. To assign a taxonomic group to each unigene, a reference database was built from UniRef90 (release of 2014–09–04)⁴², from the MMETSP project (release of 2014–07–30)¹⁴ manually curated to remove sequence redundancy, from *Tara* Oceans Single-cell Amplified Genomes (PRJEB6603). The database was supplemented with three Rhizaria transcriptomes (Collozoum, Phaeoadae and Eucyrtidium, available through the European Nucleotide Archive under the reference PRJEB21821 (<https://www.ebi.ac.uk/ena/data/view/PRJEB21821>) and transcriptomes of *Oithona nana*³³. Sequence similarities between the gene catalog and the reference database were computed in protein space using Diamond (version 0.7.9)⁴³ with the following parameters: -e 1e-5 -k 500 -a 8. Taxonomic affiliation was performed using a weighted Lowest Common Ancestor approach. For each unigene, all protein matches with a bitscore value ≥90% of the best match bitscore were kept. For each taxon, only matches with the highest bitscores were retained, and total LCA and weighted LCA (covering at least 67% of all bitscores), were further computed. In order to limit the number of false taxonomic assignments explained by the lack of reference genomes, the LCA result was corrected according to the percentage of identity of selected matches. The maximal taxonomic precision allowed was corrected as follows: >95% of identity = species, <95% of identity = genus, <80% of identity = family, <65% of identity = order, <50% of identity = class. The taxonomic assignment of unigenes is accessible at <http://www.genoscope.cns.fr/tara/>. The taxonomic assignment of eukaryotic viruses was performed as explained above but with the following modifications. First, all subject sequences with viral taxonomic identifiers were removed and replaced by viral sequences of Virus-Host DB⁴⁴ (as of 23 February 2017) to allow access to host type information. Viral unigene sequences assigned to bacteriophages or archaeal viruses were discarded from analysis. Second, we used the NCLDV nomenclature derived from the common ancestor hypothesis⁴⁵ based on seven distantly related viral families: 'Megaviridae', *Phycodnaviridae*, *Marseilleviridae*, *Iridoviridae*, *Ascoviridae*, *Asfarviridae*, and *Poxviridae*. Among these, "Megaviridae" is a recently proposed family^{46,47}. We added the following viral groups: *Pandoravirus*, *Pithovirus*, *Mollivirus* proposed to form new NCLDV families⁴⁸ as well as *Faustovirus*⁴⁹. Unclassified viroplages were classified as "dsDNA viruses, no RNA stage". Viroplages *Mavirus* and Organic Lake viroplages were classified as unclassified viroplages. RNA viruses reported in⁵⁰ were classified in their respective order or family according to their phylogenetic position. Viral groups were added for the newly described families *Chuviridae*⁵¹, *Yanvirus*, *Weivirus*, *Zhaovirus*, *Qinvirus*, and *Yuevirus*⁵⁰. Finally, the LCA result was corrected according to the percentage of identity of selected matches as follows: >95% of identity = species, <95% of identity = genus, <70% of identity = family.

Functional characterization of unigenes. Protein domain prediction was performed using the hmmssearch tool of the HMMer package (version 3.1b2)⁵² against the Pfam-A database (release 28). Only matches exceeding the internal gathering threshold (-cut_ga) were retained. Pfams often detected on the same unigenes were grouped together in a single name (i.e., *Arrestin_C*; *Arrestin_N*). These associations of Pfams followed two criteria: (1) The number of unigenes carrying the two pfams is higher or equal to 30% of the average number of unigenes carrying each Pfam. (2) The number of unigenes carrying the two pfams was higher than 30. The list of associated Pfams is given in Supplementary Data 7. The functional characterization of unigenes is accessible at <http://www.genoscope.cns.fr/tara/>.

fr/tara/. Unigenes without Pfam domains are excluded from analyses presented in Figs. 3, 6 and Supplementary Figs. 3, 4, 5, 9. The Pfam domain PF01036 was searched in unigenes and the MMETSP collection using hmmscan (from HMMer 3.1b2)⁵². NCBI sequences carrying the Pfam motif were retrieved through the PFAM portal (<http://pfam.xfam.org/>, May 2017). All-vs.-all BLAST comparisons were run at the protein level using BLAST + 2.6.0 and sequences were clustered with the MCL algorithm⁵³ using the $-\log(e\text{-value})$ as edge weights and an inflation parameter of 1.4. For each of the three largest clusters, protein sequences were aligned using MAFFT 7.31⁵⁴ and positions with more than 50% of gaps were discarded. Logo consensus sequences were created using weblogo 3 program⁵⁵. Transmembrane helices were predicted using TMHMM Server 2.05 on the consensus sequences⁵⁶. Global phylogenetic tree was constructed from a global alignment using MAFFT 7.310. The phylogenetic inference was made using approximate maximum likelihood with FastTree⁵⁷, under the gamma model of heterogeneity.

Expression and abundance of unigenes. In order to estimate the abundance and expression of each unigene in each sample, cleaned reads (from metagenomes and metatranscriptomes) were mapped against the reference catalog using the bow tool (version 0.7.4)⁵⁸. The following parameters were used: `bwa aln -l 30 -O 11 -R 1; bwa sampe -a 20000 -n 1 -N; samtools; rmdup`. Low complexity reads were removed. Reads covering at least 80% of read length with at least 95% of identity were retained for further analysis. In the case of several possible best matches, a random one was picked. Mapping results are summarized in Supplementary Data 8. Unigene expression values and genomic occurrences were computed in RPKM (reads per kilo base covered per million of mapped reads). RPKM values for each Unigenes in each sample are accessible at <http://www.genoscope.cns.fr/tara/>. The abundance or expression of each unigene was normalized and formulated in two different ways. (i) The gene expression/abundance relative to the expression/abundance of all genes from the same taxon in percentage. e.g., the expression of Pelagophyceae Ferredoxin genes (Pfam Fer2, 372 unigenes) represents 0.17% of Pelagophyceae transcriptomes. (ii) The fraction of the gene expression/abundance attributed to a particular taxonomic group. e.g., 24.3% of ferredoxin genes are expressed/present in Pelagophyceae transcriptomes. These normalized values of expression and abundance are calculated for all unigenes grouped by Pfams or GO term (Biological Processes) and a list of taxonomic groups: *Haptophyceae*, *Pelagophyceae*, *Bacillariophyta*, *Dictyochophyceae*, *O/U Stramenopiles*, *Chlorophyta*, *Dinophyceae*, *Ciliophora*, *O/U Alveolata*, *Rhizaria*, *Copepoda*, *O/U Protostomia*, *Tunicata*, *O/U Deuterostomia*, *O/U Metazoa*, *O/U Eukaryota*, *Bacteria*, root (unigenes with matches in at least two of the Eukaryota, Archaea, Bacteria, and Virus superkingdom), unknown (unigenes that have no similarities in amino acid databases), O/U = unigenes for which taxonomic affiliation ended at the indicated level or belonged to minor classes of the affiliation.

Estimation of transcriptome diversity. A total of 24 ribosomal genes, single copy, highly expressed and universally distributed⁵⁹, were selected to estimate the number of different transcriptomes in each sample: COG0049, COG0052, COG0080, COG0081, COG0087, COG0088, COG0091-COG0094, COG0096-COG0100, COG0102, COG0103, COG0184-COG0197, COG0200, COG0256, COG0522. The average number of unigenes carrying each of these COG domains was used to estimate the number of different transcriptomes. A unigene was considered to be present in a sample if at least 80% of its length was covered by sample reads with at least 95% identity. Reference genomes and their annotation used to estimate the redundancy of the gene catalog and refine transcriptome diversity estimations were downloaded from Ensembl Protists (<http://protists.ensembl.org/index.html>) for *Emiliania huxleyi*, *Thalassiosira oceanica*, *Aureococcus anophagefferens*, *Acanthamoeba castellanii* str. Neff and *Monosiga brevicollis*, from Orcae (<http://bioinformatics.psb.ugent.be/orcae/>) for *Bathycoccus prasinos* and *Micromonas pusilla* and from Genoscope (<http://www.genoscope.cns.fr/externe/GenomeBrowser/>) for *Oikopleura dioica* and *Oithona nana*. The gene catalog was aligned (BLAT v32 × 1) against predicted genes from reference genomes with a minimum of 70% of identity over at least 80% of the length of the smallest sequence of the pair (Supplementary Data 2), then fully overlapping unigenes have been removed. For each reference genome, the average number of unigenes mapping each gene and ribosomal proteins listed above were calculated. The mean of the result for each genome was used as an estimation of the catalog redundancy.

Construction of gene families. Nucleic acid homologies between all unigenes of the eukaryotic gene catalog were calculated with BLAT (v. 36) (min 70% of identity and 100 bp). The 1609 million matches obtained were clustered with MCL (v. 14–137) into 6,225,695 clusters of 3 unigenes or more, named GFs (Supplementary Fig. 6a, steps 1–2). Clusters were classified into four categories according to their percentage of unigenes with a taxonomic affiliation and/or a functional characterization. Functionally and taxonomically assigned GFs (tGFs) comprise >5% of unigenes with matches and domains; taxonomically assigned GFs (tGFs) comprise >5% of unigenes with matches but no predicted domains; new GFs (nGFs) have <5% of unigenes with matches or domains; and functionally assigned GFs (fGFs) have >5% of unigenes with domains and <5% with matches (Supplementary Fig. 6a, step 3). The most precise taxonomic affiliation carried by more than

50% of known unigenes of a given tGF or fGF was chosen to determine its taxonomic affiliation. A representative unigene for each GF with a minimum of 10 unigenes was determined by the calculation of the betweenness centrality (library Graph::Undirected, Perl) of the corresponding MCL cluster. 1,261,965 central unigenes were 6-frames translated, and similarities between them were then computed with Diamond (version 0.7.9)⁴³. The best match for each sequence pair with an $e\text{-value} < 1e^{-10}$ was selected, then all protein matches were clustered with MCL (pondered by the cluster size) (Supplementary Fig. 6a, steps 4–5). MCL clusters of GFs are named protein groups. GFs and protein groups composition and annotation are accessible at <http://www.genoscope.cns.fr/tara/>. Protein groups detailed in Fig. 5 and Supplementary Fig. 8 were analyzed for their amino acid composition. The 5 longest ORFs with a minimum of 150 amino acids found in each GF of the protein group were aligned with mafft (v. 7.310)⁵⁴ in globalpair mode and unalignlevel at 0.9. The alignment was manually curated in order to remove non-relevant ORFs, then positions with more than 50% of gaps were removed. Peptide signal sequences and cleavage sites were detected with signalP⁶⁰ and added to the alignment. The sequence logo representations were made with weblogo program⁵⁵.

All statistical analyses and graphical representations were conducted in R (v 3.1.2) with R packages ggplot2 (v 2.1.0). The PCA results shown in Supplementary Fig. 3 were obtained using the R package FactoMineR v 1.32, world maps with maps (v 3.1), phylogenetic trees with ggtree (v 1.6.11), and graph representation Fig. 5a and Supplementary Fig. 8a,e with igraph (v 1.0.1) and ggnetwork (v 0.5.1).

Code availability. Computer codes are available from the corresponding authors upon request.

Data availability. Sequencing data are archived at ENA under the accession number PRJEB4352 for the metagenomics data and PRJEB6609 for the metatranscriptomics data (see Supplementary Data 8 for details). Unigene catalog is available at ENA under accession number ERZ480625. Environmental data are available at PANGAEA (URLs for each sample are indicated in Supplementary Data 5). The Marine Atlas of Tara Oceans Unigenes (MATOU) along with functional and taxonomic annotations, unigenes abundances, expression levels and GFs are accessible at <http://www.genoscope.cns.fr/tara/>. Other relevant data are available in this article and its Supplementary Information files, or from the corresponding authors upon request.

Received: 13 October 2017 Accepted: 17 November 2017

Published online: 25 January 2018

References

- Dortch, Q. & Packard, T. Differences in biomass structure between oligotrophic and eutrophic marine ecosystems. *Deep Sea Res.* **36**, 223–240 (1989).
- Gasol, J. M., Giorgio, P. A. D. & Duarte, C. M. Biomass distribution in marine planktonic communities. *Limnol. Oceanogr.* **42**, 1353–1363 (1997).
- Barton, A. D. et al. The biogeography of marine plankton traits. *Ecol. Lett.* **16**, 522–534 (2013).
- Caron, D. A., Countway, P. D., Jones, A. C., Kim, D. Y. & Schnetzer, A. Marine protistan diversity. *Ann. Rev. Mar. Sci.* **4**, 467–493 (2012).
- Wisecaver, J. H. & Hackett, J. D. Dinoflagellate genome evolution. *Annu. Rev. Microbiol.* **65**, 369–387 (2011).
- de Vargas, C. et al. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
- Leray, M. & Knowlton, N. Censusing marine eukaryotic diversity in the twenty-first century. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, 20150331 (2017).
- Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).
- Sunagawa, S. et al. Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
- Brum, J. R. et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
- Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
- Chow, C. E. & Suttle, C. A. Biogeography of viruses in the sea. *Annu. Rev. Virol.* **2**, 41–66 (2015).
- Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
- 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).
- Lenz, P. H. et al. De novo assembly of a transcriptome for *Calanus finmarchicus* (Crustacea, Copepoda)—the dominant zooplankton of the North Atlantic Ocean. *PLoS One* **9**, e88589 (2014).

16. Alexander, H. et al. Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. *Proc. Natl Acad. Sci. USA* **112**, E5972–E5979 (2015).
17. Bertrand, E. M. et al. Phytoplankton-bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge. *Proc. Natl Acad. Sci. USA* **112**, 9938–9943 (2015).
18. Karsenti, E. et al. A holistic approach to marine eco-systems biology. *PLoS Biol.* **9**, e1001177 (2011).
19. Bork, P. et al. Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. *Science* **348**, 873 (2015).
20. Pesant, S. et al. Open science resources for the discovery and analysis of Tara Oceans data. *Sci. data* **2**, 150023 (2015).
21. Yutin, N. & Koonin, E. V. Hidden evolutionary complexity of nucleocytoplasmic large DNA viruses of eukaryotes. *Viol. J.* **9**, 161 (2012).
22. Beja, O. et al. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 1902–1906 (2000).
23. Guo, Z., Zhang, H. & Lin, S. Light-promoted rhodopsin expression and starvation survival in the marine dinoflagellate *Oxyrrhis marina*. *PLoS One* **9**, e114941 (2014).
24. 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).
25. Man, D. et al. Diversification and spectral tuning in marine proteorhodopsins. *Embo. J.* **22**, 1725–1731 (2003).
26. Arendsee, Z. W., Li, L. & Wurtele, E. S. Coming of age: orphan genes in plants. *Trends Plant. Sci.* **19**, 698–708 (2014).
27. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
28. Raven, J. A., Evans, M. C. W. & Korb, R. E. The role of trace metals in photosynthetic electron transport in O₂-evolving organisms. *Photosynth. Res.* **60**, 111–150 (1999).
29. Pierella Karlusich, J. J., Ceccoli, R. D., Grana, 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.* **7**, 750–767 (2015).
30. Lommer, M. et al. Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol.* **13**, R66 (2012).
31. Lima-Mendez, G. et al. Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
32. Guidi, L. et al. Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
33. 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).
34. 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).
35. Follows, M. J., Dutkiewicz, S., Grant, S. & Chisholm, S. W. Emergent biogeography of microbial communities in a model ocean. *Science* **315**, 1843–1846 (2007).
36. Mahowald, N. M. et al. Atmospheric iron deposition: global distribution, variability, and human perturbations. *Ann. Rev. Mar. Sci.* **1**, 245–278 (2009).
37. Behrenfeld, M. J., Boss, E., Siegel, D. A. & Shea, D. M. Carbon-based ocean productivity and phytoplankton physiology from space. *Global Biogeochem. Cycles* **19**, GB1006 (2005).
38. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4**, 170093 (2017).
39. Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
40. Zerbino, D. R., McEwen, G. K., Margulies, E. H. & Birney, E. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One* **4**, e8407 (2009).
41. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086–1092 (2012).
42. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
43. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
44. Mihara, T. et al. Linking virus genomes with host taxonomy. *Viruses* **8**, 66 (2016).
45. Iyer, L. M., Balaji, S., Koonin, E. V. & Aravind, L. Evolutionary genomics of nucleocytoplasmic large DNA viruses. *Virus Res.* **117**, 156–184 (2006).
46. Arslan, D., Legendre, M., Seltzer, V., Abergel, C. & Claverie, J. M. Distant mimivirus relative with a larger genome highlights the fundamental features of megaviridae. *Proc. Natl Acad. Sci. USA* **108**, 17486–17491 (2011).
47. Santini, S. et al. Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc. Natl Acad. Sci. USA* **110**, 10800–10805 (2013).
48. Abergel, C., Legendre, M. & Claverie, J. M. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol. Rev.* **39**, 779–796 (2015).
49. Benamar, S. et al. Faustoviruses: comparative genomics of new megavirales family members. *Front. Microbiol.* **7**, 3 (2016).
50. Shi, M. et al. Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543 (2016).
51. Li, C. X. et al. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife* **4**, e05378 (2015).
52. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
53. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
54. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
55. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
56. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).
57. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
58. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
59. Ciccarelli, F. D. et al. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
60. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).

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), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), and FRANCE GENOMIQUE (ANR-10-INBS-09), 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), ERC Advanced Grant Award (Diatomite: 294823), the LouisD foundation of the Institut de France, a Radcliffe Institute Fellowship from Harvard University to CB, JSPS/MEXT KAKENHI (Nos. 26430184, 16H06437, 16H06429, 16K21723, 16KT0020), The Canon Foundation (No. 203143100025), 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. Computations were performed using the platine, titane and curie HPC machine provided through GENCI grants (t2011076389, t2012076389, t2013036389, t2014036389, t2015036389, and t2016036389). This article is contribution number 62 of Tara Oceans.

Author contributions

E.P. and P.W. designed the study. P.W. wrote the paper with substantial input from C.B., E.P., Q.C., M.B.S., P.H., J.R., L.G., S.Su., P.B., E.K., H.O., C.d.V., and D.I. S.R., F.N., C.D., M.P., S.K.L., S.Se., and S.P. collected and managed Tara Oceans samples. J.P., A.A., and K.L. coordinated the genomic sequencing. Q.C., E.P., C.D.S., Y.S., A.K., R.B.M., G.L.M., G.Y., F.R., L.T., A.K., A.B., S.E., M.A.M., D.R., O.J., J.M.A., H.O., D.I., C.B. analyzed the genomic data. D.I., L.G. 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/s41467-017-02342-1>.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party

© The Author(s) 2018

Quentin Carradec^{1,2,3}, Eric Pelletier^{1,2,3}, Corinne Da Silva¹, Adriana Alberti¹, Yoann Seeleuthner^{1,2,3}, Romain Blanc-Mathieu⁴, Gipsi Lima-Mendez^{5,6,21,22}, Fabio Rocha⁷, Leila Tirichine⁷, Karine Labadie¹, Amos Kirilovsky^{1,2,3,7}, Alexis Bertrand¹, Stefan Engelen¹, Mohammed-Amin Madoui^{1,2,3}, Raphaël Méheust⁷, Julie Poulain¹, Sarah Romac^{8,9}, Daniel J. Richter^{8,9}, Genki Yoshikawa⁴, Céline Dimier^{7,8,9}, Stefanie Kandels-Lewis^{10,11}, Marc Picheral¹², Sarah Searson¹³, Tara Oceans Coordinators, Olivier Jaillon^{1,2,3}, Jean-Marc Aury¹, Eric Karsenti^{7,11,12}, Matthew B. Sullivan¹⁴, Shinichi Sunagawa^{10,15}, Peer Bork^{10,16,17,18}, Fabrice Not^{8,9}, Pascal Hingamp¹⁹, Jeroen Raes^{5,6}, Lionel Guidi^{12,13}, Hiroyuki Ogata⁴, Colomban de Vargas^{8,9}, Daniele Iudicone²⁰, Chris Bowler⁷ & Patrick Wincker^{1,2,3}

¹CEA - Institut de Biologie François Jacob, Genoscope, Evry, 91057, France. ²CNRS UMR Metabolic Genomics, Evry, 91057, France. ³Univ Evry, Evry, 91057, France. ⁴Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. ⁵Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, Leuven, 3000, Belgium. ⁶VIB Center for Microbiology, Herestraat 49, Leuven, 3000, Belgium. ⁷Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, Paris, F-75005, France. ⁸CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, 29680, France. ⁹Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, 29680, France. ¹⁰Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, Heidelberg, 69117, Germany. ¹¹Directors' Research European Molecular Biology Laboratory, Meyerhofstr. 1, Heidelberg, 69117, Germany. ¹²Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'océanographie de Villefranche (LOV) Observatoire Océanologique, Villefranche-sur-Mer, 06230, France. ¹³Department of Oceanography, University of Hawaii, Honolulu, 96844 Hawaii, USA. ¹⁴Departments of Microbiology and Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH 43210, USA. ¹⁵Department of Biology, Institute of Microbiology, Vladimir-Prelog-Weg 4, Zürich, 8093, Switzerland. ¹⁶Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, Heidelberg, 69120, Germany. ¹⁷Max Delbrück Centre for Molecular Medicine, Berlin, 13125, Germany. ¹⁸Department of Bioinformatics, University of Würzburg, Würzburg, 97074, Germany. ¹⁹Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, 13284, France. ²⁰Stazione Zoologica Anton Dohrn, Villa Comunale, Naples, 80121, Italy. ²¹Present address: Cellular and Molecular Microbiology, Faculté des Sciences, Université Libre de Bruxelles (ULB), Belgium. ²²Present address: Interuniversity Institute for Bioinformatics in Brussels, ULB-VUB, Boulevard du Triomphe CP 263, 1050 Brussels, Belgium. Quentin Carradec and Eric Pelletier contributed equally to this work.

Tara Oceans Coordinators

Silvia G. Acinas²³, Emmanuel Boss²⁴, Michael Follows²⁵, Gabriel Gorsky¹², Nigel Grimsley^{26,27}, Lee Karp-Boss²⁴, Uros Krzic²⁸, Stephane Pesant^{29,30}, Emmanuel G. Reynaud³¹, Christian Sardet^{12,32}, Mike Sieracki³³, Sabrina Speich^{34,35}, Lars Stemmann¹², Didier Velayoudon³⁶ & Jean Weissenbach^{1,2,3}

²³Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC), E-08003 Barcelona, Catalonia, Spain. ²⁴School of Marine Sciences, University of Maine, Orono, ME 04469, USA. ²⁵Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, 02138 MA, USA. ²⁶CNRS UMR 7232, BIOM, Avenue du Fontaulé, Banyuls-sur-Mer, 66650, France. ²⁷Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, Banyuls-sur-Mer, 66650, France. ²⁸Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstrasse 1, Heidelberg, 69117, Germany. ²⁹MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, 28359, Germany. ³⁰PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, 28359, Germany. ³¹Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland. ³²CNRS, UMR 7009 Biodev, Observatoire Océanologique, Villefranche-sur-mer, F-06230, France. ³³National Science Foundation, Arlington, VA 22230, USA. ³⁴Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, Plouzané, 29820, France. ³⁵Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, Paris Cedex 05, 75231, France. ³⁶DVIP Consulting, Sèvres, 92310, France

Titre : Rôle des protistes hétérotrophes marins dans le cycle du carbone océanique par génomique en cellule unique.

Mots clés : protistes hétérotrophes, straménopiles, génomique en cellule unique, écologie marine

Résumé : Les eucaryotes unicellulaires (protistes), ont des rôles extrêmement importants dans les cycles biogéochimiques des océans. Tout d'abord, bien qu'en moyenne moins abondants que les cyanobactéries, les protistes photosynthétiques représentent une grande part de la production primaire nette. Étant relativement plus faciles à mettre en culture, les organismes photosynthétiques sont relativement plus étudiés et leurs génomes représentent une grande fraction des génomes de protistes marins disponibles dans les bases de données. En revanche, les organismes hétérotrophes demandent un travail beaucoup plus important pour la mise en culture et sont par conséquent beaucoup moins bien connus. De plus, la majorité des génomes de protistes hétérotrophes séquencés à ce jour concernent des organismes d'intérêt pour l'homme (parasites de plantes, organismes pathogènes), mais le choix de ces organismes comme modèles d'étude ne reflète pas leur intérêt écologique.

L'objectif de cette thèse est d'étudier par une approche de génomique en cellule unique certains pico- nanoeucaryotes hétérotrophes, fortement abondants dans les océans ouverts et encore non cultivés à ce jour. Pour cela, un protocole de séquençage, d'assemblage et d'annotation de génome à partir de cellules uniques a été mis en place.

Le génome de sept lignées de straménopiles a été partiellement reconstruit et annoté, permettant de confirmer de façon robuste la phylogénie des straménopiles obtenue par les marqueurs ribosomiques, mais surtout de formuler des hypothèses quant à leur spécialisation en termes de mobilité ou de mode trophique.

En particulier, la comparaison des répertoires de gènes permettant la dégradation des carbohydrates indique des régimes alimentaires probablement différents chez les organismes étudiés.

Par ailleurs, l'utilisation combinée de ces génomes et des séquences métagénomiques de l'expédition *Tara Oceans* a permis de décrire la distribution géographique de ces organismes ainsi que la distance génétique des populations à nos génomes de références. La corrélation de ces distributions avec les paramètres environnementaux mesurés aux points d'échantillonnages montrent que la température est un facteur clé de la distribution de ces micro-organismes. De plus, l'utilisation des données métatranscriptomiques de l'expédition nous a permis de distinguer différents profils d'expression – correspondant potentiellement à différents états physiologiques – chez la lignée la plus cosmopolite étudiée.

En conclusion, cette thèse montre qu'il existe une forte diversité génomique à explorer chez les protistes marins hétérotrophes, permettant notamment d'émettre des hypothèses sur leurs modes trophiques. Elle démontre également l'intérêt de la génomique en cellule unique, en particulier sa complémentarité avec les approches métagénomiques et métatranscriptomiques pour la compréhension exhaustive des écosystèmes marins.

Title: Role of heterotrophic marine protists in the oceanic carbon cycle using single-cell genomics

Keywords: heterotrophic protists, stramenopiles, single-cell genomics, marine ecology

Abstract: Unicellular eukaryotes (protists) have important roles in the biogeochemical cycles of the ocean. First, although on average less abundant than cyanobacteria, photosynthetic protists account for a large proportion of net primary production. Since they are relatively easier to culture, photosynthetic organisms are relatively more studied and their genomes represent a large fraction of the genomes of marine protists available in databases. On the other hand, heterotrophic organisms require much more work for cultivation and are therefore much less well known. Moreover, the majority of genomes of heterotrophic protists sequenced to date concern organisms of interest to humans (plant pests, pathogenic organisms), but the choice of these organisms as study models does not reflect their ecological interest.

The objective of this thesis is to study, using a single-cell genomic approach, several heterotrophic pico- nanoeucaryotes, that are highly abundant in the open oceans and have not yet been cultivated. For this purpose, a protocol for sequencing, assembling and annotation of genomes from single cells has been developed.

The genomes of seven stramenopile lineages have been partially reconstructed and annotated, making it possible to confirm in a robust way the phylogeny of stramenopiles obtained by ribosomal markers, and, more important, to formulate hypotheses regarding their specialization in terms of mobility or trophic mode.

Particularly, the comparison of gene repertoires of carbohydrate degradation indicates likely different food spectra in the studied organisms.

In addition, the combined use of these genomes and metagenomic sequences from the *Tara* Oceans expedition made it possible to describe the geographical distribution of these organisms as well as the genetic distance between environmental populations and our reference genomes. The correlation of these distributions with the environmental parameters measured at the sampling points shows that temperature is a key factor in the distribution of these microorganisms. In addition, the use of metatranscriptomic data from the expedition allowed us to distinguish different expression profiles - potentially corresponding to different physiological states - in the most cosmopolitan lineage studied.

In conclusion, this thesis shows that there is a strong genomic diversity to be explored in heterotrophic marine protists, allowing us to make hypotheses about their trophic modes. It also demonstrates the value of single-cell genomics, in particular its complementarity with metagenomic and metatranscriptomic approaches for the comprehensive understanding of marine ecosystems.