Contribution of meta-omics data in the detection and quantification of functional traits in planktonic ecosystems

Emile Faure

SORBONNE UNIVERSITÉ

ECOLE DOCTORALE 129 SCIENCES DE L'ENVIRONNEMENT

INSTITUT DE SYSTÉMATIQUE, EVOLUTION, BIODIVERSITÉ - UMR 7205

LABORATOIRE D'OCÉANOGRAPHIE DE VILLEFRANCHE - UMR 7093

Apport des données méta-omiques dans la détection et la quantification des traits fonctionnels au sein des écosystèmes planctoniques

PRÉSENTÉE POUR OBTENIR LE GRADE DE

DOCTEUR DE SORBONNE UNIVERSITÉ PAR

EMILE FAURE

Directrice : Sakina-Dorothée Ayata

Co-directrice : Lucie Bittner

Devant un jury composé de :

SEBASTIEN MONCHY	Professeur, Université du Littoral Côte d'Opale	Rapporteur
CHRISTIAN TAMBURINI	DR CNRS, Mediterranean Institute of Oceanography	Rapporteur
THOMAS MOCK	Professor, University of East Anglia	Examinateur
INGRID OBERNOSTERER	DR CNRS, Sorbonne Université	Examinatrice
SAKINA-DOROTHÉE AYATA	MCU, Sorbonne Université	Co-directrice
LUCIE BITTNER	MCU, Sorbonne Université	Co-directrice

Soutenance prévue à Paris le 02 Décembre 2020

Abstract

Marine planktonic organisms play a crucial role in trophic networks, global biogeochemical cycles, and climate regulation. Biogeochemical models simulate planktonic ecosystems dynamics to understand and predict climate change. In most biogeochemical models, planktonic diversity is implemented either through plankton functional types (PFT), *i.e.* theoretical entities grouping planktonic organisms according to shared functional capacities (*e.g.*, calcifiers, nitrogen fixers or silicifiers), or functional traits, *i.e.* morphological, physiological or phenological features measurable at the individual level that effect growth, reproduction or survival (*e.g.* feeding modes, production of toxins or body size). These methods imply an *a priori* and restricted choice of the considered types or traits of planktonic organisms, potentially leading to oversimplified representations of planktonic diversity in models. Unprecedented amounts of meta-omics data on marine planktonic communities were recently collected at global scales, calling for the use of data-driven methodologies to determine and quantify the potential and realized functional traits of planktonic organisms *in-situ*. My objective in this thesis was therefore to determine how to use meta-omics data to quantify the distribution of functional traits in the environment. In a first part, I present how omics data can be used to describe and quantify specific, *a priori* selected traits in the global ocean. A particular focus is made on two functional traits: mixotrophy, from which the genomic basis is poorly known, and dimethyl sulfide (DMS) production, from which the genomic basis is well understood. I show how metabarcoding data on one hand and functional genomic markers on the other hand allow to decipher the biogeography of functional traits, identifying limits and advantages of the two types of data. In a second part, I present an approach allowing to detect putative protein families in metagenomics data that can be associated with functional traits, without any *a priori* choice of functional traits of interest. By quantifying the response of these emergent clusters to physico-chemical gradients in the global ocean, I show how this approach could allow to predict the functional composition of planktonic communities from environmental data in the near future. Finally, I use my results to discuss the potential of meta-omics data as a means of realistically representing the diversity of planktonic communities in biogeochemical models.

Keywords: Omics; Biogeochemistry; Functional traits; Plankton; Biogeography.

Résumé

Les organismes planctoniques marins jouent un rôle crucial dans les réseaux trophiques, les cycles biogéochimiques globaux et la régulation du climat. Les modèles biogéochimiques simule la dynamique des écosystèmes planctoniques pour comprendre et prédire le changement climatique. Dans la plupart de ces modèles, la diversité planctonique est représentée soit sous forme de types fonctionnels planctoniques (PFT), *i.e.* par des entités théoriques classant les organismes planctoniques selon leurs capacités fonctionnelles (*e.g.* organismes calcifiants, fixateurs d'azote, ou silicifiants), ou bien de traits fonctionnels, *i.e.* des caractéristiques morphologiques, physiologiques ou phénologiques mesurables au niveau individuel qui affectent la croissance, la reproduction ou la survie (*e.g.* modes trophiques, production de toxines ou taille corporelle). Un choix *a priori* et restreint des types planctoniques ou traits fonctionnels considérés est donc nécessaire, pouvant conduire à des représentations simplifiées de la diversité planctoniques dans les modèles. Des quantités inédites de données méta-omiques sur les communautés planctoniques ont récemment été collectées à l'échelle de l'océan global, appelant à l'utilisation de méthodes permettant de déterminer et quantifier les traits fonctionnels potentiels et réalisés des organismes planctoniques à partir de ces données *in-situ*. Mon objectif au cours de cette thèse fut donc de déterminer comment utiliser les données méta-omiques afin de quantifier la distribution de traits fonctionnels dans l'environnement. Dans une première partie, je présente comment les données omiques peuvent être utilisées pour décrire et quantifier dans l'océan global des traits spécifiques, choisis *a priori*. Deux traits fonctionnels sont utilisés en exemple: la mixotrophie, dont la base génomique est mal connue, et la production de diméthylsulfure (DMS), dont la base génomique est relativement bien étudiée. Je montre comment les données de métabarcoding d'une part et des marqueurs génomiques fonctionnels d'autre part permettent de décrire la biogéographie des traits fonctionnels, en identifiant les limites et les avantages des deux types de données. Dans une deuxième partie, je présente une approche permettant de faire émerger des familles protéiques putatives pouvant être associés à des traits fonctionnels au sein des données de métagénomique, sans choix *a priori* de traits fonctionnels d'intérêt. En quantifiant la réponse de ces familles aux gradients physico-chimiques dans l'océan global, je montre comment cette approche pourrait permettre de prédire la composition fonctionnelle des communautés planctoniques à partir de données environnementales dans un avenir proche. Enfin, j'utilise mes résultats pour discuter du potentiel des données méta-omiques comme moyen de représenter de manière réaliste la diversité des communautés planctoniques dans les modèles biogéochimiques.

Mots-clefs: Omiques; Biogéochimie; Traits Fonctionnels; Plancton; Biogéographie.

L'homme et la mer

Homme libre, toujours tu chériras la mer !
La mer est ton miroir ; tu contemples ton âme
Dans le déroulement infini de sa lame,
Et ton esprit n'est pas un gouffre moins amer.

Tu te plais à plonger au sein de ton image ;
Tu l'embrasses des yeux et des bras, et ton coeur
Se distrait quelquefois de sa propre rumeur
Au bruit de cette plainte indomptable et sauvage.

Vous êtes tous les deux ténébreux et discrets :
Homme, nul n'a sondé le fond de tes abîmes ;
Ô mer, nul ne connaît tes richesses intimes,
Tant vous êtes jaloux de garder vos secrets !

Et cependant voilà des siècles innombrables
Que vous vous combattez sans pitié ni remord,
Tellement vous aimez le carnage et la mort,
Ô lutteurs éternels, ô frères implacables !

Charles Baudelaire

Figure 1 – Picture taken by Joana Roussillon on board of the Atalante during the MOOSE 2018 expedition, on which I had the great pleasure to participate.

■

Acknowledgements

First of all, I would like to deeply thank the members of the jury for accepting to evaluate my work, especially the reviewers Sébastien Monchy and Christian Tamburini, but also of course Ingrid Obernosterer and Thomas Mock. I sincerely hope that you will appreciate reading this manuscript, and I am looking forward to our discussions about it in December!

Je vais maintenant passer au français pour remercier Sakina et Lucie, sans qui rien de tout cela n'aurait été possible. Sakina, nous nous sommes rencontrés pour la première fois en 2013, lors de ma première année de licence à Roscoff. Je débarquais de Paris avec pour seul et unique but de me diriger vers la biologie marine, et les maths n'étaient alors vraiment pas ma priorité. Tes cours de modélisation ont été un vrai moment charnière pour moi, m'ouvrant les yeux sur la beauté des maths appliquées au monde du vivant. Effectuer un stage de master 2 avec toi est donc apparu comme une évidence, alors même que je m'étais engagé pour un projet complètement différent au moment de recevoir ta proposition de stage. C'est à ce moment que j'ai fais la rencontre de Lucie, qui a fini de me convaincre de faire ce stage, ce que je n'ai jamais regretté un seul instant. Grâce à votre encadrement attentionné, humain et valorisant, j'ai eu l'impression durant ces trois ans et demi que tout tournait comme sur des roulettes, sans accrocs. Quand l'une attendait un heureux évènement (dédicace à Cassiopée, Atlas et Solen), l'autre prenait le relais, tel une équipe de basket à la mécanique bien huilée passant de la zone à l'indiv' sans temps mort. Votre confiance en moi, votre intérêt pour les travaux de cette thèse mais également pour mon bien être personnel ont énormément compté pour moi. Je ne vous remercierai jamais assez pour ces trois ans de thèse, si c'était à refaire, je le referais.

Une des valeurs prônée dans votre encadrement ayant été la collaboration et le partage des connaissances, j'ai eu la chance au cours de ma thèse de croiser le chemin de nombreux chercheurs, que cela soit au cours d'une des nombreuses conférences que vous m'avez permis de faire, ou au cours de visites de collaboration. J'aimerais notamment remercier Fabrice Not, qui m'a accueilli au début de ma thèse pour une semaine sans laquelle la première partie de cette thèse n'aurait pas été la même. Je remercie également Olivier

Aumont, qui m'a accueilli dès mon stage de master 2 afin de me former à l'utilisation de PISCES, puis qui a permis à mon projet de prendre en épaisseur lors de mes comités de thèse. J'aimerais aussi remercier Loïs Maignien, que j'ai pu rencontrer à Brest lors de la conférence EBAME, qui m'a ouvert les yeux sur le potentiel de l'écogénomique, et qui a grandement influencée la direction de ma deuxième partie de thèse de part ses conseils lors de mes comités. Enfin, je me dois de remercier Eric Pelletier, avec qui j'ai de nombreuses fois discuté en conférences et lors de mes comités, pour sa bienveillance et son intérêt pour mes travaux, qui m'ont beaucoup aidé à avancer.

J'aimerais maintenant remercier les collègues doctorants qui m'ont accompagnés au cours de ces trois années passées entre Jussieu et le MNHN. J'ai commencé cette aventure entouré d'Anne-Sophie et Arnaud, dans un petit bureau de l'immonde bâtiment B, dans lequel nous a ensuite rejoins Ophélie. Un énorme merci à vous trois pour avoir fait vivre notre petite équipe. J'aimerais également remercier les copains d'évolution Paris Seine, notamment Gab, Juliette, Thomas et Romain. Ma première année de thèse a été marquée par les pauses déjeuner Cocherel, les SCEP, les pauses thé chez Gab, les films de thèse que j'ai eu l'immense honneur de concocter avec vous,. . . J'aurais aimé que l'on se retrouve tous au muséum, et je suis heureux d'avoir pu garder contact avec certains d'entre vous.

Au muséum, j'ai retrouvé mon acolyte Elise, que j'ai rencontré pour la première fois à Roscoff il y a maintenant 8 ans, avant de la retrouver en master, puis en thèse pour partager le même bureau (je pense que c'est plutôt toi qui me copie que l'inverse, mais en même temps t'étais au museum avant alors y'a débat, j'avoue). Non contente d'être toujours la première pour organiser des trucs cools (quitte a parfois participer aux AG horribles et autres CU, l'art du sacrifice ça te connait), je pense qu'on peut t'attribuer le prix du meilleur public de France, ce qui fait de toi l'atout numéro uno de tout labo qui se respecte. Dommage que tu aies dû aller au collège de France 50% du temps, l'ABI c'est mieux quand t'es là. Tâche de faire ton comité un de ces quatre quand même. Au muséum j'ai également fait la rencontre de Martin, ce qui n'a pas du tout arrangé ni ma pédalite aiguë, ni mon amour pour la bouffe. Avoir quelqu'un au bureau chaque jour qui partage des passions communes, ça énerve parfois les autres (coucou Elise), mais ça permet de s'échapper dans des discussions entièrement hors boulot, et ça n'a pas de prix (contrairement aux pédales sus-citées), alors merci Martin. Je remercie également l'ensemble de l'ABI pour son accueil exceptionnel, son ambiance détendue, l'atmosphère d'entraide qui y règne, et la place de choix qui y est faite aux jeunes. Merci Sophie, Mathilde, Joël, Karen, Martine, Guillaume A., Eric, Henri, Guillaume S., et tout ceux que j'ai pu oublier, j'en suis d'avance désolé.

Je ne peux pas m'arrêter dans les remerciements des collègues sans citer l'équipe BioNano de Montpellier, qui aura été mon deuxième (ou troisième, voir quatrième si on compte les déménagements) labo au cours de cette thèse. Bien sûr, nous étions loin scientifiquement, mais le fait de me laisser un bureau et de m'accueillir aussi chaleureusement au sein de votre équipe m'a permis de vivre plus sereinement ma thèse, en m'aidant à passer du temps avec Jeanne tout en avançant dans mes recherches. Merci tout particulièrement à Christophe, Rahima, Maïda et Marion, qui sont devenus de vrais amis au fil de ces trois ans, j'ai déjà hâte à la prochaine grosse fête chez Kiki.

J'aimerais maintenant dire quelques mots de remerciements pour Aurélie et Nina, qui ont travaillé avec moi lors de leurs stages de M2 et M1. Aurélie, j'ai été impressionné par ta force, dans les moments durs que tu as dû traverser durant la période du stage, tu n'as jamais abandonné et tu as réussi à réaliser un travail que je suis fier d'avoir encadré. Nina, en seulement quelques mois au labo tu as produit tellement, merci pour ton enthousiasme débordant, je suis sûr qu'il t'emmènera loin.

Je vais maintenant passer aux copains ! Merci Duduche pour ces années de master et de thèses, personne n'allie comme toi conneries et amour pour la science, tu m'as tiré vers le haut scientifiquement avec tes 15 papiers, et humainement avec ta bonne humeur perpétuelle, t'es grand, t'es très grand. Merci Mathieu, équipier de luxe de cette aventure, on n'a pas été assez surfer mais ça nous a pas empêché de rigoler, de grimper, de randonner,... tu peux venir en post-doc à Brest après ta thèse stp ? Merci à Bastoon la fripouille, frelon parmi les frelons, abonné France Football de toujours, spécialiste du PSG s'il en est, t'es le mec le plus drôle de France et de Navarre sous tes airs de radiologue Lillois, l'oublie pas. Challah on voit Paris soulever la coupe ensemble un de ces quatre, et j'en profite d'ailleurs pour remercier les bleus pour avoir ramené la deuxième étoile à la maison. Merci à Pinol d'être redevenu mon copaing après quelques années de pause, être deux bobos parisiens dans la bande c'était royal, vamos Annie Dingo, Paris à vélo on est là. Merci à Bertino, le roi de la grotte, le Dimitri Payet de la team chill, le Zinédine des guilands, le Messi de la mauresque, en trois ans tu n'as fais que marquer des points Marseillais dans mon coeur parigo, un peu aidé par l'ami Jules, probablement en train de pédaler quelque part à l'heure qu'il est. Un énorme et sec merci à Magrino, le chopeur du Corbeau, on ne compte plus le nombre de malaises que tu as installé depuis le début de cette thèse (les physiciens qui lisent je vous vois déjà débattre de l'ordre de grandeur), et c'est comme ça qu'on t'aime ne change rien. Merci Elsa, semi-marathonienne en cheffe (même si on sait tous que t'as un marathon dans chaque jambe), futur détentrice du triple Ventoux challenge, j'espère que je serai là pour le faire avec toi ! Merci Tadpsa, chamois de la montagne, on en serait presque à instaurer des contrôles anti-dopage en

fin de rando pour vérifier que tu nous la fais pas à la Armstrong, j'attend avec impatience le jour ou je te boufferai en côte ! Merci à Raph, entraîneur adjoint de la team chill qui parfois fait même de l'ombre à Coach Bertin. Merci à Matthieu et Camille, pour avoir été l'attache roscovite précieuse de ces trois dernières années. Merci à Pepe, Florentin et Adri, cet EBAME de folie restera dans les mémoires et j'espère qu'ISME n'était que la première chorizos reunion d'une longue liste ! Merci à tout le quart de l'ambiance, Maxime, Amélie, Joana, Marylou, Armelle, Anne-So et Paul, cette campagne MOOSE 2018 restera comme un grand moment de cette thèse, faut qu'ça pègue ! Et merci enfin à toutes celles et ceux que je n'ai pas nommé ici, ou que j'aurais pu oublié, promis si c'est le cas je vous paie un coup !

Merci à Romain, Félix et Tom, vous méritez un paragraphe à vous tout seul mes trois fantastiques. Merci à Romain pour avoir continué sans relâche à organiser des trucs pendant ses trois ans ou je n'ai pas eu beaucoup de temps à vous donner. Merci Félix pour nous avoir contacté à chaque retour sur Paris, j'aurai vraiment aimé pouvoir venir te voir plus souvent à Berlin, j'espère que j'aurais l'occasion de me rattraper. Enfin merci à Tom, le joaillier de la mif, je suis trop fier de ce que tu deviens, j'ai hâte de vous inviter à Brest !

Merci ensuite aux membres du Karaba F.C., j'ai nommé Clément Léost, Jordan Aguilar, et Olivier Cotro ! Vous aussi vous avez le droit à votre paragraphe, tout simplement parce que je pense être encore plus fier de ce que j'ai pu faire avec vous cette année que de ce que j'ai fais au cours de cette thèse. La musique compte énormément pour moi (Figure 2), et vous m'avez permis de réaliser un rêve. Malgré le foutu covid qui nous prive de concert, chaque répète avec vous a été la chose que j'attendais le plus de ma semaine. Pour couronner le tout, en plus d'être des compagnons de groupes parfaits, vous êtes devenus des sacrés potes, des karaboys de première. La rouylease party de l'EP sponsorisée par Karmeliet et le Val café s'annonce incroyable, et j'espère vraiment que ce n'est que le début.

Passons maintenant à la famille ! Tout d'abord, un énorme merci à Léon. Vivre avec toi pendant ses trois ans m'a permis de réaliser à quel point tu avais grandi depuis mon départ pour Roscoff, et ça m'a fait plaisir qu'on puisse renouer des liens Faure (lol) après plusieurs années à vivre éloignés. Tu vas maintenant commencer ta thèse, et je suis sûr que tu vas cartonner, je suis hyper fier de toi. Merci également à Paul, que j'aurais aimé pouvoir voir plus souvent au cours de ces trois ans, surtout vu la vitesse à laquelle tu grandis ! Merci à Mathilde, Sélim, Camille et Olivier, j'espère qu'on pourra se voir tous ensemble ailleurs que sur zoom bientôt !

Figure 2 - 100 music albums that marked my PhD. The list is organized in 4 tier-lists, from extremely heavy rotations to reasonably heavy rotations. In each tier-list, no particular order is respected.

Merci à ma mère et Gilles, sans qui rien de tout ça ne serait arrivé. Vous m'avez appris à aimer la nature, la science, à être étonné et curieux du monde vivant, et c'est grâce à vous que j'ai été jusqu'ici. Merci à mon père, qui je pense comme moi est sans doute aussi fier des singles du Karaba F.C. que du fait que je devienne docteur, et c'est très bien comme ça ! J'ai hâte qu'on puisse refaire des concerts ensemble. Merci à ma mamie, pour ton soutiens sans faille depuis toujours, je suis fier que tu me voies devenir docteur. Merci à toute ma famille montpelliéraine, Mivette, tatine, tonton, Léda, Lucas, Marek, Violette, et maintenant Stéphane ! Merci pour votre accueil à chaque passage, et pour votre soutien. Merci enfin à Marie B, Gilles, Léa, Elie et Noémie pour avoir été une deuxième famille pour moi depuis maintenant 8 ans.

Ce qui m'amène enfin à Jeanne, par qui je me devais de finir ces remerciements. Merci d'avoir été là pour moi pendant ces trois ans, malgré la distance, malgré le stress de la thèse, malgré ma fâcheuse tendance à ne pas penser aux billets TGVmax, et à passer une heure par jour sur eurosport. Merci d'avoir illuminé mes weeks-ends, merci pour les voyages incroyables qu'on a fait, merci de m'avoir suivi dans mes délires de musique, de bières crafts, de vélo, de rando,... finalement seule l'escalade t'auras résisté ! Enfin merci d'avoir pris Mogwaï sous ton aile, elle nous aura fait du bien à tous les deux.
Sans toi je n'en serais pas là.

Contents

List of Figures

List of Tables

Chapter 1

General introduction

1.1 The role of marine plankton diversity in global biogeochemical cycles

1.1.1 Marine plankton diversity

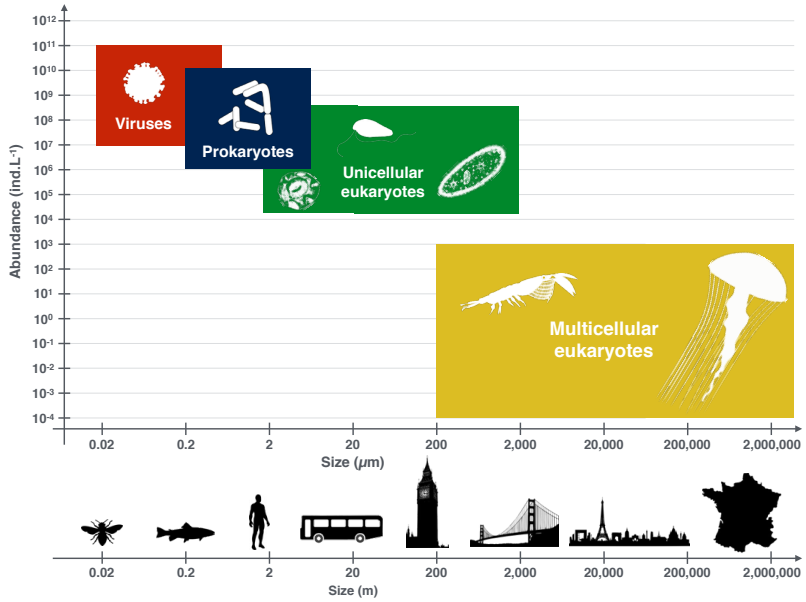

Figure 1.1 - Range of body size and abundance in logarithmic scales for viruses, prokaryotes, unicellular eukaryotes (commonly called protists) and multicellular eukaryotes (including metazoans and macro-algae). A scale of macroscopic references was added to help apprehend the extent of size diversity among planktonic organisms. Size and abundance ranges were extracted from Pesant et al. (2015), but the size scale of pluricellular eukaryotes was extended to include larger organisms like jellyfish or sargass that can reach over a meter in length.

The word *plankton* is derived from the ancient Greek *πλαγκτοζ* ("planktos"), meaning *wanderer*. It was used for the first time by Viktor Hensen at the end of the XIXth century, to describe the animals and plants floating in the see (Hensen, 1887). The term plankton now regroups all the living organisms that can not swim against the currents, in oppo-

sition to the *nekton*, corresponding to the living organisms swimming freely across water masses. Planktonic organisms range from a few nanometers to a few meters in size, and include viruses, archaea, bacteria, as well as unicellular and multicellular eukaryotes ((Pesant et al., 2015), Figure 1.1).

The *bacterioplankton* gather planktonic prokaryotic cells, whose abundance reaches about 10^{30} in the ocean subsurface (Whitman et al., 1998). It includes a vast diversity of organisms spanning across two domains of life (Figure 1.2). The bacterioplankton includes heterotrophic prokaryotes (see Box 1), that are responsible for the recycling of organic matter into inorganic nutrients in the ocean (Ducklow, 1999). It also includes small autotrophic prokaryotes (see Box 1), like *Prochlorococcus*, which was estimated to be the most abundant photosynthetic organism on earth (Partensky et al., 1999). Autotrophic species of the bacterioplankton are sometimes found in symbiosis (see Box 1) with eukaryotic organisms, for example the cyanobacteria *Synechococcus* is found in symbiosis with dinoflagellate genus like *Ornithocercus*. Such ecological relationships between planktonic organisms play key roles in marine ecosystems' dynamics (Worden et al., 2015) (Figure 1.3).

Box 1: Ecological roles and trophic modes

Autotrophs: Organisms that produce organic matter from inorganic substances.

Phototrophs: Autotrophic organisms producing organic matter from photosynthesis, using light as an energy source and inorganic nutrients.

Heterotrophs: Organisms that depend on pre-formed organic matter for nutrition.

Mixotrophs: Organisms that are both capable of auto- and heterotrophy.

Parasites: Organisms feeding strictly at the expense of an host organism, permanently or during a phase of its life cycle.

Parasitoids: Parasites that *always* kill their host.

Symbiosis: Close, prolonged association between two or more different organisms of different species that may, but does not necessarily, benefit each member.

Symbiont: Organism in a symbiotic relationship.

Phagocytosis: Nutrition mode involving the engulfment and digestion of particulate matter.

Primary production: Production of organic matter by autotrophs.

Secondary production: Production of organic matter by heterotrophs.

Among planktonic eukaryotes, the most abundant group of organisms is usually considered to be the copepods, crustaceans that are sometimes even described as the most abundant multicellular animal on earth (Bron et al., 2011). Marine planktonic unicellular eukaryotes (or protists) constitute the majority of lineages across the eukaryotic tree of life (Worden et al., 2015) (Figure 1.2). Historically, planktonic eukaryotes have been divided in two compartments based on their trophic regime. Phytoplankton corresponds to photosynthetic organisms, in opposition to the zooplankton, corresponding to heterotrophic organisms. Phytoplanktonic organisms are mostly unicellular and represent only 1% of the photosynthetic biomass on earth, but are responsible for 45% of the global net pri-

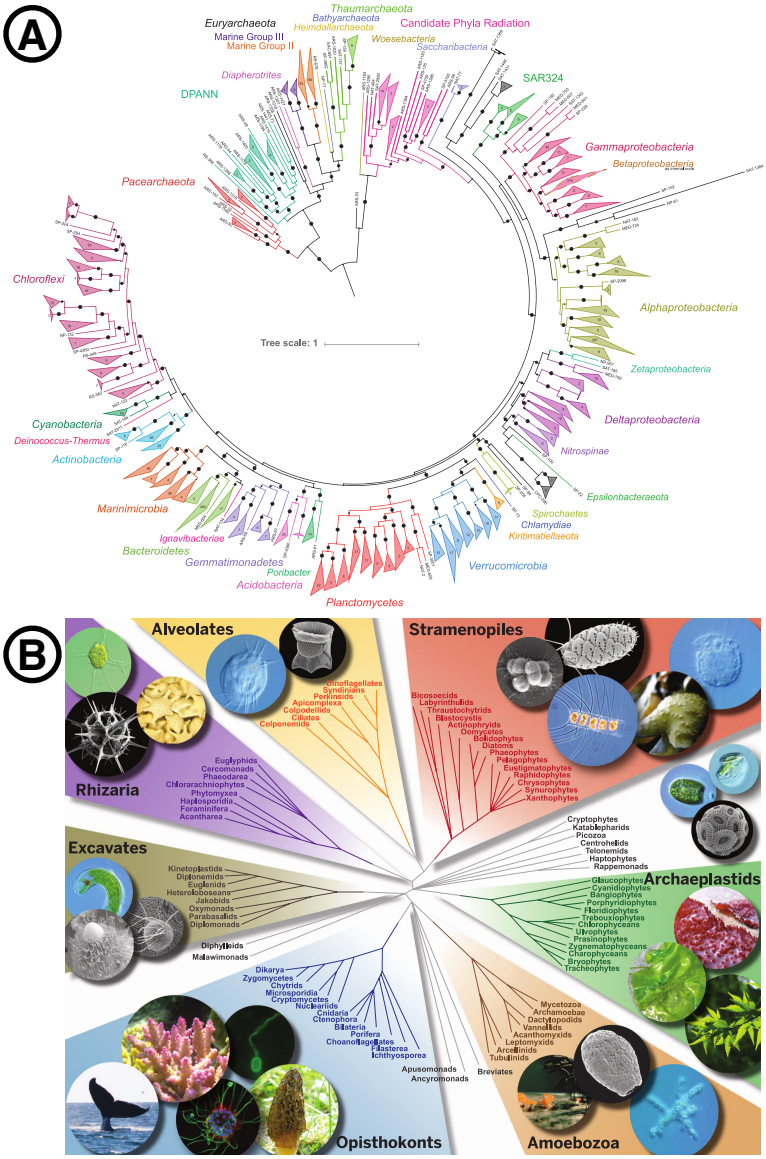

Figure 1.2 - Phylogenetic diversity of prokaryotic (A) and eukaryotic (B) marine plankton. (A) Maximum-likelihood phylogenetic tree based on 16 concatenated single-copy phylogenetic markers extracted from 2,631 re-assembled genomes of prokaryotic plankton (Tully et al., 2018). These genomes were assembled from 234 marine metagenomic samples from the Tara Oceans expedition (details on metagenome-assembled genomes are available in section 1.3.2). The black nodes on branches indicate bootstrap values >0.75, the color of polygons on branches extremities is coding for the phyla while their size is proportional to the number of genomes associated to the corresponding branch. Figure from Tully et al. (2018). (B) Diversity tree of eukaryotic plankton synthesizing information from morphological, phylogenetic (comparisons between a few marker genes from a large diversity of organisms) and phylogenomic (comparisons between parts of genomes or even full genomes of a large diversity of organisms) criteria. Seven "supergroups" are highlighted in colour, and pictures on the sides illustrate the morphological diversity of eukaryotes. Clockwise from right: archaeplastids (rhodophytes, chlorophytes, streptophytes); amoebozoa (tubulinids, arcellinids, mycetozoans); opisthokonts (fungus, microsporidians, choanoflagellates, cnidarians, bilaterians); excavates (parabasalians, oxymonads, euglenids); rhizaria (acantharians, foraminiferans, chlorarachniophytes); alveolates (ciliates, dinoflagellates); stramenopiles (labyrinthulids, synurophytes, diatoms, phaeophytes, actinophryids); unassigned [cryptomonads, katablepharids, haptophytes]. Figure from (Worden et al., 2015).

mary production (see Box 1, Field et al. (1998); Falkowski et al. (2004)). Zooplankton includes a wide diversity of organisms, from unicellular heterotrophs to jellyfish, passing by small crustaceans and planktonic larvae of nektonic organisms. They play a key role in marine ecosystems as grazers of the phytoplankton (Figure 1.3), and hence as secondary producers in marine food webs (See Box 1, Calbet et al. (2001)). However, this dichotomy modeled on the one observed between embryophytes and metazoans in terrestrial ecosystems is only poorly representing the diversity of trophic modes found among marine planktonic organisms (Flynn et al., 2013). Indeed, many marine unicellular eukaryotes are able to feed both through photosynthesis and by predating other organisms (Stoecker et al., 2017). These widely abundant organisms blurring the line between autotrophy and heterotrophy are called mixotrophs (see Box 1, Flynn et al. (2013)), and their hybrid position in the food chain constitutes a good example of the large diversity of ecological roles achieved by planktonic organisms (Figure 1.3).

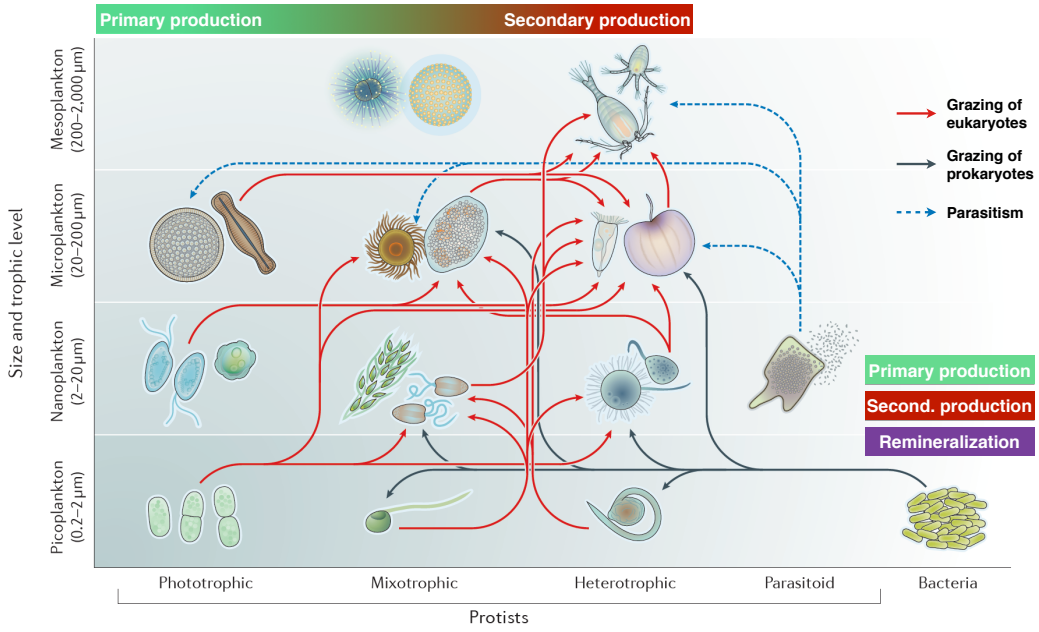

Figure 1.3 - Ecological and biogeochemical roles in marine planktonic ecosystems. Protists from 4 size ranges and 4 trophic regimes are depicted, with their ecological interactions represented by arrows. Phototrophs and mixotrophs fix atmospheric CO_2 into organic matter through photosynthesis (primary production), before being grazed on by heterotrophs (secondary production, red arrows). The organic matter then travels through higher levels of the food web, which are not represented in this diagram. Parasitoids (see Box 1) and viruses, which are not depicted here, can cause the death of their hosts (blue dotted arrows), which leads to a release of organic matter and an increase of available nutrients in the water column. Bacteria are represented as a unique simplified group in this diagram, but contain both phototrophic and heterotrophic organisms. Bacteria notably have the important role of remineralizing organic matter back to inorganic nutrients and carbon dioxide, which are then available for primary production. Figure modified from Caron et al. (2017).

The component of biodiversity that dictates the number of ecological functions achieved in an ecosystem is often described as *functional diversity* (Tilman et al., 1997), while the term *taxonomic diversity* is used to describe the biodiversity in terms of distinct

taxa/species, and the term *phylogenetic diversity* to illustrate the evolutive divergence between taxa/species (Naeem et al., 2012). Similar metrics can be computed for these three facets of diversity, like the functional,taxonomic or phylogenetic richness (*i.e.* how many functions, species or lineages are present ?) and evenness (*i.e.* are there dominant functions, species or lineages ?) of a community (Mason et al., 2005; Cadotte et al., 2010). These various facets of diversity are sometimes coupled (Galand et al., 2018), as functional capacities of organisms are shaped by their evolutive history, but a single taxonomic or phylogenetic group can also have a great diversity of ecological functions. Dinoflagellates for example, are a functionally diverse lineage of planktonic unicellular eukaryotes that include autotrophic, mixotrophic, symbiotic or even parasitic organisms, some of which producing toxins (Meng et al., 2017)). In this thesis, I will mainly focus on the links between the functional diversity of marine plankton and global biogeochemical cycles, which are driven by planktonic organisms.

1.1.2 Marine biogeochemical cycles driven by planktonic communities

The physiology of planktonic organisms and their ecological interactions have multiple impacts on global biogeochemical cycles, notably including the carbon, nitrogen, phosphorus, sulfur or iron cycles. I will now detail the role of planktonic organisms in some of these elemental cycles, and how it highlights the necessity to take planktonic communities into account when studying climate.

1.1.2.1 The carbon cycle

The ocean contains about 39,000 PgC (1 PgC = 10^{15} grams of carbon), which is more than 46 times the amount of carbon present in the atmosphere, and about 10 times the one found in terrestrial soils, permafrost and vegetation (Ciais et al., 2013; Le Quéré et al., 2018). Carbon stocks in the ocean are mainly distributed in two pools: the dissolved inorganic carbon (DIC, about 38,000 PgC), which corresponds to carbon dioxide, carbonic acid, bicarbonate and carbonate ions, and the dissolved organic carbon (DOC, about 700 PgC) (Ciais et al., 2013; Le Quéré et al., 2018). The pool of carbon corresponding to living organisms represents 3 PgC (Ciais et al., 2013; Siegenthaler and Sarmiento, 1993). The majority of the ocean carbon pool is located in intermediate and deep waters, where about 98% of the DIC is stored (Ciais et al., 2013; Siegenthaler and Sarmiento, 1993). By absorbing and stocking carbon, oceans take up to a third of anthropogenic carbon emissions, hence mitigating climate change impacts on the biosphere (Siegenthaler and Sarmiento, 1993). It is then essential to understand how carbon fluxes operate in the

ocean to be able to predict the impacts of climate change.

Fluxes of carbon within the ocean are governed by four main processes, described as *carbon pumps*: the *solubility pump*, the *biological pump* (also called the soft tissue pump), the *carbonate pump* and the *microbial pump* (Siegenthaler and Sarmiento (1993); Herndl and Reinthaler (2013); Ducklow et al. (2001); Jiao et al. (2010), Figure 1.4).

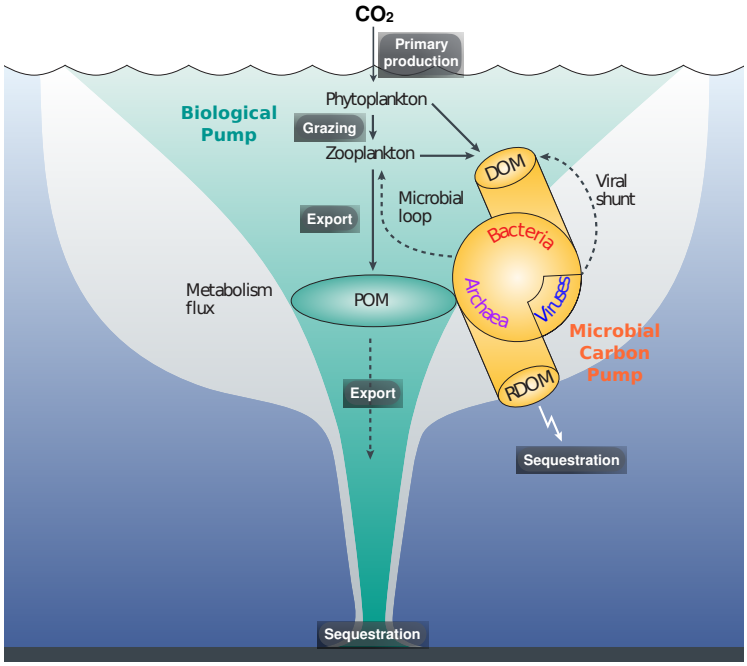

Figure 1.4 - Planktonic organisms interact to fix carbon on geological time scales in the ocean. The biological carbon pump is depicted in the green kernel, while the microbial carbon pump is represented by the yellow scheme on the right. Atmospheric carbon is either sequestered through the export of particulate organic matter (POM) produced by the biological carbon pump, or the creation of refractory dissolved organic matter (RDOM) produced by the microbial carbon pump. Diagram modified from Jiao et al. (2010).

The solubility pump is an abiotic process, where atmospheric carbon is chemically absorbed by oceanic waters before reaching the deep layers of the ocean through thermohaline circulation, where it can stay more than 1000 years before reaching surface again (Volk and Hoffert, 1985). Sometimes called the physical carbon pump, the solubility pump is very different from the biological, carbonate and microbial ones, which are all governed by biotic processes.

The biological carbon pump starts with the fixation of dissolved inorganic carbon by autotrophic organisms through primary production, *i.e.* the synthesis of organic matter from inorganic carbon (Figure 1.4). The primary production of marine planktonic organisms leads to the net fixation of 45-50 Gt C per year, which is comparable to the 45-68 Gt C fixed per year by terrestrial plants (Longhurst et al., 1995; Chavez et al., 2011). The produced organic matter is then transferred along food webs, and exported towards ocean's

depths through the sedimentation of organisms' dead bodies and fecal pellets (Ducklow et al., 2001; Legendre et al., 2015) (Figure 1.4). When the organic matter reaches the ocean floor, it can stay sequestrated on geological time scales, *i.e.* up to millions of years (Ducklow et al., 2001; Herndl and Reinthaler, 2013; Legendre et al., 2015). The biological carbon pump is estimated to cause the sinking of between 0.3 and 0.7 PgC per year at a 2000 m depth, which represents between 0.6 and 1.3% of the organic carbon produced by primary production.

The carbonate pump also refers to the sinking of biological matter towards the ocean floor, this time under the form of carbonate shells produced by calcifying planktonic organisms, such as coccolithophores or forams (Volk and Hoffert, 1985). But the magnitude of the carbonate pump is harder to estimate than the one of the biological pump, as the production of carbonate shells leads to a release of CO_2 in the surrounding waters, and eventually to the atmosphere (Legendre et al., 2015). The carbonate pump is even sometimes called the carbonate *counter-pump* (Legendre et al., 2015).

The microbial carbon pump differs from the three other pumps as it does not correspond to a vertical flux of carbon. Instead, the microbial carbon pump includes all biotic reactions allowing to switch from the most labile forms of DOC, which are short-lived (hours to days) and accessible to micro-organisms for decomposition, to the most refractory forms of DOC, which are long lived (20 to 40,000 years) and resistant to microbial decomposition (Legendre et al., 2015) (Figure 1.4). Since the decomposition of DOC by microbes leads to a production of CO_2 through a reaction called *remineralization*, the amounts of labile and refractory DOC directly influence oceanic carbon stocks (Jiao et al., 2010; Legendre et al., 2015). Hence, biological processes leading to the production of refractory DOC from labile DOC can be seen as carbon sequestration processes, analog to the three other carbon pumps.

I showed how biogenic fluxes of carbon from the atmosphere to the ocean are mainly driven by primary production. Marine primary production can be limited by multiple factors, notably nitrogen, phosphorus and trace elements like iron. I will now illustrate how planktonic communities play important roles in the biogeochemical cycles of these limiting substrates.

1.1.2.2 The nitrogen and phosphorus cycles

In the 1930s, Alfred Clarence Redfield identified that multiple elements were present in near constant ratio both in phytoplanktonic cells and in the marine water, giving his name to the famous Redfield ratio (Redfield, 1934). More precisely, he identified the

carbon:nitrogen:phosphorus ratio to always be near 106:16:1, before new observations showed that this ratio could slightly change depending on phytoplankton species and environmental conditions (Redfield, 1934; Martiny et al., 2014). Redfield also observed that when nitrogen or phosphorus was depleted, it was always also the case for the other, making it hard to identify which element was the most limiting one for primary production (Redfield, 1934). This observation led to debates between oceanographers, with most geologists identifying phosphate as the limiting nutrient, while most biologists argued for nitrogen being the main limiter of primary productivity (Gruber, 2004; Tyrrell, 1999).

Since the beginning of the century, phosphorus impact on carbon fluxes is considered to be influential on geological time scales, while nitrogen's impact on primary production is considered to be more immediate (Gruber, 2004). This is mostly explained by the fact that phosphorus enters oceanic waters only through abiotic processes (Baturin, 2003). River runoffs constitute about 80% of the total phosphorus supply into oceans, the rest being provided by volcanism, coastal abrasion, atmospheric precipitations, glacier erosion and groundwater discharge (Baturin, 2003). Inorganic phosphorus then either sediments or is assimilated by phytoplankton, and stays trapped in the Ocean for long time scales, ranging from 10,000 to 270,000 years (Baturin, 2003). On the opposite, marine nitrogen fluxes are mostly biologically driven (Gruber, 2004; Baturin, 2003).

Nitrogen is found in 5 relatively stable forms in the ocean (which is more than most other elements), and is used to synthesize structural elements of living cells and produce their metabolic energy (Gruber, 2004). These forms are dinitrogen (N_2), ammonium (NH_4^+), nitrate (NO_3^-), nitrite (NO_2^-) and nitrous oxide (N_2O), to which we can add all organic compounds containing nitrogen, like urea for instance (Gruber, 2004). Switches between all these different forms can be operated through oxidation-reduction reactions, mostly biologically regulated (Gruber, 2004). However, the most abundant chemical form of nitrogen, N_2, can not be assimilated by most organisms. Indeed, the assimilation of N_2, or nitrogen fixation, can only be achieved by organisms called *diazotrophs* (Figure 1.5). Until recently, only some *Cyanobacteria* genera like *Trichodesmium* were known to be diazotrophs, but multiple uncultured planktonic taxa like the *Cyanobacteria* UCYN-A or some *Proteobacteria* and *Planctomycetes* were identified as nitrogen fixers thanks to their genomic signature (Zehr, 2011a; Delmont et al., 2018) (more on that in section 1.3.2.2). Non-diazotrophic phytoplankton can only absorb bioavailable nitrogen (*i.e.* not dinitrogen), mainly under the form of ammonium, which is released in large quantities by bacteria through remineralization and requires a very low energetic cost to be absorbed (Gruber, 2004). Most phytoplankton can also assimilate nitrate, nitrite and urea through

enzymes reducing oxidized nitrogen to ammonium, but under higher energetic costs (Zehr, 2011a).

Processes closing the nitrogen cycle by allowing the release of N_2 back to the atmosphere are also biologically mediated. Two main metabolic reactions allow the dinitrogen release to the atmosphere: denitrification and anaerobic ammonia oxydation, or anammox (Gruber, 2004; Zehr, 2011a). These reductive processes occur almost only in regions where oxygen concentrations are very low, *e.g.* oxygen minimum zones (OMZ), benthic sediments or hydrothermal vents.

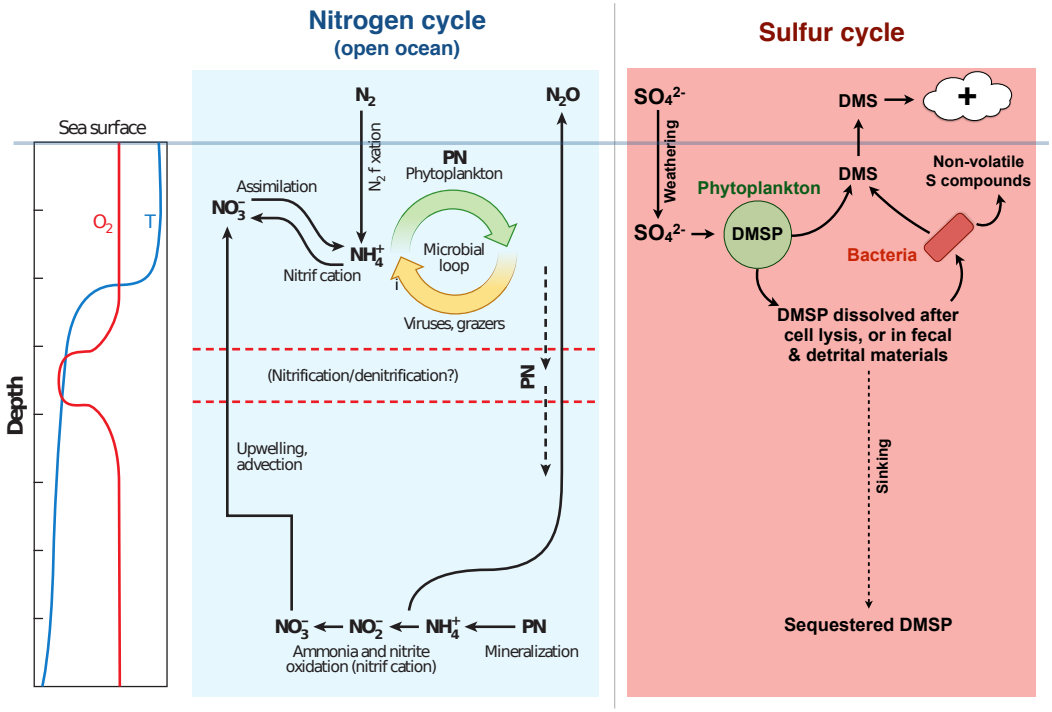

Figure 1.5 - The roles of planktonic organisms in the nitrogen and sulfur cycles. Depth, temperature and oxygen profiles are indicated in the left panel. A diagram of the open ocean nitrogen cycle is depicted in the blue panel (see section 1.1.2.2 for a detailed description). The upwelling and advection of NO_3^- and the sink of particulate nitrogen (PN) are the only processes in the diagram that are not achieved through biological processes by planktonic organisms. This cycle would be different in an oxygen minimum zone (OMZ), in which oxygen is almost absent around 200m-1000m depths. A diagram of the oceanic sulfur cycle is depicted in the red panel. Please note that this diagram focuses on the dynamics of dimethyl sulfide (DMS) and dimethylsulfoniopropionate (DMSP), and eludes all geologic parts of the marine sulfur cycle. The sulfur input to the ocean comes mostly from river runoffs, here indicated as weathering. The production of DMSP and DMS relies then exclusively on planktonic organisms. The produced DMS is ventilated to the atmosphere, which provokes the creation and densification of clouds. See section 1.1.2.3. The two left panels are extracted from Zehr (2011a), while the sulfur panel was inspired by Alcolombri et al. (2015).

The marine nitrogen cycle is then largely dominated by biologically driven processes (Figure 1.5). Many of these processes have only recently been identified (Zehr, 2011a), and we still can not culture most diazotrophs to study their metabolism in details. This makes it difficult, even now, to close the global nitrogen budget (Gruber, 2019). Most estimates

identify low nutrient areas like subtropical gyres to be concentrating nitrogen fixation (Wang et al., 2019a). In these waters, nitrogen fixation seems to be the main driver of primary production.

It was then proposed that nitrogen should be considered the *proximate limiting nutrient* in oceanic systems, or local limiting nutrient, while phosphorus should be considered as the *ultimate limiting nutrient*, limiting the total system productivity over longer time scale (Tyrrell, 1999). Still, in oligotrophic areas like the Sargasso Sea, phosphate concentrations are sufficiently low for phosphate to be the locally limiting nutrient for primary production (Wu et al., 2000). There are also zones of the ocean where both nitrogen and phosphate levels are high, while chlorophyll levels still remain low. In these high nutrient low chlorophyll (HNLC) zones, trace elements like iron become the main limiting factors of primary production (Boyd et al., 2007).

When the conditions are well adapted to primary production, *i.e.* light and high concentrations of the limiting nutrient are available, it triggers blooms of phototrophic organisms (Boyd et al., 2007). These dense populations of planktonic organisms not only affect the carbon, nitrogen or phosphorus cycles, but also many other important biogeochemical cycles. Among these cycles, the sulfur cycle is one of the most influential on the climate.

1.1.2.3 Dimethyl sulfide (DMS) production by planktonic organisms and its climatic impact

Oceanic dimethyl sulfide (DMS) is the first natural source of sulfur to the atmosphere (Charlson et al., 1987; Shaw, 1983; Simó et al., 2002). It plays an important role in marine ecology, by attracting large marine predators like fishes, but also marine birds and mammals, which seem to use it to detect planktonic blooms (Charlson et al., 1987). It is also very influential on earth climate, by facilitating the formation and condensation of clouds, thus significantly modifying the planet albedo (Alcolombri et al., 2015; Charlson et al., 1987; Shaw, 1983; Simó et al., 2002). The precursor of DMS, dimethylsulfoniopropionate (DMSP), is exclusively produced by phytoplanktonic organisms (Simó et al., 2002) (Figure 1.5). Part of the DMSP is cleaved into DMS directly by phytoplankton lineages bearing DMSP lyases, notably blooming taxa like *Phaeocystis* (Schoemann et al., 2005) or *Emiliania huxleyi* (Alcolombri et al., 2015). The rest is released into the water column where it becomes available to bacterioplankton (Levine et al., 2012; Simó et al., 2002) (Figure 1.5). Bacteria can either demethylate DMSP to produce carbon and reduced sulfur compounds, which does not lead to the production of DMS, or cleave DMSP to produce an easily accessible 3-Carbon compound and volatile DMS (Levine et al., 2012).

Both pathways are present in diverse bacterial lineages, including alphaproteobacteria, betaproteobacteria, gammaproteobacteria and epsilonproteobacteria, but are particularly found in the *Rhodobacterales* order, including the abundant *Roseobacter* genus (Curson et al., 2011). Eukaryotic and prokaryotic planktonic organisms are then responsible for the release of DMS to the atmosphere, which impacts the ecology, the geochemistry and the climate of marine ecosystems.

I reviewed evidences that planktonic organisms are the primary and secondary producers in the ocean, and that their abilities to recycle organic matter, produce carbonate shells, fix nitrogen, intake phosphorus, or build energy on sulfur compounds are all directly affecting earth climate. But how can this functional diversity be taken into account when it comes to modeling the functioning of marine ecosystems?

1.2 Functional types and traits to represent marine plankton diversity in biogeochemical models

1.2.1 Biogeochemical models and their links with plankton ecology

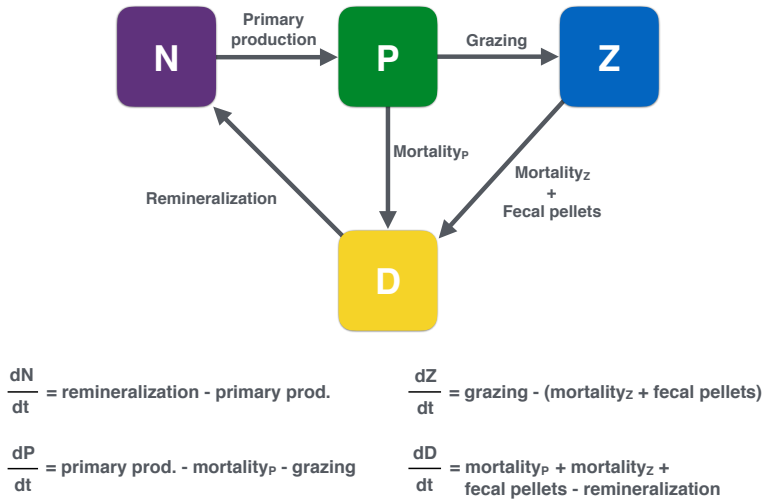

Figure 1.6 - Structure of a nutrient, phytoplankton, zooplankton and detritus model (NPZD model). State variables are represented as colored boxes. N stands for nutrients, P for phytoplankton, Z for zooplankton and D for detritus. Matter fluxes are represented by arrows. The differential equations represent the dynamics of each state variable over time.

The use of mathematical models (see Box 2) to theoretically represent planktonic communities has been common since the early works of Fleming (1939), Riley (1946) and Steele (1958). The first goals of such models were to better understand the drivers of seasonal variability in planktonic communities, first by simulating the prey-predator dynamics of

diatoms and zooplankton through a Lotka-Volterra (see Box 2) type approach (Fleming, 1939), then by representing the annual dynamics of phytoplankton concentrations in response to environmental factors like the light intensity or the nutrient limitation (Riley, 1946). To obtain a more realistic model of the phytoplankton annual dynamics, Steele (1958) added 2 state variables (see Box 2) interacting with phytoplankton populations: one following the dynamics of zooplankton populations and one corresponding to nutrients concentration.

From there, models simulating the dynamics of nutrients, phytoplankton, zooplankton and detritus, or NPZD models, became the "go-to" tool for modeling plankton communities dynamics (Gentleman, 2002) (Figure 1.6). The limitations (or even absence) of computer power first restricted these models to simulations in 0 or 1 dimension (*i.e.* across depth in the water column). But with the progress in computer power, it became possible in the late eighties to come up with basin scale and then global scale models of planktonic dynamics. The idea behind such models was to couple (see *model coupling* in Box 2) the ecological models simulating planktonic interactions with physical models of ocean circulation (see Box 2), offering the ability to simulate the processes influencing biogeochemical cycles on important temporal and spatial scales. Hence, these coupled models were called biogeochemical models, and one of the first and most famous was the one of Fasham et al. (1990). This model simulated the dynamics of 7 state variables: phytoplankton, zooplankton, bacteria, nitrate, ammonium, dissolved organic nitrogen and detritus. It was

Box 2: Theoretical ecology

Model: In this thesis, I will use the term 'model' to refer to *mathematical* models, which are mathematical representations of systems. In theoretical ecology, mathematical models are used to represent ecological systems, from simple prey-predator dynamics to global ecosystems. Such models allow to increase our understanding of the represented systems by allowing to test hypotheses on their global functioning, and sometimes predict their behavior.

State variables: Variables that define the current state of the modeled system. Examples of state variables for ecosystem models are the biomass or population size of different groups of organisms, like phytoplankton and zooplankton.

Parameters: Values that define the modeled system. Unlike state variable, parameters values stay fixed independently of the state of the system. Examples of parameters for ecosystem models are population carrying capacities, maximum growth rates, predation rates, *etc*… The step of defining parameters values is called *parameterization*.

Differential equations: Equations linking one or more function(s) to their derivative(s). Differential equations are used in ecosystem models to depict the dynamics of state variables over time. The differential equations of ecosystem models often can not be resolved analytically, and have to be solved through numerical approximations. Such approximations are computationally greedy, and responsible for most of the computing costs in ecological models.

Lotka-Volterra type model: Predator-prey model proposed independently by Alfred Lotka and Vito Volterra in the early 1900s. Originally representing two state-variables, prey and predator numbers, and using only 4 parameters: the prey growth rate, the predation rate, the efficiency of predation and the predator mortality rate.

Ocean circulation models: Physical models describing the physical and thermodynamical processes of the global Ocean.

Model coupling: The act of linking together two independent models. Two types of model coupling exist: offline coupling where outputs of one model are used as inputs in a second model, and online coupling, or full coupling, where feedbacks between the state variables of both models are defined.

coupled with a simplified vertical model of the ocean mixed layer circulation, and allowed to obtain quantitative estimates of seasonal nitrogen fluxes in the ocean (Fasham et al., 1990, 1993; Sarmiento et al., 1993).

The reduction of planktonic diversity to one, two or three state variables quickly raised the question of the potential oversimplification of biological interactions in biogeochemical models, already evoked by Riley in the 1940s (Anderson, 2010; Riley, 1946). Indeed, can we hope for realistic model predictions without simulating the dynamics of key biogeochemical actors like diazotrophs, calcifiers or remineralizing prokaryotes (Doney, 1999)? Of course, the question became even more itching with the progress in biological knowledge about planktonic communities, and the parallel increase in computing power allowing for the inclusion of more and more state variables into models. In the 2000s, the first global 3-dimensional biogeochemical models including significantly more planktonic diversity than NPZD models came out (Moore et al., 2001b; Aumont et al., 2003; Le Quéré et al., 2005; Kishi et al., 2007). These models relied on the concept of *Plankton Functional Types* (PFTs).

1.2.2 Plankton functional types and their use in biogeochemical modeling

In the early 90s, the concept of *plant functional type* was introduced in terrestrial plant ecology to group plants depending on their functional response to light and water availability (Smith and Huston, 1990; Smith et al., 1993). The same concept of *functional types* was evoked for planktonic organisms at that time, notably to describe the different size fractions of zooplankton and their different biogeochemical impacts (*i.e.* larger zooplankton grazing on larger preys, and their bigger fecal pellets sinking faster, enhancing export rates) (Armstrong et al., 1993). However, it is only in the seminal paper of Le Quéré et al. (2005) that the first operational 3D biogeochemical model explicitly relying on plankton functional types (PFTs) came out. If the paper by Le Quéré et al. (2005) was the first to clearly use the term of *plankton functional type*, a few anterior models had already been using multiple functional groups of plankton. It is notably the case of Moore et al. (2001b), who simulated the dynamics of carbon, nitrogen, phosphorus, iron, calcium carbonate and chlorophyll in 3 functional types of phytoplankton (small phytoplankton, diazotrophs and diatoms) and 1 type of zooplankton. This biogeochemical model was the first attempt at representing multiple plankton functional types and multiple limiting nutrients at global scale, but it was coupled to a grid model of ocean circulation which did not include horizontal transport (Moore et al., 2001a). The first biogeochemical model of similar complexity to use a 3D dynamic ocean circulation model was the one presented in Aumont et al. (2003), simulating the dynamics of carbon, phosphate, silicate and iron

in two functional types of phytoplankton (small phytoplankton and diatoms) and two functional types of zooplankton (microzooplankton and mesozooplankton) (Figure 1.7).

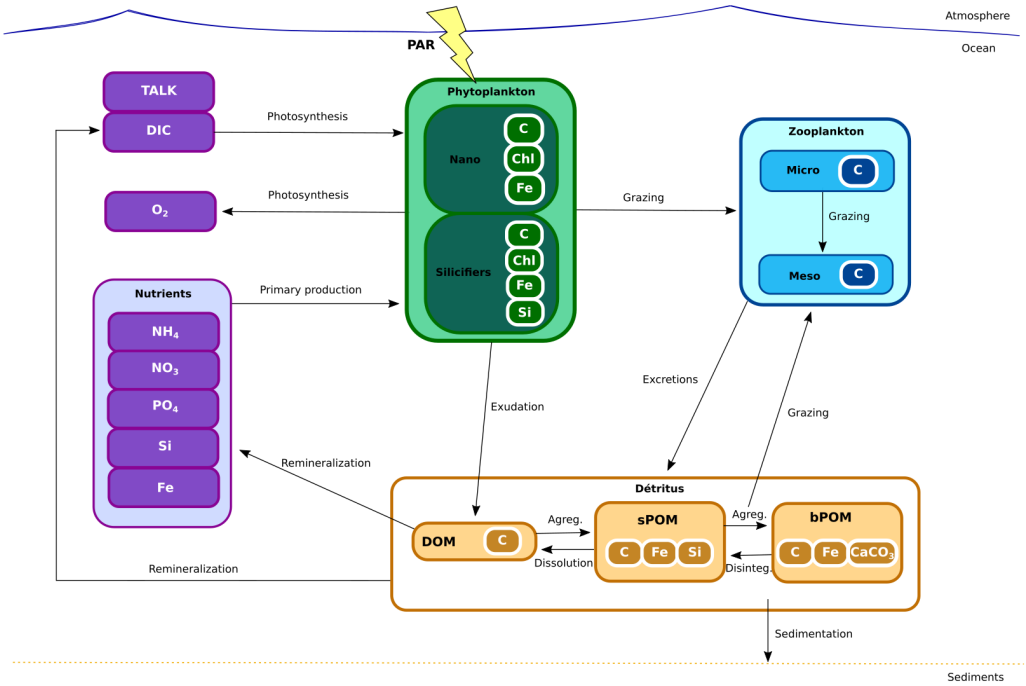

Figure 1.7 - Structure of the PISCES-version 2 model, re-drawn from Aumont et al. (2015). State variables are represented as colored boxes, with a color code similar as Figure 1.6. In this version of PISCES, 4 PFTs are represented: nanophytoplankton, autotrophic silicifiers (diatoms), microzooplankton and mesozooplankton. Acronyms : TALK = total alkalinity; PAR = photosynthetically active radiations; Chl = Chlorophyll; POM = Particulate organic matter; DOM = Dissolved Organic Matter.

In 2005, the model presented by Le Quéré et al. changed the standards for plankton diversity representation in biogeochemical models, by proposing a model with 10 different PFTs: pico-heterotrophs (*e.g.* heterotrophic bacteria and archaea), pico-autotrophs (*e.g.* cyanobacteria like *Prochlorococcus*), diazotrophs (*e.g.* *Trichodesmium*), calcifiers (*e.g.* coccolithophores), DMS producers (*e.g.* *Phaeocystis*), silicifiers (*e.g.* diatoms), mixed-phytoplankton (*e.g.* autotrophic dinoflagellates), proto-zooplankton (*e.g.* ciliates), meso-zooplankton (*e.g.* copepods) and macro-zooplankton (*e.g.* krill, jellyfish or salps). The choice of these 10 groups was motivated by 4 reasons: (1) each PFT should have a biogeochemical role, (2) the biomass of each PFT should be controlled by different physiological, environmental or nutrient requirements, (3) the behaviour of each PFT should have effects on other PFTs due to ecological interactions, and (4) each PFT should be significantly abundant in at least one part of the ocean (Le Quéré et al., 2005). From there, the use of plankton functional types in biogeochemical models became the norm, with the number of PFTs varying from 4 or 5 to sometimes hundreds depending on the models (Sinha et al., 2010; Aumont et al., 2015; Follows et al., 2007; Ward and Follows,

2016; Lévy et al., 2014).

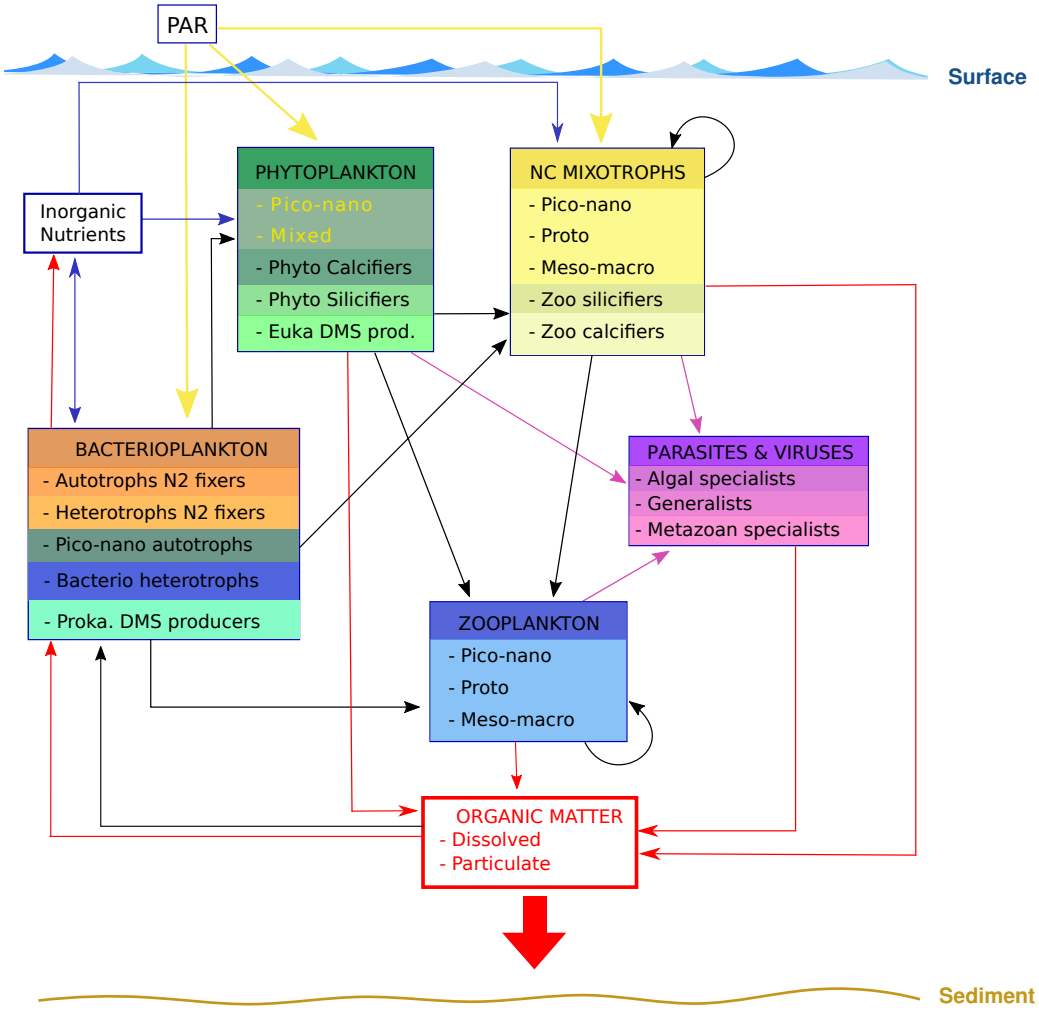

Figure 1.8 - Structure of a complex plankton functional type (PFT) model including overlooked trophic modes like mixotrophy and parasitism (Flynn et al., 2013; Worden et al., 2015), created during my master 2 internship to illustrate how the PFT approach could allow detailed representations of planktonic functional diversity. Here, 21 PFTs are represented in colored boxes, yellow arrows correspond to light uptake, blue arrows to nutrient uptake, black arrows to grazing and red arrows to organic matter transfers through sinking and remineralization. The phytoplankton PFTs written in yellow correspond to potential constitutive mixotrophs, i.e. phagotrophic algae. NC mixotrophs stands for non-constitutive mixotrophs, i.e. mixotrophs that acquire the ability to achieve photosynthesis through "stealing" chloroplasts to their preys or symbiosis. To my knowledge, no current biogeochemical model include that level of functional diversity, as even the Dutkiewicz et al. (2020) model, which has 350 PFTs, does not include parasitic and symbiotic relationships. Abreviations: PAR = photosynthetically active radiation, Euka = eukaryotes, Proka = prokaryotic, Prod = producers, Phyto = phytoplanctonic, Zoo = zooplanktonic.

Since its advent, the PFT approach has been criticized multiple times for its lack of ecological justifications (Anderson, 2005; Flynn et al., 2015). Indeed, one of the underlying assumptions behind the PFT concept is that each PFT can be modeled with a single set of parameters. But if we take the calcifiers as example, it should regroup both autotrophic calcifiers like coccolithophores (typically 5 to 100 microns in size) and heterotrophic calcifiers like forams or ostracods (both around 1 mm in size, sometimes up to several cm),

which play drastically distinct roles in the food webs. Hence, grouping together such a wide diversity of organisms asks the question of how to define the right set of parameters to model their growth in response of environmental factors (Flynn et al., 2015).

A second caveat of the PFT approach lies in the *a priori* choice of the functional types included in models, which is left to the modeler. This can lead to the exclusion of groups of organisms like mixotrophs, which are absent from the vast majority of PFT models despite their global ecological influence (Flynn et al., 2013; Stoecker et al., 2017; Caron, 2016a). Adding more and more PFTs to existing models could in theory resolve these issues, provided that *in situ* or experimental data are available to parameterize them. But it would push towards the production of increasingly complex, harder to interpret models, and we would need an unreachable number of PFTs to hope to exhaustively represent plankton diversity (Frede Thingstad et al., 2010). This way, a trade off has to be made between the quality of diversity representation in models and their complexity (Frede Thingstad et al., 2010). To better merge the biological aspects of plankton ecology with the theoretical frameworks of biogeochemical modeling, it was then proposed to switch towards approaches focused on the phenotypes of individuals, rather than functional groups or types (Flynn et al., 2015; Allen and Polimene, 2011).

1.2.3 The functional trait approach and its use in biogeochemical modeling

1.2.3.1 Concepts and definitions

During my PhD, I participated in a review of the use of functional-trait based approaches in aquatic ecology (Martini et al., under review, full version of the paper available in Appendix A). In this review, I created an interactive mental map of functional traits commonly used in aquatic ecology (Figure 1.9). I was also in charge of writing the paragraph on links between trait-based approaches and omics data (see section 1.4). Here, I will present few of the most important points and definitions of the trait-based approach, focusing on the contribution of functional trait trade-offs theory in biogeochemical modeling, focusing on how the trait-based approach can contribute to improving the representation of planktonic organisms in marine biogeochemical models.

Like the concept of *functional types*, the concepts of *traits* and *functional traits* emerged from terrestrial ecology (McGill et al., 2006). These terms were widely used in the literature in the past 20 years, sometimes to describe different concepts (Violle et al., 2007). Here, we will refer to the definitions given in Violle et al. (2007) :

- A **trait** is any morphological, physiological or phenological feature measurable at the individual level, from the cell to the whole-organism level, without reference to the environment or any other level of organization.
- A **functional trait** is any trait that impacts fitness (*i.e.* reproductive success) indirectly via its effects on growth, reproduction and survival.

Among functional traits, we can further differentiate **potential traits**, which are described from the literature, usually at the species level, and ideally covering a large variety of environmental conditions, from **realized traits**, actually measured *in situ* or in the laboratory (Reu et al., 2011).

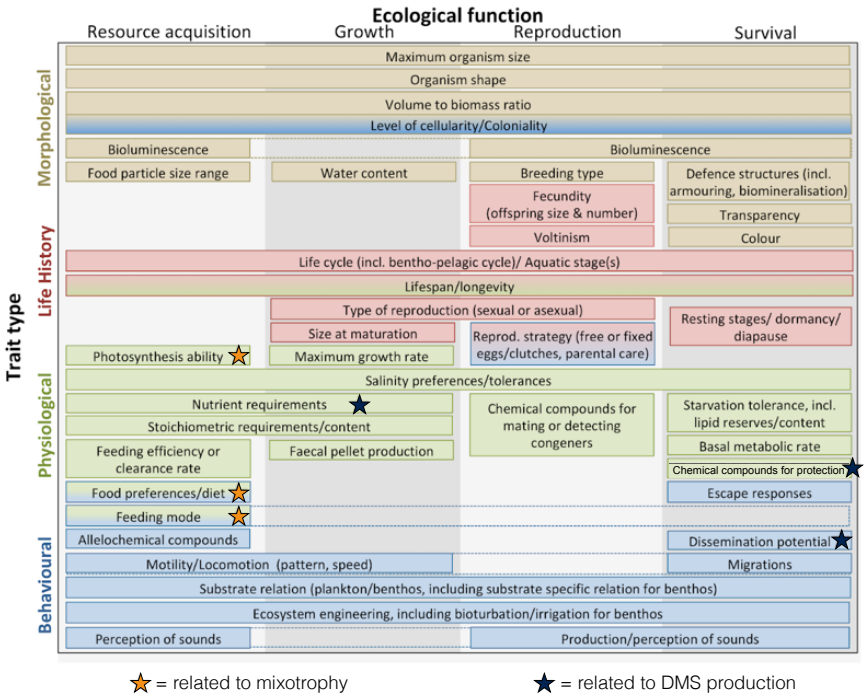

Figure 1.9 - Unified typology of aquatic functional traits, modified from Martini et al. (Appendix A). I made an interactive mental map of this topology, available at `https://github.com/EmileFaure/AquaticFunctionalTraitsMap`. *The functional traits commonly used in trait-based approaches in aquatic ecology are compiled in a common typology, where they are classified by type and ecological function (as in Litchman and Klausmeier (2008)). This typology focuses on the key functional traits that transcend taxonomic peculiarities of the different aquatic ecosystems. This typology do not explicitly include two functional traits on which I focused during my PhD: mixotrophy and DMS production. I added orange stars next to traits of the typology that are associated with mixotrophy, and blue stars next to ones associated with DMS production (see main text for further explanations on these associations).*
The interactive mental map version of this typology is under the form of a network in which blue nodes correspond to the different ecological functions depicted as columns here, while orange nodes correspond to the different trait types depicted as lines. Each trait of the typology is then represented as a grey node, and is linked to the ecological functions and trait type nodes corresponding to its classification in the typology. On this interactive mental map, it is possible to hover the mouse over a node to highlight its connections, and a simple click on a trait node will open a bibliography search about this trait in aquatic ecology studies, while a maintained click on any node will allow you to move it around.

Examples of functional traits for phyto- and zooplanktonic organisms are depicted in Figure 1.9. In this thesis, i will notably focus on two functional traits, mixotrophy and DMS production, which are not explicitly represented in Figure 1.9. Mixotrophy can be considered as implicitly included as a feeding mode, but also as a combination of traits such as the photosynthesis ability and the ability to feed through phagotrophy, or as a gradient of food preferences between full autotrophy and full heterotrophy (Figure 1.9; Berge et al. (2017)). DMS production is a harder trait to classify in a typology, as its beneficial impact on fitness still remains unclear (Levine et al., 2012). Some proposed that the impact of DMS atmospheric release increases local winds and currents, thus increasing the dissemination potential of DMS producers, others hypothesized that DMS acted as a protection against harmful UV radiation, and finally DMS production was identified as a way to create biomass in conditions of high DMSP production by phytoplankton and low requirements in sulfur in DMS producing prokaryotes (Simó, 2001; Levine et al., 2012).

The concepts of functional types and functional traits overlap in the sense that PFT are defined according to the functional traits of organisms, *e.g.* the calcifiers PFT regroups all organisms capable of producing a carbonate shell, which is a morphological and physiological feature measurable at the individual level that impacts fitness via its effect on survival. Some models even blend the concepts of functional traits and PFT, like the DARWIN model presented in Follows et al. (2007), where hundreds of PFT are randomly created from a set of functional traits. The created PFTs are then competing against each others, and only the fittest survive, allowing for the emergence of adapted functional strategies in different zones of the Ocean (Follows et al., 2007). But soon after the apparition of the first global PFT models, a new approach proposed to use functional traits to mechanistically link phytoplankton cellular-level physiology to ecosystem-level community patterns (Litchman et al., 2007). At the basis of this approach lied the concept of trade-offs between functional traits.

1.2.3.2 Traits trade-offs

Traits related to growth, reproduction and survival are often quantitatively correlated, and these correlations (or anti-correlations), described as "trade-offs", can provide a continuous view of ecological strategies among planktonic organisms (Litchman et al., 2007). For example, a strong positive correlation exists between the maximum nutrient uptake and the cell volume (Figure 1.10), or the half-saturation constant for nutrient uptake and the cell volume (Edwards et al., 2012). These trade-offs allow to define a continuous gradient of nutrient uptake strategies, for example opposing groups like diatoms with high cell volume, maximum nutrient uptake rate and half-saturation constant for nutri-

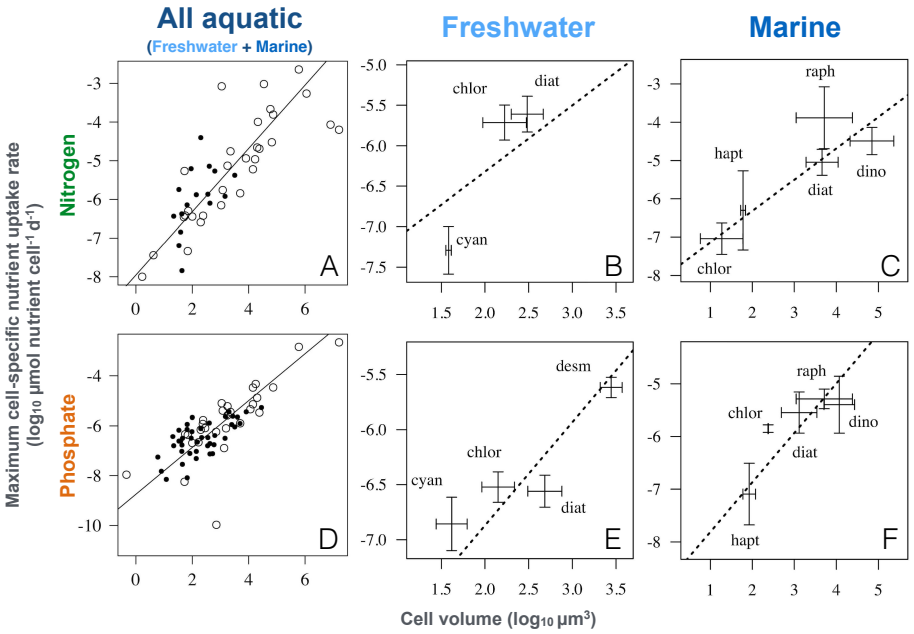

Figure 1.10 - Trade-off between cell volume and maximum cell-specific nutrient uptake rate. Graphs A, B and C represent the maximum cell-specific nitrogen uptake rate, while graphs D, E and F represent the maximum cell-specific phosphate uptake rate. In graph A and D, data from marine and freshwater species are represented by white-filled circles and black dots, respectively. The graphs B and E represent the same trade-off but only taking into account freshwater taxa, and representing the mean trait values for each taxon (+/- one standard error) instead of individual dots. Graphs C and F are the same as B and E, this time only taking into account marine taxa. The black and dashed lines were obtained from standardized major axis regressions, which all had a p-value < 0.05, showing a significant correlation between cell volume and maximum cell-specific nutrient uptake rate, independently of the ecosystem type (slopes were not significantly different between freshwater and marine taxa). cyan = cyanobacteria; desm = desmids; chlor = non-desmid chlorophytes; diat = diatoms; hapt = haptophytes; dino = dinoflagellates; raph = raphidophytes. Figure modified from Edwards et al. (2012).

ent uptake, which can be related to the "velocity" or "r" strategy in theoretical ecology (Margalef, 1978), to ones like haptophytes (including coccolithophores) with low cell volume, maximum nutrient uptake rate, and half-saturation constant for nutrient uptake (Litchman et al., 2007; Edwards et al., 2012), which can be related to the "affinity" or "K" strategy (Figure 1.10). This way, by using continuous quantitative relationships between traits, it is possible to create a multidimensional trait space in which different trait combinations are available, corresponding to different phenotypes and ecological strategies (Lamanna et al., 2014). Trait-based models typically use trade-offs between traits to define individual-level processes such as mortality, nutrient uptake rates, or metabolic costs (Kiørboe et al., 2018). By including the influence of the environmental conditions on trade offs, it is then possible to define optimal traits and trait distributions on large geographical scales, which links trait-based models to global bigeochemical ones (Kiørboe et al., 2018). The DARWIN model that I evoked earlier uses known trade-offs between functional traits to shape a limited trait-space to draw the random PFTs from, and avoid the creation of

super-dominant PFTs overrunning the ecosystem. Still, it differs from trait-based models for it does not take into account the inter-individual variations in traits in each randomly created PFT.

By focusing on continuous quantitative relationships between measurable individual features, trait-based approaches allow to incorporate the notions of plasticity and phenotypic variability between species and individual organisms (Violle et al., 2012). This key biological aspect is often absent from classic PFT modeling, where single sets of parameters are used for each PFT, and could help introduce acclimation in biogeochemical models (Flynn et al., 2015). Moreover, trade-offs between traits have not only been described in phytoplankton, but also in zooplankton (Litchman et al., 2013) and microbial populations (Litchman et al., 2015b), offering a way to describe most actors of planktonic communities.

However, trade-offs between functional traits are: (1) often not applicable to all organisms and/or all environments (*i.e.* exceptions exist to almost every trade-offs), (2) especially challenging to assess because a lot of data are required to define a trade-off as a general rule, and (3) difficult to compare and combine (how to prioritize the impact of different trade-offs on fitness ?) (Flynn et al., 2015). Trait-based approaches in general also suffer from the drawback of *a priori* choosing the functional traits included in the model, as already evoked in 1.2.2. Yet, the recent wealth in omics data is now changing our vision of planktonic diversity, which could help resolve some of these drawbacks.

1.3 Emergence of omics data to study planktonic diversity

1.3.1 Omics data and their application to plankton communities

In 1977, Frederick Sanger, Steve Nicklen and Alan Coulson presented the first rapid and reliable DNA sequencing method (Sanger et al., 1977). Only 15 years later, the sequencing of ribosomal RNA from environmental samples allowed for the first time to detect Archaea in coastal marine waters (DeLong, 1992) and in the open waters of the Pacific ocean (Fuhrman et al. (1992)). Archaea were thought to only live in extreme environments at the time, and this is an early example of how progresses made in DNA sequencing facilitated the sampling and analysis of full planktonic communities by bypassing the tedious morphological identification of species in complex samples. Only a few years later, such advances gave birth to the fields of *metagenomics* (See Box 3, Riesenfeld et al. (2004); Tringe and Rubin (2005); Venter et al. (2004)).

Before the end of the century, Sanger sequencers were used at long-term oceanographic

time-series like BATS to monitor seasonal changes in planktonic functional and taxonomic diversity (Giovannoni et al., 2014). The acquired ability to quickly analyze full planktonic communities notably led Craig Venter and his team to launch the *Global Ocean Sampling*, a large scale oceanographic cruise inspired by the circumglobal naturalist expeditions of the XIXth century (*e.g.* the Challenger expedition, 1872-1876) (Venter et al., 2004; Rusch et al., 2007). This expedition started in 2003, it used Sanger sequencing, and did not include metatranscriptomics data (See Box 3, Venter et al. (2004); Rusch et al. (2007)).

In the mid-2000s, DNA sequencing of environmental samples became an even more common method in plankton ecology with the advent of high-throughput sequencing methods (See Box 3, Riesenfeld et al. (2004); Tringe and Rubin (2005)). HTS allowed to multiply the quantity of data sequenced per day by 500,000 between 1996 and 2015, while the costs were divided by at least 250,000 over the same time period (Reuter et al., 2015). HTS methods became the reference for producing sequencing data in planktonic ecology, both for local studies including long term time-series (Gilbert et al., 2010; Galand et al., 2018; Arsenieff et al., 2020), and for larger spatial scale studies (Sunagawa et al., 2015; Acinas et al., 2019; Kopf et al., 2015). The most popular method from the 2nd generation of HTS, *Illumina* sequencing, used to produced relatively short reads, *i.e.* around 150 base pairs. In this way, Sanger sequencing, ensuring routinely the collection of 500 base pairs sequences, is still used today, especially when targeting specific genes (Levine et al., 2012). Since 2015, a 3rd generation of HTS methods were developed with the aim of produc-

Box 3: Sequencing data

High-throughput sequencing (HTS): Techniques of DNA sequencing that emerged in the late 90s and were popularized in the 2000s. In comparison to Sanger sequencing, HTS refers to sequencing techniques ensuring the production of more sequences in a relatively shorter amount of time and to a lower cost. The 2nd HTS generation (from mid-2000s) offers relatively short reads (e.g. Illumina : from 50 to 500 bp, 454 : from 300 to 600 bp) and the 3rd HTS generation (from ~ 2015) offers long reads (>1000 pb, e.g. PacBio, Nanopore). Currently in 2020, plankton sequencing data usually come from Illumina sequencing.

Omics: Molecular data obtained from HTS.

Meta-omics: Molecular data obtained from the HTS of one or multiple communities of organisms. In this thesis I will use this term to refer to metagenomics, metatranscriptomics and metabarcoding, but it can sometimes refer to other methods such as metaproteomics.

Metagenomics: Study of the DNA sequencing data of one or multiple communities of organisms in their environment. Was achieved with Sanger sequencing in the late 90s and was used to involve cloning steps. It is currently almost exclusively based on HTS.

Metatranscriptomics: Study of the RNA sequencing data of one or multiple communities of organisms in their environment.

Barcoding: Study of one or multiple molecular markers sequenced after a targeted PCR amplification of the corresponding DNA or RNA region(s). Proxy for detecting and quantifying taxa. Was achieved with Sanger sequencing in the late 90s, and sometimes still is.

Metabarcoding: Barcoding of one or multiple communities of organisms in their environment. The most commonly used markers are hyper-variable regions (labelled V1 to V9) of the16S ribosomal RNA which is universal among prokaryotes and of the 18S ribosomal RNA which is universal among eukaryotes.

Operational taxonomic unit (OTU): Cluster of sequences grouped by similarity, used as a proxy of species.

Reads: Sequences resulting from HTS.

ing long reads of multiple thousands of base pairs, while keeping the cost and speed advantages of 2nd generation sequencing (Giordano et al., 2017). For now, uses of 3rd generation sequencing methods in planktonic ecology are quite scarce, but might become the new standard in the next decade (Lombard et al., 2019).

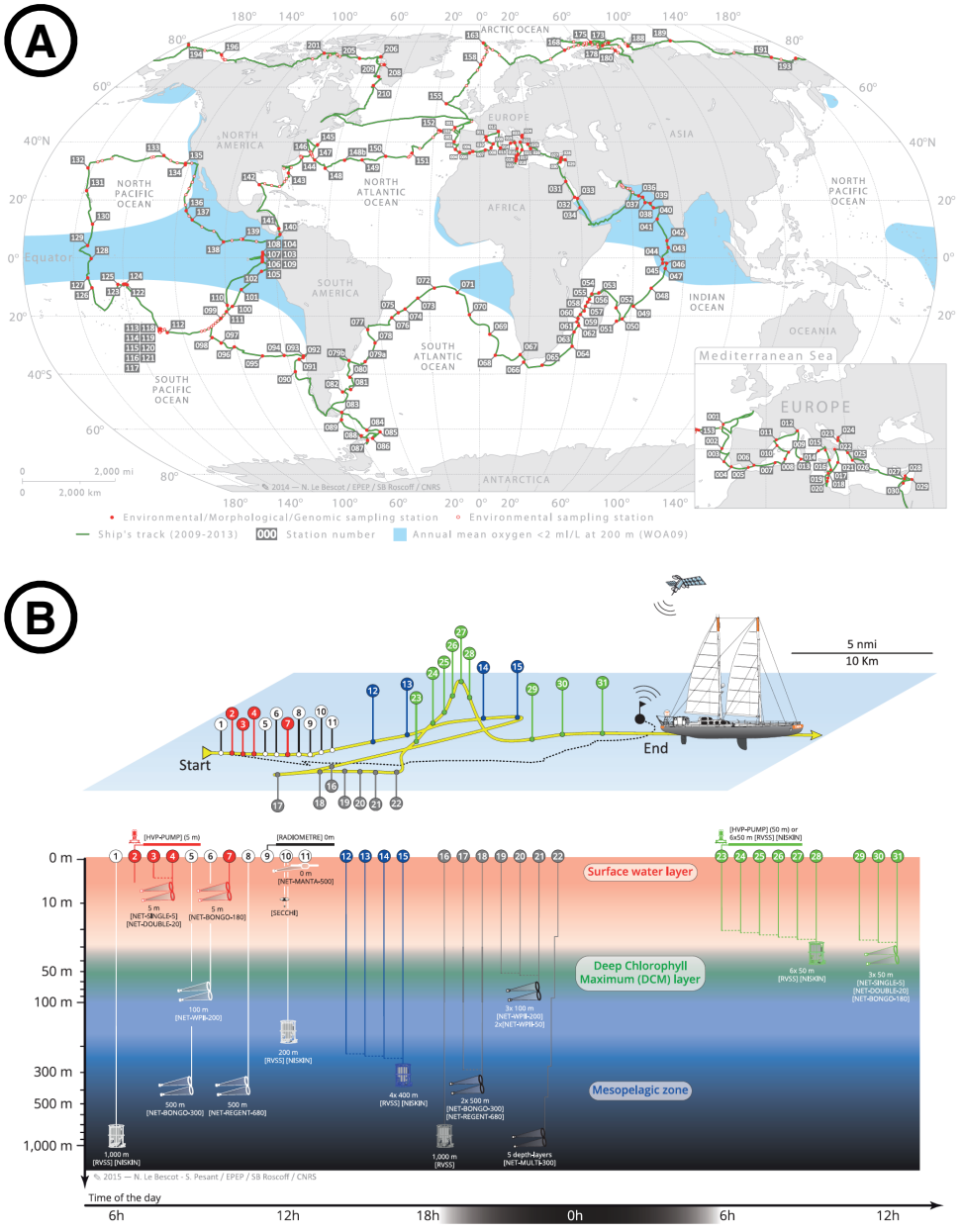

Figure 1.11 – (A) Sampling route and stations of the Tara Oceans and Tara Oceans Polar Circle campaigns. (B) Spatial representation and chronology of the sampling events conducted during a 24-48h station. The color code indicates the depth of the sampling event (surface water in red, deep chlorophyll layer in green, mesopelagic zone in blue, and other fixed depths in white for day deployments and grey for night deployments). The black dotted line indicates the deployment of an Argo drifter, used to follow the water mass during sampling. Figures from Pesant et al. (2015).

Following the footsteps of the *Global Ocean Sampling*, other large-scale cruises were launched in the 2000s and 2010s, this time taking advantage of the HTS to include metatranscriptomics (see Box 3) and cover more locations and depths of sampling (Pesant

et al., 2015; Acinas et al., 2019). Examples of such cruises are the Malaspina expedition in 2010-2011 (Duarte, 2015; Acinas et al., 2019), and the *Tara* expeditions, including *Tara Arctic* in 2006-2008, *Tara Oceans* and *Tara Oceans Polar Circle* in 2009-2013, and *Tara Pacific* in 2016-2018 (Figure 1.11;Karsenti et al. (2011); Planes et al. (2019)). By sampling plankton communities along with their physico-chemical environment, such expeditions offered the first opportunities to study the links between the environmental conditions and planktonic taxonomic and functional diversity in the global ocean (de Vargas et al., 2015; Sunagawa et al., 2015). The philosophy behind the *Tara* expeditions was described as *holistic* (Karsenti et al., 2011), as it involved meta-omics, but also physico-chemical measurements, and other methods like automatic underwater imaging, allowing to study the entire planktonic communities from viruses to pluricellular organisms (see Figure 1.1; Figure 1.11). A complementary approach to global scale plankton sampling was proposed by the *Ocean Sampling Day* (OSD) consortium, who organized the sampling of more than 300 meta-omics samples on the same day (June 21st 2014), all across the global ocean and with a unified protocol (Kopf et al., 2015). Finally, initiatives like the international census of marine microbes (ICoMM) approach the idea of *global scale sampling* by merging and homogenizing multiple independent datasets and making them available on easily accessible platforms (Amaral-Zettler et al., 2010). In the case of ICoMM, datasets come from omics but also environmental data, mass spectrometry or lipid structures data, with the goal to provide an atlas of marine unicellular organisms and their physiology (Amaral-Zettler et al., 2010).

An unprecedented amount of meta-omics data has then been sampled in the last 20 years, but it would not be useful without reference databases. Metabarcodes and/or *Operational Taxonomic Units* (OTUs, see Box 3) need to be confronted to taxonomic annotation databases such as PR2 (Guillou et al., 2013) or SILVA (Quast et al., 2013) to be annotated to a lineages. Similarly, genes and transcripts obtained from metagenomics and metatranscriptomics need to be confronted to reference databases such as KEGG (Aramaki et al., 2019), or EggNOG (Huerta-Cepas et al., 2016) to be associated with a function and/or an organism. Such reference databases play a key role for the analysis of functional traits in planktonic communities through meta-omics: functional annotation databases allow to make direct links between genes and metabolic pathways, some of which can be associated with functional traits, while taxonomic databases can be coupled with trait databases that link taxa to their potential functional traits. In the next section, I will review some concrete examples of how meta-omics data changed our understanding of planktonic ecosystems, notably illustrating the influence of reference databases.

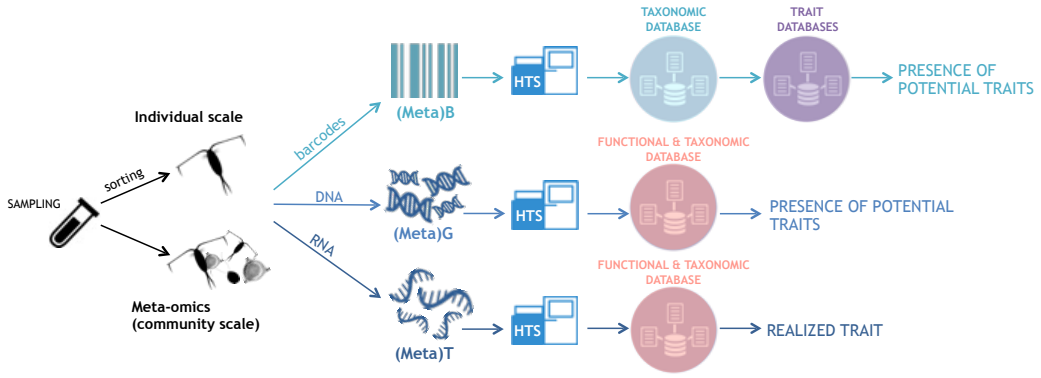

Figure 1.12 - Use of High-Throughput Sequencing (HTS) techniques to identify or measure functional traits of aquatic organisms, from Martini et al. (in review, Appendix A). B: Barcoding, G: Genomics, T: Transcriptomic. Sequencing can be done at the community scale (meta-omics) or at the individual scale after manual or automatic sorting. Metabarcodes are associated to taxa, through the use of taxonomic databases as PR2 for eukaryotes (Guillou et al., 2013) or SILVA for prokaryotes (Quast et al., 2013). The identified taxa can then be associated with potential traits through manual trait annotation or the use of trait databases, such as Traitbank which includes traits of lineages from the whole tree of life (Parr et al., 2016), or the functional traits of marine protists database (Ramond et al., 2018) (see Table 2 of Martini et al. in appendix A for a thorough list of trait databases). Genes and transcripts sequenced through metagenomics and metatranscriptomics can be associated to taxa and functions through databases like eggNOG (Huerta-Cepas et al., 2016), metagenomics giving access to potential traits while metatranscriptomics allowing to measure realized traits.

1.3.2 Planktonic functional and taxonomic diversity through the lens of meta-omics data

This section will be organized around three axes: how omics data (1) pushed us to re-scale our view of planktonic organisms' taxonomic and functional diversity, (2) provided new insights in plankton-mediated biogeochemical cycles, and (3) allowed for the reconsideration of long overlooked groups of organisms.

1.3.2.1 Quantifying taxonomic and functional diversity

As evoked in 1.3.1, one of the first contributions of omics data to plankton ecology was the discovery of archaeal 16S rDNA in open water and coastal samples (Fuhrman et al., 1992; DeLong, 1992). More recently, by analyzing about 1.7 million V9 regions of 18S rDNA sequences from 334 size-fractionated plankton samples of the *Tara Oceans* expedition, de Vargas et al. (2015) were able to detect about 110,000 OTUs, when only 11,200 species of eukaryotic plankton had been formally described morphologically at the time, according to these authors. About one third of these OTUs could not be associated to any known eukaryotic group (de Vargas et al., 2015). In parallel, Sunagawa et al. (2015) detected

35,650 prokaryotic OTU using nearly full 16S rDNA (assembled from metagetagenomic reads) of 243 size-fractionated samples of the same expedition, from which 7% could not be annotated at the phylum level. By analyzing together these 16S and 18S datasets, Ibarbalz et al. (2019) recently showed that the taxonomic diversity of planktonic Bacteria, Archaea and Eukaryota was higher at low latitudes, near the Equator, than at polar locations, like for larger marine organisms and for terrestrial ecosystems (Hillebrand, 2004; Rombouts et al., 2009).

In addition to metabarcoding data, the recent advent of metagenome-assembled genomes (MAGs) largely participated in increasing our knowledge about the taxonomic diversity of planktonic organisms (*n.b.* MAGs are also called *metagenomic species*). MAGs are near-complete genomes assembled from DNA fragments coming from metagenomes, without using reference genomes Nielsen et al. (2014). Instead, quality-filtered reads are assembled into *contigs*, and contigs reaching a minimum length (typically more than a few thousand base-pairs) are then binned together according to metrics like their abundance profiles across metagenomes (*i.e.* contigs with similar distribution across the samples) and their GC content (percentage of bases that are either guanine or cytosine), allowing to identify the groups of contigs (or "bins") coming from the same population of a single lineage (Nielsen et al., 2014; Delmont et al., 2018; Parks et al., 2017; Tully et al., 2018). The assembly of each bin can then give a more-or-less complete MAG (usually MAGs with completion estimated below 50% are discarded, completion being estimated based on the presence of sets of marker genes (Parks et al., 2015)), which can be taxonomically and functionally annotated through comparisons with reference genomes. Although the presence of a unique lineage/organism in the resulting MAG remains hypothetical (*e.g.* phylogeneticaly distinct organisms in symbiosis could for example be binned together due to high correlations in their abundances), tools like the manual refinement of bins and the computation of contamination percentages estimating the amount of badly binned contigs allow for determining high quality MAGs that can serve as proxies for taxonomic entities, in a way similar to OTUs (Delmont et al., 2018). They bring the advantage of giving access to near complete genomes instead of only ribosomal DNA, but also of not using primers and including unannotated sequences. In only the last three years, Parks et al. (2017) were able to recover nearly 8,000 MAGs from 1,550 marine and terrestrial metagenomes, while Tully et al. (2018) reconstructed 2,631 MAGs from 234 *Tara Oceans* metagenomes (Figure 1.2A), and Delmont et al. (2018) assembled 957 manually curated, high quality MAGs from 93 *Tara Oceans* metagenomes. In addition to these, 76 non-redundant and manually-curated MAGs were recently assembled from 58 bathypelagic metagenomes (Acinas et al., 2019). Finally, (Vorobev et al., 2019) proposed a similar

approach to MAGs binning, but this time aiming at building transcriptomes. They were able to detect about 12,000 groups of co-abundant genes in 365 metagenomes, among which 924 were identified as metagenomic based transcriptomes (MGTs), for they contained more than 500 *unigenes* (*i.e.* groups of transcripts coming from a unique gene) (Vorobev et al., 2019). The taxonomic assignation of unigenes among each MGT were then used to define their taxonomy and estimate their phylogenetic homogeneity. These MAGs do not only expand our view of taxonomic diversity among planktonic organisms (Figure 1.2A), but also in the tree of life, with notably 11 potential new phyla detected in the 76 bathypelagic MAGs only (Acinas et al., 2019).

Omics data also allowed to quantify the genetic diversity of plankton communities, and to link it with functional diversity through the analysis of present and expressed genes in hundreds of metagenomes and metatranscriptomes (Sunagawa et al., 2015; Acinas et al., 2019; DeLong et al., 2006; Louca et al., 2016c; Galand et al., 2018). Notably, a catalog of 47 million non-redundant genes was recently issued using 370 metagenomes of the prokaryotes-enriched size fractions from *Tara Oceans* and *Polar Circle* expeditions (Salazar et al., 2019). Comparing them with the gene available in reference databases, 39% of the genes in this catalog could not be annotated with a biological function, highlighting the gap remaining in our functional understanding of plankton genetic diversity. Similarly, 4.03 million genes were detected in 58 bathypelagic metagenomes, from which 71% had not been detected in global surveys of the upper ocean (Acinas et al., 2019; Sunagawa et al., 2015).

Thanks to such catalogs, it is now possible to quantify a part of the functional diversity of planktonic communities corresponding to genes with annotated functions, and notably to investigate its link with environmental conditions. This way, temperature was identified as the main driver of genomic functional diversity in prokaryotic plankton communities of the open ocean (Salazar et al., 2019). The quantification of both taxonomic and functional diversity of plankton communities through omics data also allowed to investigate the links between the two facets of plankton diversity on large spatial and time scales (Sunagawa et al., 2015; Louca et al., 2016c; Galand et al., 2018). This led to contrasting results, as the two facets of diversity were identified as decoupled when using the *Tara Oceans* data (Louca et al., 2016c; Sunagawa et al., 2015), but were tightly correlated in a coastal time-series station in the northwestern Mediterranean Sea (Galand et al., 2018). Finally, recent results based on a multi-omics analysis (*i.e.*, blending metagenomics and metatranscriptomics) showed that changes in gene expression rates across prokaryotic planktonic communities with similar taxonomic and genetic composition could significantly shape their functional activity (Salazar et al., 2019), adding a new

layer of complexity to the global picture.

Omics data have led to the discovery of uncharted branches in the tree of life (Rinke et al., 2013) and of millions of novel genes coding for proteins of unknown biological functions, sometimes issued from organisms of unknown phylogenetic lineages (Salazar et al., 2019; Acinas et al., 2019). This fraction of uncharacterized data has been described as the microbial dark matter, or more recently the dark side of omics, or the twilight zone, to avoid using the poor comparison with astrophysics' dark matter, which is theoretically predicted to exist but experimentally undetectable, pretty much the inverse from microbial dark matter (Lobb et al. (2015); Rinke et al. (2013); Vanni et al. (2020); Figure 1.13). The dark side of omics is present in all types of ecosystems, but is particularly significant in aquatic metagenomes, where the share of sequences of unknown function and taxonomy can reach up to 60% (Figure 1.13). I will now present some examples of how the functional characterization of unknown genes and lineages unveiled by meta-omics data led to discoveries of key planktonic actors in global biogeochemical cycles.

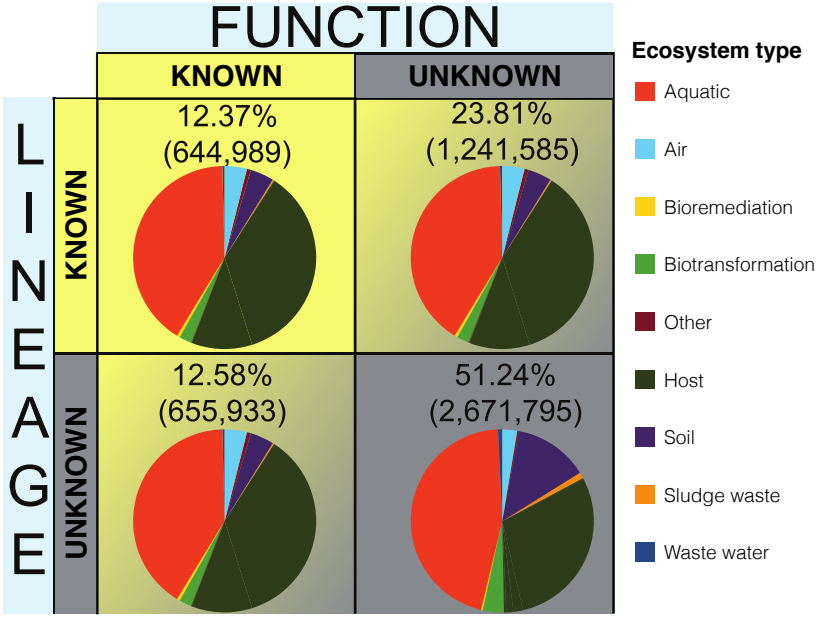

Figure 1.13 - Microbial dark matter across metagenomes from different ecosystem types. Genes inferred from the reads of 339 environmental metagenomes (Fondi et al., 2016) are divided in four categories depending on their taxonomic annotation (lines, based on the RAIphy algorithm (Nalbantoglu et al., 2011)) and functional annotation (columns, based on the Pfam database (Bateman et al., 2004)). The yellow background indicates sequences with known taxonomic and functional annotations, while the grey background indicates proteins from unknown lineages and with unknown function, hence corresponding to microbial dark matter. The number of sequences in each category is indicated, as well as the corresponding percentage, and a pie-chart indicating the distribution of the sequences across different ecosystem types. Proteins found in hosts (e.g. human gut microbiome) dominate the known/known category, while proteins from aquatic environments dominate the microbial dark matter category. For this graph, the minimum percentage of identity used for taxonomic annotation was 85% while the alignment e-value cut-off for functional annotations was of 0.1. Figure modified from Bernard et al. (2018).

1.3.2.2 Omics-based functional insights into biogeochemical cycles

As seen in section 1.1.2.2, the nitrogen cycle is largely driven by biological processes. Until the 1990s, the filamentous cyanobacteria *Trichodesmium* was considered to be the only significant diazotroph (Zehr and Kudela, 2011). Nitrogenase, the enzyme responsible for nitrogen fixation, is composed of two parts, one of which is encoded by the *nifH* gene Zehr (2011b). By amplifying *nifH* sequences in environmental samples, Zehr et al. (1998) detected nitrogenase genes in multiple cyanobacteria, but also in gamma and alpha *Proteobacteria*. Hence, by using omics data, Zehr et al. (1998) identified the cyanobacteria *Crocosphaera* and the uncultivated group of cyanobacteria called UCYN-A as globally abundant diazotrophs. By combining single cell analysis and large scale metabarcoding, UCYN-A was later identified as a ubiquitous group of symbiotic cyanobacteria, contributing to nitrogen fixation at scales comparable to *Trichodesmium* (Martínez-Pérez et al., 2016) (Figure 1.14). Even more recently, 9 non-cyanobacterial and diazotrophic MAGs were assembled from 93 *Tara Oceans* metagenomes (Delmont et al., 2018). 6 of these MAGs were annotated as Proteobacteria, while the 3 others were detected as Planctomycetes, providing the first evidence of diazotrophy among this taxonomic group (Delmont et al., 2018). Even more surprisingly, these 9 MAGs appeared as very abundant in the surface ocean, corresponding to up to 0.3% of all the sequences of a full metagenome in the Pacific Ocean (Figure 1.14).

In addition to the discovery of new diazotrophs, omics data also changed our understanding of the nitrification pathways in the global ocean (Zehr and Kudela, 2011). Indeed, Venter et al. (2004) found ammonia monooxygenase genes putatively coming from Crenarchaea in the samples from the global ocean survey, the archeal clade identified as abundant in the open and coastal ocean by Fuhrman et al. (1992) and DeLong (1992). This enzyme, responsible for the oxidation of ammonium into nitrite and nitrate, or nitrification (see section 1.1.2.2), was then considered to be only present in Bacteria like *Nitrospira* (Zehr and Kudela, 2011). Using targeted amplification of ammonia monooxygenase genes in environmental samples, Francis et al. (2005) then confirmed that ammonia-oxydizing archaea were ubiquitous and significantly abundant.

Omics data also greatly improved our understanding of the sulfur cycle (Moran et al., 2012). The first enzyme identified as involved in DMSP demethylation, *dmdA*, was described in 2006, followed a year after by the first one involved in DMSP cleavage (Howard et al., 2006; Todd et al., 2007). Only five years after that, 4 additional enzymes involved in the demethylation pathway and 10 additional enzymes involved in the cleavage pathway were described (Moran et al., 2012). Metagenomics data showed that the *dmdA* enzyme

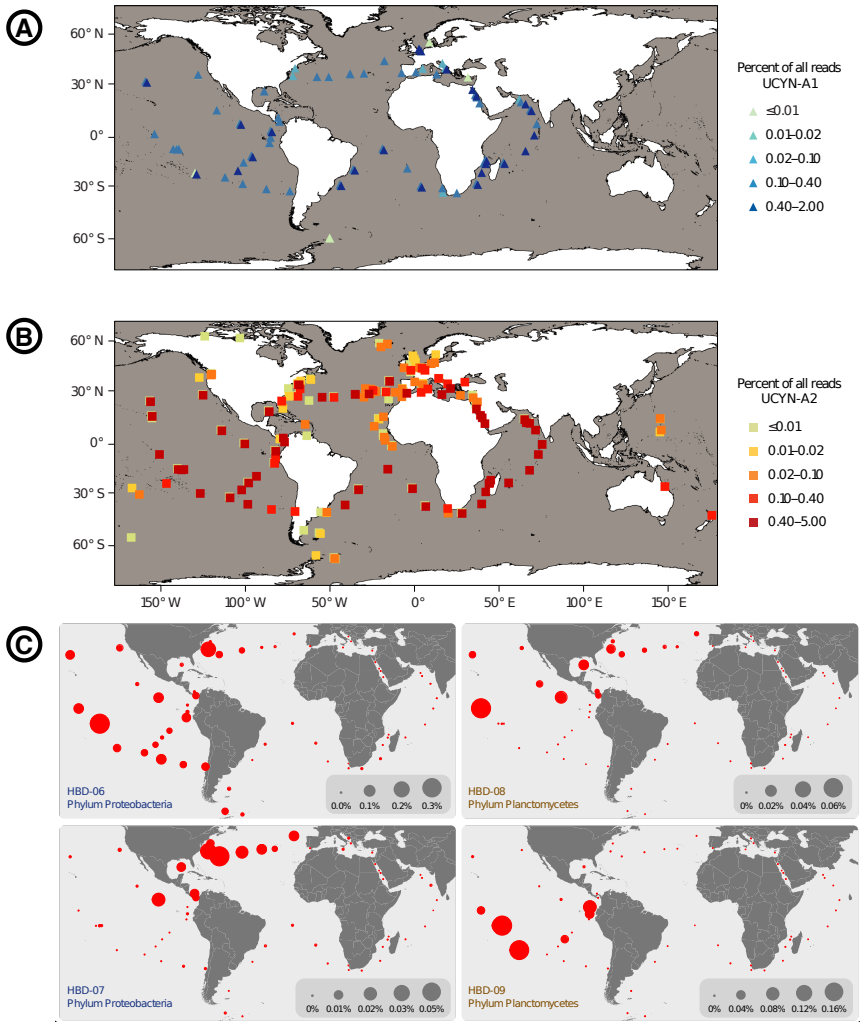

Figure 1.14 - Global abundance of diazotrophs unveiled by omics data. (A,B) Percentage of 16S rDNA reads annotated to UCYN-A1 (A) and UCYN-A2 (B) in samples from three datasets: ICoMM, OSD and Tara Oceans (See Section 1.3.1 for quick description of these datasets). Figure from Martínez-Pérez et al. (2016). (C) Maps of the sequence abundance (in % of total sample reads) of 4 diazotrophic MAGs in 61 metagenomes from the surface ocean. The two maps on the left correspond to Proteobacteria MAGs while the two maps on the right correspond to Planctomycetes MAGs. MAGs names begin with the acronym HBD, standing for heterotrophic bacterial diazotroph. Figure from Delmont et al. (2018).

was extremely abundant, and could be harbored by more than 50% of bacterioplankton cells in the surface Ocean (Moran et al., 2012; DeLong et al., 2006; Rusch et al., 2007). At the time, metagenomics also showed that cleavage-related enzymes were two orders of magnitude less abundant than demethylation-related ones, suggesting that some cleavage genes remained unknown, or that the cleavage of DMSP into DMS by bacteria might not be as important as previously thought (Moran et al., 2012). Since then, two key enzymes responsible for DMSP cleavage have been identified: the *dddK* enzyme was identified as cleaving DMSP into DMS through a previously unknown pathway in the very abundant *Pelagibacter*, while the *Alma1* enzyme was the first ever eukaryotic DMS-releasing enzyme identified, and was detected in many lineages of haptophytes and dinoflagellates (Alcolom-

bri et al., 2015). This way, missing links in the sulfur cycle highlighted by metagenomics data were resolved through targeted, biochemical extraction and purification approaches.

1.3.2.3 Rehabilitating overlooked groups of planktonic organisms

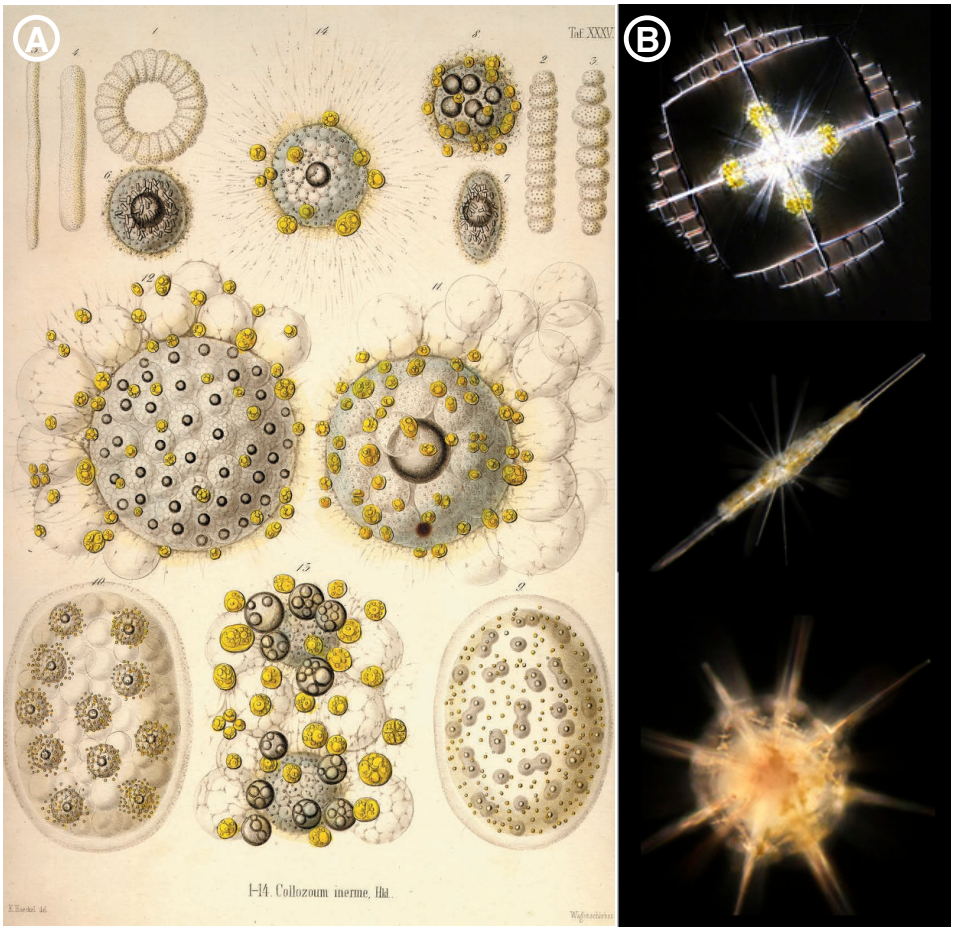

Figure 1.15 - The morphologic diversity of Rhizaria. (A) Illustrations of Collozoum inerme by Ernst Haeckel, from samples collected during the Challenger expedition (1872-76), (B) The diversity of morphology across 3 species of Acantharea (up: Lithoptera fenestra, middle: Amphilonche elongata, down: unidentified), all bearing photosynthetic symbionts. Collozoum inerme is a Collodaria species, a group of Radiolaria forming colonies, and of which every described species bears photosynthetic endosymbionts (Biard et al., 2017). Acanthareans are rhizarians with celestite skeletons, living in symbiosis with the ubiquist haptophyte Phaeocystis. Haeckel illustrations extracted from Caron (2016b), pictures taken from Decelle et al. (2012).

The advent of meta-omics data did not only lead to the discovery of novel biogeochemically impactful organisms, but also to the identification of well-known but yet overlooked lineages as key actors of the global ocean biogeochemistry. This is particularly well illustrated by the *Rhizaria* supergroup. This supergroup of unicellular and sometimes colonial eukaryotes has been known since the XIXth century, as they were described by Ernst Haeckel from samples of the *Challenger* expedition (1872-1876) (Figure 1.15). *Rhizaria* use complex pseudopodial networks to feed on preys, and harbor skeletal structures made

of opal (SiO_2), celestite ($SrSO_4$) or calcite ($CaCO_3$) (Caron, 2016b). Many species of the *Rhizaria* supergroup are also known for bearing photosynthetic endo- or ectosymbionts (Figure 1.15), and the question of their impact on global primary productivity was already asked in the 1990s (Caron et al., 1995). However, the amount of biological knowledge about these organisms remained very scarce until the last decade, mainly because their delicate skeletal structures often do not resist to sampling and preservation methods, while they remain impossible to maintain alive in culture (Caron, 2016b). Despite earlier evidences of their high abundance derived mostly from open ocean diving observations (Swanberg, 1983; Michaels et al., 1995), no global survey of their abundance and diversity had been conducted before the advent of meta-omics (Caron, 2016b). Also, they are absent of all major PFT models (*e.g.* they were not mentioned in Le Quéré et al. (2005), Aumont et al. (2003), or Follows et al. (2007)).

In 2012, omics data allowed to describe a widespread symbiotic relationship between *Phaeocystis* and the rhizarian group of *Acantharia* (Decelle et al., 2012). This symbiosis was identified through the amplification of the 18S and 28S rDNA of isolated specimens from *Acantharia*, which allowed to characterize their photosymbionts as *Phaeocystis*, a ubiquitous haptophyte genus found free-living (*i.e.* not in symbiosis) from poles to tropics, which had yet never been identified in symbiotic relationships (Decelle et al., 2012). The description of this original mode of symbiosis was soon followed by the discovery of *Rhizaria* as the second most abundant lineage in *Tara Oceans* 18S metabarcoding data, just after the *Opisthokonta* (de Vargas et al., 2015). In-situ imaging data also collected during the *Tara Oceans* expedition estimated that rhizarians might constitute up to 5.2% of the total oceanic biota carbon reservoir, confirming their greatly underestimated abundance, and asking the question of their impact on global primary production (Biard et al., 2016).

A co-occurrence based analysis of the *Tara Oceans* metabarcoding samples identified a group of 49 eukaryotic OTUs to be particularly correlated to carbon export in the global ocean, 5 of which were annotated as rhizarians (Guidi et al., 2016). Among the 5 OTUs identified as having the most influence on carbon export, two were annotated as *Collodaria*, a photosymbiotic group of *Rhizaria* (Figure 1.15A). Collodarians can either be solitary or colonials, they live in symbiosis with micro-algae of the *Brandtodinium* genus, and some species harbour an opal skeleton, while others stay "naked" (Biard et al., 2017). They contributed to 82% of the rhizarian 18S rDNA sequence sampled during the Tara Oceans cruise (Biard et al., 2017). These first insights into the ecological impacts of rhizarians, obtained through the analysis of *Tara Oceans* data, were confirmed by the identification of *Rhizaria* as important contributors to carbon export in the California Current

Ecosystem (Gutierrez-Rodriguez et al., 2019). This discovery relied on the metabarcoding analysis of sediment traps, *i.e.* oceanographic tools collecting sinking particles in the water column, in which *Radiolaria* (a branch of *Rhizaria*, including *Collodaria* and *Acantharea*) contributed to up to 90% of the 18S rDNA reads (Figure 1.16). Omics data then showed that rhizarians exhibit unique symbiotic relationships (Decelle et al., 2012), are ubiquitous and globally abundant (Biard et al., 2017), and have a significant impact on carbon export (Guidi et al., 2016; Gutierrez-Rodriguez et al., 2019; Stoecker et al., 2009). These results, combined to the recent evidences of the essential role of *Rhizaria* in the global silica cycle (Biard et al., 2018; Monferrer et al., 2020) are highlighting the necessity of their inclusion in biogeochemical models.

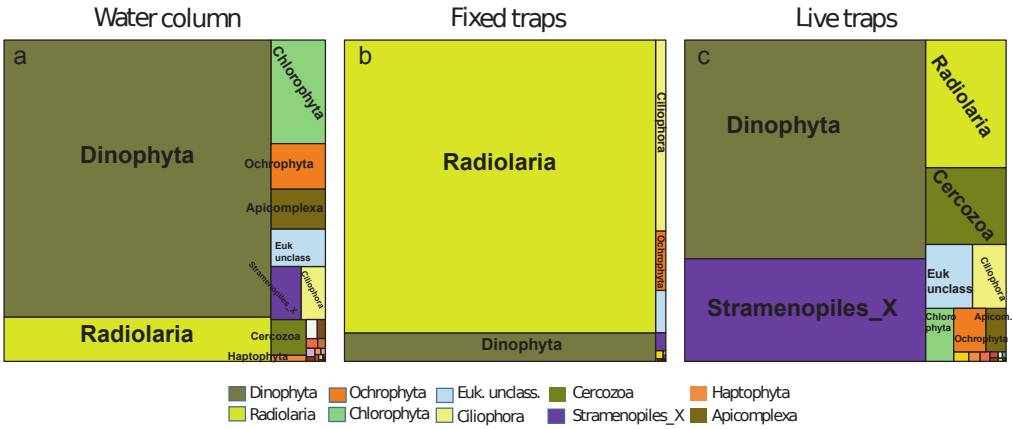

Figure 1.16 - The influence of Radiolaria on carbon export. Mean percentage of 18S rDNA reads affiliated to different plankton taxonomic groups in samples from the California Current Ecosystem. Samples were taken from the water column (a), from biologically fixed sediment traps (b, fixation with formaldehyde to minimize decomposition and consumption of organic matter), and from live sediment traps (c, no biological fixation). Radiolaria contributed to 12% of the total sequence number in the water column, 88% in fixed traps, and 9.6% in live traps. The increase of Radiolaria abundance in fixed traps compared to the water column samples demonstrate their high contribution to carbon export through sedimentation. The decrease of Radiolaria abundance in live traps was probably due to selective consumption by copepods, heterotrophic nanoflagellates, or phaeodarians in the traps, and supposes a rapid remineralization of organic matter associated with Radiolaria.

1.4 Using omics data to bridge the gap between observed and modeled diversity

1.4.1 Improving marine biogeochemical models using omics data

As shown in sections 1.2.2 and 1.2.3, traditional modeling approaches based on the representation of exchanges between a few ecosystem components (*i.e.*, nutrients, phytoplankton, zooplankton and detritus, or plankton functional types) have been criticized for their lack of ecological justifications (Anderson, 2005; Flynn et al., 2015). The proposition

of using omics data to improve the representation of planktonic diversity in biogeochemical models was raised very soon after the publication of the first results from the *Global Ocean Sampling* circumglobal cruise (Venter et al., 2004; Rusch et al., 2007; Hood et al., 2007). Even though the amount of published meta-omics data was very scarce at the time compared to what is available today, Hood et al. (2007) already stated that *"(...) it is not clear that these traditional modeling approaches will be sufficient in the face of all this emerging microbiological and genomic information; such models need to be "told" exactly what organisms and metabolisms exist in the ocean, and the rate coefficients that govern their parameterizations must be specified a priori. As such, they cannot tell you what is important and what is not."*. Hood et al. (2007) also insisted on the fact that traditional modeling approaches did not account for adaptation and evolution of planktonic organisms, which could be problematic when trying to predict the effects of climate change, and to often not explicitly represent the bacterioplankton, despite the experimental evidence of their ecological importance.

Meta-omics data then quickly appeared as an opportunity to inform traditional models (1) on the geographical distribution of organisms and metabolisms in the global ocean, (2) on the response of such organisms and metabolisms to environmental conditions, and (3) on potential new ecological theories and fundamental laws that could emerge from the unprecedented quantity of data available (Hood et al., 2007; Allen and Polimene, 2011; Mock et al., 2016). By providing information at the gene level, omics data theoretically allow for a switch from species-based or trait-based models towards metabolism-based or gene-centric models, notably inspired by systems biology, and the physics of complex systems (Reed et al., 2014; Toseland et al., 2013; D'Alelio et al., 2019; Follows and Dutkiewicz, 2011).

1.4.2 Omics-based metabolic modeling

Metabolism-based models rely on metabolic networks, which consist in a conceptual reconstruction of the metabolic pathways occurring in a (meta)genome, based on the functional annotation of its genes (Grossart et al., 2020; Budinich et al., 2017). In these networks, nodes correspond to metabolites, and arrows to metabolic reactions, driven by enzymes (Steuer et al., 2012; Budinich et al., 2017) (Figure 1.17). A metabolic network can be summarized in a stoichiometric matrix, in which columns correspond to reactions, lines correspond to metabolites, and each coefficient S_{ij} correspond to the stoichiometric coefficient of the metabolite M_i in the reaction R_j (Figure 1.17) (Budinich et al., 2017). Given a set of inputs to the network (*e.g.* a concentration of nutrients), and under the hypothesis that the rate of formation of internal metabolites is equal to their rate

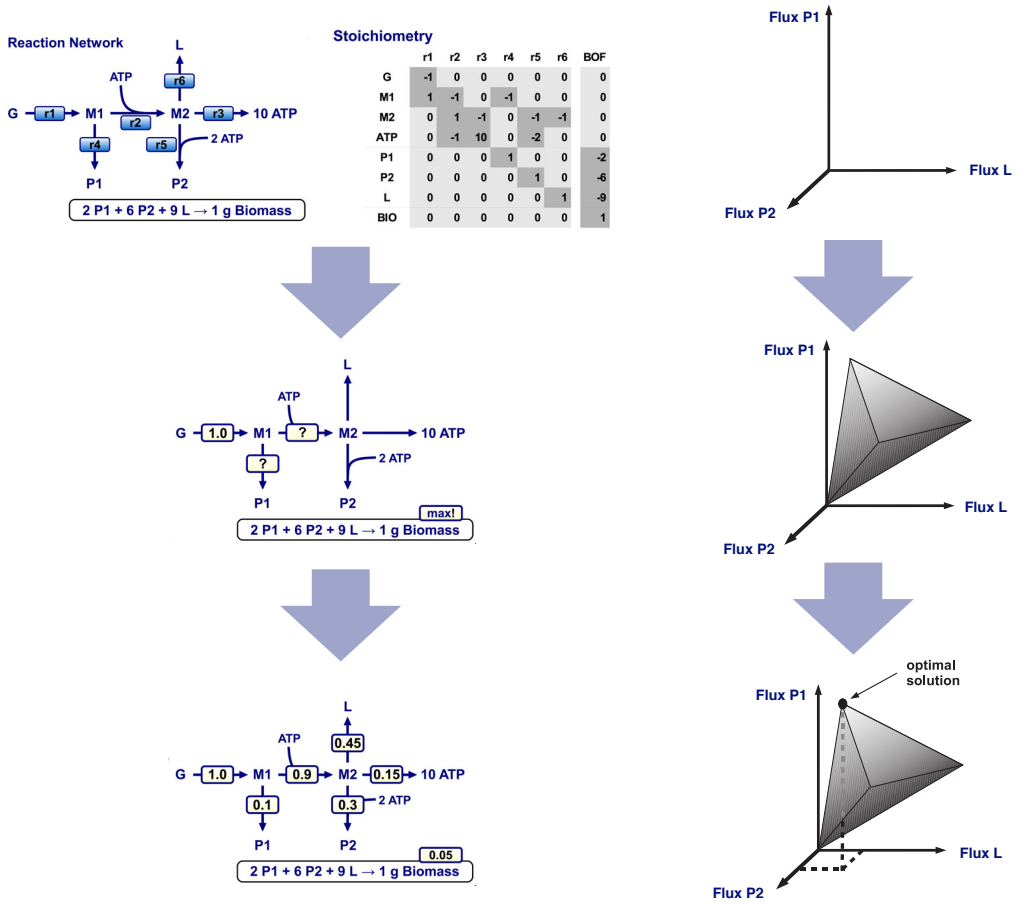

Figure 1.17 - Example of the metabolic network modeling process. A simple metabolic network is represented in the upper left, along with the corresponding stoichiometric matrix. The depicted metabolic pathway leads to the production of biomass through fluxes of three metabolites, P1, P2 and L, which are represented in a 3 dimensional space in the upper right. By initializing the flux of the first reaction r1 (here from G to M1) to 1.0, assuming steady state conditions and satisfying mass conservation constraint, it is possible to determine a space of solutions for the values of the P1, P2 and L fluxes (middle right graph), each leading to a value of biomass production. The biomass production can then be optimized in this space of solutions, as presented in the bottom graphs. Here, one optimal solution is presented, but most of the time multiple optimal solutions exist. In these cases, other cellular objectives can be optimized in addition to biomass, like minimizing the total amount of fluxes or the thermodynamic costs of reactions. Figure modified from Steuer et al. (2012).

of consumption, the stoichiometric matrix can be used to compute a constrained space of possible values for an objective function, like biomass production or growth rate (Budinich et al., 2017; Steuer et al., 2012). The optimal value for the objective function can then be obtained by solving a linear optimization problem, and corresponds to the value maximizing the "fitness" (*i.e.* the reproductive success) of the modeled organism given the conditions (Budinich et al., 2017; Steuer et al., 2012) (Figure 1.17). This approach is called *flux balance analysis* (FBA), and rely on *constraint-based models*. Metabolic networks then allow for the mechanistic modeling of intra-cellular processes, and can even be extended to model metabolisms at community scales (Budinich et al., 2017).

Moreover, tools are now available for the automatic reconstruction of metabolic networks (Konwar et al., 2015; Kanehisa et al., 2019), opening the door to their application on full communities (Budinich et al., 2017; D'Alelio et al., 2019).

But to be included in metabolic networks, genes require a functional annotation (Grossart et al., 2020), when a significant part of omics data remain poorly understood on a functional level (Section 1.3.2.1). Reed et al. (2014) wrote that *"although insightful for laboratory studies, [the metabolic network approach] is infeasible for use in conjunction with environmental genomics data because the majority of microbes are uncultured and their metabolic networks are thus unknown"*. In only 5 years, this statement became questionable with the advent of high quality MAGs for uncultured organisms, but illustrates well the necessity to have high quality genomics data and functional annotations for the metabolism-based approach to work. Metabolic networks have also been criticized for their struggle to capture temporal changes in metabolisms, like the switch between phototrophic metabolism during the day and storage-based metabolism at night in *Cyanobacteria* for example, despite the excellent quality of genomics data available (Steuer et al., 2012). But recent advances in constraint-based modeling techniques proved that metabolic diurnal cycles could be derived from metabolic networks (Reimers et al., 2017). Finally, flux balance analysis analysis have yet been rarely applied on eukaryotic organisms, at least compared to prokaryotic organisms, mainly due to the compartmentalization of eukaryotic cells, which adds a layer of complexity to the picture (Niklas et al., 2010). It is why the metabolism-based approach has not yet been applied to a full biogeochemical model, despite its great potential for a mechanistic modeling of planktonic organisms. Instead, all of the omics-informed biogeochemical models published in the last decade relied on a gene-centric approach.

1.4.3 Gene-centric approach for biogeochemical modeling

In gene-centric models, organisms are grouped according to a few genes pre-selected for their metabolic function and usually referred to as *functional genes* (Reed et al., 2014; Louca et al., 2016a). For example, the first published gene-centric biogeochemical model used 8 functional genes involved in nitrogen cycling as state variables: *amoA* for aerobic ammonia oxidation, *hzo* for anaerobic ammonium oxidation, *nor* for aerobic nitrite oxidation, *dsr* for sulfate reduction, *nap* for sulfide oxidation coupled to nitrate reduction, *sox* for aerobic sulfide oxidation, *narG* for nitrate reduction, *nirK* for nitrite reduction, *nrf* for dissimilatory nitrite reduction to ammonium and *cox* for aerobic respiration (Reed et al., 2014). The model described the rate of each gene production as dependant on nutrients and concentrations of reaction inhibitors, while the concentration of nutrients

was in turn impacted by the abundance of the different functional genes (Reed et al., 2014). It was applied to a 1 dimension (vertical) section of the Arabian Sea, allowing to reproduce observed patterns of oxygen, ammonium, nitrate and nitrite concentrations, but also observed sequence abundances of functional genes (Reed et al., 2014). A similar approach was proposed two years later, this time modeling carbon, sulfur and nitrogen cycles in a Canadian fjord through the modeling of 6 functional genes (Louca et al., 2016a). This model improved the biological realism of gene-centric approaches by not only modeling the DNA concentration of functional genes, but also their transcription from DNA to mRNA, allowing for a more mechanistic representation of enzyme production, and a better justification for comparing model outputs with metatranscriptomics data (Louca et al., 2016a).

These early attempts led to the publication of GENOME, the first 3D (*i.e.* longitude, latitude and depth) gene-centric biogeochemical model, including 20 prokaryotic gene functional groups (Coles et al., 2017), and from which the functioning is schematized in Figure 1.18. The model structure of GENOME was different from the ones of Reed et al. (2014) and Louca et al. (2016a), as instead of only following genes concentrations as state variables, it used the abundance of organisms bearing different randomly assigned genomes and transcriptomes (Figure 1.18, the terms genomes and transcriptomes referring here to the composition in functional genes and their predicted expressions). Inspired by emergent trait-based models such as DARWIN (Follows et al., 2007) (see section 1.2.3), GENOME creates random organisms to which are assigned functional genes, before being included in a global circulation model where only the fittest organisms are conserved, allowing for the emergence of the most adapted communities at each model run (Figure 1.18). It led to the observation that across different simulation runs, the community-level metabolic rates were similar in the same geographic areas, independently of the fact that the modeled organisms were different (Coles et al., 2017). The authors then proposed that genomic composition of planktonic communities in the model had more influence on biogeochemistry than the genetic composition of the individual organisms (Coles et al., 2017). This model-based hypothesis illustrates how gene-centric approaches allow to draw conclusions on functional diversity and its impact on biogeochemistry, where traditional PFT modeling would have struggled to decouple function from taxonomy.

Gene-centric approaches have also proven useful for confronting model outputs to experimental data (Reed et al., 2014; Louca et al., 2016a; Coles et al., 2017). Indeed, by modeling DNA and mRNA concentrations, gene-centric models allow for direct comparisons with metagenomics and metatranscriptomics datasets (Figure 1.18). Despite these improvements, all the gene-centric models published so far fail at answering to Hood's

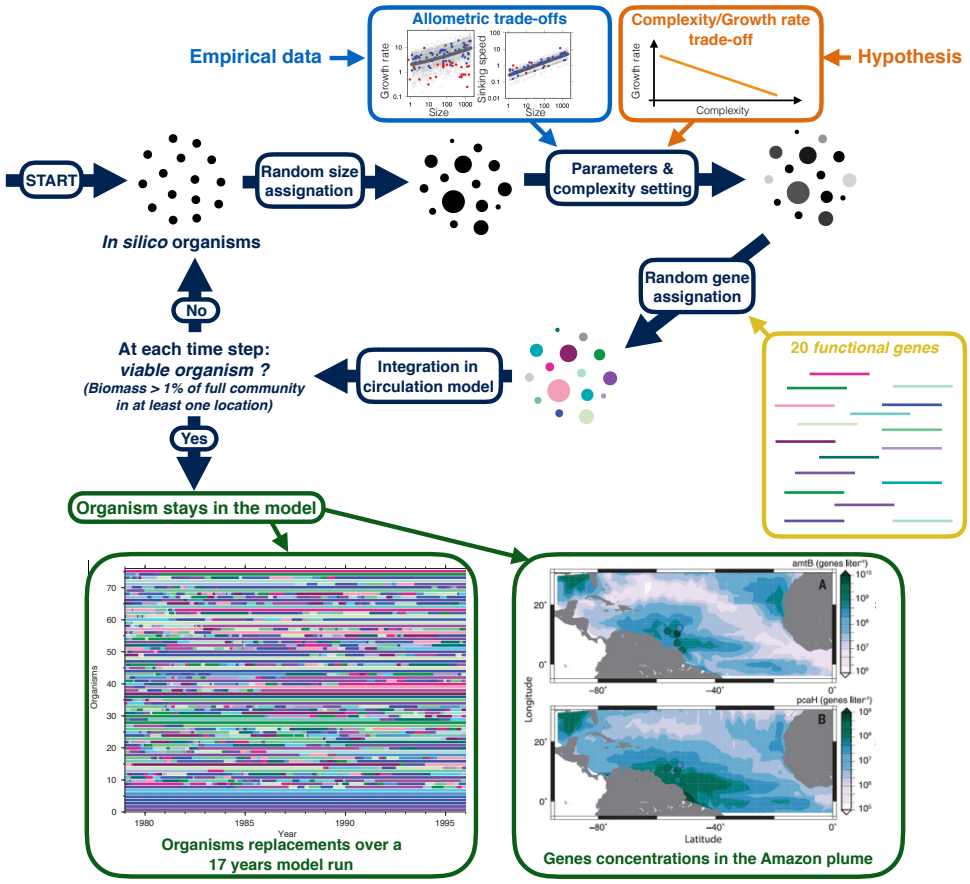

Figure 1.18 - Simplified functioning of the GENOME model, built using information and figures from Coles et al. (2017). Organisms are first randomly assigned a size from 1 µm to 2000 µm. Parameters like substrate uptake affinity, growth rate, mortality or sinking speed are then derived from the randomly selected size, according to experimental data (light blue box). Each organism is also assigned a degree of complexity, which corresponds to the number of functional genes it will be able to get. To avoid super-organisms dominating the model ecosystem, a trade-off was used to linearly decrease the maximum growth rate as the complexity increase (orange box). Functional genes are then randomly assigned to organisms, determining their impact on biogeochemistry (yellow box). Examples of such functional genes are: To avoid to keep unviable organisms in the model, the ones whose biomass did not represent more than 1% of the total community at any location were replaced by new organisms. 68 organisms co-existed at any given time during the simulations presented in Coles et al. (2017), as illustrated by the multicolored bars on the left green box. The evolution of 7 substrates were also modeled, which correspond to the plain bars in the left green box. The model notably allowed to predict the concentrations of amtB and pcaH genes in the Atlantic Ocean (coding respectively for ammonium transport and aromatic ring cleavage, right green box). On the maps in the right green box, observed concentrations of these two genes are overlaid in circles.

concerns evoked at the beginning of this section (Hood et al., 2007). Indeed, these models require to *a priori* select a set of functional genes from which the function has to be well known (*e.g* light-harvesting genes, nitrification genes, nitrogen fixation genes,...), exactly like PFT need to be *a priori* selected in traditional PFT models. This way, the gene centric approach was only applied to well known pathways, mostly present in prokaryotic organisms, and unrepresentative of the observed functional diversity of planktonic organisms (Reed et al., 2014; Louca et al., 2016a; Coles et al., 2017).

The emergence of metabolism-based and gene-centric approaches have provided a theoretical framework that allow for the use of genes as structural components of biogeochemical models (Mock et al., 2016; Stec et al., 2017; D'Alelio et al., 2019; Grossart et al., 2020; Reed et al., 2014; Louca et al., 2016a; Coles et al., 2017). The gene-centric approach has even been proven to be applicable at the scale of the entire Atlantic Ocean (Coles et al., 2017). However, it has yet failed at increasing the diversity representation in biogeochemical models, by only focusing on small numbers of functional genes (a maximum of 20 in Coles et al. (2017)) and only prokaryotic metabolic pathways. This highlights the current need for data-driven methods allowing for the automatic detection and quantification of functional genes of biogeochemical and ecological importance from meta-omics data.

1.5 How to use omics data to improve planktonic diversity representation in biogeochemical models ?

In this introduction, I have reviewed how the functionally and taxonomically diverse planktonic communities impact global biogeochemical cycles. I have then presented how these planktonic communities are currently represented in biogeochemical models, highlighting the gap between observed and modeled planktonic diversity. Finally, I showed how the advent of omics data led to further understanding of plankton diversity and contributed to the emergence of new theoretical frameworks, bearing the potential of producing biogeochemical models with realistic representations of planktonic functional diversity (Figure 1.19).

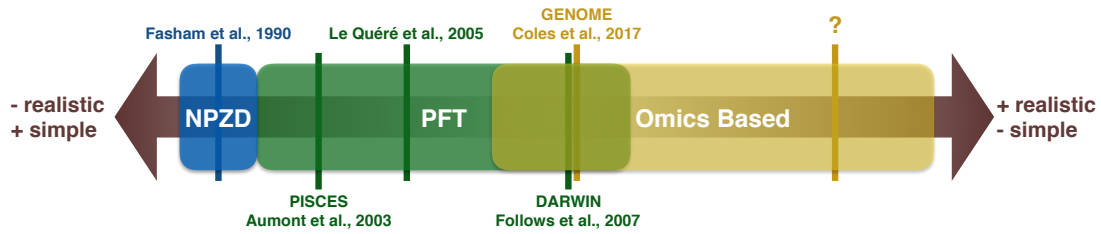

Figure 1.19 - Gradient of planktonic functional and taxonomic diversity implemented in biogeochemical models, with examples of key models discussed in this introduction: Fasham et al. (1990), Aumont et al. (2003), Le Quéré et al. (2005), Follows et al. (2007) and Coles et al. (2017). NPZD stands for nutrient, phytoplankton, zooplankton and detritus models (see section 1.2.1), PFT stands for plankton functional types models (see section 1.2.2), while omics-based models refer to gene-centric and metabolism-based approaches, as well as to potential new theoretical frameworks involving omics data and potentially allowing for a more realistic representation of planktonic diversity in biogeochemical models (see section 1.4).

In particular, I highlighted the need for data-driven methods allowing to define model structural components from observational data, to avoid *a priori* choices of the model

PFTs, traits, genes or metabolic pathways. I demonstrated the promises carried by omics data to tackle this issue, but identified multiple limitations emerging from their current use in theoretical frameworks, notably the bias towards cultivated organisms and well-described pathways in gene-centric and metabolic models, preventing us to take full advantage of the richness of omics datasets. It led me to consider 3 main research questions, which drove the analyses presented in the following chapters of this thesis.

1. Can we use meta-omics data to detect functional traits from which the genomic basis is poorly known?
2. Can we use meta-omics data to predict the distribution of functional traits/genes in the environment through statistical modeling?
3. Can the abundance and distribution of functional traits/genes be quantified in meta-omics data without any *a priori* choice of focal functions and/or species?

I will try to bring answers to these questions in the following parts and chapters of this manuscript:

In **Part I**, entitled *From genes to functional traits in the global ocean: the mixotrophy and DMS production case studies*, I will bring answers to questions 1. and 2. using two *a priori* chosen functional traits as case studies: mixotrophy, from which the genomic basis is poorly known, and DMS production, from which the metabolism is well described. This first part will be composed of two chapters:

Chapter 2 will focus on how metabarcoding can be used to describe the distribution of functional traits, using mixotrophy as an illustration. In section 1.2.2 of this introduction, I evoked how mixotrophic protists are often absent from biogeochemical models, despite growing evidence of their biogeochemical importance. Here, I propose the first ever omics-based assessment of mixotrophic protists abundance in the global ocean, showing their ubiquity and unveiling some interesting characteristics of their biogeography. This chapter will be mainly composed of a manuscript entitled *Mixotrophic protists display contrasted biogeographies in the global ocean*, published in the *ISME Journal* in 2018, of which I am first author.

Chapter 3 will focus on how genomics and transcriptomics data can be used to investigate the genomic basis of functional traits in planktonic lineages, using both mixotrophy and DMS production as illustrations. The chapter will also expose how genomic markers of functional traits can be used to derive quantitative predictions of metabolic functions, through preliminary results on the analysis of the global distribution of genomic markers of DMS production.

Part II, entitled *Data-driven approaches to identify and quantify the functional composition of planktonic communities*, will mainly be focused on questions 2. and 3., as I will present a method allowing to use omics data to quantify gene functional groups in the global ocean without *a priori*. This part will be composed of one chapter:

> **Chapter 4** will describe a data-driven method to identify gene functional groups in meta-omics datasets, quantify their abundance, and study their response to environmental gradients. The method was tested on more than 800 prokaryotic MAGs, and machine learning techniques were used to try to predict the abundance of their gene functional groups from the environmental context. This chapter will be mainly composed of a manuscript entitled *Towards omics-based predictions of planktonic functional composition from environmental data*, which I have submitted as first-author to *Nature Communications* on the 24th of April 2020, and is currently under review.

Chapter 5 will consist in a general discussion, in which I will summarize the advantages and limits of omics-driven approaches to increase diversity in biogeochemical models. I will also comment on the potential combination of omics data with other data types, such as high-throughput imaging, to improve our theoretical understanding of planktonic ecosystems. Finally, I will question if we should keep working at the functional-trait level, switch exclusively to gene-centric level, or maybe even try to come up with a new conceptual framework to improve planktonic diversity representation in biogeochemical models.

Part I

From genes to functional traits in the global ocean: the mixotrophy and DMS production case studies

Chapter 2

Metabarcoding as a tool to decipher the biogeography of functional traits

2.1 Prelude

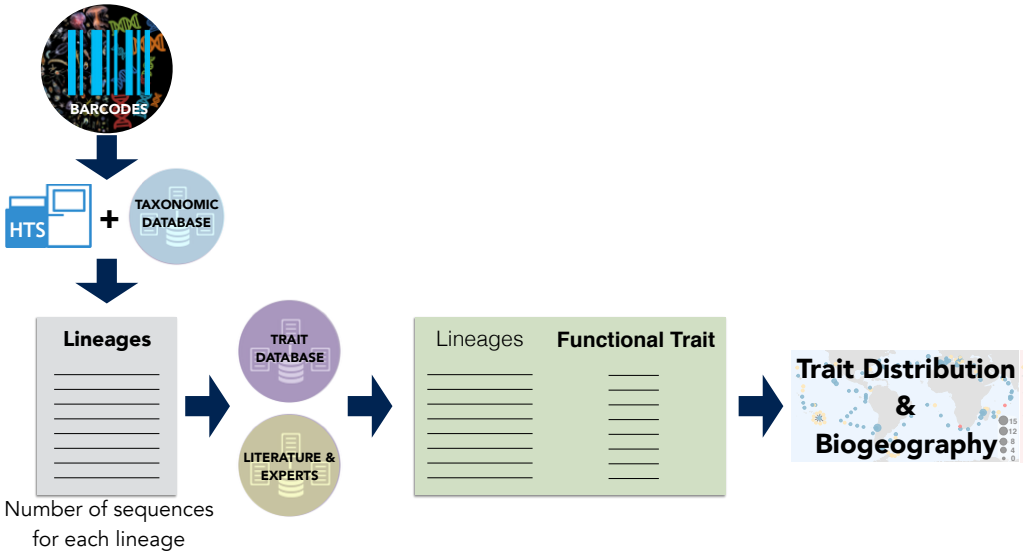

Figure 2.1 - Metabarcoding as an alternative to functional genomic markers for the detection of functional traits with poorly known genomic basis. Metabarcodes can be associated to taxonomic lineages through annotation databases such as PR2 (Guillou et al., 2013), and the number of metabarcodes associated to each lineage is approximately proportional to its biomass (Biard et al. (2017), more on this in the next section). A functional annotation can then allow to link lineages with functional traits, based on the literature and the knowledge of experts. The abundance of metabarcodes associated to each trait can then be computed.

In the introduction, I highlighted how omics data had been included in theoretical frameworks with a strong bias towards well described metabolic pathways. One of the scientific questions that I asked was *can we use meta-omics data to detect functional traits from which the genomic basis is poorly known?*. The first step towards the integration of functional traits in models through omics data is to find ways to detect them *in-situ*,

and hopefully quantify their presence in samples. Such a quantification would allow to explore the distribution of the detected trait, identify potential trade-offs with other traits, and determine its affinity for specific environmental conditions, which are key elements for the implementation of the trait in a modeling framework. The silver bullet for detecting and quantifying a trait in the environment through omics data is the access to functional genomic markers, *i.e.* one or multiple genes directly linked and/or responsible of the organism's functional trait, and of which the genomic and transcriptomic abundances can be used as a proxy of the trait presence and realization. But the detection and quantification of traits that are governed by poorly known or even unknown molecular mechanisms remains problematic (more on this issue in Chapter 3). During the first year of my thesis, I thus explored how metabarcoding could provide an alternative to functional genomic markers for the detection of traits with poorly known genomic basis, relying mostly on the manual annotation of functional traits to taxonomic lineages (Figure 2.1). In particular, I focused my work on the functional trait of mixotrophy.

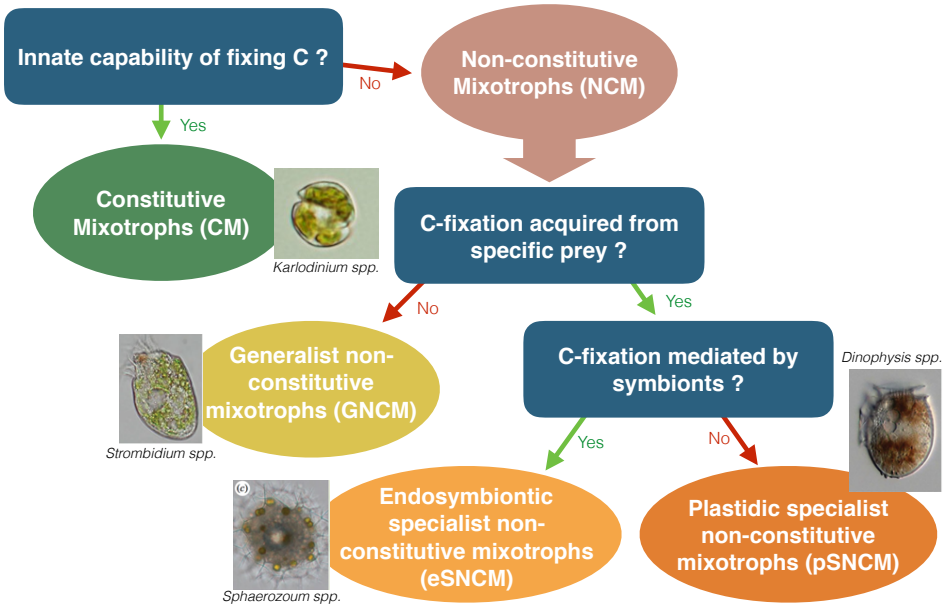

Figure 2.2 - The different types of mixotrophy, or mixotypes. Constitutive mixotrophs have an innate ability to fix carbon through photosynthesis, they are sometimes called 'algae that eat'. Non-constitutive mixotrophs do not have an innate ability to fix carbon, and can be subdivided in three subgroups depending on how they acquire this ability. Generalist non-constitutive mixotrophs (GNCM) can steal the chloroplasts from any of their preys, while plastidic specialist non-constitutive mixotrophs (pSNCM) can steal the chloroplasts from some specific preys. Finally endosymbiotic specialist non-constitutive mixotrophs (eSNCM) bear photosynthtic symbionts. Pictures from Leles et al. (2017).

Mixotrophy is a functional trait that has long been overlooked, but is now considered to be present in the majority of protistan lineages (Flynn et al., 2013). Still, the physiology and

ecology of mixotrophs remain poorly known, in part because as most protistan lineages, they remain challenging to culture in labs (Flynn et al., 2013). The first description of the global biogeography of mixotrophs was published through a review of morphological identification data in 2017, at the beginning of my PhD (Leles et al., 2017). At the time, no study had investigated mixotrophy in meta-omics data, despite their clear ecological significance (Mitra et al., 2014) and the need for information about their abundance in the open ocean (Leles et al., 2017). This lack of meta-omics based studies focusing on mixotrophs can notably be explained by the absence of genomic markers of mixotrophy, *i.e.* genes that could allow to detect mixotrophy in metagenomics and/or metatranscriptomics samples. This absence can notably be explained by the fact that multiple types of mixotrophy exist, each corresponding to organisms with distinct physiologies and behaviors (Detailed in figure 2.2), making it hard to identify specific metabolic pathways associated with this trait. Mixotrophy then appeared as a perfect case study for testing metabarcoding as an alternative to functional genomic markers, and my goal was to investigate the biogeography of mixotrophic protists in the global ocean through meta-omics data, providing the first ever list of metabarcodes corresponding to mixotrophic lineages.

The rest of this chapter will consist in a manuscript entitled *Mixotrophic protists display contrasted biogeographies in the global ocean*, published in the ISME journal in January 2019, and on which I am first author.

2.2 Mixotrophic protists display contrasted biogeographies in the global ocean

Authors: Emile Faure[1,2], Fabrice Not[3], Anne-Sophie Benoiston[2], Karine Labadie[4], Lucie Bittner[2 *], Sakina-Dorothée Ayata[1,2 *].

1-Sorbonne Université, CNRS, Laboratoire d'océanographie de Villefranche, LOV, 06230 Villefranche-sur-Mer, France

2-Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, CP 50, 57 rue Cuvier, 75005 Paris, France

3-Sorbonne Université, CNRS, UMR7144 Adaptation and Diversity in Marine Environment (AD2M) Laboratory, Ecology of Marine Plankton team, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

4-Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Evry, France, 91057 Evry, France

5- Authors contributed equally to this work

Abstract: Mixotrophy, or the ability to acquire carbon from both auto- and heterotrophy, is a widespread ecological trait in marine protists. Using a metabarcoding dataset of marine plankton from the global ocean, 318,054 mixotrophic metabarcodes represented by 89,951,866 sequences and belonging to 133 taxonomic lineages were identified and classified into four mixotrophic functional types: constitutive mixotrophs (CM), generalist non-constitutive mixotrophs (GNCM), endo-symbiotic specialist non-constitutive mixotrophs (eSNCM), and plastidic specialist non-constitutive mixotrophs (pSNCM). Mixotrophy appeared ubiquitous, and the distributions of the four mixotypes were analyzed to identify the abiotic factors shaping their biogeographies. Kleptoplastidic mixotrophs (GNCM and pSNCM) were detected in new zones compared to previous morphological studies. Constitutive and non-constitutive mixotrophs had similar ranges of distributions. Most lineages were evenly found in the samples, yet some of them displayed strongly contrasted distributions, both across and within mixotypes. Particularly divergent biogeographies were found within endo-symbiotic mixotrophs, depending on the ability to form colonies or the mode of symbiosis. We showed how metabarcoding can be used in a complementary way with previous morphological observations to study the biogeography of mixotrophic protists and to identify key drivers of their biogeography.

2.2.1 Introduction

Marine unicellular eukaryotes, or protists, have a tremendous range of life styles, sizes and forms (Caron et al., 2012), showing a taxonomic and functional diversity that remains hard to define (de Vargas et al., 2015; Pawlowski et al., 2012). This variety of organisms is having an impact on major biogeochemical cycles such as carbon, oxygen, nitrogen, sulfur, silica, or iron, while being at the base of marine trophic networks (Caron et al., 2017; Keeling and Campo, 2017; Caron, 2016a; Le Quéré et al., 2005; Amacher et al., 2009). Hence, they are key actors of the global functioning of the ocean.

Historically, marine protists have been classified into two groups depending on their trophic strategy: the photosynthetic plankton (phytoplankton) and the heterotrophic plankton (zooplankton). It is now clear that mixotrophy, *i.e.*, the ability to combine autotrophy and heterotrophy, has been largely underestimated and is commonly found in planktonic protists (Caron, 2016a; Stoecker et al., 2017; Flynn et al., 2013; Selosse et al., 2017). Instead of a dichotomy between two trophic types, their trophic regime should be regarded as a continuum between full phototrophy and full heterotrophy, with species from many planktonic lineages lying between these two extremes (Flynn et al., 2013). Mitra et al. 2016 have proposed a classification of marine mixotrophic protists into four functional groups, or mixotypes. The constitutive mixotrophs, or CM, are photosynthetic organisms that are capable of phagotrophy, also called “phytoplankton that eat” (Mitra et al., 2016). They include most mixotrophic nanoflagellates (*e.g.*, Prymnesium parvum, Karlodinium micrum). On the opposite, the non-constitutive mixotrophs, or “photosynthetic zooplankton”, are heterotrophic organisms that have developed the ability to acquire energy through photosynthesis (Stoecker et al., 2017). This ability can be acquired in three different ways: the generalist non-constitutive mixotrophs (GNCM) steal the chloroplasts of their prey, such as most plastid-retaining oligotrich ciliates (*e.g.*, Laboea strobila), the plastidic specialist non-constitutive mixotrophs (pSNCM) steal the chloroplasts of a specific type of prey (*e.g.*, Mesodinium rubrum or Dinophysis spp.), and finally the endo-symbiotic specialist non-constitutive mixotrophs (eSNCM) are bearing photosynthetically active endo-symbionts (most mixotrophic Rhizaria from Collodaria, Acantharea, Polycystinea, and Foraminifera, as well as dinoflagellates like Noctiluca scintillans).

As drivers of biogeochemical cycles in the global ocean, and particularly of the biological carbon pump (Keeling and Campo, 2017; Ducklow et al., 2001; Guidi et al., 2016), marine protists are a key part of ocean biogeochemical models (Le Quéré et al., 2005; Aumont et al., 2015; Follows et al., 2007; Reed et al., 2014). However, physiological details of mixotrophic energy acquisition strategies have only been studied in a restricted number of lineages (Stoecker et al., 2017; Johnson, 2011; Stoecker et al., 2009). They appear to be quite complex and greatly differ across mixotypes, which makes mixotrophy hard to include in a simple model structure (Flynn and Mitra, 2009; Ward and Follows, 2016; Berge et al., 2017; Ghyoot et al., 2017; Ward et al., 2011). Hence at this time, mixotrophy is not included in most biogeochemical models, neglecting the amount of carbon fixed by non-constitutive mixotrophs through photosynthesis, and missing the population dynamics of photosynthetically active constitutive mixotrophs that can still grow under nutrient limitation (Ghyoot et al., 2017; Mitra et al., 2014). This is most probably skewing climatic models

predictions (Mitra et al., 2016, 2014), as well as our ability to understand and prevent future effects of global change.

A better understanding of the environmental diversity of marine mixotrophic protists, as well as a description of the abiotic factors driving their biogeography at global scale are still needed, in particular to integrate them in biogeochemical models. Leles et al. 2017 attempted to tackle this problem by reviewing about 110,000 morphological identification records of a set of more than 60 mixotrophic protists species in the ocean, taken from the Ocean Biogeographic Information System (OBIS) database. They found distinctive patterns in the biogeography of the three different non-constitutive mixotypes (GNCM, pSNCM, and eSNCM), highlighting the need to better understand such diverging distributions (Leles et al., 2017). Environmental molecular biodiversity surveys through metabarcoding have been widely used in the past fifteen years to decipher planktonic taxonomic diversity (de Vargas et al., 2015; Stoeck et al., 2010; Bik et al., 2012; Bittner et al., 2013). Here, we exploited the global *Tara Oceans* datasets (Karsenti et al., 2011; Alberti et al., 2017; Pesant et al., 2015), and identified 133 mixotrophic lineages, that we classified into the four mixotypes defined by Mitra et al. 2016. This first ever set of mixotrophic metabarcodes allowed us to investigate the global biogeography of both constitutive and non-constitutive mixotrophs, in relation with in-situ abiotic measurements. We tested (i) if new information on marine mixotrophic protists distribution can be gained in comparison with previous morphological identifications (Leles et al., 2017); (ii) if the constitutive mixotrophs, which are not addressed in Leles et al. 2017, and the non-constitutive mixotrophs diverge in terms of biogeography; (iii) if the study of diversity and abundance of environmental metabarcodes could lead to the definition of key environmental factors shaping mixotrophic communities.

2.2.2 Materials and methods

2.2.2.1 Samples collection and dataset creation

Metabarcoding datasets from the worldwide *Tara Oceans* sampling campaigns that took place between 2009 and 2013 (Karsenti et al., 2011; Pesant et al., 2015) (data published in open access at the European Nucleotide Archive under project accession number PRJEB6610) were investigated. We analyzed 659 samples from 122 distinct stations, and for each sample, the V9-18S ribosomal DNA region was sequenced through Illumina HiSeq (Alberti et al., 2017). Assembled and filtered V9 metabarcodes (cf. details in de Vargas et al. (de Vargas et al., 2015)) were assigned to the lowest taxonomic rank possible via the Protist Ribosomal Reference (PR2) database (Guillou et al., 2013). To limit false positives, we chose to only analyze the metabarcodes (*i.e.*, unique versions of V9 sequences) for which the assignment to a reference sequence had been achieved with a similarity of 95% or higher. This represents 65% of the total dataset in terms of metabarcodes and 84% in terms of total sequences. Our dataset involved 1,492,912,215 sequences, distributed into 4,099,567 metabarcodes assigned to 5071 different taxonomic assignations, going from species to kingdom level precision.

2.2.2.2 Defining a set of mixotrophic organisms

Among these 5071 taxonomic assignations, we searched for mixotrophic protist lineages, taking into account the four mixotypes described by Mitra et al. (Mitra et al., 2016): constitutive mixotrophs (CM), generalist non-constitutive mixotrophs (GNCM), endo-symbiotic specialist non-constitutive mixotrophs (eSNCM), and plastidic specialist non-constitutive mixotrophs (pSNCM). We used the table S2 from Leles et al. (Leles et al., 2017), which is referencing 71 species or genera belonging to three non-constitutive mixotypes (GNCM, pSNCM, and eSNCM), as well as multiple other sources coming from the recent literature on mixotrophy (Caron, 2016a; Stoecker et al., 2017; Flynn et al., 2013; Mitra et al., 2016; Esteban et al., 2010; Granéli et al., 2012; Liu et al., 2010; Hansen et al., 2012; Agatha et al., 2005; Jones et al., 1993; Johnsen et al., 1999; Rhodes and Burke, 1996; Hemleben et al., 1977; Fehrenbacher et al., 2011; Spero and Parker, 1985; Faber et al., 1989; Kuile and Erez, 1984; Biard et al., 2017), and inputs from mixotrophic protists' taxonomy specialists (cf. Acknowledgments section). Within the 5071 taxonomic assignations of variable precisions, we identified 5 GNCM, 9 pSNCM, 77 eSNCM, and 42 CM lineages (detailed list available publicly under the `https://doi.org/10.6084/m9.figshare.6715754`, and all metabarcodes were tagged with their mixotypes in the PR2 database). Among these 133 taxonomic assignations that we will call "lineages", 92 were defined at the species level, 119 at the genus level, and the last 14 at higher taxonomic levels where mixotrophy is always present (mostly eSNCM groups like Collodaria). In the Chrysophyceae family, metabarcodes assigned to clades B2, E, G, H, and I were included even though we couldn't find a general proof that all species included in these clades have mixotrophic capabilities. However, if we exclude the photolithophic Synurophyceae and genera like Paraphysomonas and Spumella, which we did, a vast majority of Chrysophyceae are considered mixotrophic (Flynn et al., 2013). The final dataset included 318 054 metabarcodes assigned to the 133 mixotrophic lineages selected, as well as their sequence abundance in 659 samples (table available publicly under the `https://doi.org/10.6084/m9.figshare.6715754`).

2.2.2.3 Environmental dataset

We built a corresponding contextual dataset using the environmental variables available in the PANGAEA repository from the *Tara Oceans* expeditions (Pesant et al., 2015; Ardyna et al., 2017). The 235 environmental variables available were a priori reduced to 84, keeping only one version of each variable that was calculated twice or more using different tools, units and/or formulas. For example, the daily photosynthetically active radiations (PAR) were measured using 10 different approaches, some using *in situ* sensors, others using satellite observations combined with equations based on the diffuse attenuation coefficient (Morel et al., 2007). For similar reasons, among the conductivity, the temperature, and the salinity, only the last two were kept. Then, among the 83 remaining variables, only the ones recorded for at least half of the samples were kept. We obtained a table composed of 57 different environmental variables (Available publicly under the DOI 10.6084/m9.figshare.6715754). In this table, 8.67% of the values were missing. This table contained environmental context for 658 of our 659 samples, lacking environmental information only for the DCM of station number 120. This sample was then removed for the

statistical analysis.

2.2.2.4 Distribution and diversity of mixotrophic protists

For each mixotype, the number of metabarcodes, the total sequence abundance and the mean sequence abundance by metabarcode was computed (Table 2.1). Also, we measured each metabarcode's station occupancy, *i.e.*, the number of stations in which it was found, and station evenness, *i.e.*, the homogeneity of its distribution among the stations in which it was detected (Fig. 2.4). Diversity of mixotrophic protists was investigated through mixotype-specific metabarcode richness per station (Table 2.1). As the number of samples taken per station can impact the abundance and diversity of detected metabarcodes, richness was computed only at stations for which the maximum number of eight samples were available (40 stations over 122).

Mixotypes	CM	eSNCM	pSNCM	GNCM
Number of lineages used in this study	42	77	9	5
Number of V9 metabarcodes	26,015	288,536	2143	1360
Total sequence abundance	3,581,751	86,098,397	208.096	63.622
Mean sequence counts per metabarcode	137.7	298.4	97.1	46.8
Mean metabarcode richness per station[a] (std dev)	2162 (1115)	18502 (9238)	67 (102)	84 (111)
Number of absences/station	0/122	0/122	5/122	3/122

Table 2.1 - Detailed number of lineages found for each mixotype, as well as the number of metabarcodes, the corresponding total sequence counts over all stations, the mean sequence abundance by metabarcode, and mean metabarcode richness.
The richness was computed as the number of different metabarcodes present at each station. It was calculated for each mixotype and means are indicated in the fifth line. Absences correspond to the number of stations in which no sequences were detected for the corresponding mixotype.
CM constitutive mixotrophs, GNCM generalist non-constitutive mixotrophs, eSNCM endo-symbiotic specialist non-constitutive mixotrophs, pSNCM plastidic specialist non-constitutive mixotrophs.
[a]The mean indicated here was calculated using only stations having the maximum number of samples (see main text)

2.2.2.5 Global biogeography of mixotrophic protists

Two statistical analyses were performed to investigate mixotrophic protists biogeography. One at the metabarcode level, and one at the lineage level, *i.e.*, merging the sequence abundance of metabarcodes sharing the same taxonomical assignation. The metabarcodes abundance table was composed of 318,054 rows/metabarcodes, and 659 columns/samples, whereas the lineage abundance table was composed of 133 rows/lineages and 659 columns/samples (both datasets are available publicly at `https://doi.org/10.6084/m9.figshare.6715754`).

The two redundancy analyses led to very similar conclusions, but the biogeography of lineages appeared easier to visually represent and interpret than the one of metabarcodes. Hence, we will only present the lineage-based analysis here, before presenting and discussing the methods used and results obtained for the metabarcodes level analysis in a separate section (Section 2.2.5).

Our statistical model was designed to identify lineages (or metabarcodes) with contrasted biogeographies, and relate their presence to the environmental context. We normalized the sequence counts from the lineage abundance matrix using a Hellinger transformation (Legendre and Legendre, 1998). We used the environmental dataset and the mixotrophic lineages' abundance matrix as explanatory and response matrices, respectively, to conduct a redundancy analysis (RDA) (Legendre and Legendre, 1998). Since redundancy analyses (RDA) cannot handle missing values in the explanatory dataset, we replaced missing environmental variables by their means across all samples, to keep all the environmental variables in the analysis. This option was selected over a joint modelling approach,because of a too high collinearity among some variables, and difficulties to define a multivariate distribution fitting the whole dataset. Environmental variables were then centered prior to the analysis. For that, we made a species pre-selection using Escoufier's vectors (Escoufier, 1973), which allowed to keep only the 62 most significant mixotrophic lineages. This method selects lineages according to a principal component analysis (PCA), sorting them based on their correlation to the principal axes. We then used a maximum model (Y X) and a null model (Y 1) to conduct a two directional stepwise model selection based on the Akaike information criterion (AIC) (Borcard et al., 2011). The resulting model contained 28 response variables, among which five were qualitative (filter size, biogeographical province *sensu* Longhurst (Longhurst, 1998), season, season moment,*i.e.* early, middle or late, and sampling depth, *i.e.* surface or DCM), and 23 quantitative. The latter included: longitude, bathymetry, distance to coast, mixed layer depth, euphotic zone depth, oxygen maximum depth, ammonium at 5m, temperature, silica, oxygen, chlorophyll a, daylight duration, absorption coefficient of colored dissolved organic matter (ac-CDOM), calcite saturation state, Okubo-Weiss parameter, PO4, CO3, HCO3, photosynthetically active radiations (PAR), salinity, maximum Lyapunov exponent, optical beam attenuation coefficient at 660 nm and beam attenuation coefficient of particles. Analyses and graphs were realized with the R software version 3.4.3 (R Core Team, 2019). All scripts are available on GitHub platform (`https://github.com/upmcgenomics/MixoBioGeo`).

2.2.3 Results

2.2.3.1 Global distribution and diversity of marine mixotrophic protists

Mixotrophic protists metabarcodes were detected in all the 659 samples with a total sequence abundance of 89,951,866, representing 12.56% of the total sequence abundance in the 659 samples studied. They represented a mean of 12.64% of the total sequence abundance per sample, with a maximum of 96.96% and a minimum of 0.01%. To avoid any potential overestimation of mixotrophic lineages presence in the following results, we marked all records of less than a hundred sequences as questionable. We found both eSNCM and CM in each of the 122 stations studied (Table 2.1 and Fig. 2.3). In only two occasions the number of sequences belonging to CM was questionable, at stations for which only one sample was sequenced. GNCM were found absent in only two stations and their presence was questionable in 39 stations (Fig. 2.3). pSNCM were absent at five stations (three in the Indian Ocean, and two in the Pacific Ocean) and detected

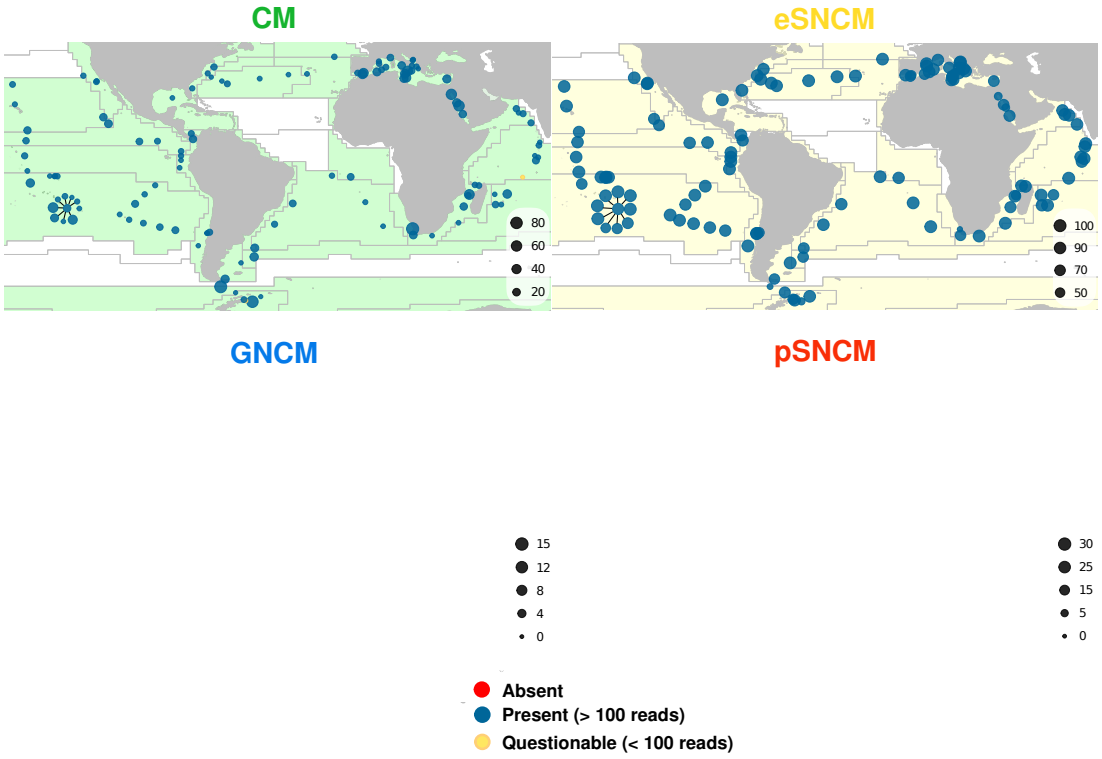

Figure 2.3 - Global distribution of mixotypes from metabarcoding data. Maps showing for each station the proportion of sequences (in %) belonging to each mixotype over the total number of mixotrophic sequences. Stations in which no sequence was found were marked as absent, ones with less than 100 sequences marked as questionable. Each Longhurst biogeographical provinces (Longhurst, 1998) is colored in the background if more than 100 sequences are detected in at least one of its stations.

with questionable presence in 54 additional stations, which were mostly located in the central Pacific and the Indian Ocean (Fig. 2.3). We found significant amounts of sequences corresponding to GNCM in the Central Pacific, Southern subtropical Atlantic, and Indian Ocean. The presence of GNCM in these areas has not yet been recorded through morphological identifications during field expeditions (Leles et al., 2017). Also, we detected more than 100 sequences of pSNCM metabarcodes at 11 stations belonging to biogeographical provinces in which no morphological identifications had been published (Leles et al., 2017; Longhurst, 1998), mostly in offshore areas of the Atlantic and Pacific Ocean (Fig. 2.3). The mean evenness of mixotrophic metabarcodes across stations was of 0.87, and 82.3% of the metabarcodes had a station evenness above 0.5 (Fig. 2.4). Station occupancy varied a lot depending on the metabarcodes, with a high density of rare metabarcodes leading to a mean of 5.14 stations over a maximum of 122, and a standard deviation of 7.7. However, three eSNCM metabarcodes were found in all the 122 stations, and three CM metabarcodes were detected in 121 stations. The maximum occupancy for a GNCM metabarcode was of 111 stations, while 92 stations was the maximum for a pSNCM metabarcode. CM and GNCM metabarcodes showed a strong tendency towards high evenness values (Fig. 2.4, means of 0.90 and 0.95, respectively), even for the most sequence abundant metabarcodes. Many eSNCM metabarcodes had high evenness values, but below average values were detected for the

most abundant ones (Fig. 2.4, global mean of 0.87). pSNCM metabarcodes had a similar mean of evenness values (0.87), but a different distribution compared to other mixotypes (Fig. 2.4). Among the 50 most abundant metabarcodes, 43 corresponded to Collodaria lineages, 47 were eSNCM and 3 were CM, all three assigned to Gonyaulax polygramma. GNCM and pSNCM metabarcodes had homogeneously low sequence abundances (Fig. 2.4 and Table 2.1).

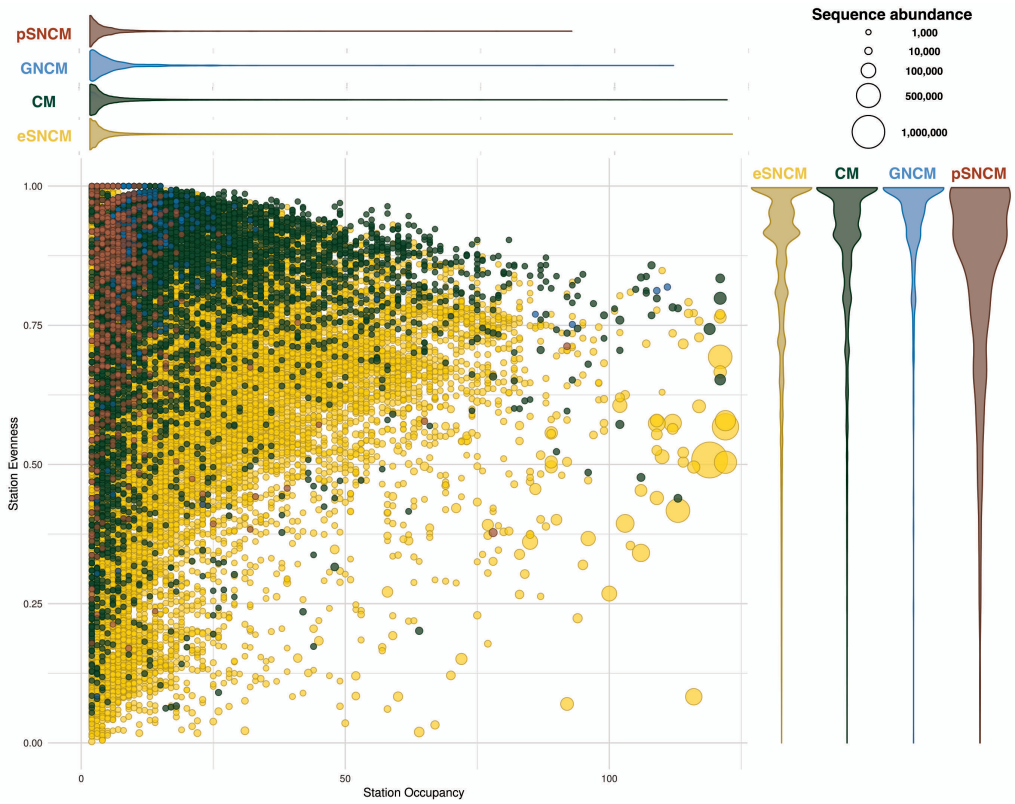

Figure 2.4 - Sequence abundance, occupancy, and spatial evenness of each mixotrophic metabarcode across sampled stations. Each metabarcode is plotted as a bubble, with its station occupancy, i.e., the number of stations in which it was found, and its station evenness, i.e., the homogeneity of its distribution among the stations in which it was detected, as coordinates. Violin plots were drawn for each mixotype on both the x and y axes. The size of each bubble is scaled to the sequence abundance found globally for the corresponding metabarcode.

2.2.3.2 Main factors affecting the biogeography of mixotrophic protists

The redundancy analysis helped to investigate further the environmental variables responsible for the mixotrophic protists' biogeography. The 62 lineages selected with the Escoufier's vector method corresponded to 20 CM, 34 eSNCM, 3 GNCM, and 5 pSNCM. Even after selection, a significant part of the lineages did not show any response to environmental data in their distribution (Fig. 2.5, *e.g.*, 19 of the 62 lineages were found between -0.01 and 0.01 on both RDA1 and RDA2). The adjusted R-squared of the RDA was of 34.89% (41.43% unadjusted), with 24.01% of variance explained on the two first axes (Fig. 2.5). The first RDA axis (14.96%) marks an opposition between samples from oligotrophic waters with low productivity (RDA1 > 0) and samples from eutrophic and productive water masses (RDA1 < 0). This axis is negatively correlated to chlorophyll concentration, particles density, ammonium concentration, absorption coefficient of

colored dissolved organic matter (acCDOM), duration of daylight, silica, CO3, oxygen, and PO4 concentration, as well as longitude. It is positively correlated to bathymetry, deep euphotic zone, deep oxygen maximum, deep mixed layer, as well as to the distance to coast. The second RDA axis (9.05%) is opposing offshore and subpolar samples (RDA2 > 0) to coastal and subtropical ones (RDA2 < 0). The axis is positively correlated to the depth of the mixed layer, the dis- tance to coast, the bathymetry, high maximum Lyapunov exponents as well as high concentrations of PO4, oxygen, CO3 and silica. It is negatively correlated to temperature, salinity, and photosynthetically active radiations (PAR).

Figure 2.5 - Impact of environmental variables on the distribution of marine mixotrophs. Triplot of the redundancy analysis (RDA) computed on the 62 Escoufier-selected lineages, after model selection. The adjusted R-squared of the analysis is of 34.89% (41.43% unadjusted). Each gray dot corresponds to a sample, i.e., one filter at one depth at one station. The blue dashed arrows correspond to the quantitative envir- onmental variables. Abbreviations: MLD mixed layer depth, O2MaxD O2 maximum depth, EuphzoneD euphotic zone depth, PAR photo- synthetically active radiations, Calcite Sat. St. Calcite Saturation State, c_660 optical beam attenuation coefficient at 660 nm, c_part beam attenuation coefficient of particles, acCDOM absorption coefficient of colored dissolved organic matter. Plain arrows correspond to mixo- trophic lineages, colors indicating mixotypes. For more readability, we do not represent all qualitative variables included in the model. That is why only the filter centroids are appearing, even though the sampling depth, season, season moment, i.e., early, middle or late, and bio- geographical province were used in the RDA calculation

Among the 20 CM lineages, seven clearly emerged from the redundancy analysis (Fig. 2.5) and showed distinct biogeographies related to environmental variables. Gonyaulax polygramma, Alexandrium tamarense, and Fragilidium mexicanum, three Dinophyceae belonging to the Gonyaulacales order, were mainly found in oligotrophic waters with a deep euphotic zone, warm temperature, high salinity, and PAR (RDA1 > 0, RDA2 < 0). The four other CMs (involving all the

Chrysophyceae included in the analysis as well as one Dinophyceae from the Kareniaceae family, Karlodinium micrum) were found mostly in productive water masses (RDA1 < 0).

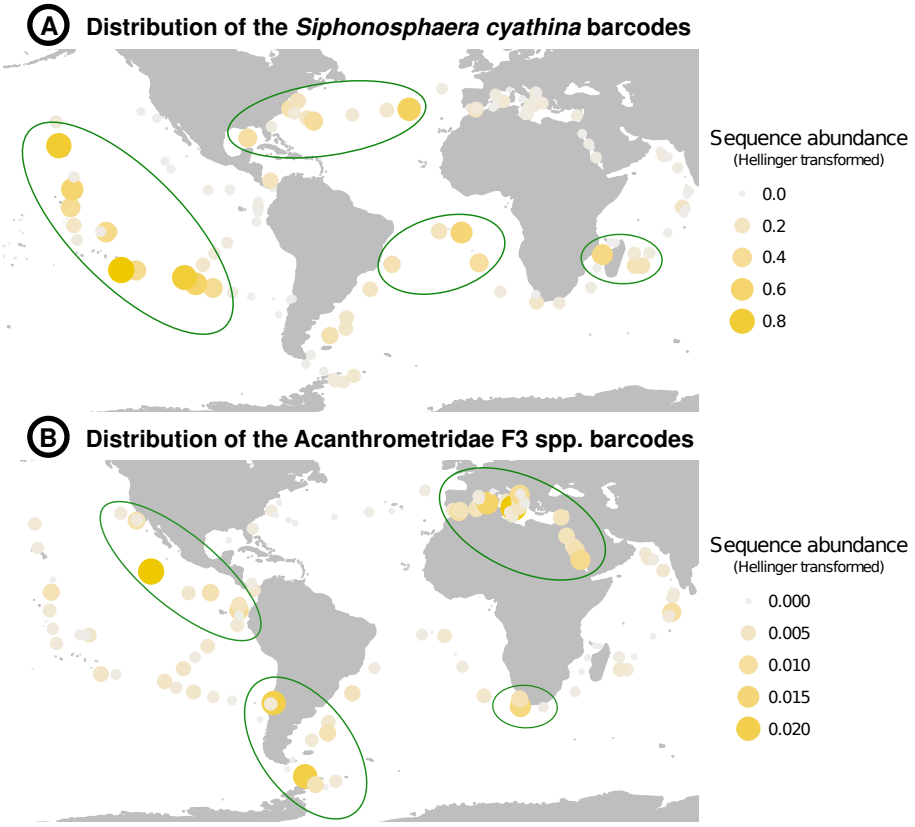

Figure 2.6 - Contrasted global distributions of metabarcodes corresponding to two eSNCM lineages. Maps of Hellinger-transformed sequence count abundances for metabarcodes assigned to the Collodaria Siphonosphaera cyathina (a) and the Acantharia Acanthrometridae F3 spp. (b). These two lineages are opposed on the first RDA axis (Fig. 2.5 and 2.8). Size and color both illustrate abundance for better readability. Ellipses were drawn to highlight high abundance zones, and reveal the differences in lineages distribution.

eSNCMs can be divided in three groups in the RDA space. The first group (RDA1 < 0) corresponds to eSNCM species dominating rich and productive environments. It includes mainly Acantharia and Spumellaria species. The second group (RDA1 > 0) dominates oligotrophic environments, and includes multiple Collodaria as well as one Dinophyceae genus (Ornithocercus). Within this group, Ornithocercus spp. is found mainly in coastal subtropical environments (RDA2 < 0), as opposed to Sphaerozoum punctatum that is found mainly in offshore subpolar regions (RDA2 > 0). Siphonosphaera cyathina lies between these two trends as it is found only in oligotrophic samples, but is not influenced by temperature or bathymetry (Figs. 2.5 and 2.6). The third group corresponds to the eSNCM lineages that can be interpreted as distributed homogeneously in regards of the environmental data we are using (*e.g.*, lineages with the shortest arrows in Fig. 2.5). These notably include the 12 Foraminifera lineages present in the RDA. Looking at filters centroids in the RDA space (Fig. 2.5), we can suppose that eSNCM lineages dominating eutrophic systems (RDA1 < 0) are smaller in size than those dominating oligotrophic ones (RDA1 > 0).

Out of the five pSNCM included in the RDA, only Mesodinium rubrum, the most abundant one, is distinctively represented in the RDA space. This suggests that the other pSNCM have homoge-

neous distributions in response to our environmental variables. Mesodinium rubrum dominates eutrophic environments, independently from the bathymetry or the temperature (RDA1 < 0, RDA2 ≈ 0). We find a similar pattern for GNCM, with only Pseudotontonia simplicidens well represented in the RDA space out of the three species included in the analysis. Like M. rubrum, Pseudotontonia simplicidens is the most abundant species in its group and it is mainly found in eutrophic waters (RDA1 < 0).

2.2.4 Discussion

2.2.4.1 Mixotrophy occurs everywhere in the global ocean

Our metabarcoding survey confirms that marine mixotrophic protists are ubiquitous in the global ocean (Leles et al., 2017), possibly extending the known range of distribution of two mixotypes (Figs. 2.3 and 2.4). Mixotrophic organisms represented more than 12% of the sequences in the complete *Tara Oceans* metabarcoding dataset, showing that they should not be understated. We found contrasted biogeographies among metabarcodes and their corresponding lineages, both within and across mixotypes (Figs. 2.4, 2.5, 2.6 and Section 2.2.5). We found constitutive mixotrophs (CM) and endo-symbiotic specialist non-constitutive mixotrophs (eSNCM) metabarcodes at all the 122 stations included in this global study (Table 2.1 and Fig. 2.4), verifying that these two mixotypes are the most abundant in the ocean (Leles et al., 2017). This dominance of eSNCM and CM in our data is also linked to the relatively high number of metabarcodes available for these two mixotypes in databases. Using 1360 generalist non-constitutive mixotrophs (GNCM) metabarcodes corresponding to only five lineages, we detected them in ten biogeographical provinces (Longhurst, 1998) where no morphological identification had been recorded before (Leles et al., 2017). GNCM metabarcodes had consistently high evenness values, and some had station occupancy records comparable to the most abundant eSNCM and CM metabarcodes (Fig. 2.4). These results support the hypothesis of a globally ubiquitous distribution of GNCM. Plastidic specialist non-constitutive mixotrophs (pSNCM) were found in five provinces in which no record existed so far from morphological identification field studies (Leles et al., 2017). However, these observations were often in a questionable range in terms of sequence abundance (Fig. 2.3), and the overall distribution of pSNCM in our data appears as very concordant with morphological observations (Leles et al., 2017). pSNCM metabarcodes had dominantly low station evenness values, which again supports the conclusions of Leles et al. 2017 that identified pSNCM as highly seasonal and spatially restricted in their distribution.

While building our set of mixotrophic lineages, some widespread and potentially mixotrophic genera did not appear, such as Ceratium spp., Tontonia spp., Amphisolenia spp., Triposolenia spp., or Citharistes spp., mainly because of a poor representation in the PR2 database. Also, we decided to only consider metabarcodes with more than 95% similarity to a reference sequence. This threshold could be too selective for some species and not enough for some others, as single similarity threshold are hardly efficient when studying whole eukaryotic populations (Wu et al., 2015; Brown et al., 2015). For example, some species appeared with low sequence abundance in

the data even though they couldnot have been sampled, such as three lacustrine species, *e.g.*, Poteriospumella lacustris. Considering these biases and the sometimes relatively low sequence counts (marked as questionable in Fig. 2.3), some of the new GNCM and pSNCM records we observed should be considered with care, as they could be over-estimated or even sometimes artefactual. However, the low number of lineages found for these two mixotypes in PR2 and in our dataset are leading us to think that we were unable to capture the whole GNCM and pSNCM communities. This supposes a global underestimation of GNCM and pSNCM abundances in our results.

Tara Oceans metabarcoding dataset is built on snapshot samples taken irregularly during a 3-year cruise, hence allowing no proper seasonal variations investigations. However, morphological identifications of mixotrophic protists revealed seasonal variations in their abundance, with Mesodinium biomass blooming in spring in coastal seas for example (Leles et al., 2017). As metabarcoding datasets have been successfully applied on time series to detect species successions across gradients of time and space (Egge et al., 2013; Gilbert et al., 2010; DeLong et al., 2006), it would be interesting to similarly investigate seasonal trends in mixotrophic communities. Our set of mixotrophic lineages and metabarcodes being publicly available, our method will be applicable to any other metabarcoding dataset, including time series. It will also be open to inputs and updates from the global scientific community.

2.2.4.2 The contrasted biogeographies of marine mixotypes

Constitutive mixotrophs

Constitutive mixotrophs (CM) have very diverse feeding behaviors, with some species requiring phototrophy to grow, others phagotrophy, and some being obligate mixotrophs (Stoecker et al., 2009). They were described in all waters of the global ocean (Arenovski et al., 1995; Safi and Hall, 1999; Moorthi et al., 2009; Unrein et al., 2010; Sanders and Gast, 2012). We found them distributed in a range of conditions almost as wide as non-constitutive mixotrophs (Figs. 2.3 and 2.5). Among highly abundant lineages, most were dominantly found in eutrophic and shallow habitats. However, a few dinoflagellates were found to be highly dominant in oligotrophic, subtropical waters, showing how wide of a range of conditions constitutive mixotrophs can grow in (Fig. 2.5). This illustrates how mixotrophy can allow organisms to dominate ecosystems even when environmental conditions are poorly adapted to purely phototrophic or heterotrophic organisms. When taken explicitly into account in biogeochemical models, marine mixotrophs increase carbon export by up to 30% (Ward and Follows, 2016). Hence, their global ubiquity supposes that the carbon export of the biological carbon pump could be underestimated in both oligotrophic and eutrophic areas (Mitra et al., 2016).

Plastidic specialist and generalist non-constitutive mixotrophs (pSNCM and GNCM)

Like Leles et al. 2017, we found pSNCM and GNCM to have quite similar biogeographies (Fig. 2.5, Section 2.2.5). Sequence abundance of most of the metabarcodes for these two mixotypes was homogeneously low (Table 2.1), but the two most abundant species, Mesodinium rubrum

(pSNCM) and Pseudotontonia simplicidens (GNCM), were found mostly in coastal and eutrophic waters, consistently with Leles et al.'s 2017 morphological observations (Fig. 2.5, Section 2.2.5). No species-level barcode is available in the PR2 database for the Tontonia genus, and only one can be found for Pseudotontonia and Laboea genera, even though morphological records of these GNCM are numerous (Leles et al., 2017). Experiments using meso- and microcosms combined with individual counts and morphological identification have found that GNCM ciliates can represent up to half of the individuals in ciliate communities of the photic zone (Mitra et al., 2016; Calbet et al., 2012; Dolan and PÉrez, 2000). A proportion we would have trouble to reach with the five lineages we were able to consider, knowing that there are 8686 different ciliate lineages available in PR2. This highlights the urgent need for supplementing 18S reference databases for mixotrophic ciliates.

Endo-symbiotic specialist non-constitutive mixotrophs (eSNCM)

Endo-symbiotic specialist non-constitutive mixotrophs (eSNCM) is by far the most widespread and abundant non-constitutive mixotype in the global ocean (Figs. 2.3 and 2.4) (Leles et al., 2017; Biard et al., 2017; Decelle et al., 2012). Their biogeography stands out, with a lot of highly abundant ubiquitous lineages, and some other specialized towards certain types of ecosystems (Fig. 2.5). They represent 95.7% of the sequence counts in our study and correspond to 90.7% of the metabarcodes (Table 2.1), which highlights their abundance and diversity. The very high number of rDNA copies present in Rhizaria orders such as Collodaria (Biard et al., 2017) might lead the eSNCM to appear more abundant in metabarcoding datasets than they ecologically are. However, in oligotrophic open oceans the Rhizaria biomass is estimated to be equivalent to that of all other mesozooplankton (Biard et al., 2016), and positively correlated to the carbon export (Guidi et al., 2016), showing how ecologically important they can be.

Investigating the divergent biogeographies of Collodaria and Acantharia

Collodaria are living either as solitary large cells or as colonies (Biard et al., 2017), which explains why they are predominantly found in macro-sized (180–2000 µm) filter samples (Fig. 2.5). All described Collodaria species so far harbor photosynthetic endo-symbionts, mostly identified as the dinoflagellate species Brandtodinium nutricula (Biard et al., 2017; Probert et al., 2014). These dinoflagellates are able to get in and out of their symbiotic state, which implies a light and/or reversible effect of the Collodarian host on its symbiont metabolism (Probert et al., 2014). Based on the same metabarcoding dataset, Collodaria were described as particularly abundant and diverse in the oligotrophic open ocean (Biard et al., 2017). In our results, Collodaria dominate oligotrophic, relatively deep waters (Figs. 2.5 and 2.6a). These Collodaria appear opposed to another set of Rhizaria (Acantharia and Spumellaria) linked to eutrophic and shallow waters (Figs. 2.5 and 2.6b, Section 2.2.5). Acantharia are found ubiquitously in the global ocean, but display particularly high sequence abundances in some specific regions (Decelle et al., 2012). Mixotrophic Acantharia live in symbiosis with the cosmopolitan haptophyte Phaeocystis, which is highly abundant and ecologically active in its free-living phase (Decelle et al., 2012). Unlike the one of Collodaria, this symbiosis is irreversible: an algal symbiont can not go back to its free-living phase (Decelle et al., 2012). Our results suppose that these specific symbiotic modes could enable Acantharia

and Collodaria to dominate different ecosystems (Figs. 2.5 and 2.6). Moreover, living in colonies as Collodaria could help to dominate oligotrophic systems, *e.g.*, by accumulating more food and nutrients through their gelatinous extra-cellular matrix (Decelle et al., 2012). Experiments and modeling studies should help to investigate the contribution of this assumption, comparing food acquisition capacity and growth rates of free individuals versus in colony.

2.2.4.3 Towards an integration of mixotrophic diversity into marine ecosystem models

The future of marine communities' modeling lies in the integration of omics datasets into modeling frameworks (Reed et al., 2014; Stec et al., 2017; Dick, 2017; Mock et al., 2016; Coles et al., 2017). The use of metabolic networks and gene-centric methods has already shown very promising results in modeling prokaryotic ecological dynamics (Reed et al., 2014; Coles et al., 2017). However, eukaryotic metabolic complexity makes these methods hard to apply on protists for now, and we still lack a universal theoretical framework on how to integrate such methods into concrete modeling (Stec et al., 2017). Mixotrophic protists are physiologically complex, and their feeding behavior can vary drastically on short time scales (Stoecker et al., 2017). It will then take a few more years of comparative genomics and transcriptomics studies before being able to model their physiology with purely gene-based approaches. Still, mechanistic models of mixotrophy exist and are quite complex (Flynn et al., 2013; Ghyoot et al., 2017), even if the one from Ghyoot et al. 2017 could be implemented in a global biogeochemical model (Shuter, 1979). Most models make the choice to represent either one or two (NCM and CM) types of organisms able to play the role of all mixotypes depending on parameterization. However, no global agreement has been reached on to what extent the different mixotypes should be modeled. This is mainly due to a lack of quantitative and comparative data on the global impact of grazing and carbon fixation by the different mixotypes (Millette et al., 2018). With our study, we show how meta-omics data can be used to identify groups of organisms distributed differently in response to the environment. It also allows the identification of ecological traits and environmental factors potentially responsible for these divergences. This information can be used to identify key species or lineages, and design controlled experiments with variations of targeted environmental factors to produce the quantitative data needed by modelers. Considering our results, we propose that host-symbiont dynamics of eSNCM should be investigated as a trait playing a potential role on Rhizaria ability to thrive in oligotrophic conditions. Particularly, the mechanisms behind holobiont formation and its potential reversibility could play major roles on eSNCM carbon fixation in various nutrient conditions. Future experiments comparing responses of Collodaria and Acantharia holobionts to different stresses in terms of grazing and carbon fixation could lead to a better understanding of the physiological differences between their two modes of symbiosis. Also, our results suggest that the metabolic flexibility of CM should allow this mixotype to grow in almost any conditions, with individual species probably spanning continuously between complete autotrophy and complete heterotrophy. The risk is then to create a "perfect beast" mixotroph dominating all systems (Flynn and Mitra, 2009). To avoid that, we need more comparative data on grazing and carbon

fixation of obligate phototrophs versus obligate heterotrophs in response to nutrient depletion and environmental fluctuation. Here again, meta-omics data could help to identify candidates for efficient experiment designs. Finally, the small number of lineages of GNCM and pSNCM in our study makes it hard to come up with strongly supported conclusions on whether they should be differentiated in models or not. They seem to share similar biogeographies using snapshot data (Fig. 2.5, Section 2.2.5), but considering that they have different abilities for conserving stolen chloroplasts over time, it might not be the case when looking at a time series analysis (Stoecker et al., 2009; Johnson et al., 2007; Schoener and McManus, 2012).

Our study uses meta-omics data to investigate the global distribution and biogeography of mixotrophic protists in the ocean. Our results, currently based on metabarcoding data, complement morphological records and will be complemented in the near future by metagenomics and metatranscriptomics studies. The latter will allow to distinguish the protists with mixotrophic capabilities from the ones with ongoing mixotrophic activity. This could lead to quantitative estimations of mixotrophic rates in environmental samples, allowing a sharpened study of mixotrophic protists ecology in the global ocean. It could also lead to a metabolic description of complex processes like kleptoplasty and endo-symbiosis, hence facilitating the modeling of mixo- trophic behaviors and its incorporation in ocean biogeo- chemical models.

2.2.5 Supplementary: Metabarcodes level redundancy analysis (RDA)

Methods

Starting with 318 054 metabarcodes in our dataset, we had to use strong selection thresholds to build a parsimonious redundancy analysis model. The Escoufier's vector method is quite robust and only asks for one threshold definition in order to select variables (Escoufier, 1973), and hence was used when working at the lineage level. However, the escouf function implemented in the R package pastecs (`https://github.com/phgrosjean/pastecs`) is not adapted to large datasets, and we couldn't apply the same method at the V9 metabarcode level. Instead, we selected metabarcodes based on rarity and variance of their abundance profiles. First, we only kept the 272 471 metabarcodes appearing in more than one station. Then, we arbitrarily selected the metabarcodes with a variance greater than 0.0001 in their Hellinger transformed abundance profiles. These 363 metabarcodes represented 0.1% of the total dataset in terms of metabarcodes, but 21.1% in terms of sequence abundance.

The environmental variables were modified following the same steps as for the lineages level RDA. Similarly, model selection was run using a two directional AIC-based stepwise selections. The resulting model contained 5 qualitative response variables (filter, biogeographical province (Longhurst, 1998), Ocean region, season moment and depth), as well as 37 quantitative response variables (CO3, HCO3, carbon flux, carbon total, density, PAR, 5m depth NO2, surface NO2, daylight duration, bathymetry, surface NO3, 5m depth NO3, salinity, iron, moon phase, acCDOM, longitude, distance to coast, chlorophyll A, latitude, SST gradient, water residence time, calcite and aragonite saturation states, Lyapunov exponent, nitracline depth, ammonium, temperature, fluorescence, mixed-layer depth, oxygen, depth of oxygen maximum as well as 5 different scatter-

ing coefficients measuring particular density).

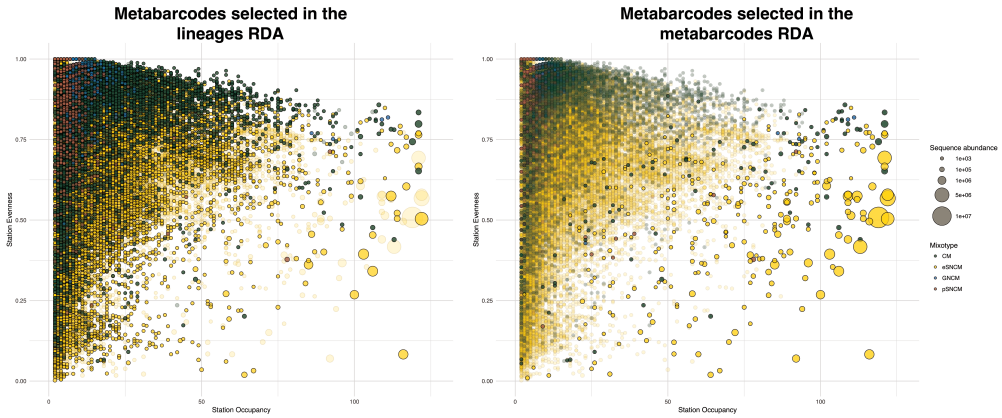

Figure 2.7 - Visualization of the metabarcodes involved in our statistical analyses. The two plots correspond to the bubble plot already presented in figure 2. On the left, bubbles corresponding to metabarcodes that were not included in the lineage RDA are blurred. On the right, bubbles corresponding to metabarcodes that were not included in the metabarcode RDA are blurred.

Results

The 363 metabarcodes selected corresponded to 92 CM, 257 eSNCM, 5 GNCM and 9 pSNCM (Figure 2.7). Even after selection, a significant part of the metabarcodes did not show any response to environmental data in their distribution (Figure 2.8 in supplementary materials). The adjusted R- squared of the RDA model was of 31.1% (versus 34.89% in the lineage level RDA), with 10.7% of variance explained by the two first axes (versus 24.01% in the lineage level RDA). The first RDA axis (5.8%) marks an opposition between samples from oligotrophic waters with low productivity (RDA1<0) and samples from eutrophic and productive water masses (RDA1>0). This axis is positively correlated to chlorophyll concentrations, carbon flux, CO3, ammonium concentration, absorption coefficient of colored dissolved organic matter (acCDOM), gradient of sea surface temperature as well as to 5 different coefficients measuring the particle density of the water. It is negatively correlated to deep nitracline, deep mixed layer, as well as to deep oxygen maximum. The second RDA axis (4.9%) is opposing subpolar samples (RDA2<0) to subtropical ones (RDA2>0). The axis is negatively correlated to the density, latitude, iron concentrations, salinity, CO3, oxygen, longitude, HCO3 and total carbon. It is positively correlated to temperature, NO2, NO3, bathymetry, photosynthetically active radiations (PAR), day length and ammonium.

Among the 92 CM metabarcodes, only a few clearly emerged from the redundancy analysis (Figure 2.8) and showed distinct biogeographies related to environmental variables. Three metabarcodes assigned to Gonyaulax polygramma, a Dinophyceae belonging to the Gonyaulacales order, were found in oligotrophic waters (RDA1<0, RDA2 0). Metabarcodes assigned to the Dinophyceae Fragilidium mexicanum and Alexandrium tamarense were also found in this area of the triplot. The other well represented CM metabarcodes (assigned to a few Chrysophyceae, a couple of Chrysochromulina species and a Dinophyceae from the Kareniaceae family, Karlodinium micrum) were found in productive water masses (RDA1>0).

eSNCMs can be divided in three groups in the RDA space. The first group (RDA1>0, RDA2 0)

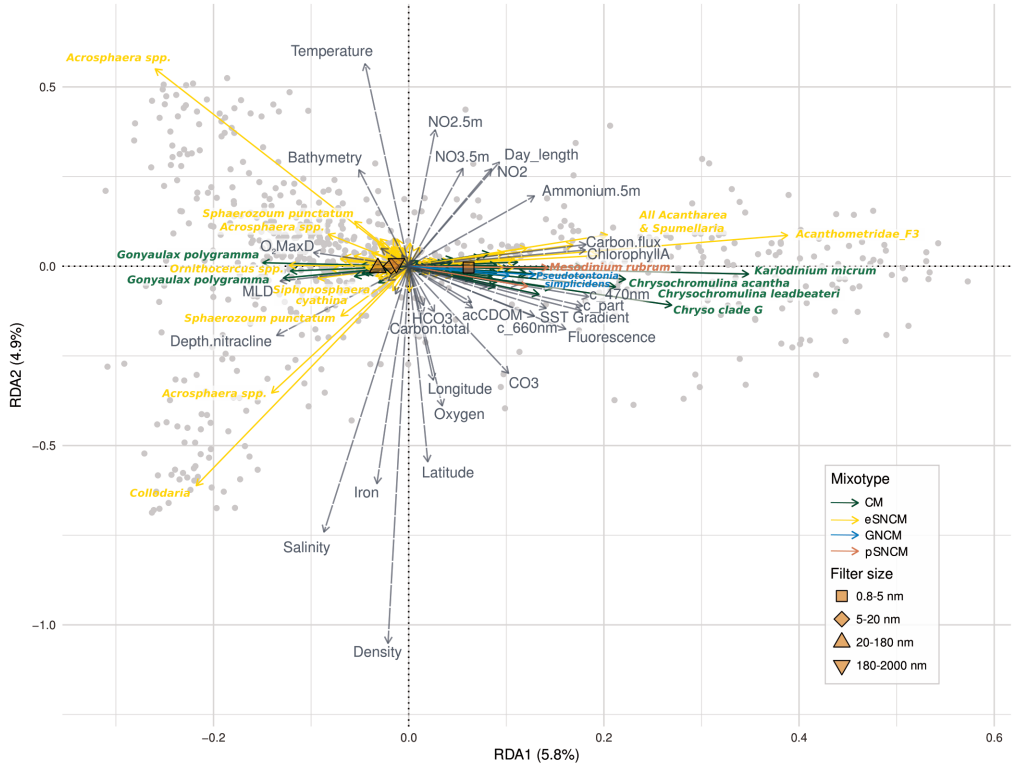

Figure 2.8 - Impact of environmental variables on the distribution of marine mixotrophs. Triplot of the redundancy analysis (RDA) computed on the 363 selected metabarcodes, after model selection, see Section 2.2.5 for details on the methods used. The adjusted R-squared of the analysis is of 31.1%. Each grey dot corresponds to a sample, i.e. one filter at one depth at one station. The blue dashed arrows correspond to the quantitative environmental variables. Abbreviations are as follows: MLD = mixed layer depth, O_2MaxD = O_2 maximum depth, c_660nm = optical beam attenuation coefficient at 660 nm, c_part = beam attenuation coefficient of particles, c_470nm = optical beam attenuation coefficient at 470 nm, acCDOM = absorption coefficient of colored dissolved organic matter, SST Gradient = Sea surface temperature gradient. Plain arrows correspond to mixotrophic metabarcodes, colors indicating mixotypes. The qualitative variable indicating filter sizes is represented through orange centroids. For more readability, we did not represent all quantitative and qualitative variables included in the model.

corresponds to eSNCM metabarcodes dominating rich and productive coastal environments. It includes only Acantharia and Spumellaria assigned metabarcodes. The second group (RDA1<0) dominates oligotrophic environments, and includes multiple Collodaria metabarcodes as well as a few metabarcodes assigned to a Dinophyceae genus (Ornithocercus). Within this group, metabarcodes can be found both in subtropical environments (RDA2>0), and in subpolar regions (RDA2<0). Some lineages like Acrosphaera spp. can even be dominant in the two conditions, showing intra-lineage variation in their biogeography. The third group corresponds to the eSNCM metabarcodes that are badly represented in the space of the redundancy analysis (*e.g.* represented with the shortest arrows in Figure 2.8), but that can be interpreted as distributed homogeneously in regards of the environmental data we are using. These notably include all the Foraminifera metabarcodes present in the RDA. Looking at filters centroids in the RDA space (Figure 2.8), we can suppose that mixotrophic organisms dominating eutrophic systems (RDA1>0) are smaller in size than those dominating oligotrophic ones (RDA1<0).

Only two pSNCM metabarcodes appeared well represented in the RDA triplot, both corresponding to Mesodinium rubrum.
Out of the five GNCM metabarcodes included in the analysis, only two were well represented in the RDA space, both assigned to Pseudotontonia simplicidens. These metabarcodes were mainly found in eutrophic waters (RDA1>0).

Discussion

The results from the metabarcode RDA are very concordant with the ones obtained through the lineage level analysis presented in the main manuscript. Indeed, if the selected environmental variables were slightly different, the global organization of the RDA space was similar in the two analyses. The mixotypes were distributed very similarly in regard of environmental variables when comparing lineage and metabarcode level analyses. There are evident similarities between the distribution of metabarcodes on Figure 2.8, and the one of their corresponding lineages on Figure 2.5. This way, the two analyses conducted led to extremely similar discussion points and conclusions.

Building a redundancy analysis model at the metabarcode level allows to investigate intra-species, intra-genus, intra-families and intra-order variabilities in terms of biogeography. Theses variabilities can't be observed when aggregating metabarcodes into broader lineages, where for example all undefined Collodaria taxa are pooled together. A good example of this problem is the Acrosphaera spp. case. On Figure 2.7, we observe that a few of the most abundant metabarcodes belonging to eSNCM lineages were not selected in the lineage RDA. These metabarcodes are all assigned to Acrosphaera spp. and Collodaria_X lineages, and can display different biogeographies (*e.g.* the three Acrosphaera spp. metabarcodes represented on Figure 2.8). By merging sequence abundances of metabarcodes with different biogeographical patterns, we attenuate their statistical signal, and it could explain why Acrosphaera spp. was not selected in our lineage RDA.
However, the overall intra-lineage variability seemed to be very limited to Collodaria lineages. Also, the opposite effect could be spotted when comparing our analyses. For example, three metabarcodes of Gonyaulax polygramma were found in our metabarcode level RDA, with arrows pointing in the exact same direction (Figure 2.8). Hence, the low intra-lineage variability led to information redundancy in the metabarcode RDA representation. In our lineage based RDA, we had 3 species of dinoflagellates with similar biogeographies selected and well represented: Gonyaulax polygramma, Fragilidium mexicanum and Alexandrium tamarense (Figure 2.5), giving more visual impact to the statistical analysis. Moreover, to build an interpretable metabarcode-level RDA, we could only include in the model 0.1% of the mixotrophic metabarcodes found in the *Tara Oceans* dataset, focusing only on the most abundant ones (Figure 2.7). It highlights the difficulty of selecting only ecologically interesting and non-redundant metabarcodes out of a complete omics dataset. Operational Taxonomical Units (OTUs) can help to answer this issue, but they can also make things worse, especially when constructed using a single similarity threshold for a whole complex eukaryotic population (Brown et al., 2015). During this project, we constructed 97%, 99%, and one difference OTUs using multiple algorithms, but always found contradictions between our original taxonomical assignations and the obtained metabarcode clusters. The algo-

rithms we used were the usearch software (Edgar, 2010) (functions -cluster_otus and -cluster_fast) and the swarm software (Mahé et al., 2015) (using the d=1 parameter: one base pair difference between sequences).

All analyses and graphs were realized with the R software version 3.4.3 (R Core Team, 2019), using the packages vegan version 2.4-5 (Oksanen et al., 2017) and ggplot2 version 2.2.1 (Wickham, 2009). For the functions implying randomness, the char2seed function from package TeachingDemos v2.10 (Snow, 2016) was used setting "Faure" as a seed. All scripts are available on GitHub (`https://github.com/upmcgenomics/MixoBioGeo`).

Acknowledgements

We would like to particularly thank Stéphane Pesant and Stéphane Audic for their work on making *Tara Oceans* datasets available. We also thank John Dolan (CNRS, LOV, Villefranche-sur-mer, France), Miguel Mendez-Sandin (Sorbonne Université, Station Biologique de Roscoff, France), and Wei-Ting Chen (National Taiwan Ocean University, Taiwan) for their essential help during the construction of the mixotrophic lineages set. We also thank Florentin Constancias for his help on the metabarcodes clustering tests conducted. Finally, we thank the three anonymous reviewers for their very constructive comments. This article is contribution number #84 of *Tara Oceans*. For the *Tara* Oceans expedition, we thank the commitment of the CNRS (in particular, Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, Fund for Scientific Research—Flanders, Rega Institute, KU Leuven, The French Ministry of Research. We also thank the support and commitment of Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Région Bretagne, Lorient Agglomération, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the *Tara* schooner and its captains and crew. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who graciously granted sampling permissions. *Tara Oceans* would not exist without continuous support from 23 institutes (`http://oceans.taraexpeditions.org`).

2.3 Conclusion: Going further than metabarcoding

In the concluding remarks of the article presented in this chapter, I evoked how metagenomics and metatranscriptomics studies should in a near future complement morphological and metabarcoding-based observations. For that, one of the strategy should be to define a set of functional genomic markers of mixotrophy, which would allow to quantify mixotrophs and their activity in metagenomics and metatranscriptomics samples. In the next chapter, I will review some of the methods available to detect such markers, and how they could help to build links between genes and functional traits.

Chapter 3

Detecting functional traits in meta-omics data through the use of genomic markers

In chapter 2, I showed how metabarcoding could be used to decipher the biogeography of mixotrophs, for which no functional genomic markers are available yet. I evoked the limits linked to such a use of metabarcoding, like quantification biases due to copy number variations, the necessity to annotate traits to taxonomic databases or the fact that metabarcodes can not reflect the level of realization of a trait (*e.g.* mixotrophic species can be detected through metabarcoding but it can not give any insights on their realized mixotrophic activity at the time of sampling). The identification of genomic markers of functional traits avoids most of these limitations, as such markers are detectable in metagenomes and metatranscriptomes, can be detected in known and unknown species as long as their sequence is conserved across distinct lineages, and can be related to measures of a trait realization. But the identification of such functional genomics markers from lab experiments and environmental data is often challenging, and multiple traits of biogeochemical importance such as mixotrophy, size, or reproduction strategy still lack genomic markers. In this chapter, I will first illustrate how the access to a known functional genomic marker allows for the quantitative exploration of a functional trait biogeography with more precision than metabarcoding, using the demethylation of DMSP as an example. Then, I will review the available methods for identifying genomic markers of functional traits with poorly known genomic basis, and illustrate them with a case study aiming at detecting markers of mixotrophy in dinoflagellates. Finally, I will evoke the important challenge of mining for functional traits in taxonomically and/or functionally unannotated data, and some recent progresses made considering this issue.

This chapter will include preliminary results partly derived from the work of two masters students I supervised during spring 2019: Nina Guerin a 1st year I supervised during 6 weeks, and Aurélie Pham, a 2nd year master student I supervised during 6 months. Nina focused on genomic markers of DMS production whereas Aurélie investigated dinoflagellate transcriptomes looking for genomic markers of mixotrophy.

3.1 Genomic markers of functional traits: simple *versus* complex traits

In medicine and agronomy, *simple* traits are defined as being inherited through Mendelian transmission patterns (trait value defined by the dominant allele of a gene inherited from the parents, whom each transmit only one allele to the descendant), while complex traits do not conform to such patterns (due to factors like incomplete dominance, polygenic or polyallelic interactions) (Peltonen et al., 2000; Sonah et al., 2015; Zak et al., 2017). But in prokaryotic and eukaryotic micro-organisms, even traits that are encoded by one single gene with a simple allelic-dominance scheme can be transmitted horizontally (*i.e.* by an other organisms with no parental link) (Koonin et al., 2001; Keeling and Palmer, 2008), questioning the applicability of this definition to the planktonic world. This is why I will classify traits as simple or complex based on the complexity of their genomic basis rather than their inheritance patterns. I will consider simple functional traits as encoded by one gene, or a few genes corresponding to enzymes involved in a single metabolic pathway (*e.g.* the production of DMS through the cleavage of DMSP which can be achieved by at least 8 enzymes, or nitrogen fixation which is encoded by the *nifH* genes), and complex traits as coded by multiple genes distributed at different loci and taking part in distinct metabolic pathways (*e.g.* body or cell size, and most reproduction and behavioral traits). In the following section, I will focus on methods aiming at the discovery of genomic markers of simple functional traits.

3.2 Identifying markers of simple functional traits

3.2.1 State of the art: biochemical extractions and genome manipulations

Genomic markers of simple functional traits can be studied through targeted wet-lab experiments, like biochemical extraction approaches or genome manipulation techniques (*e.g.* gene knock-outs or the move of DNA fragments into host strains). For example, nitrogenase, the enzyme responsible for dinitrogen fixation in microorganisms, was first described through biochemical extractions and analyses in the 60s, long before the omics era (Eady and Postgate, 1974; Hardy and Burns, 1968). The purification of nitrogenase from about 20 prokaryotic organisms allowed to describe the enzyme structure as highly conserved across diazotrophic species (Chatt et al., 1978; Zehr et al., 2003). Thanks to this observation, the hypothesis of an evolutionarily conserved nitrogenase protein complex was proposed, and the hybridization of genomes from different diazotrophic bacteria allowed to identify and sequence *nifH* genes (Mevarech et al., 1980; Ruvkun and Ausubel, 1980). Then, the omics era allowed to detect new globally abundant diazotrophs using *nifH* sequences (See section 1.3.2.2).

The production of DMS by planktonic bacteria is an other example of a simple functional trait related to plankton ecology with relatively well studied genomic markers. As evoked in section 1.1.2.3, eubacteria can either demethylate DMSP, which does not lead to the production of DMS, or cleave DMSP, which leads to the production of DMS (Moran et al., 2012). The two pathways being

concurrent, the choice between the two is often described as the DMS "bacterial switch" (Levine et al., 2012). The enzyme starting the pathway leading to DMSP demethylation was identified through the integration of transposons in the DMSP demethylating bacteria *Silicibacter pomeroyi* (Howard et al., 2006). A mutant was identified as unable to demethylate DMSP, and the position of its transposon allowed for the identification of the *dmdA* enzyme, which is now estimated to be present in up to 40% of bacterioplankton cells in the open ocean (Moran et al., 2012). Genes coding for the cleavage of DMSP into DMS were also identified through genome manipulation: DNA fragments from DMS producing bacteria were introduced in non-DMS producing hosts (*e.g. E. coli*), and the detection of DMS production in hosts allowed for the identification of the DNA fragments responsible for DMS production (Todd et al., 2007).

The experiments presented in this section would be very unlikely to work for identifying the genomic basis of complex traits. Hence, most studies investigating complex traits rely on statistical analyses of genome and transcriptome content across multiple organisms with known traits. I will present such methods in the next section, but first, I will illustrate how the access to a genomic marker for a functional trait can help to decipher its response to environmental gradients.

3.2.2 A concrete example: exploring the biogeography of the *dmdA* enzyme

3.2.2.1 Introduction

By allowing the demethylation of DMSP in marine eubacteria, *dmdA* plays a key role in the regulation of the sulfur cycle (Howard et al., 2006; Levine et al., 2012). *dmdA* transcription rates can be directly related to the state of the bacterial switch between DMSP demethylation (not leading to DMS production) and DMSP cleavage into DMS (Levine et al., 2012). A study investigating *dmdA* transcription rates at the Bermuda Atlantic Time Series (BATS) over a 10 months period identified that high temperature and UV-A dose could lead to more DMSP cleavage into DMS, while colder temperature led to more DMSP demethylation (Levine et al., 2012). These transcription rates were obtained from qPCR using adapted primers (Levine et al., 2012). In spite of the referencing of *dmdA* in functional annotation databases such as KEGG (Aramaki et al., 2019; Salazar et al., 2019), these findings are yet to be confirmed by a global scale study. Salazar et al. (2019) demonstrated that the transcriptomic abundance of *dmdA* was negatively correlated to the ones of assimilatory sulfate reduction marker genes, using the *Tara Oceans* and *Tara Polar Circle* datasets, and proposed that the DMSP demethylation pathway could be concurrent with assimilatory sulfate reduction pathways for sulfur integration in the metabolism. However they did not relate these observations to particular environmental conditions or to transcriptomic abundances of DMSP cleavage enzymes. Here, the same global scale meta-omics and metadata from the *Tara* expeditions will be used, but this time to focus on the biogeography of *dmdA*, and identify the main abiotic drivers of its expression at global scale.

3.2.2.2 Material & methods

The dataset from *Tara Oceans* and *Tara Polar circle* included 187 metatranscriptomes and 370 metagenomes sampled from 126 globally distributed sampling stations. These samples were obtained from the surface (5-10m), the deep chlorophyll maximum (DCM, 20-200m) and the mesopelagic layer (200-1000m), and corresponded to prokaryote and virus enriched size fractions (0.22-1.6 μm and 0.22-3 μm). 9 epipelagic samples could not be classified as surface nor DCM, and were annotated as mixed layer (25-200m). Metagenomics and metatranscriptomics data were obtained through Illumina sequencing following the protocols described in Pesant et al. (2015) and Alberti et al. (2017). Metagenomic reads were quality-filtered, assembled, gene-coding sequences were predicted and dereplicated which led a set of 46,775,154 non-redundant genes catalog, named the OM-RGC.v2 (available at `https://www.ocean-microbiome.org`; detailed methods in Salazar et al. (2019)). This catalog was functionally annotated using BlastKOALA (Aramaki et al., 2019) and eggNOG-mapper (Huerta-Cepas et al., 2017).

Metagenomic and metatranscriptomic abundance profiles were determined for each sample by mapping the quality-filtered metagenomics and metatanscriptomics reads to the OM-RGC.v2 catalog, and normalizing the mapped read counts by the median abundance of 10 universal single-copy phylogenetic marker genes (see details in Salazar et al. (2019)). Finally, the profiles were converted to variance-stabilized integer counts by dividing each profile by its maximum value, multiplying the result by 10^9, and applying a $\log_2$ transformation. These normalized abundances were used to compute metagenomic and metatranscriptomic profiles at the gene level but also at the functional level (grouping genes according to their KEGG or eggNOG functional annotations). Among the 187 metatranscriptomes and 370 metagenomes, 129 came from the same samples, *i.e.* collected at the same location and depth using the same size fraction. For these 129 samples, an expression profile was computed as the difference between the $\log_2$-transformed metatranscriptomic and metagenomic profiles (available at `https://www.ocean-microbiome.org`). Here, I focused on the 129 metagenomic, metatranscriptomic and expression profiles available for the K17486 KEGG ortholog group, corresponding to the *dmdA* enzyme.

The environmental context of the 129 samples, was retrieved from `https://www.ocean-microbiome.org`. It corresponded to 37 variables: ID of the sample, station label, sampling layer (surface, deep chlorophyll maximum, mesopelagic or mixed layer), sample located in the polar or non polar area, upper threshold of the size fraction (1.6 or 3 μm), date of sampling, latitude, longitude, nominal depth (in meters), ocean region, temperature, oxygen, chlorophyll A, total carbon, salinity, sea surface temperature gradient, fluorescence, CO_3, HCO_3, water density, PO_4, NO_3, Si, Photosynthetically active radiation (PAR), Alkalinity, Ammonium at 5m depth, Depth of the mixed layer, Lyapunov, NO_2, Depth of the O_2 minimum, NO_2/NO_3, Nitracline, depth of the maximum Brunt-Vaisala frequency (which is a proxy for the depth of the mixed layer), iron at 5 meter depth, depth of the O_2 maximum, Okubo Weiss parameter (values below/above 0 indicate that the sample is inside/outside an eddy) and water residence time.

For further statistical analysis, this environmental dataset was scaled, centered, and consequently

7 variables were removed: sample ID, date of sampling and station label because they had no value as environmental drivers of *dmdA* abundance; total carbon, NO_2/NO_3 and NO_3 because they showed too high colinearity with other variables; and PAR because it had more than 50% of missing values (92 NAs over the 129 samples). Each of the remaining missing values in the dataset was replaced by the mean of the concerned variable in the 5 nearest samples in terms of environmental profile.

All of these operations were achieved using the PreProcess command from the *caret* package (Kuhn, 2008) in R version 3.5.3 (R Core Team, 2019), through options center, scale, knnImpute, corr and nzv. The R code for this project is available at `https://github.com/EmileFaure/DmdA`.

3.2.2.3 Results & Discussion

Distribution of the dmdA enzyme in the global ocean

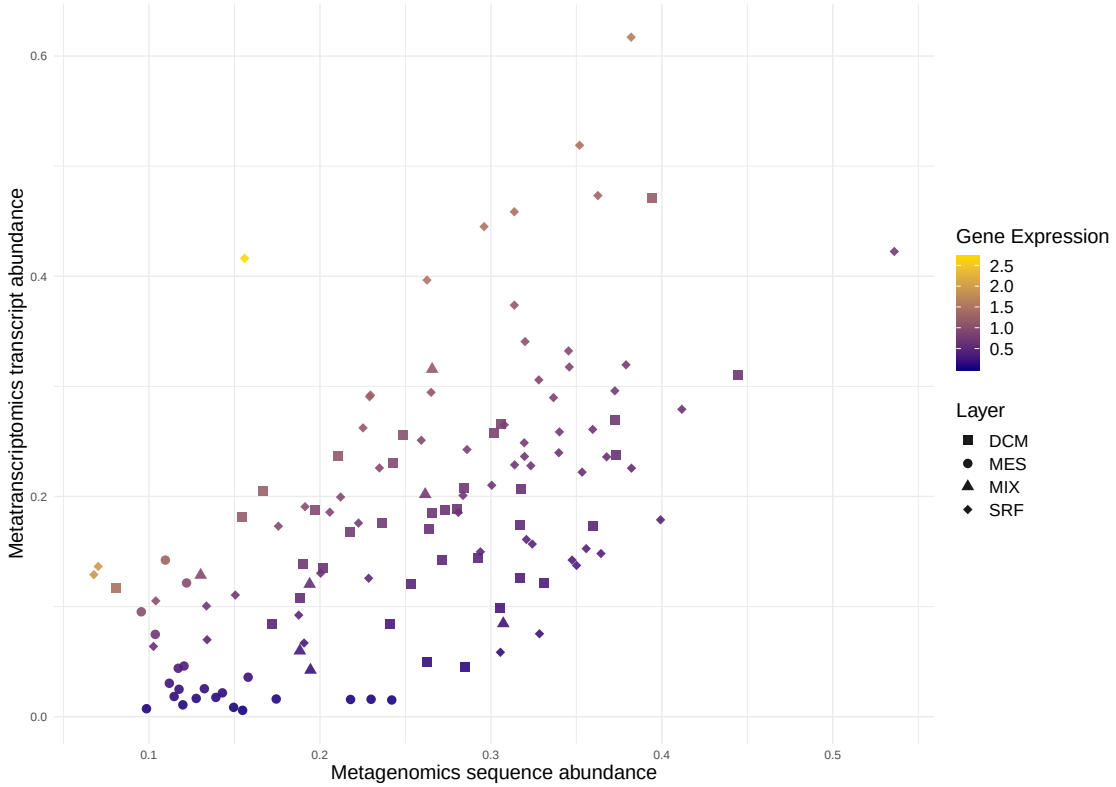

Figure 3.1 - dmdA gene versus transcript abundance. Each dot corresponds to one of the 129 samples (a sample corresponding to a station and a given depth layer, indicated by the point shape). Abundances are normalized, but the log transformation was not applied here to keep positive abundances. Dots were colored according to gene expression (reflecting the ratio between log2-transformed transcript abundance and gene abundance).

The metagenomic and metatranscriptomic abundance profiles of the samples had a Pearson correlation coefficient of 0.58 (Figure 3.1). Samples from the mesopelagic layer all had a normalized metagenomic sequence abundance below 0.25 (before $\log_2$ transformation to keep positive values; mean of 0.14), and a metatranscriptomic transcript abundance below 0.15 (mean of 0.04). Other depth layers exhibited higher abundance values (mean of 0.28 and 0.24 for gene and transcript

abundance at the surface; 0.27 and 0.18 at the DCM; 0.22 and 0.14 in the mixed layer; Figure 3.1). Expression levels were comprised between a minimum of 0.04, reached in a mesopelagic sample of the North Pacific Ocean (station TARA_109), and a maximum of 2.67, reached in a surface sample of the southern Ocean (station TARA_084). This 2.67 maximum of expression appeared as an outlier: it was 4.5 standard deviations above the median, and was only seconded by a value of 1.94 (decrease of 0.72 (27%)), which was reached in a surface sample of the Arctic Ocean (station TARA_208) (Figure 3.1, Figure 3.2). Overall, maximums of gene expression did not correspond to maximums of gene abundance, but rather to locations with high to moderate transcript abundance and low to moderate gene abundance (Figure 3.1).

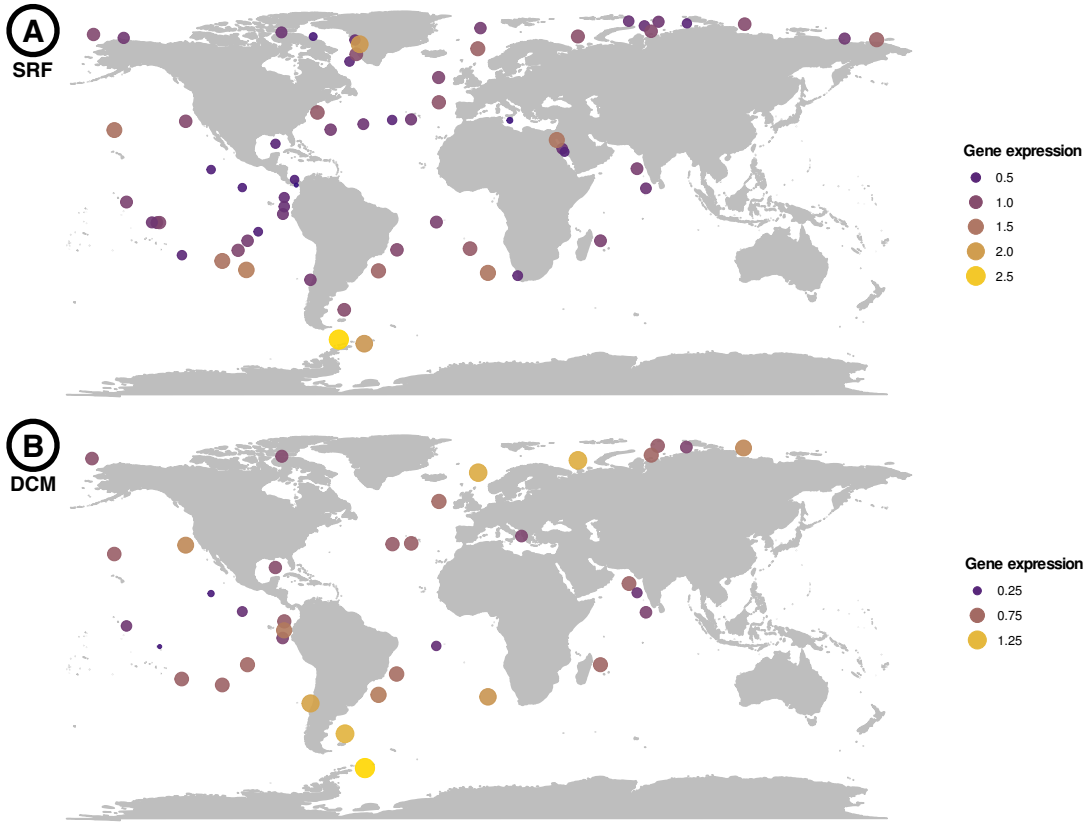

Figure 3.2 - Maps of dmdA gene expression at the surface (A) and at the deep chlorophyll maximum (B). Each sample is represented as a dot, dot size and color both representing the level of gene expression. Abundance values are normalized but the log2-transformation was not applied to display positive abundances.

The *dmdA* enzyme had non-zero sequence and transcript abundances in all the 129 metagenomes and 129 metatranscriptomes, corresponding to 68 stations from the *Tara* expeditions, which illustrates its ubiquity in the global ocean. Gene expression appeared higher in polar waters than in subtropical ones (Figure 3.2), which is coherent with the observations from Levine et al. (2012), whom related DMSP demethylation with colder temperature.

Environmental factors driving the dmdA distribution

To better identify the drivers of expression patterns, I computed a multivariate analysis of gene abundance, transcript abundance and gene expression. In a redundancy analysis (RDA), gene abundance, transcript abundance and gene expression were used as interest variables, while

environmental data served as explanatory variables. The complete RDA, including the 30 environmental variables, was significant (F = 10.504, p-value < 0.001). A bi-directional stepwise model selection based on the Akaike Information Criteria was then performed to select the most parcimonious model. The selected model contained 8 environmental variables: nominal depth, oxygen, temperature, depth of O_2 maximum, HCO_3, Density, Chlorophyll A and latitude. Both axis of the RDA axes were significant (p-value < 0.001, F = 320.1 and F = 35.34). The adjusted R^2 value of the RDA was of 72.1%, 66.49% of the variance was explained by the first axis, while 7.34% was explained by the second (Figure 3.3).

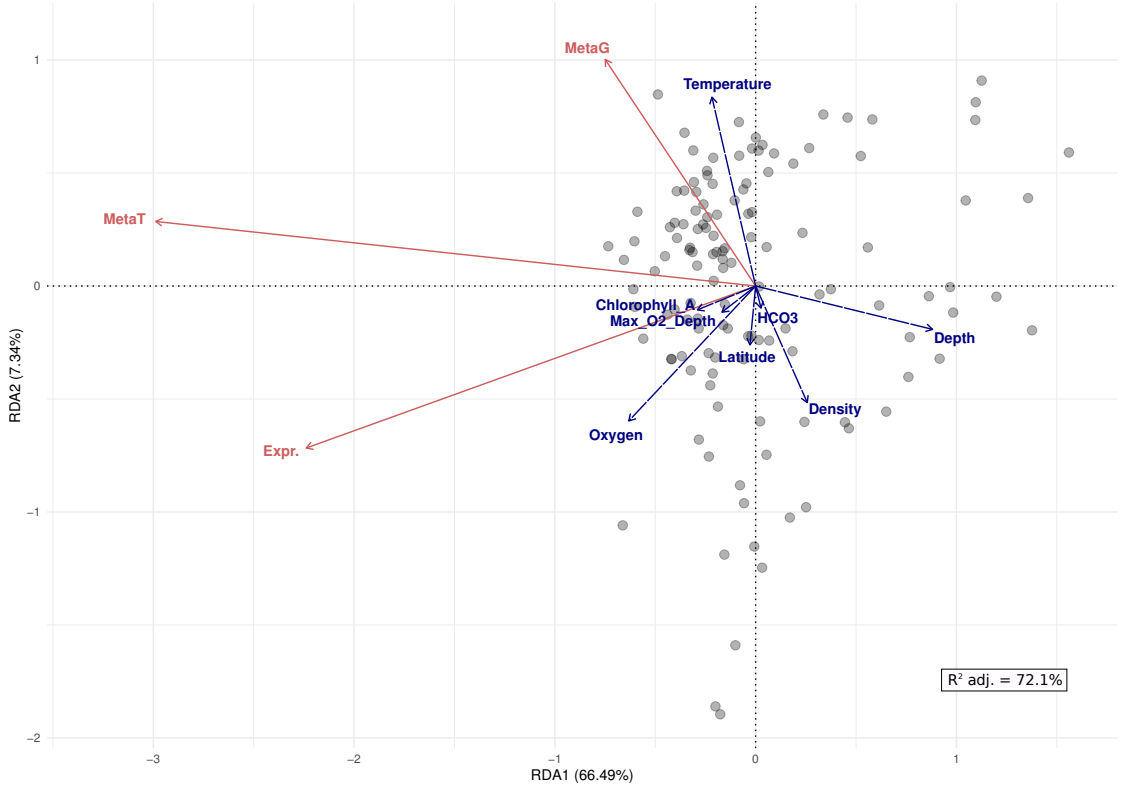

Figure 3.3 - Triplot of the RDA. Grey dots in the background correspond to the 129 samples. Blue arrows correspond to the 8 selected environmental variables, and red arrows correspond to the three interest variables: gene abundance (MetaG), transcript abundance (MetaT) and gene expression (Expr.).

The first axis of the RDA opposes samples taken at high depths (RDA>1) from samples taken closer to the surface (RDA<1) (Figure 3.3). It confirms that gene abundance, gene expression and especially transcript abundance are all higher in surface samples. This seems coherent considering that DMSP is produced by phytoplanktonic organisms, mostly found in the euphotic zone (Moran et al., 2012), and that mesopelagic samples exhibited very low transcript abundance (Figre 3.1). The second axis of the RDA opposes samples from subtropical, warm waters (RDA2>0), and samples from subpolar, cold and dense waters (RDA2<0) (Figure 3.3). The RDA confirms that gene abundance is higher in subtropical waters, and identify it as only poorly correlated to gene expression, which is mostly high in chlorophyll rich, oxygenated waters (Figure 3.3). The mismatch between gene abundance and expression could be explained by the presence of organisms that are able to demethylate DMSP but are rather using the cleavage pathway in warmer conditions.

Again, this is coherent with the results from Levine et al. (2012), whom identified the bacterial switch to be in favor of the demethylation pathway over the cleavage one in cold waters, and our results confirm the applicability of their findings at the global scale. The high adjusted R^2 value suggested that *dmdA* abundance was strongly linked to the environmental context, which led me to explore its predictability from environmental variables using a machine learning approach.

Predictions of dmdA abundance from the environmental context

The richness of meta-omics data offers the potential to predict the distribution of functional traits *via* their genomic markers in the environment through statistical modeling (Tang and Cassar, 2019). The large quantities of data collected by global meta-omics datasets allows for the construction of large *training* sets, *i.e.* subsets of data that need to include a wide enough range of conditions to be representative of the global dataset, so that statistical models can be trained on them, while retaining a part of the samples as *test* sets, which are used to compare models predictions with observations and test the models performance. Reaching this ability of predicting the abundance for a large set of traits would (1) help attaining an unprecedented level of precision in our knowledge of the abiotic drivers of functional diversity in planktonic communities, (2) provide quantitative insights to improve the construction and validation of biogeochemical models, through the identification of general ecological laws governing functions distribution in the global ocean.

Here, to test the predictability of *dmdA* expression, genes and transcripts abundances from environmental data, elastic net regressions were used, *i.e.* a combination of lasso and ridge regressions, allowing to penalize uninformative predictors by shrinking their regression coefficient towards 0. Unlike random forest regressions or neural networks, this method only uses linear relationships between the interest variable and predictors to produce predictions, which is less likely to lead to overfitting (*i.e.* an over-adaptation of the model to the training set leading to poor capacities of predictions over additional data). Independent elastic net regressions were computed for each of the abundance and expression profiles: one with gene abundance as the interest variable, one with transcript abundance and one with gene expression. For each elastic net regression, environmental variables were used as predictors, with categorical variables coded as dummy variables (*i.e.* columns of 0 or 1 for each categorical level). Training sets corresponding to 105 samples (*i.e.* 80% of the data) were randomly selected. Elastic net regression models were trained on these training sets with a 3 times repeated 10 fold cross-validation process. Regularization (λ, or penalty coefficient) and mixing (α, the mix level between a Lasso approach and a Ridge one, leading to different penalizations of coefficients) parameters, were optimized for each regression model by selecting the pair of parameters minimizing cross validation error across all possible combinations of 10 random values of α and λ (higher numbers of combinations were tested, not leading to better R^2). Finally, I used the best models selected after cross validation and parameters optimization to predict gene abundance, transcript abundance and gene expression values from the test sets (*i.e.* 20% of the samples that were not selected in training sets). Models performance at predicting *dmdA* abundance and expression was measured by computing R^2 value

from comparisons between test set observations and model predictions.

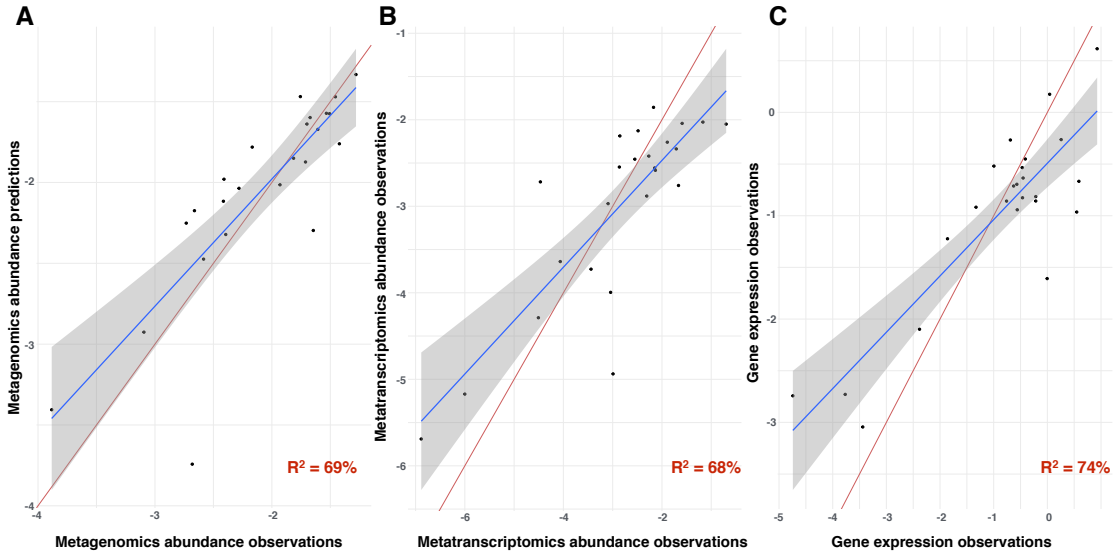

Figure 3.4 - Comparison of observations from the test sets and predictions from the elastic net regressions on (A) gene abundance, (B) transcript abundance and (C) gene expression. Red lines indicate perfect predictions (y=x). Blue lines are simple linear regression lines, with the corresponding 95% confidence interval indicated in grey. The R^2 values indicated in red are the one of each elastic net regression model.

The three models led to predictions of *dmdA* abundance and expression with R^2 values comprised between 68 and 74% (Figure 3.4).

In the gene abundance elastic net regression, 13 environmental variables had non-zero regression coefficients, with the 5 most influential ones being: Ocean region Southern Ocean (-0.76), polar (-0.32), temperature (0.26), nominal depth (-0.23) and chlorophyll A (0.14). These results confirm the observations from the RDA, identifying metagenomic abundance of *dmdA* as higher in warm, subtropical waters.

In the transcript abundance elastic net regression, 26 variables had non-zero coefficients, with the 5 most influential being: mesopelagic layer (-0.89), Ocean region [MS] Mediterranean Sea (-0.79), mixed water layer (-0.66), Ocean region [SO] Southern Ocean (0.52) and nominal depth (-0.45). Here again, observations from the RDA are confirmed, with depth appearing as the most influential factor on transcript abundance. The elastic net regression identifies the Mediterranean Sea as a zone of particularly low *dmdA* transcript abundance, but this should be taken with precaution as only two samples from the Mediterranean sea are present in the dataset.

In the gene expression elastic net regression, 25 variables had non-zero coefficients, with the 5 most influential being: Ocean region [MS] Mediterranean Sea (-0.94), Ocean region [SO] Southern Ocean (0.60), oxygen (0.46), mesopelagic layer (-0.42) and nominal depth (-0.35). Gene expression then appears particularly high in the Southern Ocean, and low in the Mediterranean Sea. The positive influence of oxygen on gene expression supposed from the RDA results is confirmed, as is the negative correlation with depth.

Using 129 samples from the global ocean, I was able to produce predictions of *dmdA* gene abundance, transcript abundance, and expression from the environmental context, with R^2 above 65%.

One obvious way of increasing the statistical power of our models would be the addition of more metagenomes and metatranscriptomes to the dataset, allowing to increase the size of the training and test sets. For example, the fact that only two samples were issued from the Mediterranean Sea and 3 from the Southern Ocean limit the capacities to draw general conclusions on these areas despite their significant influence as predictors. In the future it would then be interesting to reproduce this approach on other global datasets like the Ocean Sampling Day (Kopf et al., 2015) or Malaspina (Duarte, 2015). An elegant way to confirm the predictive abilities of models such as the ones presented here would be to use samples from an expedition as a training set, and the ones from an other expedition as the test set. The principal difficulty in this case would be to homogenize the values of gene and transcript abundances across datasets to be able to compare them. Applying the same normalization steps across datasets to maintain homogeneity would be quite trivial, but discrepancies in sequencing methods and/or size fractions might be hard to deal with.

By investigating the *Tara Oceans* and *Tara Polar Circle* datasets, the principal environmental drivers of a single genomic marker coding for DMSP demethylation were identified in the global ocean. Gene and transcript abundances of *dmdA* were correlated, but governed by different drivers: gene abundance was mainly linked with temperature, while transcript abundance was mostly explained by depth (Figure 3.3). This could be explained by the presence of organisms bearing the *dmdA* enzyme at high depth, were DMSP is unavailable, making the transcription of the demethylation enzyme useless and/or avoided. This genomic signal could notably be due to the presence of genetic material from dead organisms sinking in the water column, which do not appear in transcriptomic samples (Singh et al., 2009). These discrepancies led the gene expression measures to be quite poorly correlated with gene abundance, and oxygen appeared as one of their most important drivers. Machine learning allowed to predict *dmdA* gene abundance, transcript abundance and gene expression from the environmental context with a good accuracy (Figure 3.4), demonstrating how the activity of this pathway with a strong influence on the sulfur cycle could be determined from purely physico-chemical data. It is particularly interesting to note that the best predictors of the different elastic net regressions were identified either as variables that are fixed through time (*e.g.* depth or Ocean region), or as routinely measured variables for which high quantities of data are available at the global scale (*e.g.* temperature, oxygen, chlorophyll A). These variables being already described in the majority of biogeochemical models, the results here suggest the potential of predicting DMSP demethylation at global scale using data issued from models and observations, allowing to target the use of such predictions as inputs or validation tools in biogeochemical models.

3.2.2.4 Perspectives

In the next 6 months (starting in October 2020), I plan to reproduce this approach on other marker genes linked to the sulfur cycle, and especially on DMSP lyases, which cleave DMSP to DMS (Moran et al., 2012). The access to both *dmdA* and DMSP lyases abundances would allow to quantitatively describe the state of the bacterial switch at a global scale for the first time. But the

cleavage pathway is trickier to study than the demethylation one, as seven different DMSP lyases have been described in prokaryotes, and only one (*dddL*) corresponds to a KEGG ortholog group (K16953). Moreover, this KEGG ortholog group is absent from the *Tara Oceans* and *Polar circle* functional profiles.

Nina Guérin's internship aimed at overcoming some of the difficulties encountered when studying DMSP lyases in global scale datasets, by (1) constituting a database of available DMSP lyase sequences, (2) using an alignment algorithm (Nina used Diamond, Buchfink et al. (2015)) to find matches to the database in environmental samples. The database she computed should allow to extract DMSP lyases gene abundance, transcript abundance and gene expression in the near future using the same *Tara* dataset as described in the study presented in this section. This will allow to provide the first meta-omics based study of the environmental factors governing the DMSP bacterial switch, including both the demethylation and the cleavage pathways.

3.3 Exploring the genomic basis of complex functional traits

3.3.1 State of the art: linkage, association methods and comparative transcriptomics

The investigation of the links between genotypes and complex traits constitutes a whole field of research, mainly driven by medical and agronomic studies (Members of the Complex Trait Consortium, 2003; Visscher et al., 2017). The goal of such studies is often to identify portions of genomes, genes or sets of genes associated with multigenic traits like plant height and weight in agronomy (Sonah et al., 2015), disease susceptibility in medicine (Members of the Complex Trait Consortium, 2003), and less frequently with behavioral or life-history traits in planktonic ecology (Routtu et al., 2014). These studies aim at finding statistical links between compositional variations in genomes (often focusing on single nucleotide polymorphisms, or SNPs) and functional traits. This can be achieved by comparing the genetic variants of segregating biparental populations of organisms over multiple generations, to identify quantitative trait loci (QTL) linked with variations in trait values across individuals (Members of the Complex Trait Consortium, 2003; Sonah et al., 2015). For example, the genomic basis of sediment browsing in the planktonic freshwater crustacean *Daphnia magna* was investigated through the genotyping of 185 F2 (*i.e.* second generation of offspring) recombinant individuals obtained through in-lab culture and breeding (Arbore et al., 2016). The *Daphnia* genus are considered as keystone species in many ponds and lake ecosystems, and *Daphnia magna* and *Daphnia pulex* are often used as model organisms (Czypionka et al., 2019). An SNP-based genetic map of the *D. magna* genome was released in 2014 Routtu et al., which led to the identification of markers for different sediment browsing strategies (Arbore et al., 2016), but also of markers of diapause termination (diapause being a dormancy phase in the life cycle of many invertebrates to avoid unfavourable conditions) (Czypionka et al., 2019).
In addition to QTL mapping, it is also possible to compare whole genomes of numerous organisms

from which trait values are measured, to identify statistically significant associations between portions of genomes and trait values, in which case we talk about genome wide association studies, or GWAS (Visscher et al., 2017). QTL and GWAS approaches can even be used in conjunction, with QTL allowing to validate and quantify the influence of candidate genes identified through GWAS (Sonah et al., 2015). GWAS has notably been used to identify markers of infectivity in *Pasteuria ramosa*, a model bacterial pathogen of *D. magna* (Andras et al., 2020). Despite their ability to decipher the genomic basis of complex functional traits, these approaches remain exclusively applied to model planktonic organisms, for which genomes of multiple individuals are available. A first reason is that QTL methods are based on vertical heredity of traits, and recombination events between generations (*i.e.* the exchange of genetic material between organisms leading to the production of offsprings with different traits than their parents). These methods are centered and developed on eukaryotic multicellular organisms (Metazoa, Plants), and seem quite unadapted for studying prokaryotes, in which horizontal transfers and clonal reproduction are common. Also, both QTL and GWAS methods rely on the genotyping of numerous single individuals with known trait values and/or known demographic history, often implying intensive culture and the use of reference omic sequences to accurately detect SNPs (Visscher et al., 2017). Thus, studies focusing on planktonic functional traits rely so far mainly on comparative transcriptomics.

Comparative transcriptomics studies compare populations of the same species or strain exposed to varying abiotic and/or biotic conditions, with the aim to identify up-regulated and down-regulated genes (Caron et al., 2017; Marchetti et al., 2012; McKie-Krisberg et al., 2018; Liu et al., 2016; Lv et al., 2019). For example, experiments of mixotrophic algae (here two prasinophytes, *Micromonas polaris* and *Pyramimonas tychotreta*) in different nutrient conditions allows to compare transcriptomes of algae eating through phagotrophy (*i.e.* low nutrients conditions) with transcriptomes of photosynthetically active ones (*i.e.* high nutrients conditions) (McKie-Krisberg et al., 2018). Similarly, transcriptomes of three mixotrophic protists (the haptophyte *Prymnesium parvum* and two chrysophytes: *Dinobryon* sp. and *Ochromonas* sp.) have been sequenced across gradients of light (Liu et al., 2016), allowing to identify potential marker genes of phagotrophy in mixotrophic lineages. Finally, a genome-wide transcriptomic analysis of the marine diatom *Thalassiosira pseudonana* in different conditions (combinations of silicon limitation, nitrogen limitation and iron limitation) allowed to identify genes involved in the biogenic production of the silica cell-wall structures typical of diatoms (Mock et al., 2008).
But comparative transcriptomics do not solve all the issues evoked earlier, as sequencing comparable transcriptomes across a gradient of trait values requires the focal species to be cultivable. Also, the potential markers detected through comparative transcriptomics need to be investigated through targeted experimental designs similar to the ones evoked in paragraph 3.2 in order to be mechanistically validated. The accumulation of sequenced transcriptomes and the creation of databases like the marine microbial eukaryote transcriptome sequencing project (MMETSP, grouping almost 800 transcriptomes, Keeling et al. (2014)) now allow public access to hundreds of transcriptomes from marine planktonic species, allowing for large scale analysis of the omic basis of functional traits.

Meng et al. (2018) illustrated this by using the MMETSP resource to investigate the genomic basis

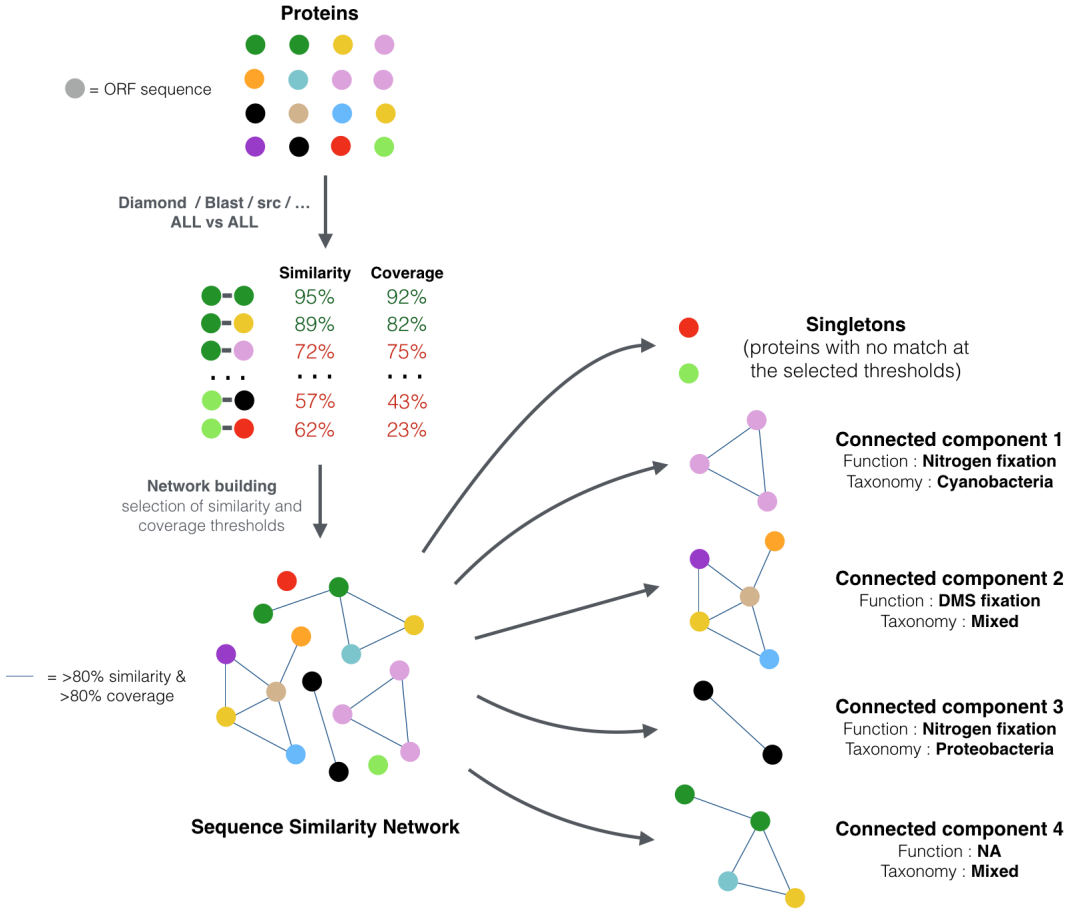

Figure 3.5 - Building a sequence similarity network. Pairwise values of similarity and coverage are computed on a set of sequences, here open reading frames (ORF) sequences as the focus is on proteins, represented as circles with colors coding for their taxonomy. The sequence similarity network is then built by linking together all the pairs of proteins that satisfy a certain similarity and coverage threshold. The network is composed of singletons, i.e. proteins that did not match to any other in the network with the selected thresholds, and connected components, i.e. subgraphs in which at least one path allows to directly or indirectly connect two proteins. Each protein in the network can be associated with a functional and/or taxonomic annotation, which then allows to investigate the links between similarity, function and taxonomy at the scale of the full network, or in connected components. Examples of possible annotations at the connected component level are given here for illustrative purposes.

of functional diversity in dinoflagellates. They based their approach on a sequence similarity network built from 46 transcriptomes of dinoflagellates strains from 28 distinct genera. A sequence similarity network (SSN) is a graph in which nodes are proteins, and links represent the similarity (*i.e.* the percentage of common amino acids) and coverage (*i.e.* the length of overlap between two proteins relative to total protein length) between each pair of proteins (Figure 3.5). SSNs are composed of singletons, *i.e.* proteins that are not linked to any other protein in the network, and connected components (CCs), *i.e.* subgraphs in which nodes are directly or indirectly (*i.e.* through other nodes) connected together, but disconnected from the rest of the network (Figure 3.5). The first step to build an SSN is then to compute pairwise similarity and coverage measures on a set of sequences, often through an all against all alignment (*e.g.* using Diamond (Buchfink et al., 2015)

or BWA-MEM (Li, 2013)). By applying stringent thresholds on the obtained similarity and coverage measures to build the links of the network, one can create a SSN composed of connected components potentially sharing similar functions, and assimilable to protein families (Atkinson et al., 2009; Cheng et al., 2014; Meng et al., 2018; Lopez et al., 2015). Statistics can then be computed from the network, such as connected components number, sizes, density (number of observed connections relative to the number of potential connections), or functional and taxonomical homogeneity which can be derived from the functional and taxonomic annotations of the proteins of the network. In Meng et al. (2018), each sequence has been taxonomically and functionally annotated, but also tagged with the functional traits of its organism of origin: mixotrophy, toxicity, kleptoplasty, symbiosis, parasitism, or DMSP production. The sorting of connected components allowed to identify CCs coming exclusively from toxic organisms, or symbiotic ones for example, which makes them good candidates as potential genomic markers of functional traits. 5 connected components were only composed of proteins from toxic dinoflagellate species, corresponding to 49 protein coding domains among which none were functionally annotated in the Gene Ontology. In this study, the annotation of mixotrophy was limited to 'yes' or 'no', with no distinction of the types of mixotrophy, and connected components associated with mixotrophic species were not investigated in details. Hence, the results of Meng et al. (2018) suppose that a more particular focus on mixotrophy using the same data might allow to discover markers of mixotrophy in dinoflagellates, and maybe even extend them to other mixotrophic lineages available in MMETSP.
From January to June 2019, I supervised Aurélie Pham for her Master 2 internship, which focused on finding markers of mixotrophy in dinoflagellates. The preliminary results that I will present in the next sections are mainly derived from her work.

3.3.2 A concrete example: Markers of mixotrophy in dinoflagellate transcriptomes

3.3.2.1 Introduction

During the past decade, the historic dichotomy classifying planktonic unicellular eukaryotes as either *phytoplankton* or *zooplankton* has been replaced by a new vision based on a distribution of protists along a continuum from full autotrophy to full heterotrophy, in which most of the planktonic organisms display mixotrophic abilities (Flynn et al., 2013; Mitra et al., 2016; Stoecker et al., 2017; Caron, 2016a). This led to multiple efforts aiming at better understanding the biogeography of mixotrophs, and trying to identify their potential affinity with particular environmental conditions (Leles et al., 2017, 2019; Faure et al., 2019). These studies have either used morphological identification data (Leles et al., 2017, 2019) or metabarcoding data (Faure et al., 2019), all identifying mixotrophs as ubiquitous and abundant, but also highlighting strong limitations in their approaches. The main identified limit lies in the necessity of using databases of mixotrophic species and/or the available literature to define the focal set of mixotrophic lineages, while the lack of routine protocols to measure mixotrophic capacities led to misconceptions about the trophic mode of most primary producers in the global ocean (Leles et al., 2019). It led these studies to

focus on a maximum of 133 mixotrophic lineages, when more than 4300 species of phytoplankton are morphologically described, and OTU-based estimates suppose a 3 to 8 times higher number (de Vargas et al., 2015). There is then a strong discrepancy between the statement that most protists display mixotrophic capacities (Flynn et al., 2013) and the number of lineages that are included in biogeographical studies (Leles et al., 2017, 2019; Faure et al., 2019). The discovery of genomic markers of mixotrophy would tackle this issue by allowing to directly detect and potentially quantify mixotrophic activities in metagenomes and metatranscriptomes, which are now available at global scales (de Vargas et al., 2015; Ibarbalz et al., 2019; Richter et al., 2019).

The search for genomic markers of mixotrophy has mainly focused on constitutive mixotrophy, *i.e.* the ability to eat through phagotrophy in inherently photosynthetic organisms, essentially through comparative transcriptomics approaches (McKie-Krisberg et al., 2018; Liu et al., 2016). Genes upregulated during the assimilation of prey chloroplasts in the plastidic-specialist non-constitutive mixotroph dinoflagellate *Nusuttodinium aeruginosum* were also identified earlier this year (Onuma et al., 2020). But none of the genes identified as upregulated in these studies seems to establish as a good candidate for the detection of mixotrophy in environmental samples. In parallel, Burns et al. (2018) proposed a gene-based predictive model of phagotrophy and photosynthesis using complete genomes of 35 eukaryotic lineages. Their approach consisted in identifying protein clusters based on sequence similarity that were enriched in organisms either capable of phagotrophy or photosynthesis, allowing them to identify a set of 474 proteins associated with phagocytosis, and one of 243 proteins associated with photosynthesis (Burns et al., 2018). Using these protein sets, they were able to correctly predict constitutive mixotrophy in the haptophyte *Prymnesium parvum*, and the absence of mixotrophy in the strictly phototrophic diatom *Phaeodactylum tricornutum* (Burns et al., 2018). These results demonstrate the potential lying in the use of annotated databases of whole-genomes to identify genomic markers of mixotrophy, but remains limited for further application on marine mixotrophic plankton by the presence of only 4 genomes of marine planktonic organisms among the 35 reference genomes selected for the study. This low representation of marine plankton among the references is explained by the fact that Burns et al. (2018) mainly aimed at detecting phagocytose and photosynthetis marker genes in archaea, and not in protists.

A study also using similarity-based protein clusters to identify genomic markers of metabolic functions focused on the *Alveolata* lineage of dinoflagellates (Meng et al., 2018). These protists are known for their high functional diversity, as the lineage include strict autotrophs (*e.g. Pelagodinium beii*), constitutive mixotrophs (*i.e.* phagotrophs that display an innate capability to achieve photosynthesis, *e.g. P. parvum*), non-constitutive mixotrophs (*i.e.* heterotrophs that acquire photosynthetic capacity through the stealing of chloroplasts from any prey - generalists -, specific preys - plastidic-specialists - or the bearing of endosymbionts - endo-symbiotic specialists -, *e.g.* the plastidic specialist non-constitutive mixotroph *Dinophysis acuminata*), and strict heterotrophs (*e.g. Polykrikos kofoidii*) (Jeong et al., 2010; Mitra et al., 2016). As I evoked in section 3.3.1, Meng et al. (2018) used 46 dinoflagellates transcriptomes to compute a sequence similarity network, in which they were able to retrieve connected components corresponding to highly similar clusters of proteins, potentially coding for the same functions (Meng et al., 2018; Atkinson et al., 2009;

Lopez et al., 2015). They were then able to identify the clusters only composed of proteins from toxic or mixotrophic dinoflagellate species, constituting putative markers of the corresponding functions (Meng et al., 2018). However, they annotated mixotrophy without differentiating the different types of mixotrophy (*i.e.* constitutive *versus* non-constitutive), and did not focus on the functional annotations in clusters identified as putative markers of mixotrophy.
Here, I will use a dataset of 47 dinoflagellate transcriptomes including 19 from mixotrophic species to try to detect protein clusters that could serve as markers of mixotrophy in dinoflagellates.

3.3.2.2 Material & methods

Trophic modes (strict autotrophy, strict heterotrophy, constitutive mixotrophy, generalist non-constitutive mixotrophy, plastidic-specialist non-constitutive mixotrophy or endosymbiotic-specialist non-constitutive mixotrophy) from 798 transcriptomes of 705 species were annotated through bibliographic research. Among these 798 transcriptomes, 650 came from MMETSP (Keeling et al., 2014), while transcriptomes of 45 species came from the Roscoff culture collection (`http://roscoff-culture-collection.org/`), 6 came from the OCEANOMICS database (`http://www.oceanomics.eu/`), 1 from the Pasteur culture collection (`https://webext.pasteur.fr/cyanobacteria/`), and 4 from the Meng et al. (2018) study. One transcriptome of *Alexandrium minutum* that was extracted from Le Gac et al. (2016) was finally added to the dataset, constituting the only dinoflagellate transcriptome of our analysis that was not already included in Meng et al. (2018). Within the 798 microbial eukaryotic transcriptomes, I identified 105 transcriptomes of constitutive mixotrophs (from 23 different genera), 1 of generalist non-constitutive mixotroph, 4 of plastidic-specialist non-constitutive mixotrophs, and 3 of endo-symbiotic non-constitutive mixotrophs (full list available in Appendix B). Among the 47 dinoflagellate transcriptomes corresponding to 43 distinct species from 27 genera, 18 came from constitutive mixotrophs and 1 came from a plastidic-specialist non-constitutive mixotroph (*D. acuminata*). The methods described in Meng et al. (2018) were then used to build a sequence similarity network (SSN) of the 47 transcriptomes: protein coding domains were detected and functionally annotated through TransDecoder (v5.5.0, Haas et al. (2013)) and InterProScan (v5.24-63.0, Jones et al. (2014)), before being aligned in *all versus all* mode using the DIAMOND software to retrieve similarity and coverage statistics for each transcripts pair (Buchfink et al., 2015). As only one transcriptome was added to the dataset in comparison to Meng et al. (2018), the same thresholds were used to perform this analysis: edges with a similarity higher than 60% and a coverage of more than 80% were conserved in order to build the SSN using the R (R Core Team, 2019) package igraph (Csardi et al., 2006). Meng et al. (2018) selected these parameters to maximize the number of connected components with more than 30 vertices and the number of connected components involving a unique functional annotation.

3.3.2.3 Results and discussion

The SSN was composed of 2,901,054 proteins, including 728,916 singletons (25.1%) and 304,026 connected components, ranging from 2 to 43,480 proteins in size (Table 3.1). In comparison,

Table 3.1 - Metrics of the dinoflagellates transcriptomes sequence similarity network. CC stands for connected component, CM for constitutive mixotroph.

	Number of CCs (%)	**Proteins in CCs (%)**
Total CCs	304,026 (100%)	2,172,138 (100%)
CCs with at least one sequence from a CM	143,676 (47.2%)	803,736 (36.9%)
CCs with at least one sequence from a pSNCM	15,212 (5.0%)	33,346 (3.1%)
CCs with at least one sequence from *A. minutum* (Le Gac et al., 2016)	46,473 (15.3%)	107,672 (4.9%)
CCs with at least one sequence from the MMETSP *A. minutum*	6,357 (2.1%)	10,190 (0.5%)
CCs only composed of sequences from CMs	56,791 (18.7%)	153,340 (7%)
CCs only composed of sequences from mixotrophs	60,864 (20%)	163,310 (7.5%)

the SSN built in Meng et al. (2018) (without the *A. minutum* transcriptome from Le Gac et al. (2016)) was composed of 2,790,387 proteins including 1,514,476 singletons (54.3%) and 350,267 connected components ranging from 2 to 1600 proteins in size. The important differences in numbers of singletons and maximum size of connected components clearly questions our choice of using the same thresholds as Meng et al. (2018). Indeed, even though only one transcriptome was added to the SSN, it ranked second in number of proteins in connected components, and significantly changed the structure of the network, notably leading to the creation of 4 gigantic connected components of more than 10,000 proteins each. These gigantic connected components included proteins from all the 47 transcriptomes, and their functional homogeneity was poor: *e.g.* more than 40 distinct functional annotations were found in the biggest CC. This indicates that the thresholds of similarity and coverage used were probably too low and led to the construction of chimeric connected components composed of functionally and evolutionary unrelated proteins. The important effect of the addition of the *A. minutum* transcriptome on the network structure can be explained by its important size (110,667 proteins, which makes it the fifth biggest in the data set), and its high quality. Indeed, this transcriptome was obtained from the separate sequencing of 18 strains of *A. minutum* (Le Gac et al., 2016), and contained more than 10 times more protein coding domains than the *A. minutum* transcriptome already available in the MMETSP collection (which contained 10,572 proteins). Considering that transcriptomes corresponding to different strains of the same species were included separately in the dataset (in Meng et al. (2018), transcriptomes from different strains of the same species were kept separated when their numbers of reads were sufficient to create independent high-quality[1] transcriptomes, *e.g.* two strains of *Brandtodinium nutricula* and of *Kryptoperidinium foliaceum*), implementing the sequence similarity network with 18 separated transcriptomes corresponding to each strains of *A. minutum* might lead to better results. As evoked earlier, the impact of this added transcriptome on the network structure could also be mitigated by a modification of the similarity and coverage thresholds, but

[1] in Meng et al. (2018), high-quality transcriptomes are defined as having more than 30,000 transcripts, with 50% of the whole transcriptome in transcripts longer than 400 base pairs and read re-mapping rate over 50%

increasing these thresholds would lead to the creation and exclusion of singletons that were taken in account in Meng et al. (2018), and the break of interesting connected components identified as putative markers into multiple smaller ones. To avoid this, it could be interesting to reproduce the study using a community detection algorithm, like Louvain (Blondel et al., 2008), which allow to detect communities of highly connected nodes in large networks. This way, large connected components can be subdivided in smaller communities (Watson et al., 2019). The Louvain algorithm has the advantage of being fast, and to not rely on parameter choices by the user (Blondel et al., 2008), when other algorithms like MCL ask the user to choose parameter values that are not trivial to define (Watson et al., 2019).

Table 3.2 - Composition of the 4 connected components identified as potential markers of mixotrophy in dinoflagellates. Species with names in green are constitutive mixotrophs while Dinophysis acuminata, indicated in red, is a plastidic-specialist non-constitutive mixotroph. Functional annotations are from InterProScan v5.24-63.0 (Jones et al., 2014).

Connected Component	Number of Proteins	Species	Order	Functional Annotation(s)
CC 1	12	*Alexandrium andersonii*	Gonyaulacales	Unknown
		Alexandrium catenella	Gonyaulacales	
		Alexandrium monilatum	Gonyaulacales	
		Alexandrium tamarense	Gonyaulacales	
		Lingulodinium polyedra	Gonyaulacales	
		Heterocapsa sp.	Peridiniales	
		Heterocapsa trinquera	Peridiniales	
CC 2	12	*Alexandrium andersonii*	Gonyaulacales	Unknown
		Alexandrium catenella	Gonyaulacales	
		Ceratium fusus	Gonyaulacales	
		Heterocapsa rotundata	Peridiniales	
		Heterocapsa sp.	Peridiniales	
		Heterocapsa trinquera	Peridiniales	
CC 3	13	*Alexandrium catenella*	Gonyaulacales	Unknown
		Alexandrium tamarense	Gonyaulacales	
		Prorocentrum minimum	Prorocentrales	
		Lingulodinium polyedra	Gonyaulacales	
		Heterocapsa sp.	Peridiniales	
		Heterocapsa trinquera	Peridiniales	
CC 4	15	*Ceratium fusus*	Gonyaulacales	Unknown
		Lingulodinium polyedra	Gonyaulacales	
		Amphidinium carterae	Gymnodiniales	
		Amphidinium massartii	Gymnodiniales	
		Prorocentrum minimum	Prorocentrales	
		Dinophysis acuminata	Dinophysiales	

In our SSN, 56,791 connected components were only composed of proteins from constitutive mixotrophs, corresponding to almost a fifth of the total number of CCs (Table 3.1). These connected components had a mean size of 2.7 proteins, and showed low taxonomic richness (*i.e.* low number of distinct species represented in the CC), with a maximum of 7 mixotrophic species found in the same connected component (over 18 mixotrophic species in the dataset). The four

connected components associated only to mixotrophic species that had the highest species richness were selected as the best candidates for being potential markers of mixotrophy (Table 3.2). Two of them were composed of 12 proteins, one of 13 and one of 15. One had proteins from 7 different constitutive mixotrophs, two others from 6 and one last component blended proteins from 5 constitutive and 1 plastidic-specialist non-constitutive mixotrophs (Table 3.2). Among the CCs only composed of sequences from constitutive mixotrophs, two mixed species from two different orders, *Gonyaulacales* and *Peridiniales*, while the third one also included a *Prorocentrales* species (Figure 3.2). A phylogeny of 47 dinoflagellates transcriptomes based on 1043 orthologous protein sets identified *Gonyaulacales*, *Peridiniales* and *Prorocentrales* as a clade (Stephens et al. (2018); Figure 3.6). The three CCs identifed here could then constitute markers of constitutive mixotrophy at the level of this clade. But they could also be phylogenetic markers of this clade with no functional relation to mixotrophy. The 3 CCs did not contain any proteins from the 11 transcriptomes of non-mixotrophic species belonging to the *Gonyaulacales*/*Peridiniales*/*Prorocentrales* clade in our dataset, but they also did not include proteins from some transcriptomes of mixotrophic *Peridiniales* (*e.g.* *Scrippsiella trochoidea*) and *Gonyaulacales* (*e.g.* *Ceratium fusus*) (Table 3.2, Figure 3.6). It is then hard to conclude with certainty on the potential of these CCs as markers of constitutive mixotrophy without further wet-lab explorations (as evoked in section 3.1), especially since they could not be associated with any function in reference databases.

The fourth CC identified as a potential marker of mixotrophy blended 15 proteins from 4 different orders, *Gonyaulacales*, *Gymnodiniales*, *Prorocentrales* and *Dinophysiales* (Table 3.2). It is less likely for the proteins of this connected component to be limited to phylogenetic markers than it was for the three other candidate CCs, because these four orders belong to a large clade also including the *Peridiniales*, *Suessiales* and *Noctilucales* orders (Figure 3.6), from which no proteins appear in the CC. It is however legitimate to question the credibility of a genomic marker that would detect both constitutive and plastidic-specialist non-constitutive mixotrophy, as organisms from the two trophic modes differ strongly in terms of physiology (Mitra et al., 2016). For example, this CC could not be a marker of kleptoplasty, which can not be performed by constitutive mixoitrophs. However, both types of mixotrophic organisms share the ability to eat through phagocytosis, and this CC could then be tested as a marker of phagotrophy in dinoflagellates. Here again, all the 15 proteins from the connected component were 'known unknowns', *i.e.* they were found in the InterProScan database (Jones et al., 2014), but could not be associated to a biological function. They could then constitute interesting proteins to target for wet lab experiments focusing on phagocytosis, to check for their influence on mixotrophic abilities in dinoflagellates.

Despite the fact that mixotrophic abilities were associated with more than a hundred transcriptomes of MMETSP through a literature review, no information were available in the metadata of these transcriptomes indicating the mode of feeding of the cultivated organisms. Considering that constitutive mixotrophs often have varying feeding behaviors, adapting their rate of mixotrophy to their environment (McKie-Krisberg et al., 2018; Liu et al., 2016; Lv et al., 2019), it is difficult to tell whether the transcriptomes of mixotrophic species in the MMETSP database should carry markers of mixotrophy or not. This way, even with an improved version of the SSN, and eventually the addition of more transcriptomes to the 798 ones that were annotated, the potential detection

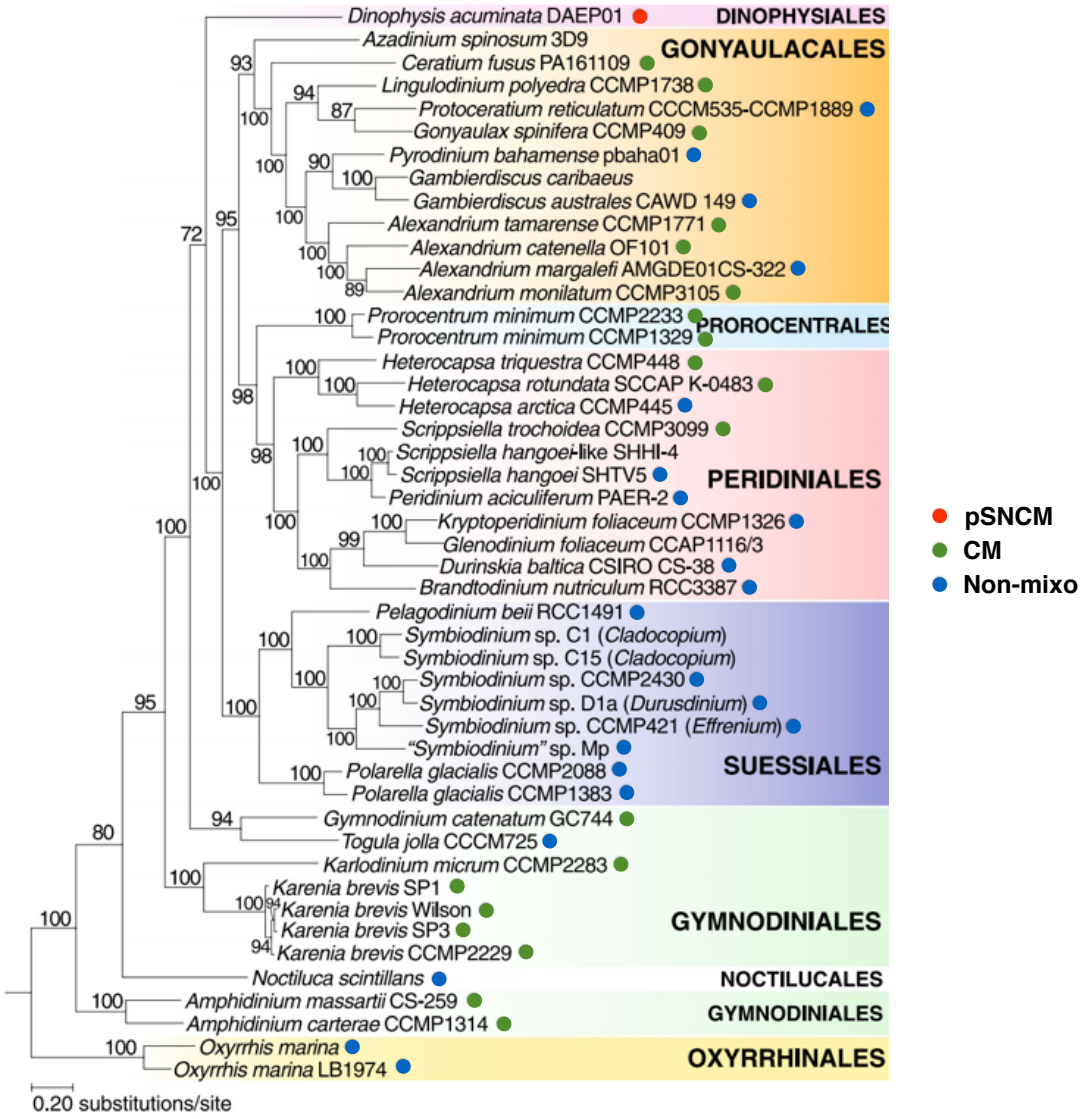

Figure 3.6 - Maximum-likelihood phylogeny of dinoflagellates inferred using 1043 orthologous protein sets issued from 47 transcriptomes of dinoflagellates, modified from Stephens et al. (2018). Bootstrap support values are indicated at each split. Branch length is based on the rate of substitution per site. Colored dots were added next to every species that was represented in our sequence similarity network (strains were not taken into account except for the Symbiodinium genus, in which species could not be attributed). Red dots indicate plastidic-specialist non-constitutive mixotrophs (pSNCM), green dots indicate constitutive mixotrophs (CM) and blue dots indicate species for which no proofs of mixotrophy were found in the literature.

of mixotrophy markers would still be questionable regarding the data used. This illustrates well how the fact that mixotrophy is a continuum between autotrophy and heterotrophy hardens the analysis of its genomic basis.

3.3.2.4 Conclusion

The more reliable way to find markers of mixotrophy would be to reproduce a SSN-based approach using transcriptomes of mixotrophic organisms issued from comparative transcriptomics studies at a fine scale (*i.e.* including species sharing similar mixotrophic behaviours, and exposing them

to similar sets of conditions). The goal should be to produce quantitative data that would allow to sort connected components/transcript families based on expression level, implying the need for replicates (this analysis showed the impact of sequencing depth) in different conditions favoring certain trophic modes. Studies are already showing the way by sequencing transcriptomes of mixotrophic species along gradients of lights and nutrients, giving us access to transcriptomes of mixotrophs with varying levels of photosynthetic and phagotrophic capacity (McKie-Krisberg et al., 2018; Liu et al., 2016; Lv et al., 2019). By building a sequence similarity network using transcriptomes coming out of such experimental designs, and by analyzing it through the lens of metadata indicating rates of phagotrophy and photosynthesis, it should be possible to find connected components grouping proteins from multiple species in the same feeding behavior state. It could allow to expand from species or strain-specific putative markers detected by comparative transcriptomics to higher trophic levels ones. The presence of such markers could then be tested in other, less studied mixotrophic species, or in environmental metagenomics and metatranscriptomics samples.

3.4 Next challenge: linking unknown functions and uncultivated organisms to functional traits

Until now, I have only presented approaches based either on wet lab investigations of specific organisms, or on the statistical investigation of large datasets of full genomes/transcriptomes. This often implies that the focal organisms can be cultivated in labs. However, up to 99% of microbial species remain impossible to cultivate (Rappé and Giovannoni, 2003; Rinke et al., 2013; Watson et al., 2019; Mangot et al., 2017), so one important challenge is to associate functional traits to poorly known and not yet culturable organisms.

As illustrated in section 3.2.2, the mining of genomic markers in metagenomes and metatranscriptomes can allow to draw quantitative hypotheses on the distribution of functional traits without distinguishing cultivated and uncultivated organisms. In a more organism-centered way, the recent identification of hundreds of MAGs from the *Tara Oceans* metagenomics data allowed to detect nitrogen fixation genes in abundant yet uncultivated Planctomycetes and Proteobacteria (See section 1.3.2.2, Delmont et al. (2018)), genomes of uncultured picoeukaryotes and giant viruses were retrieved from targeted single-cell genomics (Mangot et al., 2017; Needham et al., 2019), and transcriptomes of uncultured eukaryotes were determined from metagenomic samples (Vorobev et al., 2019). Similarly, Lannes et al. (2019) were able to detect carbon fixation pathways in marine ultrasmall prokaryotes, without even relying on the assembly of genomes or transcriptomes. Instead, they filtered sequences from metagenomes of the viral size fraction from *Tara Oceans*, in order to only keep sequences affiliated to prokaryotes. This way, they identified ultrasmall prokaryotes to collectively harbor (*i.e.* without proof of presence in a single genome) the dicarboxylate/4-hydroxybutyrate pathway and the 4-hydroxybutyrate pathway, which are both energy efficient pathways leading to autotrophic carbon fixation (Lannes et al., 2019).

Still, 40 to 60% of the open reading frames (ORF) detected in microbiome analyses are of unknown

function (Vanni et al., 2020; Bernard et al., 2018). To explore the function of such ORFs, one method is to identify their remote homologs, also called distant homologs, or transitive homologs (Lopez et al., 2015; Watson et al., 2019). The idea behind this method is to find proteins in environmental samples that are indirectly homologous to proteins from functional annotation databases, *e.g.* an unannotated homolog to a protein that has a match in functional annotation databases (Watson et al., 2019). Lobb et al. (2015) found 15.3% of the 484,121 ORFs analyzed in their study to be distant homologs of structurally characterized proteins, and were able to identify hundreds of novel enzymes. Very recently, a new database and tool called AGNOSTOS came out, that references known and unknown genes as clusters based on their sequence similarity, allowing to very rapidly identify distant homologs of query proteins (Vanni et al. (2020); more on this in the general discussion).

These are examples of the methods available to decipher the functional potential of uncultivated organisms and functionally unannotated genes. However, in matters of functional traits, only traits with well known genomic markers can be detected in meta-omics samples and in MAGs. We can hope that in a near future, comparative transcriptomics, methods like GWAS and wet lab experiments will allow to better describe the genomic basis of complex traits like cell size or mixotrophy. Only then, investigating these multigenic markers in environmental samples and uncultivated organisms will be possible. This is why I decided to test a different approach for the next part of my thesis, trying to focus on detecting functional clusters of genes of interest without any *a priori* selection based on their functional annotations.

Part II

Data-driven approaches to identify and quantify the functional composition of planktonic communities

Chapter 4

Towards omics-based predictions of planktonic functional composition from environmental data

4.1 Prelude

In the introduction, I highlighted how current biogeochemical modeling approaches did not allow to define model structural components from observational data, and always relied on *a priori* choices of the model PFTs, traits, genes or metabolic pathways. During the first part of my thesis, I did not address this particular issue, as I focused my attention on two particular traits, namely mixotrophy and DMS production. As presented in chapter 3, the investigation of *a priori* chosen functional traits often implies to rely on genomic markers detected in cultivated organisms. I presented how meta-omics data led to the detection of such markers in unexpected taxa, and how they allowed to better understand the biogeography of functional traits independently from taxonomical assignations. But such studies do not take full advantage of the richness of meta-omics data, as they focus on one or a few *a priori* selected genes and discard the rest, including functionally unannotated genes, even though full metagenomes and metatranscriptomes contain information on the functional potential of planktonic communities as a whole (Vanni et al., 2020).

In the second part of my thesis, my goal was then to design an approach to extract functionally homogeneous clusters of proteins from meta-omics data without any *a priori* based on their functional and/or taxonomic annotation (Figure 4.1). One of the main objectives was to be able to compute the abundance of each cluster in the environment, to be able to describe and understand their biogeography. It allowed me to identify and confirm the main drivers of functional composition in planktonic prokaryotic communities, but also to highlight proteins, functions and MAGs particularly associated with environmental gradients in the global ocean.

In the following section, I will present a data-driven method applicable to any set of sequences, allowing to build protein functional clusters and quantitatively link their abundance to the environment without *a priori* selection of taxa or metabolic functions, while including all unannotated

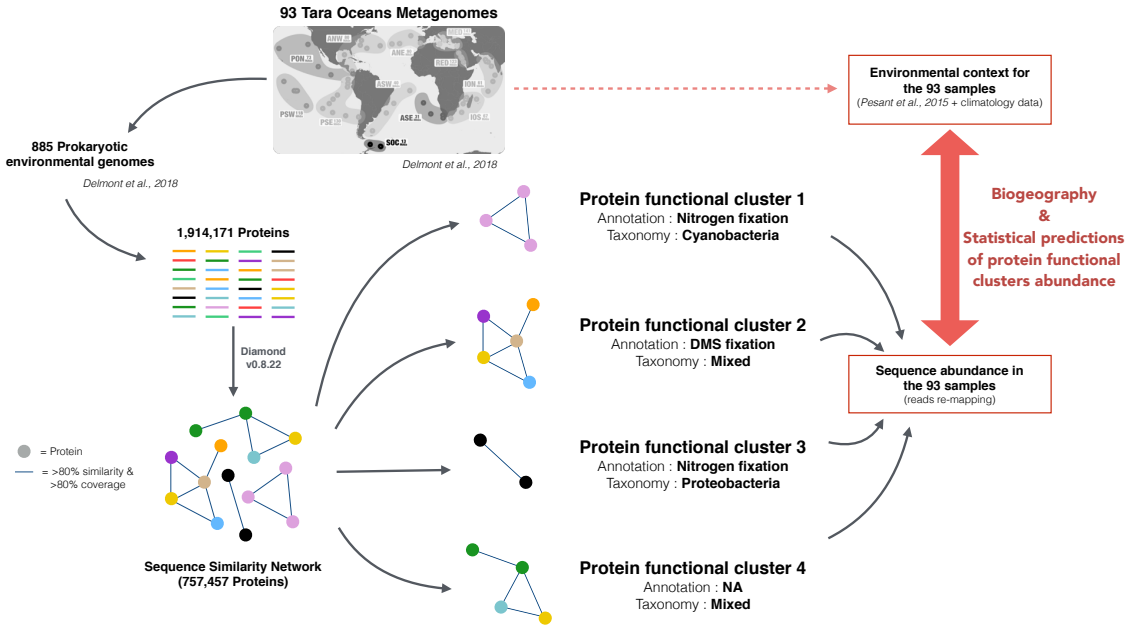

Figure 4.1 - Analyzing the biogeography of functionally homogeneous protein clusters obtained from metagenome-assembled genomes. Nearly 2 millions of proteins from 885 prokaryotic MAGs are assembled into a sequence similarity network, i.e. a graph in which nodes are proteins and links reflect the similarity and coverage between each pair of proteins. In this toy example, the sequence similarity network is composed of 4 connected components, or groups of nodes connected together directly or indirectly, and disconnected form the rest of the network. Each connected component is defined as a protein functional cluster, and examples of taxonomic and functional annotations are indicated to illustrate the kind of information that can be retrieved from the approach. The abundance of each protein functional cluster can be computed in environmental metagenomes, through the mapping of environmental reads to each protein. These abundances can finally be statistically related to the environmental context.

proteins. I decided to apply this method on the 885 prokaryotic MAGs produced by Delmont et al. (2018), which were manually curated and appeared to me as more reliable than fully-automatically binned ones (*e.g.* Parks et al. (2017) or Tully et al. (2018)). The same approach could in theory be applied to larger datasets, and even to full *Tara Oceans* gene catalogs, but computational limitations and the necessity to benchmark the approach with a more easy to handle dataset pushed me to first focus on the MAGs. They appeared as good candidates as they mostly correspond to uncultivated organisms, with poorly described functional potentials despite high abundances across the global ocean.

The rest of this chapter will consist in a manuscript entitled *Towards omics-based predictions of planktonic functional composition from environmental data*, currently undergoing modifications after a first round of revisions in *Nature Communications*.

4.2 Towards omics-based predictions of planktonic functional composition from environmental data

Authors: Emile Faure[1,2], Sakina-Dorothée Ayata [1,2 *], Lucie Bittner [2 *].

1-Sorbonne Université, CNRS, Laboratoire d'océanographie de Villefranche, LOV, 06230 Villefranche-sur-Mer, France

2-Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, CP 50, 57 rue Cuvier, 75005 Paris, France

* Authors contributed equally

Abstract: Marine microbes play a crucial role in climate regulation, biogeochemical cycles, and trophic networks. Unprecedented amounts of data on planktonic communities were recently collected, sparking a need for innovative data-driven methodologies to quantify and predict their ecosystemic functions. We reanalysed 885 marine prokaryotic metagenome-assembled genomes through a network-based approach and detected 233,756 protein functional clusters, from which 15% were functionally unannotated. We investigated all clusters' distributions across the global ocean through machine learning, identifying biogeographical provinces as the best predictors of protein functional clusters' abundance. The abundances of 2,444 clusters were predictable from the environmental context, including 183 functionally unannotated clusters. We analyzed the biogeography of these 2,444 clusters, identifying the Mediterranean Sea as an outlier in terms of protein functional clusters composition. Applicable to any set of sequences, our approach constitutes a step towards quantitative predictions of functional composition from the environmental context.

4.2.1 Introduction

Planktonic organisms play an essential role in biogeochemical cycles through the capture and export of carbon into the deep ocean, nitrogen fixation, remineralization of organic matter, or the production of dimethyl-sulfur, hence impacting global climate (Falkowski et al., 1998; Guidi et al., 2016; Whitman et al., 1998; Ferrera et al., 2015; Sunagawa et al., 2015). The understanding and modeling of such biogeochemical functions is key for predicting the global functioning of oceanic ecosystems, and especially their response to climate change (Le Quéré et al., 2005; Litchman et al., 2015a; Follows et al., 2007). These biogeochemical functions are usually modeled by simulating the dynamics of plankton functional types (PFT) that are theoretical entities grouping planktonic organisms according to shared functional capacities (*e.g.* calcifiers, nitrogen fixers or silicifiers) (Le Quéré et al., 2005). This approach allows to incorporate the functional diversity of marine plankton into biogeochemical models (Follows et al., 2007; Aumont et al., 2015; Coles et al., 2017; Leles et al., 2016), but often relies on a priori and restricted choices of the considered types of planktonic organisms and of their physiological rates or parameters (Flynn et al., 2015). For example, prokaryotic organisms are often lacking an explicit representation in global PFT models (Aumont et al., 2015; Leles et al., 2016), even though more than 1030 prokaryotic cells inhabit the

ocean's subsurface (Whitman et al., 1998). To tackle this issue, recent works proposed to switch towards data-driven modeling of planktonic communities and their impact on the environment, notably through the use of high-throughput sequencing data (Coles et al., 2017; Mock et al., 2016; Louca et al., 2016a).

Next generation sequencing technologies have led to significant advances in the knowledge of the taxonomic and functional diversity of planktonic organisms (Sunagawa et al., 2015; Louca et al., 2016c). Bioinformatics workflows allow the assembly of metagenome-assembled genomes (MAGs), which are near-complete genomes retrieved from DNA fragments coming from environmentally sequenced individuals of one or a few closely related populations (Parks et al., 2017; Tully et al., 2018; Delmont et al., 2018; Nielsen et al., 2014). MAGs can be taxonomically annotated using multi marker gene approaches, and organism-level functional profiles can be drawn from their genomic content (Parks et al., 2017; Tully et al., 2018; Delmont et al., 2018). Reads from environmental meta-omics datasets can also be mapped to their reconstructed sequences to obtain abundance measurements both at MAG and single protein level (Delmont et al., 2018; Salazar et al., 2019). MAGs can be considered as representative of the genetic potential of natural populations, hence allowing to retrieve genomes of cultivable, uncultivable or even unknown species present in the environment. They constitute a promising tool for investigating as a whole the functional potential of known and unknown planktonic life forms.

Recently, a genomics-based model revealed that the gene content of planktonic communities is more relatable to biogeochemical gradients than taxonomic content (Coles et al., 2017). In another study, omics data were used to quantitatively estimate global nitrogen fixers abundance through machine learning algorithms (Tang and Cassar, 2019). It illustrates how quantitative, data-driven biogeochemical models can be built from global omics datasets. However, these studies focused only on a relatively small number of well-described genes (*e.g.* nif or amtB genes, involved in dinitrogen and ammonium fixation, respectively) (Coles et al., 2017; Tang and Cassar, 2019), far from exploiting the rich functional diversity observed in omic samples. This way, the large proportion of unknown sequences detected in environmental meta-omics datasets, that is to say the open reading frames (ORFs) which can not be linked to any biological functions (usually around 40% for prokaryotes, and about 50% for eukaryotes), is as yet untapped (Ferrera et al., 2015; Sunagawa et al., 2015; Salazar et al., 2019; de Vargas et al., 2015; Acinas et al., 2019; Carradec et al., 2018). Besides, many meta-omics studies have either focused on semi-quantitative diversity and interactions surveys at global scales (de Vargas et al., 2015; Lima-Mendez et al., 2015), on specific taxonomic groups (*e.g.* Collodaria (Biard et al., 2017)) or on particular biological functions (such as nitrogen fixation or mixotrophy (Delmont et al., 2018; Tang and Cassar, 2019; Faure et al., 2019)). A recent study has grouped protein sequences of marine planktonic prokaryotes according to their annotated metabolic pathways to investigate their differential abundance and expression, mainly focusing on pre-selected biogeochemical functions such as photosynthesis or nitrogen fixation (Salazar et al., 2019). By investigating the response of biogeochemistry-related protein groups to environmental conditions, significant differences in terms of presence and expression were identified between polar and non-polar areas, and between mesopelagic and surface depths (Salazar et al., 2019). These results highlight the potential of function-clustering

based approaches for deciphering global ocean biogeochemistry, but could be further extended by skipping any sequence pre-selection step requiring database-dependent metabolic pathways annotations.

In this study, we followed a similar approach while avoiding any a priori choices of particular genes or metabolic pathways. We used 51 quantitative and qualitative environmental variables to detect both known and unknown protein clusters that are sensitive to environmental gradients. We re-analysed 885 high quality MAGs from marine prokaryotic plankton belonging to the Bacteria (n=820) and Archaea (n=65) domains, assembled by Delmont et al. (Delmont et al., 2018) using 93 *Tara Oceans* picoplanktonic metagenomes from the surface of the global ocean. With these almost 2 million sequences, we built functional clusters of proteins using a sequence similarity network, *i.e.* a graph in which nodes are protein sequences, and edges represent the similarity and coverage between each pair of sequences (Atkinson et al., 2009; Forster et al., 2015; Meng et al., 2018; Bittner et al., 2010; Lopez et al., 2015). Such approaches allow for the construction of sequence clusters putatively homogenous in function (Atkinson et al., 2009), and were recently used to investigate the genomic basis of functional diversity in prokaryotes (Cheng et al., 2014), in a lineage of eukaryotes (Meng et al., 2018), or in natural microbial communities (Lopez et al., 2015). Particularly, we are here interested in knowing if the abundance of some protein clusters could be predicted from environmental data in the oceanic ecosystem. For example, is the distribution of biogeochemistry-related protein clusters more sensitive to environmental gradients than the one of other clusters? We thus explored the biogeography of environment-related protein clusters in light of their potential functional and / or taxonomic annotation, in order to identify the ones being specific to certain environmental conditions, such as oligotrophic or particularly cold waters.

We introduce here a data-driven, large-scale, fast and automatable approach, potentially applicable to any set of environmental sequences, which involves (1) the network-based construction of sequence clusters, putatively homogeneous in function, (2) the functional annotation of these clusters, (3) the calculation of environmental abundance values for each of these protein clusters through environmental reads re-mapping, and (4) the description of statistical relationships between cluster abundances and environmental gradients through machine learning and constrained ordination methods. We then present the first biogeographical analysis of known and unknown prokaryotic protein functional clusters identified as sensitive to environmental gradients in the global ocean, with no a priori choice of specific functions or taxa.

4.2.2 Results

4.2.2.1 From sequence similarity network to protein functional clusters

We analysed the 1,914,171 proteins from 885 prokaryotic MAGs from marine plankton, recovered from 12 geographically bound assemblies of metagenomic sets corresponding to a total of 93 *Tara Oceans* samples (Delmont et al., 2018). 39.6% of the MAGs' proteins (757,457) were involved in our sequence similarity network, *i.e.* they had at least one similarity relationship with another protein that satisfied the chosen threshold of 80% similarity and 80% coverage (see Methods).

51.1% of the network proteins could be annotated to 4,922 unique molecular function IDs in the KEGG database (Aramaki et al., 2019), associated with 327 distinct metabolic pathways (a full list of these pathways is displayed in Table S1, available in appendix C). 85.2% of the network proteins were annotated to 17,009 eggNOG functional descriptions (Huerta-Cepas et al., 2017, 2016).

PFC size		Functional scores					Taxonomy scores				
		Homogeneity		Unknowns quantification			Homogeneity		Unknowns quantification		
Mean	3.24	Mean homogeneity score with EggNOG annotations (Number of NA values)	0.94 (35,818)	EggNOG annotations	PFCs only composed of annotated proteins (% of total PFCs)	181,595 (77.7%)	PFCs associated to only 1 Phylum (% of total PFCs) *(% of PFCs with at least one Phylum annotation)*	221,541 (94.8%) *(97.5%)*	Phylum level	Only proteins from annotated MAGs (% of total PFCs)	220,839 (94.5%)
										Only proteins from unannotated MAGs (% of total PFCs)	6,367 (2.7%)
					PFCs with at least one annotated protein (% of total PFCs)	197,938 (84.7%)	PFCs associated to only 1 Class (% of total PFCs) *(% of PFCs with at least one Class annotation)*	192,095 (82.2%) *(96.8%)*	Class level	Only proteins from annotated MAGs (% of total PFCs)	186,331 (79.7%)
										Only proteins from unannotated MAGs (% of total PFCs)	35,338 (15.1%)
Minimum	2				PFCs only composed of unknown proteins (% of total PFCs)	35,818 (15.3%)	PFCs associated to only 1 Order (% of total PFCs) *(% of PFCs with at least one Order annotation)*	144,265 (61.7%) *(93.8%)*	Order level	Only proteins from annotated MAGs (% of total PFCs)	135,046 (57.8%)
										Only proteins from unannotated MAGs (% of total PFCs)	79,921 (34.2%)
Maximum	1072	Mean homogeneity score with KEGG annotations (Number of NA values)	0.99 (113,321)	KEGG annotations	PFCs only composed of annotated proteins (% of total PFCs)	91,103 (39.0%)	PFCs associated to only 1 Family (% of total PFCs) *(% of PFCs with at least one Family annotation)*	100,801 (43.12%) *(95.3%)*	Family level	Only proteins from annotated MAGs (% of total PFCs)	88,404 (37.8%)
										Only proteins from unannotated MAGs (% of total PFCs)	128,010 (54.76%)
					PFCs with at least one annotated protein (% of total PFCs)	120,435 (51.5%)	PFCs associated to only 1 Genus (% of total PFCs) *(% of PFCs with at least one Genus annotation)*	21,921 (9.4%) *(91.9%)*	Genus level	Only proteins from annotated MAGs (% of total PFCs)	13,544 (5.8%)
					PFCs only composed of unknown proteins (% of total PFCs)	113,321 (48.5%)	PFCs associated to only 1 MAG (% of total PFCs)	7,146 (3.1%)		Only proteins from unannotated MAGs (% of total PFCs)	209,892 (89.8%)

Table 4.1 – Metrics computed on the 233,756 protein functional clusters (PFC) from the sequence similarity network of MAGs proteins. Functional scores are based on the functional annotation of MAGs proteins, with a functional homogeneity score of 1 meaning that all proteins in a PFC share the same annotation, while a score of 0 indicates that all proteins have different annotations (see Methods for details). By "unknown proteins" we refer both to sequences with no match in databases (KEGG and/or eggNOG) and to sequences existing in databases but with no functional and/or taxonomic annotation. Taxonomy scores are based on taxonomic annotations of MAGs available from Delmont et al. 2018. This way, the 6,367 PFCs with only proteins from MAGs unannotated at the phylum level were only composed of proteins coming from the 45 Bacteria MAGs of unidentified phylum. Detailed functional and taxonomic annotations for each protein sequence are available online, as well as detailed sizes and functional/taxonomy scores for each PFC (see Data availability section).

The sequence similarity network involved 233,756 connected components (CCs), *i.e.* groups of nodes (here proteins) connected together by at least one path and disconnected from the rest of the network. According to KEGG and eggNOG databases, 15.3% and 48.5% of the CCs remained without any functional annotation (*i.e.* all sequences from the CC were unmatched in the databases, or had a match but were not yet linked to any biological function, Table 1), and 14.8% were functionally unannotated for both databases. We ranked the functional homogeneity of CCs involving at least one functional annotation from 0 (all annotations in the CC are different) to 1 (all annotations in the CC are the same), and found mean homogeneity scores of 0.99 over 1 for KEGG annotations and 0.94 over 1 for eggNOG ones (see Methods for score calculation details). Only 88 (0.04%) CCs had an homogeneity score below 0.5 in both annotation databases, all with sizes below 5 proteins. 177 PFCs (0.07%) had a score below 0.8 in both databases, all under 12 proteins in size. These CCs were kept in the analysis while tagged as poorly homogenous. We thereafter

considered each CC as a protein functional cluster (PFC), numeroted from #1 to #233,756.

To check for the influence of taxonomic relationships between the MAGs on our PFCs, we computed different metrics based on MAGs taxonomic annotations provided by Delmont et al. 2018 (Table 1). This taxonomic annotation based on 43 single-copy core genes allowed to annotate 100% of the MAGs at the domain level, and 95% of the MAGs at the phylum level, the remaining 5% corresponding to Bacteria MAGs of unidentified phyla (Delmont et al., 2018). Only 1,330 PFCs (0.6%) mixed proteins from the Archaea and Bacteria domains. PFCs were very homogeneous at the phylum level, then the homogeneity decreased at lower taxonomic rank, meaning that PFCs studied here were generally not specific from a single class, order, family, genus or MAG (Table 1). 7,834 PFCs (3.4%) were only composed of proteins with no functional annotation in KEGG and eggNOG databases, and no taxonomic annotation under the phylum level. Their sizes ranged from 2 to 30 proteins (mean of 2.62). Their 20,552 proteins came from Euryarcheota MAGs (12,458; 60.6%), Bacteria MAGs of unidentified phylum (2,742; 13.3%), Candidatus Marinimicrobia MAGs (2,451; 11.9%), Proteobacteria MAGs (1,528; 7.4%), Acidobacteria MAGs (1,031; 5%), Verrucomicrobia MAGs (103; 0.5%), Planctomycetes MAGs (89; 0.4%), Bacteroidetes MAGs (79; 0.4%), Chloroflexi MAGs (59; 0.3%) and Candidate Phyla Radiation MAGs (12; 0.05%). We hereafter considered these functionally and taxonomically unknown PFCs as “microbial dark matter” (Rinke et al., 2013; Bernard et al., 2018). Their nucleotidic sequences are available in separate supplementary files (see Data availability section). The abundance of microbial dark matter PFCs was significantly different from the abundance of other PFCs in 85 samples over 93 (Wilcoxon test, p-value <0.05). The median abundance of microbial dark matter PFCs was higher than the one of other PFCs in 36 of these 85 samples, and lower in the 49 others. Further details on dark matter PFCs’ abundance are available in the *Detection of the rare biosphere* section.

4.2.2.2 Identification of protein functional clusters highly related to environmental gradients

To identify the PFCs that responded the most to environmental gradients, we first selected the 228,914 clusters with non-zero variance abundance profiles (*i.e.* at least 10% of distinct abundance values across all samples, and less than a 95 to 5 ratio between the most and the second most observed abundance value). We then built random forest regression models for each of these 228,914 clusters. We used the sequence abundances as response variables, or labels, and 51 environmental variables as explanatory variables (see Methods for details of model training and tuning). About a fifth of the random forest regression models showed a clear statistical signal: 44,653 models (19.5%) had R^2 values over 0.25, corresponding to PFCs linked to environmental conditions, and 2,444 (1.1%) had values over 0.5 (Figure 4.2A), corresponding to PFCs highly linked to environmental gradients. The mean R^2 value over all models was 0.09, with a maximum of 0.91 (Figure 4.2A). Longhurst biogeographical provinces (Longhurst, 1998) were detected as the most important predictor in 90,235 models (39.4%), and were in the top 3 most important predictors in 166,639 models (72.8%) (Figure 4.2B). Among models with biogeographical provinces as the best predictor, the mean R^2 reached 0.15. Temperature was in the top 3 most important

PFC size		Functional scores					Taxonomy scores				
		Homogeneity		Unknowns quantification			Homogeneity		Unknowns quantification		
Mean	5.54	Mean homogeneity score with EggNOG annotations (Number of NA values)	0.95 (192)	EggNOG annotations	PFCs only composed of annotated proteins (% of selected PFCs)	2,054 (84%)	PFCs associated to only 1 Phylum (% of selected PFCs) *(% of selected PFCs with at least one Phylum annotation)*	2,289 (93.7%) *(94.2%)*	Phylum level	Only proteins from annotated MAGs (% of selected PFCs)	2,392 (97.9%)
										Only proteins from unannotated MAGs (% of selected PFCs)	14 (0.6%)
					PFCs with at least one annotated protein (% of selected PFCs)	2,252 (92.1%)	PFCs associated to only 1 Class (% of selected PFCs) *(% of selected PFCs with at least one Class annotation)*	2,100 (85.9%) *(92.6%)*	Class level	Only proteins from annotated MAGs (% of selected PFCs)	2,121 (86.8%)
										Only proteins from unannotated MAGs (% of selected PFCs)	177 (7.2%)
Minimum	2				PFCs only composed of unknown proteins (% of selected PFCs)	192 (7.9%)	PFCs associated to only 1 Order (% of selected PFCs) *(% of selected PFCs with at least one Order annotation)*	1,725 (70.6%) *(92.6%)*	Order level	Only proteins from annotated MAGs (% of selected PFCs)	1,454 (59.5%)
										Only proteins from unannotated MAGs (% of selected PFCs)	582 (23.8%)
		Mean homogeneity score with KEGG annotations (Number of NA values)	0.996 (828)	KEGG annotations	PFCs only composed of annotated proteins (% of selected PFCs)	1,199 (49.1%)	PFCs associated to only 1 Family (% of selected PFCs) *(% of selected PFCs with at least one Family annotation)*	604 (24.7%) *(84.4%)*	Family level	Only proteins from annotated MAGs (% of selected PFCs)	480 (19.6%)
										Only proteins from unannotated MAGs (% of selected PFCs)	1,728 (70.7%)
Maximum	196				PFCs with at least one annotated protein (% of selected PFCs)	1,616 (66.1%)	PFCs associated to only 1 Genus (% of selected PFCs) *(% of selected PFCs with at least one Genus annotation)*	48 (1.96%) *(29.6%)*	Genus level	Only proteins from annotated MAGs (% of selected PFCs)	8 (0.3%)
					PFCs only composed of unknown proteins (% of selected PFCs)	828 (33.9%)	PFCs associated to only 1 MAG (% of selected PFCs)	102 (4.2%)		Only proteins from unannotated MAGs (% of selected PFCs)	2,282 (93.4%)

Table 4.2 - Metrics computed on the 2,444 protein functional clusters detected as particularly linked to environmental gradients (hlePFCs). Functional scores are based on the functional annotation of MAGs proteins, with a functional homogeneity score of 1 meaning that all proteins in a PFC share the same annotation, while a score of 0 indicates that all proteins have different annotations (See Methods). By "unknown proteins" we refer both to sequences with no match in databases (KEGG and/or eggNOG) and to sequences existing in databases but with no functional and/or taxonomic annotation. Taxonomy scores are based on taxonomic annotations of MAGs from Delmont et al. 2018

predictors in 11,856 models (5.2%). Models with temperature as the best predictor had a mean R^2 of 0.29, which is the highest value of all quantitative variables.

We focused on the 2,444 PFCs associated with models showing R^2 values over 0.5, hereafter called "hlePFCs" for highly linked to environmental conditions protein functional clusters. 207 KEGG pathways were associated with the 2,444 hlePFCs, *i.e.* 63% of the pathways identified on the full network were detected in hlePFCs (c.f. Table S1 in Appendix C for a detailed list). Among commonly detected pathways (arbitrary threshold of at least 1000 detections in the similarity sequence network, more details in the *Links between models R^2 and their associated metabolic pathways* section), the 5 most conserved pathways after selection were methane metabolism (2.63% of the PFCs associated to this pathway were hlePFCs), ribosome (2.43%), carbon fixation in photosynthetic organisms (2.28%), carbon fixation pathways in prokaryotes (2.01%) and pentose phosphate pathway (1.99%).

The functional homogeneity of the 2,444 hlePFCs was similar to the one of the total 233,756 PFCs (Table 1, Table 2). In parallel, proportions of functionally unannotated hlePFCs were lower and the proportion of hlePFCs only composed of unannotated proteins in both functional databases was divided by two (from 14.8% to 7.5%).

The proportions of hlePFCs homogenous at the phylum, class and order level were comparable to

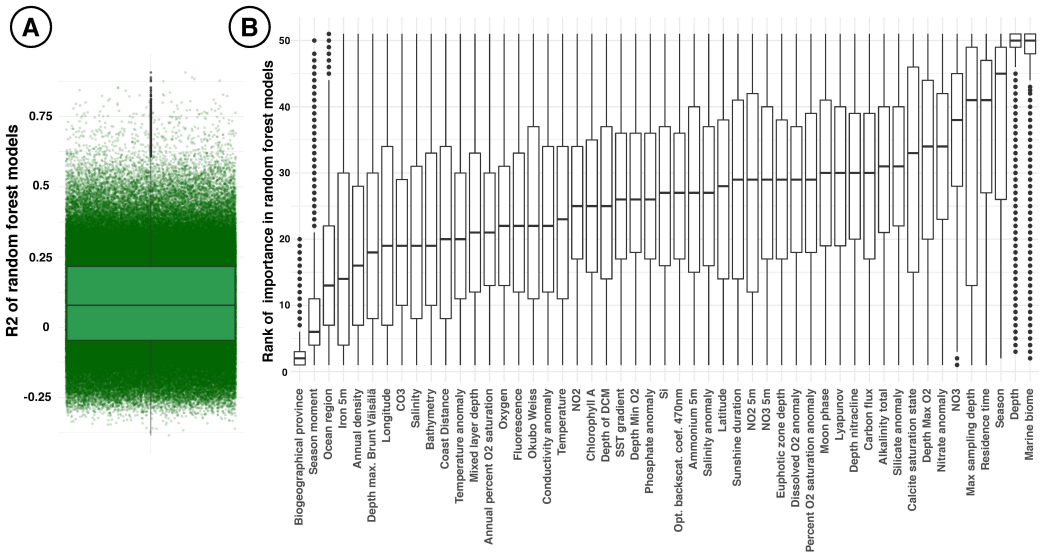

Figure 4.2 - (A) Boxplot distribution of R^2 values of all the 228,914 random forest models. Each point corresponds to a model, the y axis corresponding to the R^2 values, and the x axis position being randomized for visual representation. The mean R^2 value over all models was of 0.09, with a maximum of 0.91. (B) Rank of importance of each environmental variable in models with positive R^2 values. Ranks were attributed from 1 for the most important to 51 for the least important variable in each model.

the one of the total PFCs and superior or equal to 70% (Table 1, Table 2). The trend differed at the family and genus level, where only 84.4% and 29.6% of the hlePFCs with at least one protein annotated were associated to a single taxonomic annotation, while this proportion was of 95.3% and 91.9% in the total PFCs (Table 1, Table 2). This way, hlePFCs tended to mix more taxa at the family and genus levels than the rest of the PFCs, while retaining a high functional homogeneity. The proportion of taxonomically unannotated hlePFCs was lower than the one of total PFCs at the phylum, class and order levels, but was higher at the family and genus levels (Table 1, Table 2). 24 of the 7,834 PFCs (*i.e.* 0.3%) tagged as microbial dark matter (*i.e.* PFCs without any functional annotation and without taxonomic information under the phylum level) were selected among the 2,444 hlePFCs. These 24 hlePFCs corresponded to 49 proteins belonging to 12 unique MAGs, annotated as Proteobacteria (n = 5), Euryarchaeota (n = 3), Candidatus Marinimicrobia (n = 2), and Bacteria unannotated at Phylum level (n = 2).

4.2.2.3 Global biogeography of the protein functional clusters highly linked to environmental gradients

The canonical correspondence analysis (CCA) achieved on the 2,444 hlePFCs to investigate their biogeographical repartition had an R^2 value of 72.9%, and was significant (p-value < 0.001). The first axis (20.62% of explained variance) opposed warm and oligotrophic waters (CCA1>0) to cold and nutrient rich waters (CCA1<0) (Figure 4.3). The second axis (17.03%) mostly opposed samples from the Mediterranean Sea (CCA2>0) to the rest of the samples. At the exception of Mediterranean samples, samples from geographically close biogeographical provinces appeared close to each other in the CCA space, with samples from the Southern Ocean and the Atlantic

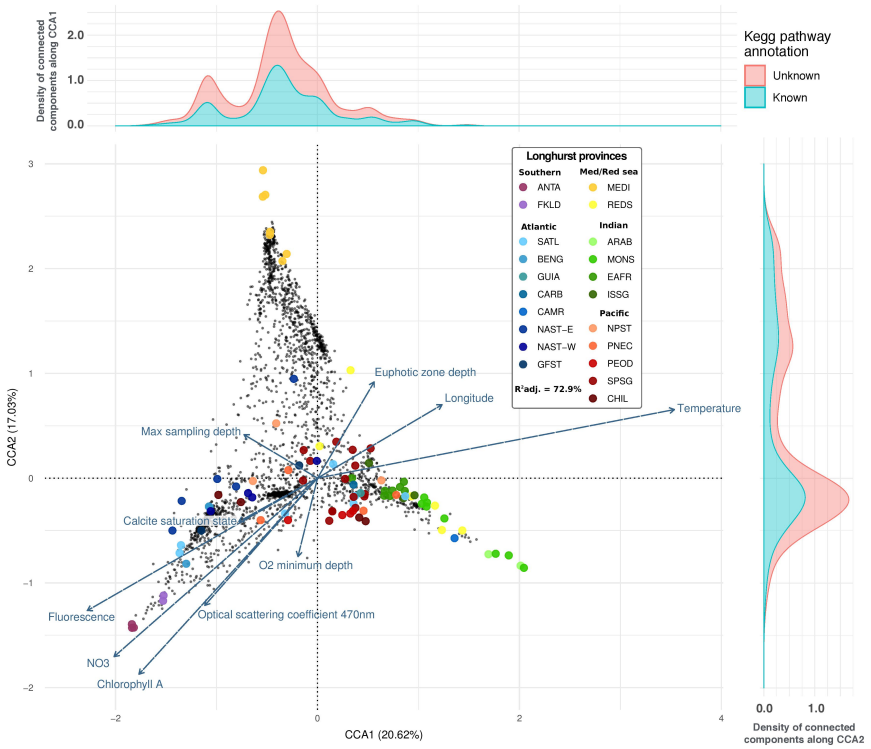

Figure 4.3 - Canonical correspondence analysis (CCA) on abundances of the 2,444 protein functional clusters highly linked to environmental variables (hlePFCs). hlePFCs are represented as black dots, quantitative environmental variables as arrows, and samples as circles colored according to their biogeographical province (correspondence between 4 letters codes used here and full biogeographical provinces names, as well as descriptions of all other environmental variables are displayed in Table S2, available in Appendix C). For simplification issues, other qualitative variables (Season moment and Ocean region) were not represented. On the right and upper panels, density plots are represented along each axis, illustrating the density of functionally annotated and unannotated hlePFCs based on KEGG annotations (i.e. functionally annotated hlePFCs contain at least one functionally annotated protein; functionally unannotated hlePFCs contain only functionally unannotated proteins). The mean hlePFC density was of 0.27 along CCA1 (standard deviation = 0.34, maximum = 1.38), and 0.21 along CCA2 (standard deviation = 0.23, maximum = 0.94). The mean difference in density between functionally annotated and unannotated hlePFCs along CCA1 was of 0.06 (standard deviation = 0.07, maximum = 0.28). The mean density difference between annotated and unannotated hlePFCs along CCA2 was 0.04 (standard deviation = 0.03, maximum = 0.13). Similar observations were done using eggNOG annotations densities (Figure 4.4).

zones on the left, Pacific Ocean in the middle and Indian Ocean on the right (Figure 4.3). The two closest samples from the Mediterranean ones in the CCA space were from the closest Atlantic station to the strait of Gibraltar (station TARA_004) at the entrance of the Mediterranean Sea, and the second closest Red Sea station to the Suez canal mouth (station TARA_032) (Figure 4.3).

Combining the CCA results with the functional annotation of hlePFCs, we identified several metabolic pathways enriched in particular environmental conditions (Figure 4.5). Carotenoid biosynthesis was enriched in the Mediterranean Sea (CCA2>0), as well as flagellar assembly and oxidative phosphorylation pathways, while mismatch repair and DNA replication pathways were enriched in nutrient rich, cold waters (Figure 4.5; Figure 4.7). Pathways related to biogeochemical functions (*e.g.* carbon fixation pathways in prokaryotes or methane metabolism) or linked to ecological interactions between organisms (*e.g.* biosynthesis of antibiotics, quorum sensing or ABC

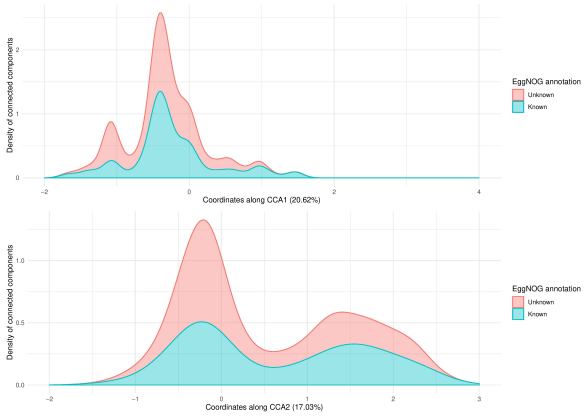

Figure 4.4 - Density plots illustrating the proportions of functional known and unknown hlePFCs along CCA1 and CCA2 axes, based on eggNOG annotations (i.e. a known hlePFC corresponds to a protein cluster in which at least one sequence has an eggNOG assignment; an unknown hlePFC corresponds to a protein cluster without any eggNOG assignment). The mean hlePFC density was of 0.26 along CCA1 (standard deviation = 0.32, maximum = 1.27), and 0.21 along CCA2 (standard deviation = 0.21, maximum = 0.82). The mean difference in density between functional known and unknown hlePFCs along CCA1 was of -0.0006 (standard deviation = 0.1, maximum = 0.34). The mean density difference between known and unknown hlePFCs along CCA2 was 0.03 (standard deviation = 0.1, maximum = 0.33).

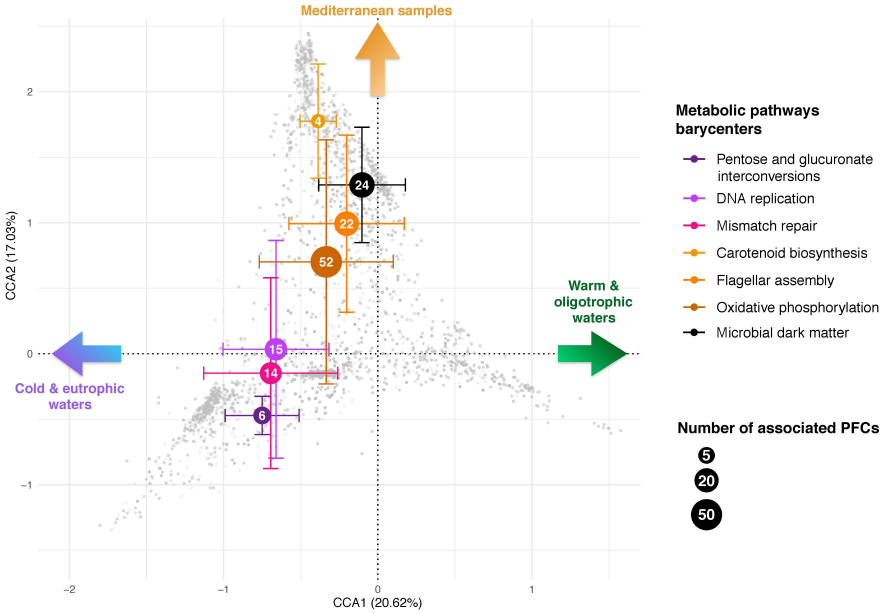

Figure 4.5 - Barycenters of the positions of protein functional clusters highly linked to the environment (hlePFCs) associated to 6 selected metabolic pathways and to microbial dark matter in the canonical correspondence analysis (CCA) space. These barycenters were selected for their peripheral positions in the CCA space. Pathways with barycenters in shades of pink were related to cold rich waters only, while those with barycenters in shades of orange were mostly associated to Mediterranean samples. The barycenter of microbial dark matter hlePFCs (i.e. hlePFCs without functional annotation and taxonomical assignment below the phylum level) was represented in black, and was strongly associated to Mediterranean samples. Error bars correspond to the standard deviations of hlePFCs positions on x and y axes. Size of barycenters represent their number of associated hlePFCs, with the exact corresponding values written in white in each barycenter. Colored arrows indicate the environmental conditions associated with the different zones of the CCA space (See Figure 4.3).

transporters) were present homogeneously in the CCA space (Figure 4.7). Nitrogen metabolism was an exception as it was only associated with hlePFCs that were quite central in the CCA space (Figure 4.7), corresponding mainly to Pacific and North Atlantic samples. No particular niche showed an overabundance of functionally unannotated hlePFCs (Figure 4.3), while the 24 microbial dark matter hlePFCs were strongly associated to Mediterranean samples (Figure 4.5, Figure 4.9B).

We then examined the position of hlePFCs associated with different levels of taxonomic annotations in the CCA space. hlePFCs containing sequences from the phylum Candidatus Marinimicrobia were over-abundant in Mediterranean samples (Figure 4.6A). It was the only phylum with a strong association to a particular niche in the CCA space (cf. the larger standard deviation bars for other phyla on Figure 4.6A). hlePFCs containing proteins of Opitutae, Dehalococcoidetes and Betaproteobacteria classes were mostly positioned on the positive side of CCA2, corresponding to Mediterranean samples, but with standard deviation bars spanning to negative values (Figure 4.6B). hlePFCs containing Betaproteobacteria proteins were rare in warm and oligotrophic samples from the Indian Ocean (Figure 4.6). hlePFCs containing proteins from two classes of cyanobacteria (Prochlorales and Chroococcales) were particularly abundant in warm and oligotrophic waters (CCA1 > 0 and CCA2 < 0).
191 hlePFCs were highly overabundant in the Mediterranean Sea (CCA2>2.0), corresponding to 422 proteins from 34 different MAGs of 6 classes. 26 of these MAGs originated from the same assembly performed by Delmont et al. 2018 on Mediterranean samples, and accounted for 410 proteins. This predominance of MAGs from one particular assembly was not observed for another assembly along CCA2 (Figure 4.8). A similar yet less marked pattern was observed along CCA1, with 350 and 344 of the 992 proteins of hlePFCs correlated to cold and rich waters (CCA1 < -1) coming from MAGs of the Atlantic South-East and Atlantic South-West assemblies, respectively (Figure 4.8).

Our analysis allowed to identify environmental variables driving the abundance of functionally unannotated hlePFCs. For example, PFC #90,349 was composed of 9 unannotated proteins coming from 4 different MAGs (3 Flavobacteriales, 1 Gammaproteobacteria), and had a strong response to high temperature (Figure 4.9A), as well as other environmental variables (R^2 value of 0.504 for the associated random forest model). Conversely, PFCs #102,286 (2 proteins coming from the same Saprospiraceae), #210,456 (2 proteins from two distinct Flavobacteriaceae), and #161,812 (2 proteins from the same Flammeovirgaceae) were highly linked to cold temperature (Figure 4.9A). PFCs #172,160, #176,586 and #177,371 were microbial dark matter hlePFCs overabundant in Mediterranean samples.
Positions of all functionally and/or taxonomically unannotated hlePFCs in the CCA space, the most important drivers of their abundance according to random forest models, the nucleotidic sequences of their proteins and their MAGs of origins are all publicly accessible (link in Data availability section).

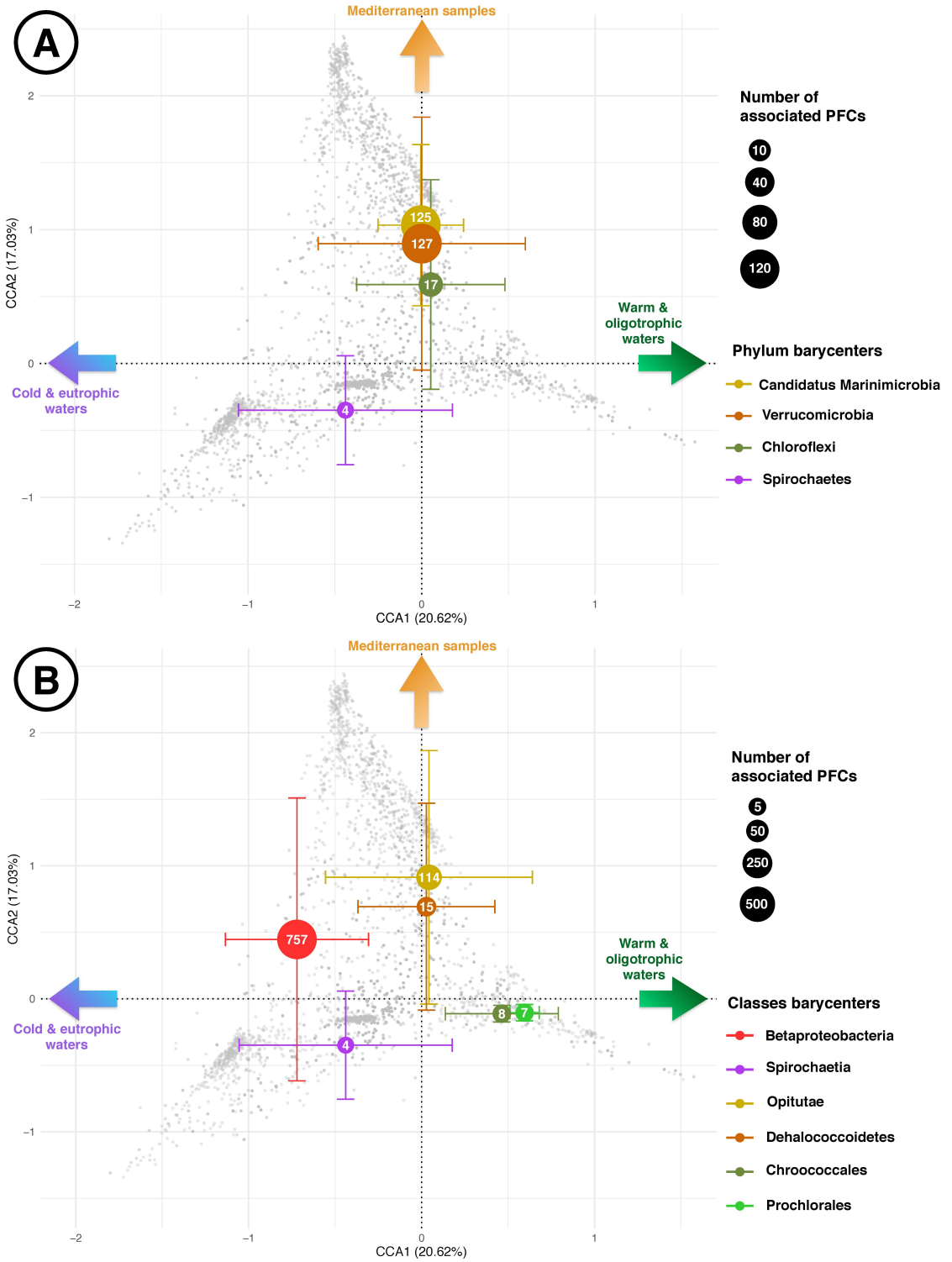

Figure 4.6 - Distribution in the canonical correspondence analysis (CCA) space of the barycenters of protein functional clusters highly linked to the environment (hlePFCs) associated to particular taxa: (A) 4 selected phylum and (B) 6 selected classes. These taxa were selected because they had the most peripheral barycenters' positions in the CCA space. Error bars correspond to the standard deviations of hlePFCs positions on CCA1 and CCA2 axes for each taxa. The size of barycenters represents the number of associated hlePFCs for each taxa, with the exact corresponding values written in white in each barycenter. Colored arrows indicate the environmental conditions associated with the different zones of the CCA space (c.f. Figure 4.3).

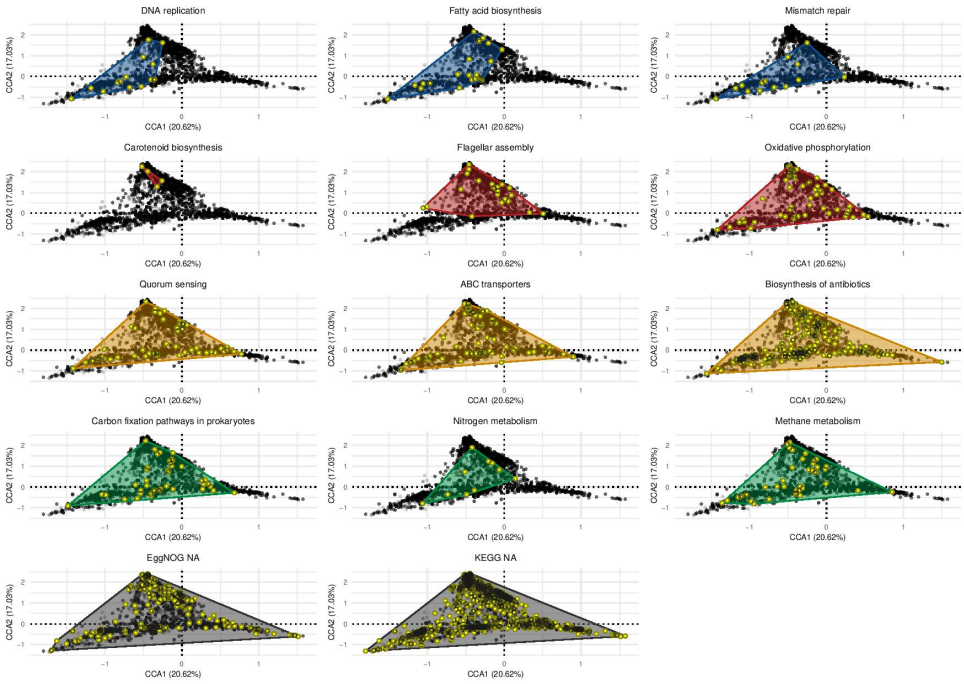

Figure 4.7 - Convex hulls englobing all hlePFCs associated to different pathways in the CCA two-dimensional space. On each graph, black dots represent hlePFCs that are not associated to the focal pathway, while yellow dots represent hlePFCs containing at least one sequence associated to the focal pathway. Convex hulls were drawn in different colors depending on the type of pathway. Pathways associated to cold and rich waters are included in blue convex hulls, while the ones associated to mediterranean samples are in red convex hulls. We also selected three pathways linked to inter-organisms interactions, in orange hulls, and three pathways related to biogeochemical functions in green hulls. Finally, two black convex hulls englobing all functional unknown PFCs were represented, one for KEGG annotations and another for eggNOG ones.

4.2.2.4 Detection of the rare biosphere

By defining microbial dark matter protein functional clusters (PFCs) as made of proteins functionally unannotated and taxonomically unannotated under the phylum level, we selected quite abundant PFCs, probably missing most of the " rare biosphere " (Lynch and Neufeld, 2015) (See main text). To check for the impact of our choice of definition, we investigated the abundance of microbial dark matter PFCs using other thresholds of taxonomic annotation to define microbial dark matter.w 12,895 PFCs were only composed of proteins with no functional annotation in KEGG and eggNOG databases, and had no taxonomic annotation under the class level. The abundance of these 12,895 PFCs was significantly different from the one of the rest of PFCs in 88 samples over 93 (Wilcoxon test, p-value <0.05). Their median abundance was lower in 32 of these 88 samples, and higher in the 56 others. 20,166 PFCs were only composed of proteins with no functional annotation in KEGG and eggNOG databases, and had no taxonomic annotation under the order level. The abundance of these 20,166 PFCs was significantly different from the one of the rest of PFCs in 84 samples over 93 (Wilcoxon test, p-value <0.05). Their median abundance was lower in 65 of these 84 samples, and higher in the 19 others. 32,737 PFCs were only composed of proteins with no functional annotation in KEGG and eggNOG databases, and had no taxonomic annotation under the family level. The abundance of these 32,737 PFCs was

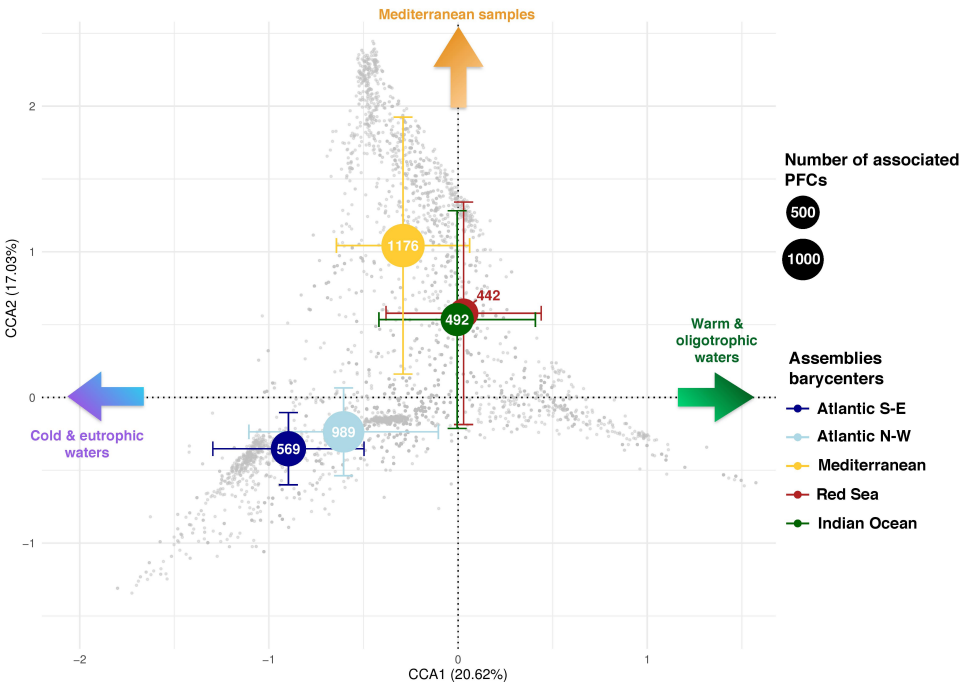

Figure 4.8 - Barycenters of the protein functional clusters highly linked to environmental gradients (hlePFCs) from 5 selected assemblies in the canonical correspondence analysis (CCA) space. These assemblies were selected because they had the most peripheral barycenters' positions in the CCA space. Error bars correspond to the standard deviations of hlePFC positions on CCA1 and CCA2 axis for each assembly. Size of barycenters represent the number of associated hlePFCs for each assembly, with the exact corresponding values written in white in each barycenter, except for the Red Sea assembly for which it is written in red next to the barycenter. Colored arrows indicate the environmental conditions associated with the different zones of the CCA space (See Figure 4.3).

significantly different from the one of the rest of PFCs in all samples (Wilcoxon test, p-value <0.05), and their median abundance was lower in all samples but one, a surface sample from the Indian Ocean (TARA_064). This way, step by step considering PFCs taxonomically unannotated under the class, order and family level as microbial dark matter led to a decrease of median abundance of microbial dark matter PFCs at each step. It then seems that the " rare biosphere " was better detected when including unidentified lineages of known class, order or family, than when using only unidentified lineages of known phylum. As stated in the main text, this can be explained by the lack of knowledge about the abundants Archaea and Candidatus Marinimicrobia phyla. Although, we could also argue that using MAGs might not be the best way to observe the " rare biosphere ". Indeed, the binning of contigs into MAGs relying mainly on co-abundance profiles (Delmont et al., 2018; Parks et al., 2017; Tully et al., 2018) , it is likely that organisms with very low and erratic abundances over samples could be missed in the binning steps.

4.2.2.5 Links between models R2 and their associated metabolic pathways

The 228,914 random forest produced models with R 2 values ranging from -0.38 to 0.91. We investigated the mean R 2 values of models grouped by their associated metabolic pathways. The maximum mean value of 30% was obtained for the calcium signaling pathway, but only 1 model was associated to it. Among pathways associated to at least 100 proteins, values ranged from

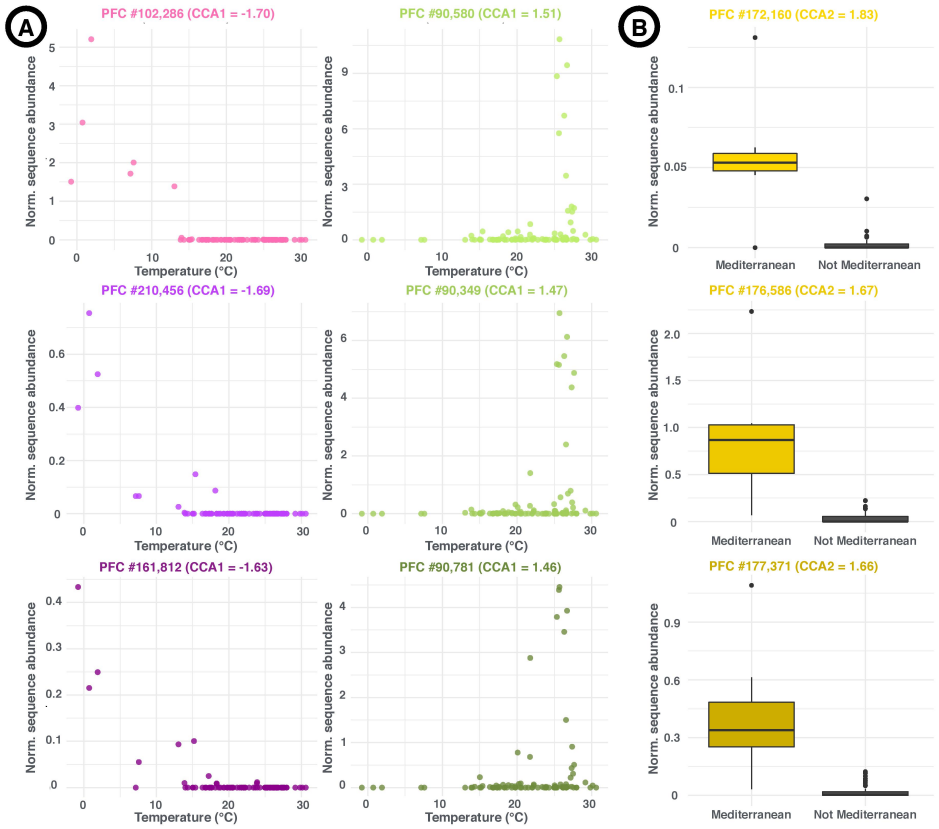

Figure 4.9 - (A) Relationships between normalized sequence abundance and temperature for 6 selected protein functional clusters highly linked to the environment (hlePFCs) that were only composed of functionally unannotated sequences. The three graphs on the left, in purple, correspond to the three hlePFCs functionally unannotated in both KEGG and eggNOG databases that had the lowest positions on the first axis of the canonical correspondence analysis (CCA1) (cold and nutrient rich waters). The three graphs in the middle, in green, correspond to the three hlePFCs functionally unannotated in both KEGG and eggNOG databases that had the highest positions on CCA1 (oligotrophic and warm waters). (B) Relationships between normalized sequence abundance and location of sampling, whether in the Mediterranean Sea or not, for 3 microbial dark matter hlePFCs (only functionally unannotated sequences and no taxonomic annotations under the phylum level). These 3 hlePFCs had the highest positions among microbial dark matter PFCs on the second axis of the canonical correspondence analysis (CCA2) (correlated to Mediterranean samples).

5.6% to 17.1%, showing that no common pathway could be associated to particularly high R 2 values. 16 negative values were observed, mostly corresponding to human-related metabolic pathways, *e.g.* rheumatoid arthritis, cocaine addiction, thyroid hormone signaling pathway, or pancreatic secretion. We then investigated the metabolic pathways associated to PFCs highly linked to environmental gradients (hlePFCs), to check if some pathways were more associated to hlePFCs than others. 121 pathways were not detected at all in hlePFCs (Table S1 in appendix C). The most selected pathway was biosynthesis of vancomycin group antibiotics , with 4.65% of its associated PFCs being hlePFCs. But this pathway was associated to only 43 PFCs in total, and 2 of them only were hlePFCs. Similarly, other rare pathways appeared as well selcted even though they did not make much sense for planktonic communities, like prostate cancer (66 associated PFCs, 4.54% in hlePFCs) or cardiac muscle contraction (182 associated PFCs, 4.40% in hlePFCs). This is why we decided to focus on more abundant pathways only. A threshold of 200 associated PFCs was too low, as pathways like Parkinson disease (associated to 215 PFCs) or Alzheimer

disease (associated to 371 PFCs were still among the most selected pathways). At a threshold of 500 associated PFCs, we had 98 unique pathways, with still some unrelated to prokaryotic communities like thermogenesis (591 associated PFCs, 2.71% in hlePFCs) or pathways in cancer (555 associated PFCs, 1.26% in hlePFCs). But results were already close to the ones presented in the main text (using a threshold of 1000 associated PFCs), with methane metabolism , ribosome , carbon fixation in photosynthetic organisms and pentose phosphate pathway in the top 10 of most selected pathways among hlePFCS. Keeping a threshold of 500 associated pathways, we could even add the photosynthesis pathway to this list, which was the 7th most selected pathway (641 associated pathways, 2.18% in hlePFCs). Using a threshold of 1000 associated pathways as presented in the main text, we were able to focus only on the 62 most common pathways, which were all making sense in a planktonic ecology context.

4.2.3 Discussion

4.2.3.1 Functional composition of prokaryotic plankton communities is driven by interactions between multiple environmental factors rather than by single variables

Building statistical models including 51 environmental variables to test for their effect on protein functional clusters (PFCs) abundance, we were able to quantify the impact of each variable both globally and in each individual random forest model. Our results hence give access to the most influential predictors of 228,914 protein functional cluster's abundances (see Data availability section), while pushing towards a consideration of other variables in addition to temperature and oxygen when studying prokaryotic communities functional composition. Indeed, water temperature is commonly presented as the most influential determinant of the taxonomic and functional composition of prokaryotic communities (Sunagawa et al., 2015; Salazar et al., 2019; Delmont et al., 2017). Here, we found temperature to be one of the best quantitative predictors of PFC abundance. Still, it was determined as less important than other quantitative variables like iron, carbonate or oxygen concentrations, but also salinity or bathymetry. However, when temperature was the most important variable in a model, it highly increased abundance predictions accuracy (mean R^2 values three times higher than the one over all models), showing how influential this variable is on at least some ecosystemic functions. Among all environmental predictors, Longhurst biogeographical provinces (Longhurst, 1998) were by far the most important variable in our random forest models, and were well distinguished on the canonical correspondence analysis (CCA) triplot (Figure 4.3). Longhurst provinces represent homogeneous areas both in terms of physico-chemical and ecological conditions (Longhurst, 1998). Here we thus suggest that functional composition is more impacted by interactions between multiple variables, than by one or a few variables like it has been previously suggested (Sunagawa et al., 2015; Salazar et al., 2019). This result adds to a similar observation made by a recent global biogeographical analysis of planktonic communities, finding Longhurst biogeographical provinces to match the distribution of viruses, bacteria and eukaryotes smaller than 20 µm (Richter et al., 2019). Sampling depth had

among the lowest impacts on our regression models, confirming the weak differences in functional composition between surface and deep chlorophyll maximum samples of picoplankton (Salazar et al., 2019).

4.2.3.2 Identifying protein functional clusters and metabolic pathways associated with particular environmental conditions

The identification of biogeographical provinces as best predictors of PFC's abundance can be interpreted as a consequence of the Baas Becking hypothesis 'everything is everywhere, but the environment selects' (Wit and Bouvier, 2006; Fondi et al., 2016), which implies that all microbes are potentially ubiquitous, but dominant taxa depend on the environmental niche. A precedent study observed this pattern at the proteic level, by comparing protein families sampled in different ecosystems such as sea water, sludge water or soils, and showing that the ecosystem type had more impact on protein families composition than geographical distance (Fondi et al., 2016). Similarly, by identifying Longhurst biogeographical provinces to be the best predictors of PFCs abundance, we verify that environmental niches are the most determinant drivers of marine prokaryotic communities functional composition. However, the fact that 80.5% of our PFCs showed poor responses to environmental conditions challenges the extent of applicability of the Baas Becking hypothesis at the proteic level, at least within a single ecosystem. It could be explained by the high decoupling observed between functional diversity and taxonomic diversity among marine prokaryotic communities (Louca et al., 2016c). Indeed, functional redundancy among prokaryotes can lead to stable functional diversity even with high taxonomic variability (Louca et al., 2016b). In our analysis, we chose to focus on changes in communities functions because it could lead to more stable abundance measures than when relying on taxonomic entities, and provide thus more valuable information on the ecosystem functioning and associated biogeochemical functions (Louca et al., 2016c; Salazar et al., 2019; Louca et al., 2016b).
Still, 19.5% of the PFCs were linked to environmental gradients, and 1.1% showed very strong responses to particular environmental conditions. Among these 2,444 PFCs identified as highly linked to environmental gradients (hlePFCs), we observed a clear distinction between the ones associated with polar nutrient-rich waters and those abundant in tropical nutrient-poor ones, which is coherent with classical observations in marine ecology (Faure et al., 2019; Ibarbalz et al., 2019). Metabolic pathways like DNA replication or mismatch repair could be associated with eutrophic conditions and colder waters (Figure 4.5), potentially reflecting the higher growth potential and metabolic activity of micro-organisms in the eutrophic conditions of the polar summer (Alonso-Sáez et al., 2008). No particular metabolic pathway could be associated with warm and oligotrophic waters, but proteins from two classes of Cyanobacteria were overrepresented in hlePFCs abundant in these waters, which may reflect that cyanobacteria are particularly abundant in tropical, nutrient-poor waters (Flombaum et al., 2013). Among commonly observed pathways, methane metabolism, carbon fixation in photosynthetic organisms and carbon fixation pathways in prokaryotes were three of the 4 most selected pathways in hlePFCs, and were correlated to a wide range of physico-chemical conditions (Figure 4.7). Hence, such key biogeochemical func-

tions are quite ubiquitously present in the global ocean, but can be achieved by different actors and protein families depending on the environmental conditions. One exception was the nitrogen metabolism pathway, which was mainly associated with Pacific and North Atlantic samples, corroborating results obtained with the same set of MAGs (Delmont et al., 2018).

More surprisingly, we identified Mediterranean samples as clear outliers, with an important part of hlePFCs showing higher abundances in Mediterranean samples than elsewhere. These samples were notably characterized by a relative over-abundance of metabolic functions like carotenoid biosynthesis and flagellar assembly, and an over-abundance of proteins from MAGs of the Candidatus Marinimicrobia phylum, which is only composed of poorly known and yet uncultivable bacteria of potentially high biogeochemical impact (Hawley et al., 2017). Our strongest hypothesis to explain these patterns lies in the fact that the Mediterranean Sea is a semi-enclosed sea that experienced multiple isolation and colonization events (Patarnello et al., 2007). For some pelagic species, the strait of Gibraltar constitutes a phylogeographic barrier causing genetic contrasts between Atlantic and Mediterranean populations (Patarnello et al., 2007; Lowe et al., 2012). Here, we identified most Atlantic samples to be closer to Pacific ones than to Mediterranean ones in terms of hlePFCs composition, at the exception of one which came from the mouth of the strait of Gibraltar. Also, the Mediterranean Sea was the only biogeographical zone exhibiting a strong over-abundance of locally assembled proteins. This way, hlePFCs overabundant in Mediterranean samples shared only very few links with proteins from MAGs of other assemblies in our sequence similarity network, highlighting their functional and taxonomical originality. We then propose that the strait of Gibraltar and the Suez canal could shape the genetic and functional structure of some planktonic prokaryotic populations, as it is observed in some eukaryotic species (Patarnello et al., 2007; Lowe et al., 2012).

4.2.3.3 Mining the unknown to identify potential key organisms and proteins

The main originality of our approach is its ability to take into account both annotated and unannotated sequences. It enables the identification of PFCs composed of functionally unannotated sequences, of taxonomically unannotated sequences, and of both, leading here to the inclusion of at least 15% more proteins than methods excluding functionally unannotated sequences. By including 7,834 PFCs corresponding to 20,552 protein sequences that could not be annotated under the phylum level nor to a biological function, we propose an original way to highlight the response of microbial dark matter abundance to environmental gradients. While a previous study estimated that the inclusion of microbial dark matter sequences could increase by up to 58% the amount of analysed sequences (Bernard et al., 2018), we provide here a pragmatic bioinformatic pipeline which helps to extend our knowledge in environmental microbiology.

It is often proposed that most of the unidentified microbial diversity could come from rare organisms, described as the 'rare biosphere', and which are considered as diversity reservoirs able to respond rapidly to environmental changes (Logares et al., 2014; Lynch and Neufeld, 2015). Our results partly corroborate this theory (see *Detection of the rare biosphere*), but we also found the 7,834 microbial dark matter PFCs to be relatively overabundant in 41% of our samples. This can

be explained by the fact that 72.5% of the proteins from our microbial dark matter PFCs came from Candidatus Marinimicrobia and Euryarcheota MAGs, two yet poorly studied and uncultivable phyla identified as highly abundant in the global ocean, and potentially impacting biogeochemistry (Hawley et al., 2017; Santoro et al., 2019; Francis et al., 2005).

Our analysis allowed to describe the biogeography of 183 functionally unannotated hlePFCs, which might participate in metabolic pathways involved in functional responses to peculiar environmental conditions. They included 24 microbial dark matter hlePFCs, mainly found in Mediterranean samples and related to Candidatus Marinimicrobia MAGs. Our method being applicable to any set of sequences, we predict that an accumulation of similar results on multiple datasets will help identify recurrent unannotated protein clusters linked to specific environmental niches (Wyman et al., 2018). It could further help targeting wet lab studies towards the description of unknown proteins particularly adapted to specific conditions, like subtropical nutrient-poor waters or oxygen minimum zones (Hawley et al., 2017). However, functionally unannotated hlePFCs sometimes contained proteins from only one MAG, and in this case their response to environmental gradients could be a reflection of the global abundance of this MAG instead of a real functional level response. We then advise future wet lab investigations to mainly select PFCs involving proteins from different MAGs. To pave the way for such further analyses, we have provided all nucleotide sequences for each microbial dark matter PFC, as well as the statistics associated with their response to environmental gradients (see Data availability).

4.2.3.4 Towards more global quantitative studies of meta-omics at the function level

Statistical models in this study were based on the abundances of each PFC in 93 metagenomic samples. For each random forest model, 75% of the samples (*i.e.* 70 samples) were used as a training set. Even though each model was run 10 times on 10 distinct training sets, it remains a relatively low amount of samples to do abundance predictions and extrapolations at the global ocean scale (as a way of comparison, 181 samples allowed to predict diatoms abundance from environmental data in a chinese river (Shin et al., 2019)). Hence, machine learning models were not used to provide extrapolated predictions in this study, but to detect protein functional clusters highly linked to environmental gradients and the main drivers of their biogeography. However, as more and more omics datasets are collected in the global ocean (Salazar et al., 2019; Acinas et al., 2019; Vorobev et al., 2020; Planes et al., 2019), we assume that similar approaches could be conducted with much more samples in the near future, which should increase models' performances. By using less than 100 samples, we were nonetheless able to obtain 2,444 models with R^2 values over 0.5. It highlights the potential of such quantitative approaches for predicting the abundance of key protein families in the global ocean. Moreover, our dataset was only composed of metagenomics samples, when it is hypothesized that a big part of prokaryotic communities response to environmental change comes from variations in gene expression (Moran et al., 2013). This assumption was recently disputed (Salazar et al., 2019), but applying our method to metatranscriptomes in the future would allow to use the environmental context to predict protein

expressions instead of metagenome sequence abundances, which could help improve the accuracy of models predictions.

In the future, biogeochemical modeling should benefit from our ability to quantify and predict biological functions using environmental and omics data (Coles et al., 2017; Mock et al., 2016; Stec et al., 2017; Tang and Cassar, 2019). Through our quantitative and data-driven analysis, we have shown one illustration of how metagenomics data can be used without a priori choices of taxon or metabolic function to (1) identify key environmental drivers of planktonic communities functional composition, such as biogeographical provinces, (2) detect potential key protein families and organisms with original biogeographies and (3) investigate the microbial dark matter response to environmental fluctuations. We have then paved the way for more quantitative analysis taking advantage of the richness of global omics datasets, both at the functional and taxonomic level, which should in the long term increase our ability to better predict future global climate.

4.2.4 Methods

4.2.4.1 Samples collection and metagenome-assembled genomes (MAGs)

We focused our study on the 885 non-eukaryotic MAGs made publicly available (Delmont et al., 2018). The whole bioinformatic workflow designed to build these MAGs, as well as all the links leading to the fasta files and Anvi'o profiles for each MAG can be found at `http://merenlab.org/data/tara-oceans-mags/`. These MAGs were built from 93 *Tara Oceans* metagenomes retrieved from 61 surface samples and 32 deep chlorophyll maximum samples collected worldwide in the global ocean, using a size filter targeting free-living microorganisms (0.2-3 µm). Original metagenomes are available under the European Bioinformatics Institute (EBI) repository with project ID ERP001736. To date, the work achieved by Delmont et al. 2018 constitutes the only database of manually refined MAGs constructed using the *Tara Oceans* project data. Automated binning efforts provided larger numbers of MAGs and focused on multiple size fractions (Tully et al., 2018), but are subject to higher binning errors, causing sometimes obvious contigs misplacement (as discussed here: `https://bjtully.github.io/posts/2018/10/re-visiting-tmed-mags/`). Further information on the MAGs' genomic features, such as their completion or GC content, can be found in the supplementary table 5 of Delmont et al. 2018.

4.2.4.2 Gene detection and quantification

Prodigal v2.6.3 (Hyatt et al., 2010) was run to retrieve the nucleotide and protein sequences of each detected gene for each of the 885 MAGs. By concatenation, one nucleotide and one protein fasta files were finally created, containing each in total 1,914,171 sequences. The nucleotide sequences were then used for the mapping and quantification step (hereafter developed) whereas the protein sequences were used for building the sequence similarity network (cf. next paragraph).

The nucleotide file was used as an index to quantify the MAGs' genes abundance in the 93 metagenomes used by Delmont et al. for the MAGs binning process (Delmont et al., 2018). For this,

we mapped metagenome reads to the MAGs gene catalog using the quant function from Salmon v.0.11.3 (Patro et al., 2017) in quasi-mapping mode, with the following parameters '–libType A –meta –incompatPrior 0.0 –seqBias –gcBias –biasSpeedSamp'. To normalize the obtained read counts, we divided them by the gene length, and by the total of sequenced reads per sample, then multiplied them by 10e9. The obtained value is analogous to an RNA-seq transcripts per million value (TPM), except that TPM calculation is based on the total amount of reads that mapped to the transcripts index, while we used here the total amount of reads that has been sequenced in each sample (*e.g.* mapped + unmapped). In fact, the underlying assumption behind TPM and other RNA-seq orientated normalizations is that all compared samples should come from similar tissues, hence displaying a comparable number of mapped reads, which is incompatible with environmental metagenomics. Indeed, *Tara Oceans* samples contain variable quantities of biological matter coming from different sampling in the global ocean, leading them to have very variable amounts of total sequenced and mapped reads. Typically, if a sample has a high total number of sequenced reads but a low number of mapped reads, it will still display high abundance values for the few mapping reads when using the classic TPM normalization, while it would not be the case with our method.

4.2.4.3 Building a Sequence Similarity Network (SSN) from 885 prokaryotic MAGs

A Sequence Similarity Network (SSN) is a graph object in which vertices correspond to sequences and edges represent the similarity and coverage between pairs of sequences (Atkinson et al., 2009; Meng et al., 2018; Lopez et al., 2015; Rizzolo et al., 2019). Diamond v0.8.22 was used in blastp mode to compute the percentage of similarity between every pair of proteins detected in the MAGs, using options '-e 1e-3 -p 30 –sensitive'. A sequence similarity network was built with the diamond output using 80% identity and 80% coverage threshold. This coverage threshold is commonly used in SSN studies (Meng et al., 2018; Lopez et al., 2015; Rizzolo et al., 2019) and we also tested 4 other similarity thresholds: 70%, 75%, 85% and 90%. We selected the intermediary 80% identity threshold to minimize the amount of singletons, while maximizing the functional homogeneity between linked proteins.

4.2.4.4 Extracting, annotating and quantifying protein functional clusters in the Sequence Similarity Network (SSN)

A SSN is made of singletons (vertice or sequence without any homology with other sequences) and connected components (subgraphs composed of at least two vertices disconnected from the rest of the network). In our case, a connected component (CC) corresponds to a group of at least two protein sequences that are linked together (directly or via neighbors), and that have no link with other groups of sequences in the SSN. We assume that the proteins contained in a CC potentially share a similar molecular function (Forster et al., 2015; Meng et al., 2018; Atkinson et al., 2009; Rizzolo et al., 2019). The term "protein family" is often used to describe such clusters of homologous proteins, but as this term is related to the description of evolutionary relationships,

we here prefer the use of protein functional clusters (PFC).

Our SSN was composed of 233,756 protein functional clusters, including 757,457 proteins (*i.e.* 1,156,714 singletons were excluded from the analysis). These proteins were functionally annotated using eggNOG mapper v4.5.1 (Huerta-Cepas et al., 2016, 2017) and KofamScan (Aramaki et al., 2019). EggNOG emapper was run using the diamond mode and the –no_annot flag. It produced a table containing seed orthologous sequences for 677,684 of our proteins (89.5%), the rest of them not being similar enough from any sequence in the eggNOG database. The annotation phase was then launched on these 677,684 proteins, using the seed orthologous sequences table as input to the emapper function, and the –annotate_hits_table flag. We obtain an annotation table with GO IDs, KEGG IDs, and eggNOG descriptions. KoFamScan was launched with default options and -mapper flag. The KEGG API was then used to retrieve KEGG pathways ID and descriptions for each KEGG ID identified by KoFamScan in our protein catalog. To assess for the functional homogeneity in our protein functional clusters, we computed an homogeneity score F_{hom}:
$N_{\text{annot}} > 1 \Rightarrow F_{\text{hom}} = 1 - \frac{N_{\text{annot}}}{N_{\text{prot}}}$
$N_{\text{annot}} == 1 \Rightarrow F_{\text{hom}} = 1$

With N_{annot} the number of unique annotation terms found in the PFC (either KEGG IDs or eggNOG terms), and N_{prot} the number of proteins in the PFC.
As multiple eggNOG terms can exist for similar functions (*e.g.* 'UBA-ThiF-type NAD FAD binding protein' and 'UBA-THIF-type NAD FAD binding'), they can lead to artifactually low homogeneity scores. For this reason, protein functional clusters with low homogeneity scores obtained with the EggNOG database were tagged as poorly homogeneous but were kept in the analysis.
Statistics on functionally unannotated PFCs presented in Table 1 and Table 2 include both (1) query sequences that did not match to any reference in public databases, and (2) query sequences that match to one or multiple references in public databases, but could not yet be associated to any biological function.
To assess taxonomic diversity in our PFCs, we used the taxonomic annotation of the 885 MAGs provided by Delmont et al. 2018. This taxonomic annotation was inferred from 43 single-copy core genes through a combined use of CheckM (Parks et al., 2015), RAST (Aziz et al., 2008) and manual BLAST searches (See Delmont et al. (2018) for further details).
We computed a mean abundance for each protein functional cluster in each of the 93 metagenomes, using relative protein abundances (see Gene detection and quantification). We obtained an abundance table composed of 233,756 rows, corresponding to protein functional clusters, and 93 columns, corresponding to the 93 *Tara Oceans* metagenomes used in the study.

4.2.4.5 Environmental dataset

For each of the 93 *Tara Oceans* metagenomes, we retrieved the environmental context from Faure et al. (`https://figshare.com/articles/Data_MixoBioGeo_Faure_et_al_2018/6715754`) 2019. To complete this environmental dataset, we added 10 climatology variables retrieved from the World Ocean Atlas (Boyer et al., 2018): temperature, salinity, density, conductivity, dissolved oxygen, percent oxygen saturation, apparent oxygen utilization, silicate, phosphate and nitrate. For tem-

perature, salinity and conductivity we retrieved the mean and the mean seasonal anomaly at each sampling point (precision of 1°) over the 2005-2012 period. Only the mean was retrieved for density. For the 6 other variables, we retrieved the mean and the mean seasonal anomaly at each sampling point (precision of 1°) over all available years. In total, we obtained 74 environmental variables, that we reduced to 51 by getting rid of near zero variance variables and too highly correlated ones, using options "nzv" and "corr" from the preProcess function of the caret package (Kuhn, 2008) in R (R Core Team, 2019). A detailed description of these variables is available in Table S2 (Appendix C). We then scaled and centered the 51 selected environmental variables, and used a k-nearest neighbours approach to replace NA values (6.6% of the data) by the mean of the concerned variable in the 5 nearest samples in terms of global environmental profile (knnImpute option from caret's preProcess function (Kuhn, 2008)).

4.2.4.6 Identification of protein functional clusters varying along environmental gradients

Among the 233,756 protein functional clusters, we detected 4,842 (2,1%) clusters with near zero variance using caret preProcess function (Kuhn, 2008), *i.e.* they had less than 10% of abundance values across all samples that were distinct, and a ratio between the most common abundance value and the second most common one that was higher than 95 to 5. These clusters were removed from further statistical analysis. We built a random forest regression model for each of the remaining 228,914 protein functional clusters, using the environmental variables as predictors of cluster relative abundance. To suppress eventual biases linked to over/underfitting due to training set selection, each model was launched 10 times using 10 different training sets built using 75% of the 93 samples available. For each iteration, *i.e.*, for each pair of training set and protein functional cluster, 500 trees were built. For each model, we computed the mean prediction error over the 10 iterations, as well as the mean R^2, and the mean rank of importance in the model for each environmental predictor. The mean R^2 was used to discriminate protein functional clusters following significant environmental gradients from the ones showing no response to the environmental context. Specifically, we considered every model with R^2 values over the arbitrary threshold of 0.5 to be very tightly linked to environmental gradients. Different thresholds ranging from 0.25 to 0.75 were tried, thresholds higher than 0.5 tended to select a few hundreds of PFCs, mainly the ones overabundant in the Mediterranean Sea, while too low thresholds tended to diminish the R^2 value and readability of the canonical correspondence analysis (cf next section). All random forest models were launched using the randomForest R package (Liaw and Wiener, 2002).

4.2.4.7 Biogeography of protein functional clusters (PFCs) linked to environmental gradients

We used a canonical correspondence analysis (CCA) to describe in a more integrated way the relationships between PFCs and environmental variables. The CCA used the relative abundance

table of all PFCs linked to environmental gradients (mean R^2 of random forest regressions > 0.5) as response variables, and 13 selected environmental variables as explanatory variables: biogeographical province, ocean region, season moment (*i.e.* early / middle / late), temperature, depth of the euphotic zone, longitude, maximum sampling depth, optical backscattering coefficient at 470nm, depth of the O2 minimum, calcite saturation state, fluorescence, NO3, chlorophyll A. The 13 environmental variables were selected through a backward and forward stepwise selection based on the AIC criterion (Legendre and Legendre, 1998).
Using positions of PFCs in the two first dimensions of the CCA space (37.65% of variance), we computed a barycenter position for each metabolic pathway detected among PFCs (Figure 4.6). Similarly, we computed barycenters for phyla, classes and genomic assemblies in the CCA space (Figure 4.9, 4.7, 4.8). We also represented pathways distributions along each CCA axis using boxplots, to help identify pathways that were the most characteristic of certain environmental conditions (then represented in Figure 4.6). Finally, convex hulls englobing all PFCs associated to a pathway were drawn for a selection of pathways corresponding to (1) pathways enriched in cold and eutrophic waters, (2) pathways enriched in Mediterranean samples, (3) pathways linked to inter-organisms interactions, (4) pathways associated to biogeochemical functions, and (5) pathways composed of only unknown sequences (Figure 4.7).

Data availability

Instructions on how to build or download the MAGs used in this study are available at `http://merenlab.org/data/tara-oceans-mags/`. All other data used in this study are available at `https://figshare.com/s/b33fc72a62db44b7192f`. All bash and R codes necessary to reproduce our analysis are available at `https://github.com/EmileFaure/MAGsProteinFunctionalClusters`.

Ackowledgements

We would like to particularly thank Loïs Maignien, Olivier Aumont and Eric Pelletier for the insightful discussions concerning this study. We also thank the Meren Lab and all the people involved in *Tara Oceans* for producing the data we used, and making them publicly available. We would also like to thank all participants of the GOBITMAP and GREENOCEAN workshops for their useful advice. Finally, EF would like to thank Raphaël Berthier and Jean-Olivier Irisson for their advice concerning machine learning, as well as François Duchenne, Elise Kerdoncuff, Benoît Pérez and Eric Bapteste for their helpful comments.

Funding

This work was funded mainly by our salary as French State agents and therefore by French taxpayers' taxes. EF acknowledges a 3-year Ph.D. grant from the "Interface Pour le Vivant" (IPV) doctoral program of Sorbonne Universite. SDA acknowledges the CNRS for her two sabbatical years as visiting researcher at ISYEB in 2018-2020. Additional support was provided by the Institut des Sciences du Calcul et des Données (ISCD) of Sorbonne Université through the support of the sponsored junior team FORMAL (From ObseRving to Modeling oceAn Life). EF acknowledges

the financial help of the Korean Institute of Ocean Science and Technology to attend the IMBeR 2019 conference, which led to very helpful discussions concerning this work.

Competing interest

Authors declare no conflict of interest.

Authors contributions

EF, SDA and LB conceived the study. EF processed and analyzed the data, with inputs from SDA and LB. EF, SDA and LB wrote the manuscript.

4.3 Conclusion: bridging the gap between observations and modeling

The approach presented in this chapter and the preliminary results of section 3.2.2 both show promising results of gene or protein functional clusters quantification and prediction using meta-omics and environmental data. While section 3.2.2 focused on the *a priori* selected *dmdA* gene, I presented in this chapter a method which enables to create protein functional clusters without *a priori*, and analyze their biogeography. I showed how such protein functional clusters could be related to functional traits and to taxonomy, the two of them not being completely decoupled by this approach. Indeed, with the similarity and coverage thresholds that I used I showed that metabolic functions could be associated to multiple distinct protein functional clusters, which could correspond to proteins coding for the same function in distinct phylogenetic groups. This specificity of our approach could be seen as a drawback compared to other options like the creation of clusters based uniquely on functional annotation (*e.g.* by Kegg IDs like in Salazar et al. (2019)), which theoretically allow to entirely decouple function from taxonomy (*e.g.* all sequences with the same KEGG ID are found in the same cluster independently from their taxonomy in Salazar et al. (2019)). However, these other approaches (1) do not include functional unknowns (or in separate analysis) and (2) are heavily dependant on the used database, whereas our approach includes unknowns and only depends on the similarity and coverage thresholds used in the sequence similarity network, for which multiple sets of values can easily be compared, producing functional clusters of different granularities. I believe it makes our approach more inclusive, and thus more adapted to answer diverse questions by varying the thresholds of similarity and coverage to push towards a decoupling of taxonomy and function (low thresholds), a coupling of both (high thresholds), or an intermediate solution. In the following general discussion, I will come back on the findings presented in chapters 2, 3 and 4, discussing their relevance for trait-based approaches and particularly trait-based models, discussing their limits and proposing future ways of improvements. I will propose ideas on how to concretely integrate these data-driven approaches in trait-based modeling studies.

Chapter 5

General discussion and perspectives

5.1 Detecting and quantifying functional traits using meta-omics data

5.1.1 Summary of the principal results

Throughout this thesis, I demonstrated that the richness of meta-omics data allows to decipher trait/environment associations, providing new insights on functional traits biogeographies. I achieved this using 3 main methods: the annotation of metabarcoding data to detect *a priori* selected traits in chapter 2, the quantification of *a priori* selected genomic markers in chapter 3, and the quantification of gene functional clusters without any *a priori* choice of function or species in chapter 4. All these methods were applied on data from the *Tara Oceans* project, but remain applicable to any other meta-omics datasets.

In the introduction, I raised 3 scientific issues that served as the guidelines of my thesis work:

1. *Can we use meta-omics data to detect functional traits from which the genomic basis is poorly known?*

   In chapter 2, I showed how metabarcoding data allowed to detect mixotrophs in the global ocean, despite the absence of genomic markers of mixotrophy. The results of this chapter demonstrate that metabarcoding data can help to study the biogeography of any functional trait, as long as a manual or database-based annotation of traits to metabarcodes is possible. However, this approach comes with quantitative limitations inherent to metabarcoding data, on which I will come back later in this discussion (Section 5.1.2.1). This is why in chapter 3, I presented how network-based approaches combined to wet lab investigations might allow to detect genomic markers of functional traits. But the limited results obtained when trying to identify genomic markers of mixotrophy during my thesis combined to the few available in the literature so far tend to show that we might not be ready yet for the quantification of some key complex traits in meta-omics data (See section 5.1.2.2).

2. *Can we use meta-omics data to predict the distribution of functional traits/genes in the environment through statistical modeling?*

   Overall, the results presented in this thesis show meta-omics data as an efficient tool to quantitatively describe the global distribution of plankton functional traits. In Chapter 2, I showed how metabarcoding data could be used to decipher the biogeography of plankton functional traits, by providing the first ever omics-based biogeography of mixotrophic protists. In chapter 3, I showed how machine learning methods could be used to predict the gene abundance, transcript abundance and gene expression of an enzyme with a key biogeochemical role, only using environmental data as predictors. In chapter 4, I showed that similar methods could be applied on hundred thousands of gene functional groups in just a few days of computing, allowing to automatically identify gene functional groups responding tightly to environmental gradients. Around 80% of the gene functional groups were only poorly predictable from the environment, thus this second question remain open for many genes and functions. Still, I provided in this thesis some clear evidence of the possibility to use machine learning to model the response of gene functional groups to environmental fluctuations.

3. *Can the abundance of gene functional groups be quantified in meta-omics data without any a priori choice of focal functions and/or species?*

   One of the main criticisms made to plankton functional types and gene functional groups approaches lies in the necessity to *a priori* define functional types or groups without any guarantee to represent the actual functional diversity of natural planktonic communities. Focusing on the search for specific, pre-defined traits in meta-omics data seemed to reproduce the same flaw. Hence in chapter 4 I presented a network-based approach allowing to define gene functional groups without any *a priori* choice of function or species. Through this approach I showed that it was possible to make gene functional groups emerge from meta-omics samples, and to quantify their abundance in the global ocean.

I will now highlight the remaining challenges regarding the detection and quantification of functional traits in meta-omics data. Then, in the second part of this discussion, I will provide some insights and ideas on how the richness of meta-omics data could be used to inform new kinds of trait-based models.

5.1.2 Main challenges remaining to detect and quantify functional traits in meta-omics data

5.1.2.1 The metabarcoding approach to study functional traits and its limits

Metabarcoding offers a relatively affordable and direct access to lists of present lineages in meta-omics samples, and remain widely used in biodiversity surveys of natural planktonic communities (Santoferrara, 2019). In Chapter 2, I manually annotated the mixotypes of 133 lineages, which allowed to study their distribution in the global ocean. A similar approach was conducted by

Ramond et al. (2018), who annotated a wider set of 30 biological traits to 2,007 taxonomic lineages detected in metabarcoding data obtained from 277 water samples taken off the coastal North-East Atlantic Ocean. This annotation allowed to (i) identify 6 functional groups corresponding to lineages sharing similar traits, and (ii) quantify the functional diversity of coastal protistan communities, identifying it as coupled to taxonomic diversity (Ramond et al., 2019). These results and the ones presented in Chapter 2 both show the annotation of traits to lineages identified in metabarcoding data as a promising tool to move towards more realistic diversity representations in trait-based studies.

The main difficulty in applying this methodology lies in the annotation of potential traits to metabarcodes and/or OTUs, which is time-consuming, and prone to oversights and mistakes. However, this drawback might become less and less problematic in the years to come, as metabarcoding databases such as EukRef (Del Campo et al., 2018) and PR2 (Guillou et al., 2013) start to include potential functional trait information at the species level. The mixotrophy annotation provided in chapter 2 should notably appear in the next PR2 database iteration, and Ramond et al. (2018) recently published a metabarcoding database including the annotation of 30 functional traits to more than 2000 protistan lineages (*i.e.* maximum and minimum size, kleptoplastidy, coloniality, production of DMS,...). Integrating information from the many available trait databases (a complete list of trait databases including traits from marine and freshwater species is available in table 2 from Martini et al., in appendix A), which are mostly not yet related to metabarcodes but only to taxonomic references, should allow for an increase in the amount of annotated traits in metabarcoding databases. Although these trait databases might help to automatize the metabarcode to trait annotation process, they are unfortunately marred with multiple flaws. The first one is a clear bias in species representation, with for example only 9% of all UK demersal (*i.e.* living near or on the bottom) marine fauna annotated with 8 fundamental functional traits in online databases (body size, diet, feeding method, reproductive timing, fecundity, larval dispersal, adult dispersal and longevity), with a clear difference in annotations between fish (median of 6 annotated traits) and invertebrates (median of 1 annotated trait, with body size being the most frequently annotated one) (Tyler et al., 2012). The second main flaw of trait databases lies in the lack of a common vocabulary across the different fields using trait-based approaches, and in the heterogeneity of data formats and units (Schneider et al. (2019), see also paragraph 2.1 and 2.2 in appendix A), making it difficult to compare and use multiple databases. Finally, metadata in trait databases often poorly describe the methods and conditions in which the traits were measured. But all these flaws are starting to be tackled by international initiatives like the open traits network (Gallagher et al., 2019), and some of the major problems in using metabarcoding data to quantify functional traits might rather come from the omics data themselves than from the databases.

The ability of metabarcoding to give a quantitatively accurate view of planktonic community is quite debated (Lamb et al., 2019). In the past ten years, some studies showed strong relationships between biomass and metabarcoding abundance (Hirai et al., 2015; Lindeque et al., 2013), but others highlighted biases due to sequencing errors and to sometimes important rDNA copy number variations across and even within species (*e.g.* radiolarians tend to have very high rDNA copy numbers which could cause them to be overabundant in metabarcoding samples, Biard et al.

(2017); Decelle et al. (2014)). Recently, a meta-analysis assessed the ability of metabarcoding data to quantitatively estimate biomass using the data from 22 different projects, and found the metabarcode abundance to be loosely related to biomass (slope of 0.52 in a linear regression), but with very high variability (+/- 0.34 variance in slope) (Lamb et al., 2019). In addition to this quite poor quantitative representation of diversity, it is important to note that metabarcode abundance can only reflect biomass, and not the realization of functional traits. For example, metabarcoding data can only tell whether lineages known to be capable of mixotrophy are present, but can not give any indication on whether the lineages were feeding through mixotrophy at the moment of sampling.

Other potential biases linked to metabarcoding that are commonly discussed in the literature include false positives (detection of a lineage that was not present in the sample) due to contamination or phylogenetically mixed OTUs, and false negatives (non-detection of a present lineage), due to low abundance, incomplete extractions, primer mismatches or incomplete reference databases (Santoferrara, 2019). Hence, metabarcoding can only give a biased estimate of species biomass, and only for species that are present in metabarcodes databases, while metagenomics and metatranscriptomics offer access to gene and transcript abundance, which are potentially translatable to trait realization (see Chapter 3). However, we have also shown that for complex traits like mixotrophy, the way towards detection of genomic markers could be difficult.

5.1.2.2 Tackling the need for more genomic markers of functional traits

In chapter 3, I showed that enzymes coding for a simple functional trait could be quantified in meta-omics data, and linked to environmental gradients. The *dmdA* enzyme was a good example to display this, as it is very well conserved across organisms, and has a well known role in DMSP catabolism (Curson et al., 2011), giving it the potential to be quantitatively linked to a concrete functional trait. Similarly, the biology of the dinitrogen fixing *nifH* enzyme is well known, and its abundance in meta-omics data was recently used to build a data-driven model of global diazotrophs abundance (Tang and Cassar, 2019). These results illustrate well the benefits of having access to well defined genomic markers for functional traits.

Trait-centred annotation tools

Recently, a new annotation tool called DRAM (Distilled and Refined Anotation of Metabolism) was designed to specifically detect genomic markers of metabolic traits in genomes and metagenomes (Shaffer et al., 2020). Crossing the functional annotations from 5 databases (Figure 5.1), this tool automatically gives access to the list of metabolic traits in a genome or metagenome, along with the FASTA files of the corresponding marker genes (Figure 5.1). The emergence of such trait-centred annotation tools should be of great help for the detection of functional traits in meta-omics data, allowing results from data-driven statistical models, like the one focusing on diazotrophs presented in Tang and Cassar (2019) or the ones presented in Chapter 3, to be obtained on more functional traits.

However, as I have exposed in chapter 3, genomic markers of complex traits are hard to detect

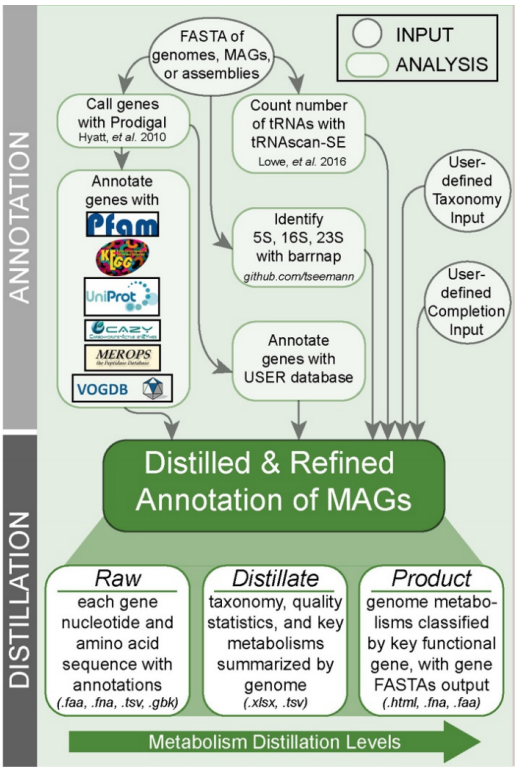

Figure 5.1 - Towards trait-centred annotation algorithms: computational workflow of the DRAM (Distilled and Refined Annotation of Metabolism) algorithm, extracted from Shaffer et al. (2020). The user is asked to provide FASTA files of assembly-derived contigs, either binned into genome-resolved data (e.g. MAGs or isolate genomes) or unbinned (e.g. metagenome contigs). The user can also input taxonomic annotations and genome or MAGs completion files into DRAM, which will then be included in output files. Prodigal (Hyatt et al., 2010) is used to call genes in the input file(s), before 5 databases are used to annotate them. Users can also add any annotation database of their choice. In parallel, tRNA are detected using tRNAscan-SE (Lowe and Chan, 2016), as well as 5S, 16S and 23S rRNA fragments using barrnap (`https://github.com/tseemann/barrnap`). All these data are then compiled into three files: the 'raw' file corresponds to the 'classic' output one would obtain from an annotation software, the 'distillate' file contains pathway-centric annotation and genome statistics, and the 'product' file is an html document indicating metabolic pathway coverage, electron transport chain component completion and presence/absence of specific functions associated to functional traits (e.g. methanogenesis).

and remain under-investigated. They are also hard to quantify considering that multiple loci and genes are involved. Indeed, even if a panel of genes from multiple loci are identified as markers of a specific trait, should we use the mean or median transcript abundance over all loci as a proxy for trait realization? Or maybe build a statistical model to link loci abundance to trait value? These questions remain open, and would need high quantities of data from specifically targeted wet-lab studies in order to be answered.

High-throughput measurements of traits

We thus need to improve our ability to sequence meta-omics samples in controlled environments (*e.g.* microcosms or mesocosms) with high numbers of replicates to obtain accurate quantitative relationships between traits and omics content over time and changing conditions (Faust, 2019). High-throughput cultivation, *i.e.* simultaneous and automated cultivation of tens of microbial communities through minibioreactor arrays (Auchtung et al., 2015) or microfluidic flow

cells (Pousti et al., 2018), was successfully applied on mock communities of prokaryotes from the human gut, leading to accurate, data driven models of microbial interactions (Venturelli et al., 2018; Hekstra and Leibler, 2012). In parallel, recent progresses in microfluidic greatly improved our capacity to sort cells from aqueous environment based on their physiological properties, leading to the emergence of omics datasets sequenced from targeted single planktonic cells sorted *in-situ* (Needham et al., 2019). Through high-throughput, in-lab sorting, it is possible to sort 200 to 500 cells per hour according to their functional properties based on the use of isotopic labels on selected elements (Lee et al., 2019). Such high-throughput cell sorting techniques, associated with high-throughput cultivation and measurement of trait values should improve our ability to build omics datasets with high number of replicates, allowing to decipher the genomic basis of multiple complex functional traits such as mixotrophy, body size, body shape, ability to remineralize organic matter or coloniality. But such approaches will remain biased by the lack of inclusion of yet unculturable organisms.

What about unculturable organisms?

Usually between 40 and 60% of sequences from metagenomes can not be linked to a function or species, mostly issued from unculturable organisms (Bernard et al., 2018; Vanni et al., 2020). In chapter 4, I did not discard the unknown sequences, and even particularly focused my attention on them, using a sequence similarity network to group them with known and other unknown sequences into protein functional clusters. Sequence similarity networks were used to directly infer functions of unknown sequences, but it now appears evident that more will be needed to decipher the globality of the dark side of omics (Arnold, 2018; Vanni et al., 2020). Very recently, a new bioinformatic tool was proposed with the aim to build a link between known and unknown genes from meta-omics data (Figure 5.2, Vanni et al. (2020)). Their approach combines gene clustering based on similarity and the use of metadata like co-expression, phylogenetic relatedness, biogeographical distribution or wet-lab experimental knowledge (Figure 5.2). In 1,749 microbial metagenomes from marine and human environment where only 44% of the genes could be annotated through Pfam, AGNOSTOS was able to annotate 70% of them, while identifying key lineages responsible for most of the remaining unknown part of genes, *e.g.* the newly described *Cand. Riflebacteria* and *Cand. Patescibacteria* (Vanni et al., 2020). This shows how the accumulation of data and observations on cultivated organisms and uncultivated organisms can be combined into frameworks allowing to make sens of the unknown, and ultimately get a better grasp of planktonic functional diversity.

We might not be ready yet for the use of omics data to implement complex traits in mechanistically realist biogeochemical models, or even to make accurate quantitative predictions of complex traits realization. But the results of this thesis and the promising works that I have discussed in this section emphasize the important progresses made in this domain in the past decade. Putting a strong focus on the biology and genomics of complex traits, which should also help to better understand the dark side of omics, appears essential to better understand and model functional diversity. In parallel, we must bridge the gap between observations and biogeochemical models, to be able to better integrate our knowledge on diversity into modeling frames.

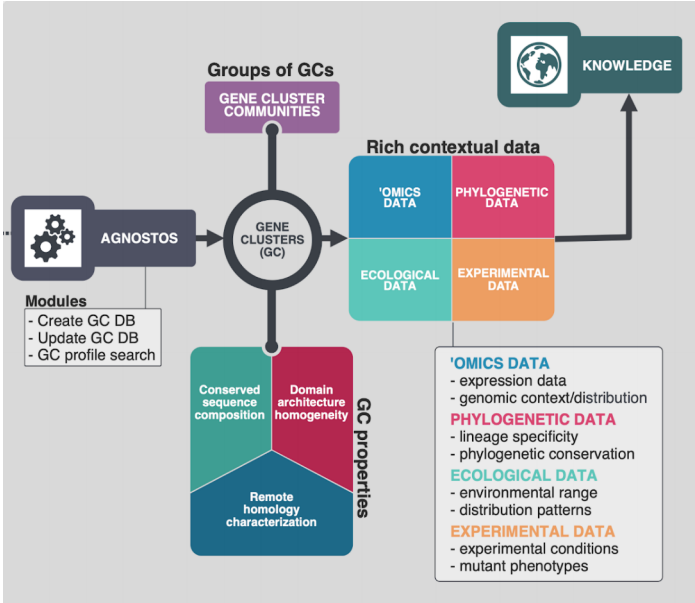

Figure 5.2 - Computational workflow of AGNOSTOS, a tool for shedding light on the dark side of omics, from (Vanni et al., 2020). The idea is to bridge the known coding sequences with unknown coding sequences using sequence composition, domain architecture and remote homology (i.e. indirect homology) to build a database (DB) of gene clusters (GC). The clusters are then complemented by meta-data such as levels of expression in metatranscriptomes, abundance in metagenomes, phylogenetic origin or environmental range to diffuse (or not) functional and/or taxonomic annotations from known sequences to unknown sequences in each cluster. Overall, the pipeline uses as many information as we have on the known coding sequence space to draw information on the unknown coding sequence space.

5.2 How to bridge the gap between observations and biogeochemical models ?

In the introduction of this thesis, I highlighted the gap between observed (*i.e.* measured *in-situ*) and modeled functional diversity of plankton communities and its potential impacts on model predictions. In this section, I will present a few leads on how to bridge this gap using meta-omics data.

5.2.1 Using omics data to validate trait-based models

Trait-based models are hard to confront with most types of *in-situ* observations, as traits transcend taxonomy to focus on function, which is hard to measure at community-level. New types of data such as meta-omics, but also high-throughput imaging and satellite imagery now offer chances to make in-situ measures at the trait-level in planktonic communities (*e.g.* *dmdA* measures of Chapter 3, or automated detection of planktonic organisms attributes through machine learning treatment of in-situ images (Luo et al., 2018)). But most of these data come from cruises in which samples are taken in 'snapshots', *i.e.* at one point in space and time, while global biogeochemical models are usually run over at the scale of decades (*e.g.* 20 years runs for Coles et al. (2017), 10 years runs for Follows et al. (2007)). This discrepancy between the time scales of observations

and models makes it hard to compare direct measures with model outputs. Coles et al. (2017) managed to do it, taking advantage of using gene functional groups as model agents to compare the abundance of two well known genes in their model and in meta-omics samples. But the approach remained limited in the sense that the model had for goal to give a broad overview of prokaryotic plankton functional diversity over the whole Amazon river plume region, while the scarce number of snapshot meta-omics samples available for model validation only covered a very small spatio-temporal range (each sample corresponding to a maximal temporal range of a few days spent in a space of a few square kilometers). Hence, it appears to me that the best way to validate model outputs with omics data would be to derive general statistical relationships from observations, similar to general ecological laws, that could then be used as guidelines to check model outputs, instead of 'point to point' comparisons.

Chapters 3 and 4 illustrate well how the richness of meta-omics datasets can allow to derive strong statistical relationships from observations at global scale. Such strong relationships, whether it be between markers of different traits or between markers of a functional trait and the environment, offer a chance to better validate model outputs with observations. Indeed, the verification of the observed relationship in the model could serve as a kind of validation. To go further, we could imagine using correlative models based on these relationships, inspired from the niche modelling approach, that could give access to global maps of traits distribution on spatio-temporal scales similar to classic biogeochemical models (Figure 5.3). For example, the study by Tang and Cassar (2019) provide such a data-driven global scale model of diazotrophs abundance, and could help to validate the outputs of biogeochemical models including diazotrophs, and even help to identify the zones of good and poor performances from both models.

5.2.2 Using omics data to model individual functional traits

Trait-based models rely on the description of trade-offs between functional traits, that define the parameters of individual-level processes such as mortality, or the acquisition and allocation of resources (Kiørboe et al., 2018). Such trade-offs are often defined empirically, thus on limited numbers of species, and would benefit from more mechanistic descriptions, allowing for a more general applicability (Kiørboe et al., 2018). Omics data could help to better define trade-offs in two ways: (i) by providing more general empirical evidence of trade-offs than wet-lab experiments through global scale sampling, and (ii) by allowing to compute metabolic networks leading to the identification of trade-offs in a mechanistic manner, through metabolic modeling.

In this thesis, I mainly focused on building statistical relationships between functional marker genes and the environment, but similar approaches can be used to decipher links between genes responsible for different functions. For example, Salazar et al. (2019) identified *dmdA* transcript abundance to be significantly anti-correlated to *cysN* and *cysD* transcript abundances, two enzymes involved in the assimilatory sulfate reduction pathway. This supposes a trade-off between the use of sulfate and DMSP as a source for sulfur in prokaryotes (Salazar et al., 2019). Hence, when marker genes are available for different traits, meta-omics data offer the chance to test for trade-offs between them using publicly available, large scale datasets.

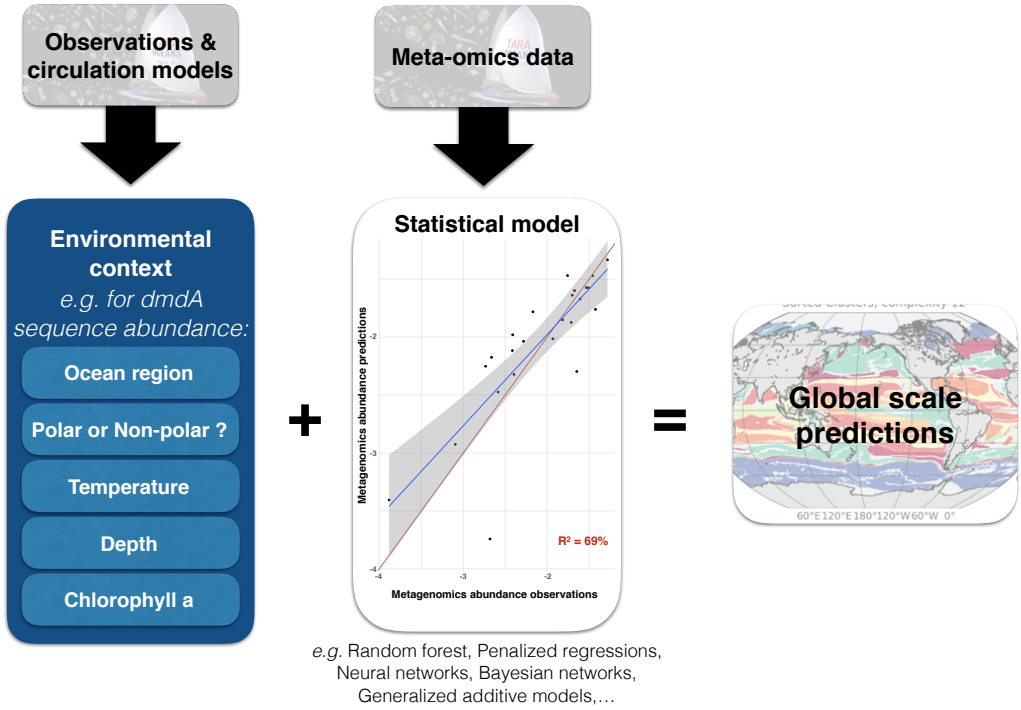

Figure 5.3 - Concept of correlative models, illustrated with the dmdA metagenomics sequence abundance predictions obtained in chapter 3. Observations from oceanographic cruises, satellite imaging, remotely operated underwater vehicles, and global circulation models can be used to obtain global maps of environmental variables. These data can then be used as predictors of sequence abundance for specific genes or organisms through machine learning algorithms or bayesian network approaches. The statistical model used as an illustration here was presented in chapter 3, and allowed to predict dmdA sequence abundance in metagenomics samples using 5 variables: Ocean region, is the sample from polar or non polar area, temperature, depth and chlorophyll a. chapter The illustration used for global scale predictions was taken from Sonnewald et al. (2020).

Models using genome-scale metabolic reconstruction include flux-balance analysis models (Steuer et al., 2012; Budinich et al., 2017), as presented in the introduction, and more recently a new generation of resource allocation models (Reimers et al., 2017; Sharma and Steuer, 2019). These two types of models are similar in the sense that they are based on genome-scale reconstruction of metabolic networks, and the evolutionary optimality principle, *i.e.* the assumption that all metabolic fluxes are set to obtain the maximal growth rate, or maximal fitness (Sharma and Steuer, 2019). Resource allocation models are based on the idea that for a cell in a given environmental condition, the acquired resources can be allocated either to non-enzymatic components, or to the translation of catalysing enzymes, which will define the rates of metabolic reactions in the cell (Sharma and Steuer, 2019). All reactions modeled in the cell are then bounded by the availability of the corresponding catalyzing enzyme, which is defined by the amount of resources spent in the translation of the enzyme. The model can then be optimized in order to select the resource allocation option that leads to the maximal growth rate (Reimers et al., 2017; Sharma and Steuer, 2019). The modeled cells can adopt different allocation strategies depending on the environmental conditions, and such strategies 'automatically' emerge from the model structure. Resource allocation models can thus predict the physiological behaviour of bacterial cells in dif-

ferent environments (Sharma and Steuer, 2019), in the presence of multiple competing nutrients (Wang et al., 2019b), and model outputs allow to identify intra-cellular trade-offs leading to the choice of one strategy or the other (Sharma and Steuer, 2019; Reimers et al., 2017).

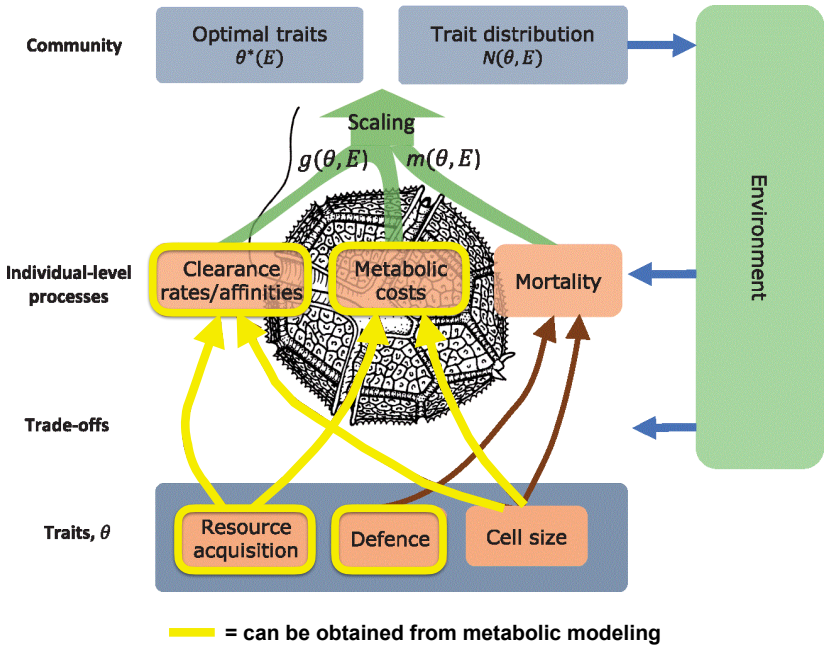

Figure 5.4 - The principles of trait-based approaches, and what metabolic modeling can bring to the table, modified from Kiørboe et al. (2018). The figure depicts how individual-level processes can be derived from traits through the definition of trade-offs between traits. Large scale community-level functional attributes can then be defined as a function of the available traits and the environmental context. Yellow boxes and arrows were added to the original figure to highlight where metabolic modeling could be of help. Please note that the modifications made on this figure only apply to well known unicellular organisms, at least for now (see main text). Metabolic modeling allows to detect resource acquisition traits from genomic content, and to derive all related metabolic costs and affinities through mechanistically realist modelling of intra-cellular trade-offs. Most defence traits, like the production of toxin, or the building of siliceous or calcareous shells can be derived from genomic contents. However, their influence on mortality does not depend exclusively on metabolism, but mainly on the impact of such defence strategies on ecological interactions like predation, which can hardly be derived from metabolic modeling. In most metabolic models, cell size and density are fixed, thus cell size was not highlighted as derivable from metabolic modeling here. Nevertheless, the trade offs linked with the impact of different cell sizes on metabolic costs and nutrient affinities can already be explored through metabolic modeling. Metabolic modeling could then help to define a data-driven set of optimal traits, and the corresponding trait distribution.

Kiørboe et al. (2018) proposed that three categories of functional traits were essential to any trait-based approach: resource acquisition traits, defense traits, and size-related traits (Figure 5.4). Figure 5.4 summarizes what metabolic models, *i.e.* models based on genome-scale metabolic reconstruction, including flux-balance analysis and resource allocation models, can bring to such classic trade-based approaches. But as stated in the introduction, such models are heavily limited by data availability. For example, the biochemical resource allocation model approach presented by Sharma and Steuer (2019) needs not only the list and nucleotidic composition of each catalyzing enzyme of the modeled organism, but also quantitative information on their activity. While the

authors advocate that "reasonable estimates for all required parameters exist" (Sharma and Steuer, 2019), it might not be the case for unicellular eukaryotes, and it is obviously not the case for multicellular organisms, for which these types of approaches have not yet been adapted. It then appears necessary to focus on better understanding the genomic and metabolic basis of functional diversity for these promising modeling approaches to become more widely applicable, through approaches like the ones presented in chapter 3 and 4, or the creation and general use of tools like AGNOSTOS (Vanni et al., 2020).

I have discussed here the potential contributions of omics data to the modeling of functional traits. But as shown in the introduction, most of the global biogeochemical models including explicit representations of planktonic diversity rely on functional entities such as plankton functional types or gene functional groups. In the next paragraph, I will discuss how omics and meta-omics data might be used to better represent plankton functional diversity in such models.

5.2.3 Using omics data to improve plankton functional diversity representation in models

Until now, plankton functional type (PFT) models have always been built through *a priori* selections of the model agents, often with a strong focus on their biogeochemical impacts, rather than their biological and ecological attributes (Flynn et al., 2015). The gene functional groups approach allowed to bypass the association that often remain between PFTs and taxonomy, but remain for now based on an *a priori* selection of the modeled gene groups (Coles et al., 2017). In theory, the advent of omics data offer ways of including a more realistic view of plankton functional diversity in ecosystem models, through the detection of *in-situ* functional traits and metabolic functions (Chapter 2 and 3) or even gene functional groups (Chapter 4), but also through the automatic biogeochemical characterization of genomes (Shaffer et al., 2020), and the functional characterization of uncultivated fractions of planktonic communities (Vanni et al., 2020). It potentially offers the opportunity to build plankton ecosystem models with data-driven definition of agents. But increasing diversity and agent number rhymes with increasing model complexity, and trait-based approaches were introduced to reduce complexity compared to species-based approaches, when the original PFT models with 10 to 15 PFT were already deemed as too complex (Frede Thingstad et al., 2010). Here I will discuss some ideas on how omics data could be used in a reasonable way to better represent planktonic diversity within biogeochemical models.

In 'classic' PFT models, each PFT is represented by a single differential equation. This equation represent the sum of source terms (*e.g.* reproduction, growth), minus the sum of sink terms (*e.g.* mortality, predation). Such equations include the definition of a growth rate, often dependent on parameters like maximum growth rate, temperature dependence of growth, and for autotrophs light dependence of growth (Aumont et al., 2003). Such parameters are derived from wet-lab experiments based on one or a few species deemed as representative of the concerned PFT (Le Quéré et al., 2005). Hence, when a proposition of adding more diversity (*e.g.* more variables, *i.e.* more PFTs) into biogeochemical models is made, the parameterization is automatically raised

as problematic. In parallel, the fact that parameters in PFT models remain quite heavily biased towards culturable organisms is quite problematic as well, and the idea of generalizing such parameters across a whole group of taxonomically diverse organisms (*e.g.* the proto-zooplankton PFT in Le Quéré et al. (2005), which notably includes cilliates and dinoflagellates) is biologically doubtful (Flynn et al., 2015).

Deriving parameters from genome-scale metabolic networks

Already evoked in the precedent section, genome-scale reconstruction of metabolic networks and their associated modeling approaches offer the possibility to derive environment dependent metabolic rates directly from omics data (Budinich et al., 2017; Steuer et al., 2012; Sharma and Steuer, 2019; Reimers et al., 2017). Hence, in a more classic modeling context, relying on metabolic modeling for the parameterization of growth rates, nutrient exchange rates and biogeochemistry-related rates seems possible. Of course, it would probably only be applicable on prokaryotes and simple traits for now (*e.g.* DMS production or diazotrophy, See precedent section). Still, the advent of MAGs theoretically allows to apply genome-scale reconstruction of metabolic networks on a great diversity of plankton organisms, and the combination of such networks with innovative annotation tools like DRAM (Shaffer et al., 2020) or AGNOSTOS (Vanni et al., 2020) could be used in conjunction to automatically derive the sets of parameters representing the physiology of a very wide diversity of organisms. In fact, considering the poor representation of prokaryotes in many PFT models (*e.g.* no prokaryotes explicitly included in Aumont et al. (2003), where the remineralization rate varies only with depth), even a small step towards more data-driven prokaryotic diversity representation could be significant (Coles and Hood, 2016), notably concerning remineralization estimates (Miki et al., 2008).

Using network-based approaches to define ecological interactions

PFT models also include parameters and functional response curves (*e.g.*, Holling type (Holling, 1959; Gentleman et al., 2003)) representing ecological interactions between the different functional types, mostly focusing on predation in current models, but that could ideally include more complex interactions like parasitism and symbiosis as well (Worden et al., 2015). Reconstructions of food webs in models depend mostly on parameters such as maximum grazing rates and preference for different kinds and sizes of preys, which are hard to measure and thus often broadly estimated from wet-lab experiments and allometric relationships (Aumont et al., 2003; Le Quéré et al., 2005). The definition of such parameters can become particularly problematic when the number of PFTs increase. One way to limit this problem is the use of allometric relationships to define size-dependant parameters (*e.g.* Ward et al. (2012)). However, such models still rely on the classical PFT approach.

Meta-omics data could help to resolve this issue through the use of co-occurrence networks and machine learning. Co-occurrence networks are graphs in which nodes correspond to biological entities (*e.g.* lineages, OTUs or amplicon sequence variants), and links to a measure reflecting the co-occurrence between them (Figure 5.5). The construction of such networks in the light of the environmental context can lead to the identification of significant positive and negative associations between biological identities, which can be interpreted as ecological interactions (Faust and

Raes, 2012; Lima-Mendez et al., 2015; Berry and Widder, 2014). For example, a co-occurrence network based on metabarcoding and metagenomics data from *Tara Oceans* allowed to draw hypothetic ecological interactions between more than 9,000 taxa, including some unknown symbiotic relations which were validated in parallel by imagery data (Lima-Mendez et al., 2015). But interpretations of interactions are not straight forward, and a positive interaction could be interpreted as cross-feeding, co-aggregation in biofilms or niche overlap among other possibilities, while a negative interaction could be explained by a predator-prey relationship, but also amensalism, competition, *etc.* (Faust and Raes, 2012).

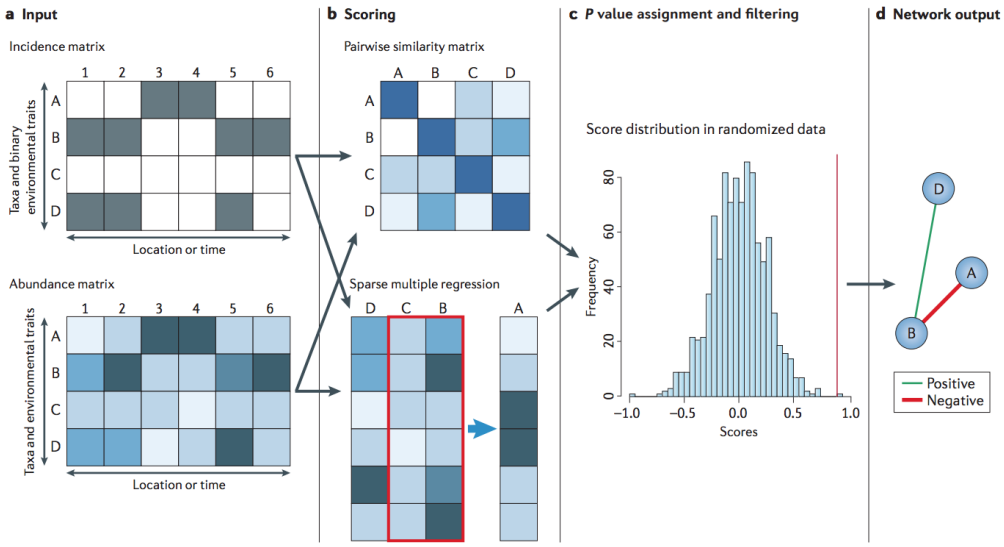

Figure 5.5 - Co-occurrence networks to decipher ecological interactions in microbial populations, extracted from Faust and Raes (2012). (a) Co-occurence networks are built from presence/absence or abundance matrices, with biological entities in rows, and samples in columns. Adding environmental factors in rows along with biological identities allows to better distinguish in the network if interactions are due to similar responses to environmental gradients, or to 'true' direct or indirect ecological interactions. (b) From the incidence or abundance matrix, it is possible to base the construction of the network on a matrix of similarity or dissimilarity, but the use of multiple regressions is preferable as it allows to directly detect associations between more than two biological entities at the time. Cross-validated penalized regressions (similar to the ones conducted on dmdA data in Chapter 3) are often preferred to classic regressions to avoid over-fitting. Species with the highest co-occurrence are given the highest score. (c) The scoring step is repeated multiple times on randomized data, to obtain a Gaussian of score distribution. It allows to compute for each relationship the probability of obtaining a score equal or higher to the observed one by chance, or p-value. (d) Pairs of biological entities with p-value inferior to a selected threshold are drawn as links to form the co-occurrence network. Link width and color can reflect the strength and sign of the relationship.

Overall, the inference of ecological interactions from co-occurrence networks is based on the hypothesis that ecological interactions are the first driver of species occurrences, which is of course debatable (Faust and Raes, 2012; Freilich et al., 2018). Hence, multiple studies aimed at testing the ability of co-occurrence networks to detect real non-trophic and trophic interactions. Berry and Widder (2014) used a simulated microbial community to build a co-occurrence network, and showed that it was able to retrieve ecological interactions when following some guide rules. These rules included for instance: the removal of infrequent taxa, the inclusion of more than 25 samples (and as many as possible), only coming from similar environments (at least from the same

ecosystem type), or the use of corrections on relative abundances to avoid apparent correlations. One important flaw of this study is that they assumed all species of the simulated community to be efficiently sampled and quantified, which is probably often not the case in 'real' data. Freilich et al. (2018) used a different approach, as they relied on a heavily-studied ecological interaction network derived from long-term observations as reference, and tested the ability of co-occurrence data to retrieve the full network. They found a very poor overlap between links of the 'real' network (from which the quality and completeness is difficult to assess) and the one derived from the co-occurrence network (Freilich et al., 2018). However, their construction of the co-occurrence network appeared quite poor: they did not remove infrequent species, used only presence/absence instead of absolute or even corrected relative abundance, and most importantly included only the tide level as an environmental factor despite using samples from 49 different sites, sampled unevenly and possibly in different seasons (not precised in the manuscript) during a 15 years period (Freilich et al., 2018). It is then still difficult to estimate whether co-occurrence networks can or can not reflect ecological interactions. But here again, genome-scale reconstruction of metabolic networks might be of help. Indeed, Freilich et al. (2011)[1] demonstrated that metabolic models offered the opportunity to predict competition and cooperation between prokaryotes. They used 118 genome-scale metabolic networks, and predicted the metabolic interactions for each of the 6903 corresponding species pairs based on their metabolic needs, using methods described in (Stolyar et al., 2007). Thus, co-occurrence networks ability to retrieve ecological interactions could be tremendously increased by using full genomes or MAGs abundance instead of metabarcoding data or presence/absence. It would allow to integrate information from metabolic modeling and full genome trait-centered annotations into the classic network inference frameworks.

How many plankton groups should be included ?

Meta-omics data could then improve planktonic diversity representation into models by providing the ability to make data-driven choices of model actors, and by facilitating the parameterization steps. But even in a case where the parameterization of a 'realistic' diversity of PFTs would not be problematic, the model complexity would cause (i) computational power issues and (ii) difficulties to understand the model behaviour, making it difficult to use the model as a tool for answering ecological questions. It then appears absolutely necessary for any biogeochemical model including a data-driven description of planktonic diversity to include an aggregation step, like illustrated in Figure 5.6. The idea would be to start with complete genomes, MAGs or MGTs, and combine information from annotation tools, metabolic models, co-occurence networks and omics-based biogeographies to detect genomes homogeneous in function and sharing similar niches, that could then be aggregated into functional types. Biochemical resource allocation models allow to aggregate metabolic networks into simplified, coarse-grained networks, which could help to obtain parameters at the aggregated functional group level (Sharma and Steuer, 2019). Algorithms to automatically reduce metabolic networks to coarse-grained models also exist (Erdrich et al., 2015). Of course, the method for aggregating the functional groups based on the blending of such different types of information would not be straight forward. I would advocate

[1] Shiri Freilich, first author of Freilich et al. (2011), and Mara Freilich, first author of Freilich et al. (2018) are different persons.

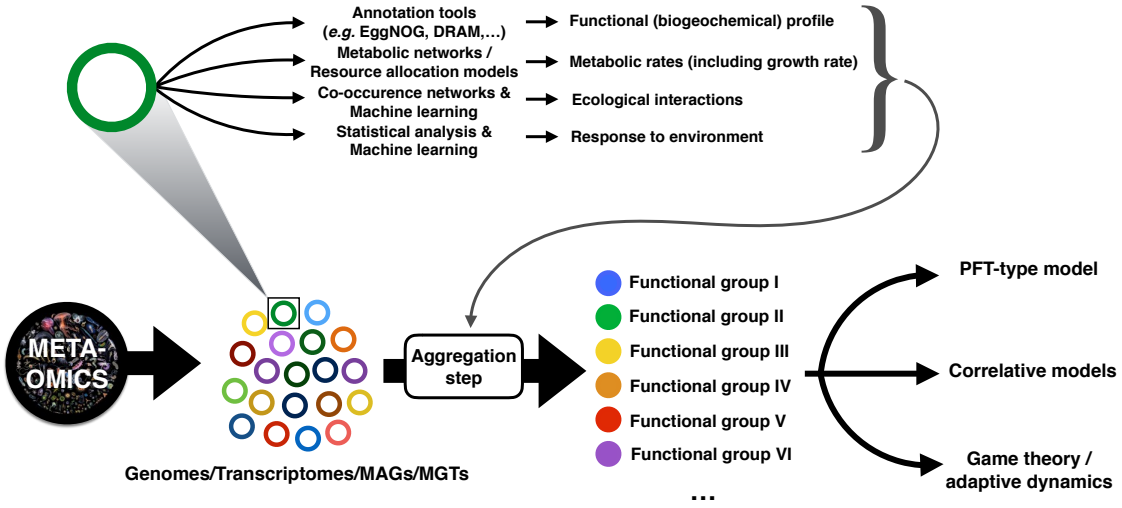

Figure 5.6 - Conceptual scheme of a solution for integrating multiple types of data and models to improve the representation of planktonic diversity in models. This scheme focuses on near complete genomes and transcriptomes, which can be obtained from culture but also from meta-omics data for metagenome-assembled genomes (MAGs) and metagenomics-based transcriptomes (MGTs). Each genome and/or transcriptome can be functionally annotated, notably via tools specifically targeting biogeochemical functions like the recent DRAM algorithm (Shaffer et al., 2020). Metabolic rates can be extracted from them through metabolic network reconstruction, or resource allocation models like the biochemical resource allocation model proposed by Sharma and Steuer (2019). Co-occurrence networks can be used on the genomes and transcriptomes abundance to determine potential antagonistic or mutualistic interactions between them (Lima-Mendez et al., 2015). Finally, I demonstrated in this thesis how statistical analyses and machine learning could help to decipher the functional and taxonomic response to environmental gradients. This information could serve as a way to aggregate genomes and transcriptomes that are sharing similar metabolims, ecological interactions and response to environment through multivariate network analysis (see main text). To my knowledge, this has not yet been done. Such aggregated genomes and transcriptomes would be analog to functional groups, and could be integrated into different types of models, PFT like models but also correlative models like presented in figure 5.7 (replacing gene clusters), and models involving adaptive dynamics to simulate evolutive adaptation.

for the use of hierarchical classifiers, unsupervised clustering and similarity networks for each separated type of information. For example, functional similarity scores could be computed based on the number of complete KEGG pathways shared between MAGs, and these scores could be used to compute a functional similarity network of MAGs from which functional clusters could be derived (using connected components as in Chapter 4, or using Louvain clustering for example). In parallel, clusters of MAGs with similar response to the environmental context could be obtained from an unsupervised clustering of their coordinates in an RDA multidimensional space. This step could be followed by a comparison of the genome clusters found for each information type, with the aim to detect MAGs that appear in the same cluster for multiple information types. Ideally, depending on the thresholds applied during the aggregation step, the model obtained could have a more or less coarse representation of diversity, enabling to answer a variety of different ecological questions, but always with a data-driven start.

Challenges and limits

The modeling approach proposed in this section would have flaws like the heavy computational power needed, the probably heavy dependence on the data used as inputs, on the tools used for the pre-aggregation steps, and on the methods used for the aggregation step itself. As discussed in the precedent paragraph, the detection of ecological interactions through co-occurrence networks is still heavily debated and should definitely be improved. Also, after discussing with experts of metabolic network modeling about the current state of the field, it appears that such an approach could maybe be applicable on prokaryotes, focusing on well known metabolic reactions, but probably not yet on eukaryotes. Considering the recent advances presented in this discussion and the only recent first applications of systems biology approaches in planktonic ecology, I believe that we can hope for important progresses in the domains of genome-scale metabolic modeling and network-based community ecology in the years to come, which will be key for integrating realistic planktonic diversity into biogeochemical models. But integrating such a realistic diversity in 'classic' PFT modeling frameworks would not allow to take full advantage of the available data, as it would lead to a 'fixed' representation of biodiversity, not including any representation of acclimation or evolutive adaptation in biogeochemical models (Flynn et al., 2015). Considering that many biogeochemical models are used over long time scale for climate predictions (Ciais et al., 2013), the integration of multiple modeling types seems essential for a better understanding of the role of microbial diversity in future oceans (Song et al., 2014; Flynn et al., 2015; D'Alelio et al., 2019).

5.2.4 Integrating multiple types of models to understand the role of microbial diversity in biogeochemical cycles

The idea of integrating multiple types of modeling was proposed in most review papers focusing on the future of biogeochemical models (Song et al., 2014; Flynn et al., 2015; D'Alelio et al., 2019; Allen and Polimene, 2011). The four most frequently evoked types of models are: trait-based models, Lotka-Volterra derived models (PFT models being mixed between trait-based and generalized Lotka-Volterra models), correlative/regression-based/species distribution models, and game theory/adaptive dynamics models. The results presented in this thesis notably pushes towards the creation of biogeochemical models inspired from correlative models (Figure 5.7).

Chapters 3 and 4 focused on the predictability of functional gene clusters abundance from the environmental context, providing some encouraging results. Tang and Cassar (2019) provided other compelling evidence of the possibility to use omics and environmental data to predict the abundance of functional gene clusters. What lacks from both this study and my results is a way to quantitatively link functional genes to biogeochemically meaningful trait realization measures. A recent study focusing on plant traits was able to achieve this through a bayesian network approach (Guadagno et al., 2020). Their approach relied on transcriptomic data from controlled experiments that were coupled to measures of three traits in specimens of the globally cultivated crop *Brassica rapa*: assimilation rate, photosystem efficiency and stomatal conductance. Traits measures and leaf tissue samples for transcriptomic analysis were taken every four hours during

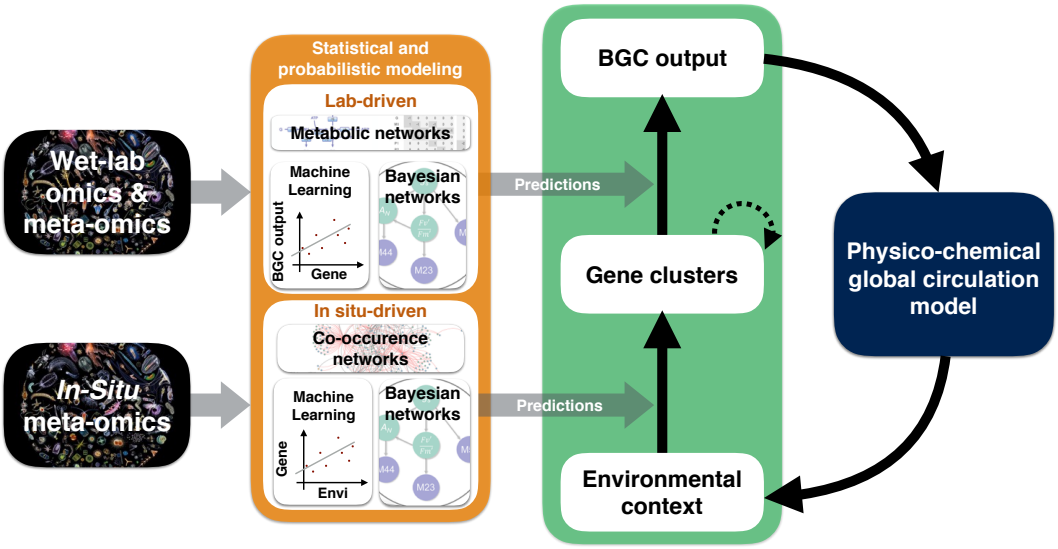

Figure 5.7 - Linking environmental context to biogeochemical (BGC) outputs through gene predictions. In-situ meta-omics and environmental data allow to link the environmental context with genomic and/or transcriptomic abundance. Here I propose to treat these abundance at the functional gene cluster level but it could also be possible to use non-clustered sequence abundance. Wet lab experiments allow to link genomic and transcriptomic abundances to biogeochemical fluxes (BGC output). The dotted arrow indicates the potential use of genomic and transcriptomic abundance of some functional gene clusters to predict the one of other related clusters in the case of clusters with significantly correlated or anti-correlated abundances. This could improve predictions while including some kind of ecological interactions in the model (e.g. gene clusters that come from antagonistic species should be anti-correlated in co-occurrence networks, and would not be predicted in high proportions in the same samples). The environmental context could be obtained from a general circulation model (cf figure 5.3), and the biogeochemical outputs could in turn influence elemental fluxes in the model, creating a correlative biogeochemical model. Illustrations were taken from Guadagno et al. (2020), Guidi et al. (2016) and (Steuer et al., 2012).

a period of 48h, sequencing was achieved using Illumina HiSeq and reads were aligned to a previsouly published reference genome of *B. rapa* (Wang et al., 2011). These coupled measures were used to build a bayesian network, *i.e.* a graph where nodes are variables (here including both trait measures and transcript abundances), and links are their conditional dependencies. Such a network then describes the probabilistic relationships between transcript abundances and realized traits. The network built in Guadagno et al. (2020) was able to predict trait realization from the abundance of transcripts, while capturing the uncertainty of both the biological system and the measures. Such an approach could be used on planktonic organisms, especially with the recent progresses in microfluidics and high-throughput cultivation (Needham et al., 2019; Lee et al., 2019; Faust, 2019). The combination of such an approach with metabolic networks modeling, machine learning, and the methods used in this thesis could allow for the construction of probabilistic biogeochemical models including a biological layer (*e.g.* the *Gene clusters* layer in Figure 5.7). This biological layer could be removed to directly try to predict the trait realization from the environment, as in Tang et al. (2019), where they attempted to predict nitrogen fixation rate from environmental conditions without predicting diazotrophs abundance. But the presence of a biological layer allows (i) to confront model outputs to observations from meta-omics datasets in order to control model performance, and (ii) to take into account the positive and negative

correlations between gene clusters to predict the trait output, which allows to include some kind of biological feedback in the model and could greatly improve its performance.

Such models would be reasonable in terms of computational power, especially compared to models based on hundreds of differential equations. They would allow to take advantage of all the richness of the data available, as there is no need to pre-select markers or to use aggregation steps. In fact, such a method could even include gene clusters of unknown functions, if they appear as statistically linked to a trait and/or environmental conditions. Genomic markers of both simple and complex traits might then emerge from the construction of such models. Finally, this approach would allow to take into account the effects of environmental variations on trait realization, and to estimate the uncertainty of model outputs at every step, which is not doable with a classic 'PFT' model. This approach would then be quite complementary to more classic trait-based and PFT approaches.

An important drawback of such an approach resides in the absence of explicit representations of ecological interactions such as grazing, and in the absence of representation of population stocks, which can be important for decision makers. Also, many technical challenges regarding the construction of such models remain. Notably, in order to derive trait values from transcriptomic composition on a significant set of different traits, we would need large amounts of data from specifically designed wet-labs experiments, which do not exist for now. Our results of Chapter 4 also suppose that some functional genes might not be well predicted by the environment, at least through the metagenomics data that we used. More studies using similar approaches should then be conducted before focusing our efforts on the actual modeling process.

Among the four types of models that I evoked at the beginning of this section, adaptive dynamics models are the only ones that I have not yet discussed. Multiple clear evidences exist of rapid adaptation (*e.g. inter-generational change (evolution) to inherited traits involving changes to the DNA sequence*; Flynn et al. (2015)) of planktonic organisms to environmental changes, and the inclusion of this strong adaptive ability in climatic models is a key challenge (Mock et al., 2017; Ward et al., 2019; Cermeño et al., 2016; Sauterey et al., 2015). As for now, state-of-the-art global biogeochemical models do not take into account the adaptive potential of planktonic organisms (Flynn et al., 2015; Ward et al., 2019). Indeed, despite existing efforts to account for acclimation, or *reversible intra-generational change through changes in expression of inherited traits* (Flynn et al., 2015), such as the use of varying stoichiometric ratios to simulate photo-acclimation (Ayata et al., 2013), very few models manage to include proper adaptation in their formulation. This is notably due to the fact that most biogeochemical models do not model individuals, and so the inheritance of traits is not explicitly modeled (Flynn et al., 2015). Some methods manage to bypass that, notably by allowing for the variation of parameters to maximize growth rate in response to environmental changes (Toseland et al., 2013), but such approaches allow for large jumps in trait values and reversibility of changes, which are characteristics of acclimation more than adaptation (Flynn et al., 2015). The adaptive dynamics theory provides a clear modeling framework for the explicit representation of adaptation in ecological models (Kisdi and Geritz, 1999).

Adaptive dynamics are based on a simple idea: a resident population with fixed traits and a

fixed growth rate is invaded by a mutant, with slightly different parameters. If the mutant has a positive invasion fitness, he invades the population, and becomes representative of the new resident population. The environment being shaped by the traits of the resident population, the invasion fitness of the mutant will depend on its per-capita growth rate in the conditions given by the traits of the resident population (Kisdi and Geritz, 1999; Sauterey et al., 2015). In simple systems, the invasion fitness can be obtained through analytical resolution, but this is not doable in biogeochemical models with many state variables (Sauterey et al., 2015). The method can still be applied to biogeochemical models through the numerical resolution of mutant's growth rate (Sauterey et al., 2015, 2017).

A similar approach could be conducted using resource allocation models. Toseland et al. (2013) showed how such models allowed to predict growth rates in specific environments, and using a similar model in an adaptive dynamics context should be feasible. The main difference between Toseland et al. (2013) and an adaptative dynamics approach is that in Toseland et al. (2013), the assumption is that 'everything is everywhere' and so the parameters giving the optimal growth rate are always selected, while in adaptive dynamics, the system must be at steady state with fixed parameters when the mutant arrives, and the success of the mutant is entirely dependant on who was there before him (Sauterey et al., 2015). In a model like the one presented in Figure 5.6, with data-driven metabolic networks of organisms serving as the basis for parameterization, we could even imagine modeling the appearance of mutant by introducing random mutations at the genome level.

5.3 Perspectives

During most of my PhD, I tried to build links between meta-omics data, functional traits and the environmental context. If my initial goal was to find new ways of doing 'concrete' biogeochemical modeling, I then realized that multiple methods for the integration of sequencing data into models already existed, and that the limiting factors were actually more often on the biological side of things, where a significant part of unknown remains. Hence, I focused on the statistical exploration of the functional and taxonomic composition of planktonic communities, trying to disentangle and organize the complex richness of biological data, a step that I consider as key for later improving diversity representation in biogeochemical models. Thus, as a post-doc, I will keep working on large scale meta-omics data, this time focusing on prokaryotes from the Antarctic, sampled during the Antarctic circumglobal expedition (ACE, `https://spi-ace-expedition.ch/`).

In particular, I hope to take advantage from the multiple datasets available from the expedition (*e.g.* meta-omics datasets focusing on viruses, prokaryotes, and protists are available, as well as metadata on the environmental context of sampling including detailed measures of primary production and trace metals) to dig into the linking between genomic composition and trait realization (upper part of the green box in Figure 5.7). Actually, I have indicated in Figure 5.7 that this step should take place in labs, but I hope to be able to show that inter-disciplinary cruises on highly-equipped boats like ACE provide sufficient lab power to focus on this question *in-situ*.

Data from the cruise notably include in-situ mass spectrometry measures of net primary production, which I hope to relate to prokaryotic metagenomes. Other data types could actually allow to investigate this question, like in-situ high throughput imaging which gives information on traits such as shape, size, defense structure, coloniality or transparency (Martini et al., under review, Appendix A), or like satellite imagery, which is starting to be used to automatically quantify plankton functional groups at the ocean surface through machine learning (El Hourany et al., 2020). Building links between these different kinds of high-throughput data is the first step towards their common integration into modeling frameworks, which I believe will be the future of biogeochemical modeling.

Figure 5.8 - Furseals and the Akademik Treshnikov vessel during the third leg of the Antarctic Circumpolar Expedition (ACE), on South Georgia Island.

Bibliography

Acinas, S. G., Sánchez, P., Salazar, G., Cornejo-Castillo, F. M., Sebastián, M., Logares, R., Sunagawa, S., Hingamp, P., Ogata, H., Lima-Mendez, G., Roux, S., González, J. M., Arrieta, J. M., Alam, I. S., Kamau, A., Bowler, C., Raes, J., Pesant, S., Bork, P., Agustí, S., Gojobori, T., Bajic, V., Vaqué, D., Sullivan, M. B., Pedrós-Alió, C., Massana, R., Duarte, C. M., and Gasol, J. M. (2019). Metabolic Architecture of the Deep Ocean Microbiome. *bioRxiv*, page 635680. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Agatha, S., Strüder-Kypke, M. C., Beran, A., and Lynn, D. H. (2005). Pelagostrobilidium neptuni (Montagnes and Taylor, 1994) and Strombidium biarmatum nov. spec. (Ciliophora, Oligotrichea): phylogenetic position inferred from morphology, ontogenesis, and gene sequence data. *European Journal of Protistology*, 41(1):65–83.

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., Albini, G., Aury, J.-M., Belser, C., Bertrand, A., Cruaud, C., Silva, C. D., Dossat, C., Gavory, F., Gas, S., Guy, J., Haquelle, M., Jacoby, E., Jaillon, O., Lemainque, A., Pelletier, E., Samson, G., Wessner, M., Team, G. T., Bazire, P., Beluche, O., Bertrand, L., Besnard-Gonnet, M., Bordelais, I., Boutard, M., Dubois, M., Dumont, C., Ettedgui, E., Fernandez, P., Garcia, E., Aiach, N. G., Guerin, T., Hamon, C., Brun, E., Lebled, S., Lenoble, P., Louesse, C., Mahieu, E., Mairey, B., Martins, N., Megret, C., Milani, C., Muanga, J., Orvain, C., Payen, E., Perroud, P., Petit, E., Robert, D., Ronsin, M., Vacherie, B., Acinas, S. G., Royo-Llonch, M., Cornejo-Castillo, F. M., Logares, R., Fernández-Gómez, B., Bowler, C., Cochrane, G., Amid, C., Hoopen, P. T., Vargas, C. D., Grimsley, N., Desgranges, E., Kandels-Lewis, S., Ogata, H., Poulton, N., Sieracki, M. E., Stepanauskas, R., Sullivan, M. B., Brum, J. R., Duhaime, M. B., Poulos, B. T., Hurwitz, B. L., Coordinators, T. O. C., Acinas, S. G., Bork, P., Boss, E., Bowler, C., Vargas, C. D., Follows, M., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Karp-Boss, L., Karsenti, E., Not, F., Ogata, H., Pesant, S., Raes, J., Sardet, C., Sieracki, M. E., Speich, S., Stemmann, L., Sullivan, M. B., Sunagawa, S., Wincker, P., Pesant, S., Karsenti, E., and Wincker, P. (2017). Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition. *Scientific Data*, 4:170093.

Alcolombri, U., Ben-Dor, S., Feldmesser, E., Levin, Y., Tawfik, D. S., and Vardi, A. (2015). Identification of the algal dimethyl sulfide–releasing enzyme: A missing link in the marine sulfur cycle. *Science*, 348(6242):1466–1469.

Allen, J. I. and Polimene, L. (2011). Linking physiology to ecology: towards a new generation of plankton models. *Journal of Plankton Research*, 33(7):989–997. Publisher: Oxford Academic.

Alonso-Sáez, L., Sánchez, O., Gasol, J. M., Balagué, V., and Pedrós-Alio, C. (2008). Winter-to-summer changes in the composition and single-cell activity of near-surface Arctic prokaryotes. *Environmental Microbiology*, 10(9):2444–2454. _eprint: https://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1462-2920.2008.01674.x.

Amacher, J., Neuer, S., Anderson, I., and Massana, R. (2009). Molecular approach to determine contributions of the protist community to particle flux. *Deep Sea Research Part I: Oceanographic Research Papers*, 56(12):2206–2215.

Amaral-Zettler, L., Artigas, L. F., Baross, J., Bharathi, L., Boetius, A., Chandramohan, D., Herndl, G., Kogure, K., Neal, P., Pedrós-Alió, C., and others (2010). A global census of marine microbes. *Life in the world's Oceans: Diversity, Distribution and Abundance*, pages 223–245. Publisher: Blackwell Publishing Ltd Oxford.

Anderson, T. R. (2005). Plankton functional type modelling: running before we can walk? *Journal of Plankton Research*, 27(11):1073–1081. Publisher: Oxford Academic.

Anderson, T. R. (2010). Progress in marine ecosystem modelling and the "unreasonable effectiveness of mathematics". *Journal of Marine Systems*, 81(1):4–11.

Andras, J. P., Fields, P. D., Du Pasquier, L., Fredericksen, M., and Ebert, D. (2020). Genome-wide association analysis identifies a genetic basis of infectivity in a model bacterial pathogen. *Molecular Biology and Evolution*.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2019). KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *bioRxiv*, page 602110.

Arbore, R., Andras, J., Routtu, J., and Ebert, D. (2016). Ecological genetics of sediment browsing behaviour in a planktonic crustacean. *Journal of Evolutionary Biology*, 29(10):1999–2009. Publisher: John Wiley & Sons, Ltd.

Ardyna, M., Ovidio, F., Speich, S., Leconte, J., Chaffron, S., Audic, S., Garczarek, L., Pesant, S., Consortium, C. T. O., and Expedition, P. T. O. (2017). Environmental context of all samples from the Tara Oceans Expedition (2009-2013), about mesoscale features at the sampling location. PANGAEA.

Arenovski, A. L., Lim, E. L., and Caron, D. A. (1995). Mixotrophic nanoplankton in oligotrophic surface waters of the Sargasso Sea may employ phagotrophy to obtain major nutrients. *Journal of Plankton Research*, 17(4):801–820.

Armstrong, R. A., Drange, H., Parslow, J. S., Powell, T. M., Taylor, A. H., and Totterdell, I. J. (1993). Trophic Resolution. In Evans, G. T. and Fasham, M. J. R., editors, *Towards a Model of Ocean Biogeochemical Processes*, NATO ASI Series, pages 71–92, Berlin, Heidelberg. Springer.

Arnold, F. H. (2018). Directed Evolution: Bringing New Chemistry to Life. *Angewandte Chemie International Edition*, 57(16):4143–4148. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201708408.

Arsenieff, L., Le Gall, F., Rigaut-Jalabert, F., Mahé, F., Sarno, D., Gouhier, L., Baudoux, A.-C., and Simon, N. (2020). Diversity and dynamics of relevant nanoplanktonic diatoms in the Western English Channel. *The ISME Journal*, 14(8):1966–1981. Number: 8 Publisher: Nature Publishing Group.

Atkinson, H. J., Morris, J. H., Ferrin, T. E., and Babbitt, P. C. (2009). Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLOS ONE*, 4(2):e4345.

Auchtung, J. M., Robinson, C. D., and Britton, R. A. (2015). Cultivation of stable, reproducible microbial communities from different fecal donors using minibioreactor arrays (MBRAs). *Microbiome*, 3(1):42.

Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., and Gehlen, M. (2015). PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. *Geoscientific Model Development*, 8(8):2465–2513.

Aumont, O., Maier-Reimer, E., Blain, S., and Monfray, P. (2003). An ecosystem model of the global ocean including Fe, Si, P colimitations. *Global Biogeochemical Cycles*, 17(2). _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2001GB001745.

Ayata, S.-D., Lévy, M., Aumont, O., Sciandra, A., Sainte-Marie, J., Tagliabue, A., and Bernard, O. (2013). Phytoplankton growth formulation in marine ecosystem models: Should we take into account photo-acclimation and variable stoichiometry in oligotrophic areas? *Journal of Marine Systems*, 125:29–40.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*, 9(1):75.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Research*, 32(suppl_1):D138–D141. Publisher: Oxford Academic.

Baturin, G. N. (2003). Phosphorus Cycle in the Ocean. *Lithology and Mineral Resources*, 38(2):101–119.

Berge, T., Chakraborty, S., Hansen, P. J., and Andersen, K. H. (2017). Modeling succession of key resource-harvesting traits of mixotrophic plankton. *The ISME Journal*, 11(1):212–223.

Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P., and Bapteste, E. (2018). Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biology and Evolution*, 10(3):707–715.

Berry, D. and Widder, S. (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology*, 5. Publisher: Frontiers.

Biard, T., Bigeard, E., Audic, S., Poulain, J., Gutierrez-Rodriguez, A., Pesant, S., Stemmann, L., and Not, F. (2017). Biogeography and diversity of Collodaria (Radiolaria) in the global ocean. *The ISME Journal*, 11(6):1331–1344.

Biard, T., Krause, J. W., Stukel, M. R., and Ohman, M. D. (2018). The Significance of Giant Phaeodarians (Rhizaria) to Biogenic Silica Export in the California Current Ecosystem. *Global Biogeochemical Cycles*, 32(6):987–1004. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018GB005877.

Biard, T., Stemmann, L., Picheral, M., Mayot, N., Vandromme, P., Hauss, H., Gorsky, G., Guidi, L., Kiko, R., and Not, F. (2016). *In situ* imaging reveals the biomass of giant protists in the global ocean. *Nature*, 532(7600):504–507.

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., and Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, 27(4):233–243.

Bittner, L., Gobet, A., Audic, S., Romac, S., Egge, E. S., Santini, S., Ogata, H., Probert, I., Edvardsen, B., and de Vargas, C. (2013). Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. *Molecular Ecology*, 22(1):87–101.

Bittner, L., Halary, S., Payri, C., Cruaud, C., de Reviers, B., Lopez, P., and Bapteste, E. (2010). Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. *Biology Direct*, 5:47.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Borcard, D., Gillet, F., and Legendre, P. (2011). *Numerical ecology with R*. Use R! Springer.

Boyd, P. W., Jickells, T., Law, C. S., Blain, S., Boyle, E. A., Buesseler, K. O., Coale, K. H., Cullen, J. J., Baar, H. J. W. d., Follows, M., Harvey, M., Lancelot, C., Levasseur, M., Owens, N. P. J., Pollard, R., Rivkin, R. B., Sarmiento, J., Schoemann, V., Smetacek, V., Takeda, S., Tsuda, A., Turner, S., and Watson, A. J. (2007). Mesoscale Iron Enrichment Experiments 1993-2005: Synthesis and Future Directions. *Science*, 315(5812):612–617. Publisher: American Association for the Advancement of Science Section: Review.

Boyer, T., Locarnini, R. A., Baranova, O., Garcia, H. E., Mishonov, A. V., Paver, C., Reagan, J. R., Seidov, D., Smolyar, I., Weathers, K. W., and Zweng, M. (2018). The World Ocean Atlas 2018: Improvements and Uses of Climatological Mean Fields. *AGU Fall Meeting Abstracts*, 13.

Bron, J. E., Frisch, D., Goetze, E., Johnson, S. C., Lee, C. E., and Wyngaard, G. A. (2011). Observing copepods through a genomic lens. *Frontiers in Zoology*, 8(1):22.

Brown, E. A., Chain, F. J. J., Crease, T. J., MacIsaac, H. J., and Cristescu, M. E. (2015). Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? *Ecology and Evolution*, 5(11):2234–2251.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60. Number: 1 Publisher: Nature Publishing Group.

Budinich, M., Bourdon, J., Larhlimi, A., and Eveillard, D. (2017). A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. *PLOS ONE*, 12(2):e0171744. Publisher: Public Library of Science.

Burns, J. A., Pittis, A. A., and Kim, E. (2018). Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. *Nature Ecology & Evolution*, 2(4):697–704.

Cadotte, M. W., Davies, T. J., Regetz, J., Kembel, S. W., Cleland, E., and Oakley, T. H. (2010). Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecology Letters*, 13(1):96–105. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1461-0248.2009.01405.x.

Calbet, A., Garrido, S., Saiz, E., Alcaraz, M., and Duarte, C. M. (2001). Annual Zooplankton Succession in Coastal NW Mediterranean Waters: The Importance of the Smaller Size Fractions. *Journal of Plankton Research*, 23(3):319–331. Publisher: Oxford Academic.

Calbet, A., Martínez, R. A., Isari, S., Zervoudaki, S., Nejstgaard, J. C., Pitta, P., Sazhin, A. F., Sousoni, D., Gomes, A., Berger, S. A., Tsagaraki, T. M., and Ptacnik, R. (2012). Effects of light availability on mixotrophy and microzooplankton grazing in an oligotrophic plankton food web: Evidences from a mesocosm study in Eastern Mediterranean waters. *Journal of Experimental Marine Biology and Ecology*, 424-425:66–77.

Caron, D. A. (2016a). Mixotrophy stirs up our understanding of marine food webs. *Proceedings of the National Academy of Sciences*, 113(11):2806–2808.

Caron, D. A. (2016b). The rise of Rhizaria. *Nature*, 532(7600):444–445. Number: 7600 Publisher: Nature Publishing Group.

Caron, D. A., Alexander, H., Allen, A. E., Archibald, J. M., Armbrust, E. V., Bachy, C., Bell, C. J., Bharti, A., Dyhrman, S. T., Guida, S. M., Heidelberg, K. B., Kaye, J. Z., Metzner, J., Smith, S. R., and Worden, A. Z. (2017). Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology*, 15(1):6–20.

Caron, D. A., Countway, P. D., Jones, A. C., Kim, D. Y., and Schnetzer, A. (2012). Marine Protistan Diversity. *Annual Review of Marine Science*, 4(1):467–493.

Caron, D. A., Michaels, A. F., Swanberg, N. R., and Howse, F. A. (1995). Primary productivity by symbiont-bearing planktonic sarcodines (Acantharia, Radiolaria, Foraminifera) in surface waters near Bermuda. *Journal of Plankton Research*, 17(1):103–129. Publisher: Oxford Academic.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Jaillon, O., Aury, J.-M., Karsenti, E., Sullivan, M. B., Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., Vargas, C. d., Iudicone, D., Bowler, C., and Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, 9(1):1–13. Number: 1 Publisher: Nature Publishing Group.

Cermeño, P., Chouciño, P., Fernández-Castro, B., Figueiras, F. G., Marañón, E., Marrasé, C., Mouriño-Carballido, B., Pérez-Lorenzo, M., Rodríguez-Ramos, T., Teixeira, I. G., and Vallina, S. M. (2016). Marine Primary Productivity Is Driven by a Selection Effect. *Frontiers in Marine Science*, 3. Publisher: Frontiers.

Charlson, R. J., Lovelock, J. E., Andreae, M. O., and Warren, S. G. (1987). Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature*, 326(6114):655–661. Number: 6114 Publisher: Nature Publishing Group.

Chatt, J., Dilworth, J. R., and Richards, R. L. (1978). Recent advances in the chemistry of nitrogen fixation. *Chemical Reviews*, 78(6):589–625.

Chavez, F. P., Messié, M., and Pennington, J. T. (2011). Marine Primary Production in Relation to Climate Variability and Change. *Annual Review of Marine Science*, 3(1):227–260. _eprint: https://doi.org/10.1146/annurev.marine.010908.163917.

Cheng, S., Karkar, S., Bapteste, E., Yee, N., Falkowski, P., and Bhattacharya, D. (2014). Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. *Frontiers in Ecology and Evolution*, 2. Publisher: Frontiers.

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Quéré, C., Myneni, R., S, P., and Thornton, P. (2013). 2013: Carbon and Other Biogeochemical Cycles. In Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 465–570. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Coles, V. J. and Hood, R. R. (2016). Approaches and Challenges for Linking Marine Biogeochemical Models with the "Omics" Revolution. In Glibert, P. M. and Kana, T. M., editors, *Aquatic Microbial Ecology and Biogeochemistry: A Dual Perspective*, pages 45–63. Springer International Publishing, Cham.

Coles, V. J., Stukel, M. R., Brooks, M. T., Burd, A., Crump, B. C., Moran, M. A., Paul, J. H., Satinsky, B. M., Yager, P. L., Zielinski, B. L., and Hood, R. R. (2017). Ocean biogeochemistry modeled with emergent trait-based genomics. *Science*, 358(6367):1149–1154.

Csardi, G., Nepusz, T., and others (2006). The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9.

Curson, A. R. J., Todd, J. D., Sullivan, M. J., and Johnston, A. W. B. (2011). Catabolism of dimethylsulphoniopropionate: microorganisms, enzymes and genes. *Nature Reviews Microbiology*, 9(12):849–859.

Czypionka, T., Fields, P. D., Routtu, J., Berg, E. v. d., Ebert, D., and Meester, L. D. (2019). The genetic architecture underlying diapause termination in a planktonic crustacean. *Molecular Ecology*, 28(5):998–1008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.15001.

D'Alelio, D., Eveillard, D., Coles, V. J., Caputi, L., Ribera d'Alcalà, M., and Iudicone, D. (2019). Modelling the complexity of plankton communities exploiting omics potential: From present challenges to an integrative pipeline. *Current Opinion in Systems Biology*, 13:68–74.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Luke, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans Coordinators, Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., Boss, E., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E. G., Sardet, C., Sullivan, M. B., and Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237):1261605–1261605.

Decelle, J., Probert, I., Bittner, L., Desdevises, Y., Colin, S., de Vargas, C., Gali, M., Simo, R., and Not, F. (2012). An original mode of symbiosis in open ocean plankton. *Proceedings of the National Academy of Sciences*, 109(44):18000–18005.

Decelle, J., Romac, S., Sasaki, E., Not, F., and Mahé, F. (2014). Intracellular Diversity of the V4 and V9 Regions of the 18S rRNA in Marine Protists (Radiolarians) Assessed by High-Throughput Sequencing. *PLOS ONE*, 9(8):e104297. Publisher: Public Library of Science.

Del Campo, J., Kolisko, M., Boscaro, V., Santoferrara, L. F., Nenarokov, S., Massana, R., Guillou, L., Simpson, A., Berney, C., Vargas, C. d., Brown, M. W., Keeling, P. J., and Parfrey, L. W. (2018). EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLOS Biology*, 16(9):e2005849. Publisher: Public Library of Science.

Delmont, T. O., Kiefl, E., Kilinc, O., Esen, O. C., Uysal, I., Rappe, M. S., Giovannoni, S., and Eren, A. M. (2017). The global biogeography of amino acid variants within a single SAR11 population is governed by natural selection. *bioRxiv*, page 170639.

Delmont, T. O., Quince, C., Shaiber, A., Esen, O. C., Lee, S. T., Rappé, M. S., McLellan, S. L., Lücker, S., and Eren, A. M. (2018). Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, 3(7):804–813.

DeLong, E. F. (1992). Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences*, 89(12):5685–5689. Publisher: National Academy of Sciences Section: Research Article.

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-U., Martinez, A., Sullivan, M. B., Edwards, R., Brito, B. R., Chisholm, S. W., and Karl, D. M. (2006). Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science*, 311(5760):496–503.

Dick, G. J. (2017). Embracing the mantra of modellers and synthesizing omics, experiments and models. *Environmental Microbiology Reports*, 9(1):18–20.

Dolan, J. R. and PÉrez, M. T. (2000). Costs, benefits and characteristics of mixotrophy in marine oligotrichs. *Freshwater Biology*, 45(2):227–238.

Doney, S. C. (1999). Major challenges confronting marine biogeochemical modeling. *Global Biogeochemical Cycles*, 13(3):705–714. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/1999GB900039.

Duarte, C. M. (2015). Seafaring in the 21st century: the Malaspina 2010 circumnavigation expedition. *Limnology and Oceanography Bulletin*, 24(1):11–14. Publisher: Wiley Online Library.

Ducklow, H. W. (1999). The bacterial component of the oceanic euphotic zone. *FEMS Microbiology Ecology*, 30(1):1–10. Publisher: Oxford Academic.

Ducklow, H. W., Steinberg, D. K., and Buesseler, K. O. (2001). Upper ocean carbon export and the biological pump. *OCEANOGRAPHY-WASHINGTON DC-OCEANOGRAPHY SOCIETY-*, 14(4):50–58.

Dutkiewicz, S., Cermeno, P., Jahn, O., Follows, M. J., Hickman, A. E., Taniguchi, D. A. A., and Ward, B. A. (2020). Dimensions of marine phytoplankton diversity. *Biogeosciences*, 17(3):609–634.

Eady, R. R. and Postgate, J. R. (1974). Nitrogenase. *Nature*, 249(5460):805–810. Number: 5460 Publisher: Nature Publishing Group.

Edgar, R. (2010). Usearch. Technical report, Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States).

Edwards, K. F., Thomas, M. K., Klausmeier, C. A., and Litchman, E. (2012). Allometric scaling and taxonomic variation in nutrient utilization traits and maximum growth rate of phytoplankton. *Limnology and Oceanography*, 57(2):554–566. _eprint: https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.4319/lo.2012.57.2.0554.

Egge, E., Bittner, L., Andersen, T., Audic, S., de Vargas, C., and Edvardsen, B. (2013). 454 pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: a test for marine haptophytes. *PloS One*, 8(9):e74371.

El Hourany, R., Bowler, C., Mejia, C., Crépon, M., and Thiria, S. (2020). A neural-based bio-regionalization of the Mediterranean Sea using satellite and Argo-float records. 22:12004. Conference Name: EGU General Assembly Conference Abstracts.

Erdrich, P., Steuer, R., and Klamt, S. (2015). An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. *BMC Systems Biology*, 9(1):48.

Escoufier, Y. (1973). Le Traitement des Variables Vectorielles. *Biometrics*, 29(4):751.

Esteban, G. F., Fenchel, T., and Finlay, B. J. (2010). Mixotrophy in Ciliates. *Protist*, 161(5):621–641.

Faber, W. W., Anderson, O. R., and Caron, D. A. (1989). Algal-foraminiferal symbiosis in the planktonic foraminifer Globigerinella aequilateralis; II, Effects of two symbiont species on foraminiferal growth and longevity. *Journal of Foraminiferal Research*, 19(3):185–193.

Falkowski, P. G., Barber, R. T., and Smetacek, V. (1998). Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science*, 281(5374):200–206.

Falkowski, P. G., Katz, M. E., Knoll, A. H., Quigg, A., Raven, J. A., Schofield, O., and Taylor, F. J. R. (2004). The Evolution of Modern Eukaryotic Phytoplankton. *Science*, 305(5682):354–360. Publisher: American Association for the Advancement of Science Section: Review.

Fasham, M. J. R., Ducklow, H. W., and McKelvie, S. M. (1990). A nitrogen-based model of plankton dynamics in the oceanic mixed layer. *Journal of Marine Research*, 48(3):591–639.

Fasham, M. J. R., Sarmiento, J. L., Slater, R. D., Ducklow, H. W., and Williams, R. (1993). Ecosystem behavior at Bermuda Station "S" and ocean weather station "India": A general circulation model and observational analysis. *Global Biogeochemical Cycles*, 7(2):379–415. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/92GB02784.

Faure, E., Not, F., Benoiston, A.-S., Labadie, K., Bittner, L., and Ayata, S.-D. (2019). Mixotrophic protists display contrasted biogeographies in the global ocean. *The ISME Journal*.

Faust, K. (2019). Towards a Better Understanding of Microbial Community Dynamics through High-Throughput Cultivation and Data Integration. *mSystems*, 4(3).

Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550. Number: 8 Publisher: Nature Publishing Group.

Fehrenbacher, J. S., Spero, H. J., and Russell, A. D. (2011). Observations of living non-spinose planktic foraminifers Neogloboquadrina dutertrei and N. pachyderma from specimens grown in culture. *AGU Fall Meeting Abstracts*, 41.

Ferrera, I., Sebastian, M., Acinas, S. G., and Gasol, J. M. (2015). Prokaryotic functional gene diversity in the sunlit ocean: Stumbling in the dark. *Current Opinion in Microbiology*, 25:33–39.

Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, 281(5374):237–240. Publisher: American Association for the Advancement of Science Section: Report.

Fleming, R. H. (1939). The Control of Diatom Populations by Grazing. *ICES Journal of Marine Science*, 14(2):210–227. Publisher: Oxford Academic.

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincon, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K. W., Lomas, M. W., Veneziano, D., Vera, C. S., Vrugt, J. A., and Martiny, A. C. (2013). Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proceedings of the National Academy of Sciences*, 110(24):9824–9829.

Flynn, K. J. and Mitra, A. (2009). Building the "perfect beast": modelling mixotrophic plankton. *Journal of Plankton Research*, 31(9):965–992.

Flynn, K. J., St John, M., Raven, J. A., Skibinski, D. O., Allen, J. I., Mitra, A., and Hofmann, E. E. (2015). Acclimation, adaptation, traits and trade-offs in plankton functional type models: reconciling terminology for biology and modelling. *Journal of Plankton Research*, 37(4):683–691.

Flynn, K. J., Stoecker, D. K., Mitra, A., Raven, J. A., Glibert, P. M., Hansen, P. J., Graneli, E., and Burkholder, J. M. (2013). Misuse of the phytoplankton-zooplankton dichotomy: the need to assign organisms as mixotrophs within plankton functional types. *Journal of Plankton Research*, 35(1):3–11.

Follows, M. J. and Dutkiewicz, S. (2011). Modeling Diverse Communities of Marine Microbes. *Annual Review of Marine Science*, 3(1):427–451. _eprint: https://doi.org/10.1146/annurev-marine-120709-142848.

Follows, M. J., Dutkiewicz, S., Grant, S., and Chisholm, S. W. (2007). Emergent Biogeography of Microbial Communities in a Model Ocean. *Science*, 315(5820):1843–1846.

Fondi, M., Karkman, A., Tamminen, M. V., Bosi, E., Virta, M., Fani, R., Alm, E., and McInerney, J. O. (2016). "Every Gene Is Everywhere but the Environment Selects": Global Geolocalization of Gene Sharing in Environmental Samples through Network Analysis. *Genome Biology and Evolution*, 8(5):1388–1400.

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S., Lopez, P., Stoeck, T., and Bapteste, E. (2015). Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biology*, 13:16.

Francis, C. A., Roberts, K. J., Beman, J. M., Santoro, A. E., and Oakley, B. B. (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proceedings of the National Academy of Sciences*, 102(41):14683–14688. Publisher: National Academy of Sciences Section: Biological Sciences.

Frede Thingstad, T., Strand, E., and Larsen, A. (2010). Stepwise building of plankton functional type (PFT) models: A feasible route to complex models? *Progress in Oceanography*, 84(1-2):6–15.

Freilich, M. A., Wieters, E., Broitman, B. R., Marquet, P. A., and Navarrete, S. A. (2018). Species co-occurrence networks: Can they reveal trophic and non-trophic interactions in ecological communities? *Ecology*, 99(3):690–699. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecy.2142.

Freilich, S., Zarecki, R., Eilam, O., Segal, E. S., Henry, C. S., Kupiec, M., Gophna, U., Sharan, R., and Ruppin, E. (2011). Competitive and cooperative metabolic interactions in bacterial communities. *Nature Communications*, 2(1):589. Number: 1 Publisher: Nature Publishing Group.

Fuhrman, J. A., McCallum, K., and Davis, A. A. (1992). Novel major archaebacterial group from marine plankton. *Nature*, 356(6365):148–149. Number: 6365 Publisher: Nature Publishing Group.

Galand, P. E., Pereira, O., Hochart, C., Auguet, J. C., and Debroas, D. (2018). A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *The ISME Journal*, 12(10):2470–2478. Number: 10 Publisher: Nature Publishing Group.

Gallagher, R., Falster, D. S., Maitner, B., Salguero-Gomez, R., Vandvik, V., Pearse, W., Schneider, F., Kattge, J., Alroy, J., Ankenbrand, M. J., Andrew, S., Balk, M., Bland, L., Boyle, B., Bravo-Avila, C., Brennan, I., Carthey, A., Catullo, R., Cavazos, B., Chown, S., Fadrique, B., Feng, X., Halbritter, A. H., Hammock, J., Hogan, J. A., Holewa, H., Iversen, C., Jochum, M., Kearney, M., Keller, A., Mabee, P., Madin, J., Manning, P., McCormack, L., Michaletz, S., Park, D., Penone, C., Perez, T., Pineda-Munoz, S., Poelen, J. H., Ray, C., Rossetto, M., Sauquet, H., Sparrow, B., Spasojevic, M. J., Telford, R. J., Tobias, J. A., Violle, C., Walls, R., Weiss, K. C. B., Westoby, M., Wright, I., and Enquist, B. (2019). The Open Traits Network: Using Open Science principles to accelerate trait-based science across the Tree of Life. preprint, EcoEvoRxiv.

Gentleman, W. (2002). A chronology of plankton dynamics in silico: how computer models have been used to study marine ecosystems. *Hydrobiologia*, 480(1):69–85.

Gentleman, W., Leising, A., Frost, B., Strom, S., and Murray, J. (2003). Functional responses for zooplankton feeding on multiple resources: a review of assumptions and biological dynamics. *Deep Sea Research Part II: Topical Studies in Oceanography*, 50(22):2847–2875.

Ghyoot, C., Flynn, K. J., Mitra, A., Lancelot, C., and Gypens, N. (2017). Modeling Plankton Mixotrophy: A Mechanistic Model Consistent with the Shuter-Type Biochemical Approach. *Frontiers in Ecology and Evolution*, 5.

Gilbert, J. A., Field, D., Swift, P., Thomas, S., Cummings, D., Temperton, B., Weynberg, K., Huse, S., Hughes, M., Joint, I., Somerfield, P. J., and Mühling, M. (2010). The Taxonomic and

Functional Diversity of Microbes at a Temperate Coastal Site: A 'Multi-Omic' Study of Seasonal and Diel Temporal Variation. *PLoS ONE*, 5(11):e15545.

Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., Tischler, G., Jackson, D. K., Keane, T. M., Li, J., Yue, J.-X., Liti, G., Durbin, R., and Ning, Z. (2017). De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Scientific Reports*, 7(1):3935. Number: 1 Publisher: Nature Publishing Group.

Giovannoni, S., Vergin, K., and Carlson, C. (2014). Twenty-five Years of Omics at BATS. page 19.

Granéli, E., Edvardsen, B., Roelke, D. L., and Hagström, J. A. (2012). The ecophysiology and bloom dynamics of Prymnesium spp. *Harmful Algae*, 14(Supplement C):260–270.

Grossart, H.-P., Massana, R., McMahon, K. D., and Walsh, D. A. (2020). Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnology and Oceanography*, 65(S1):S2–S20. _eprint: https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.1002/lno.11382.

Gruber, N. (2004). The Dynamics of the Marine Nitrogen Cycle and its Influence on Atmospheric CO2 Variations. In Follows, M. and Oguz, T., editors, *The Ocean Carbon Cycle and Climate*, NATO Science Series, pages 97–148, Dordrecht. Springer Netherlands.

Gruber, N. (2019). Consistent patterns of nitrogen fixation identified in the ocean. *Nature*, 566(7743):191–193. Number: 7743 Publisher: Nature Publishing Group.

Guadagno, C. R., Millar, D., Lai, R., Mackay, D. S., Pleban, J. R., McClung, C. R., Weinig, C., Wang, D. R., and Ewers, B. E. (2020). Use of transcriptomic data to inform biophysical models via Bayesian networks. *Ecological Modelling*, 429:109086.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J. R., Coelho, L. P., Espinoza, J. C. I., Malviya, S., Sunagawa, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Poulain, J., Searson, S., Stemmann, L., Not, F., Hingamp, P., Speich, S., Follows, M., Karp-Boss, L., Boss, E., Ogata, H., Pesant, S., Weissenbach, J., Wincker, P., Acinas, S. G., Bork, P., de Vargas, C., Iudicone, D., Sullivan, M. B., Raes, J., Karsenti, E., Bowler, C., and Gorsky, G. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532(7600):465–470.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., and Christen, R. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, 41(D1):D597–D604.

Gutierrez-Rodriguez, A., Stukel, M. R., Lopes dos Santos, A., Biard, T., Scharek, R., Vaulot, D., Landry, M. R., and Not, F. (2019). High contribution of Rhizaria (Radiolaria) to vertical

export in the California Current Ecosystem revealed by DNA metabarcoding. *The ISME Journal*, 13(4):964–976. Number: 4 Publisher: Nature Publishing Group.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N., and Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8):1494–1512. Number: 8 Publisher: Nature Publishing Group.

Hansen, P., Moldrup, M., Tarangkoon, W., Garcia-Cuetos, L., and Moestrup, O. (2012). Direct evidence for symbiont sequestration in the marine red tide ciliate Mesodinium rubrum. *Aquatic Microbial Ecology*, 66(1):63–75.

Hardy, W. F. and Burns, R. C. (1968). Biological Nitrogen Fixation. *Annual Review of Biochemistry*, 37(1):331–358. _eprint: https://doi.org/10.1146/annurev.bi.37.070168.001555.

Hawley, A. K., Nobu, M. K., Wright, J. J., Durno, W. E., Morgan-Lang, C., Sage, B., Schwientek, P., Swan, B. K., Rinke, C., Torres-Beltrán, M., Mewis, K., Liu, W.-T., Stepanauskas, R., Woyke, T., and Hallam, S. J. (2017). Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients. *Nature Communications*, 8(1):1–10. Number: 1 Publisher: Nature Publishing Group.

Hekstra, D. R. and Leibler, S. (2012). Contingency and Statistical Laws in Replicate Microbial Closed Ecosystems. *Cell*, 149(5):1164–1173. Publisher: Elsevier.

Hemleben, C., Be, A. W. H., Anderson, O. R., and Tuntivate, S. (1977). Test morphology, organic layers and chamber formation of the planktonic foraminifer Globorotalia menardii (d'Orbigny). *Journal of Foraminiferal Research*, 7(1):1–25.

Hensen, V. (1887). Über die Bestimmung des Planktons oder des im Meere treibenden Materials an Pflanzen und Thieren.

Herndl, G. J. and Reinthaler, T. (2013). Microbial control of the dark end of the biological pump. *Nature Geoscience*, 6(9):718–724.

Hillebrand, H. (2004). On the Generality of the Latitudinal Diversity Gradient. *The American Naturalist*, 163(2):192–211. Publisher: The University of Chicago Press.

Hirai, J., Yasuike, M., Fujiwara, A., Nakamura, Y., Hamaoka, S., Katakura, S., Takano, Y., and Nagai, S. (2015). Effects of plankton net characteristics on metagenetic community analysis of metazoan zooplankton in a coastal marine ecosystem. *Journal of Experimental Marine Biology and Ecology*, 469:36–43.

Holling, C. (1959). The components of predation as revealed by a study of small-mammal predation of the European pine sawfly.

Hood, R. R., Laws, E. A., Follows, M. J., and Siegel, D. A. (2007). Modeling and Prediction of Marine Microbial Populations in the Genomic Era. *Oceanography*, 20(2):155–165. Publisher: Oceanography Society.

Howard, E. C., Henriksen, J. R., Buchan, A., Reisch, C. R., Bürgmann, H., Welsh, R., Ye, W., González, J. M., Mace, K., Joye, S. B., Kiene, R. P., Whitman, W. B., and Moran, M. A. (2006). Bacterial Taxa That Limit Sulfur Flux from the Ocean. *Science*, 314(5799):649–652. Publisher: American Association for the Advancement of Science Section: Report.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, 34(8):2115–2122.

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

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., Coelho, L. P., Endo, H., Gasol, J. M., Gregory, A. C., Mahé, F., Rigonato, J., Royo-Llonch, M., Salazar, G., Sanz-Sáez, I., Scalco, E., Soviadan, D., Zayed, A. A., Zingone, A., Labadie, K., Ferland, J., Marec, C., Kandels, S., Picheral, M., Dimier, C., Poulain, J., Pisarev, S., Carmichael, M., Pesant, S., Acinas, S. G., Babin, M., Bork, P., Boss, E., Bowler, C., Cochrane, G., de Vargas, C., Follows, M., Gorsky, G., Grimsley, N., Guidi, L., Hingamp, P., Iudicone, D., Jaillon, O., Kandels, S., Karp-Boss, L., Karsenti, E., Not, F., Ogata, H., Pesant, S., Poulton, N., Raes, J., Sardet, C., Speich, S., Stemmann, L., Sullivan, M. B., Sunagawa, S., Wincker, P., Babin, M., Boss, E., Iudicone, D., Jaillon, O., Acinas, S. G., Ogata, H., Pelletier, E., Stemmann, L., Sullivan, M. B., Sunagawa, S., Bopp, L., de Vargas, C., Karp-Boss, L., Wincker, P., Lombard, F., Bowler, C., and Zinger, L. (2019). Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell*, 179(5):1084–1097.e21.

Jeong, H. J., Yoo, Y. D., Kim, J. S., Seong, K. A., Kang, N. S., and Kim, T. H. (2010). Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. *Ocean Science Journal*, 45(2):65–91.

Jiao, N., Herndl, G. J., Hansell, D. A., Benner, R., Kattner, G., Wilhelm, S. W., Kirchman, D. L., Weinbauer, M. G., Luo, T., Chen, F., and Azam, F. (2010). Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nature Reviews Microbiology*, 8(8):593–599. Number: 8 Publisher: Nature Publishing Group.

Johnsen, G., Dalløkken, R., Eikrem, W., Legrand, C., Aure, J., and Skjoldal, H. R. (1999). Ecophysiology, bio-optics and toxicity of the ichtyotoxic Chrysochromulina leadbeateri (Prymnesiophyceae). *Journal of Phycology*, 35(6):1465–1476.

Johnson, M. D. (2011). Acquired Phototrophy in Ciliates: A Review of Cellular Interactions and Structural Adaptations. *Journal of Eukaryotic Microbiology*, 58(3):185–195.

Johnson, M. D., Oldach, D., Delwiche, C. F., and Stoecker, D. K. (2007). Retention of transcriptionally active cryptophyte nuclei by the ciliate Myrionecta rubra. *Nature*, 445(7126):426–428.

Jones, H. L. J., Leadbeater, B. S. C., and Green, J. C. (1993). Mixotrophy in marine species of Chrysochromulina (Prymnesiophyceae): ingestion and digestion of a small green flagellate. *Journal of the Marine Biological Association of the United Kingdom*, 73(02):283.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240. Publisher: Oxford Academic.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, 47(D1):D590–D595. Publisher: Oxford Academic.

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., Pesant, S., Reynaud, E. G., Sardet, C., Sieracki, M. E., Speich, S., Velayoudon, D., Weissenbach, J., Wincker, P., and the Tara Oceans Consortium (2011). A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biology*, 9(10):e1001177.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., Deschamps, P., Dyhrman, S. T., Edvardsen, B., Gates, R. D., Gobler, C. J., Greenwood, S. J., Guida, S. M., Jacobi, J. L., Jakobsen, K. S., James, E. R., Jenkins, B., John, U., Johnson, M. D., Juhl, A. R., Kamp, A., Katz, L. A., Kiene, R., Kudryavtsev, A., Leander, B. S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A. M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M. A., Murray, S., Nadathur, G., Nagai, S., Ngam, P. B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M. C., Rengefors, K., Romano, G., Rumpho, M. E., Rynearson, T., Schilling, K. B., Schroeder, D. C., Simpson, A. G. B., Slamovits, C. H., Smith, D. R., Smith, G. J., Smith, S. R., Sosik, H. M., Stief, P., Theriot, E., Twary, S. N., Umale, P. E., Vaulot, D., Wawrik, B., Wheeler, G. L., Wilson, W. H., Xu, Y., Zingone, A., and Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biology*, 12(6):e1001889.

Keeling, P. J. and Campo, J. d. (2017). Marine Protists Are Not Just Big Bacteria. *Current Biology*, 27(11):R541–R549.

Keeling, P. J. and Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618. Number: 8 Publisher: Nature Publishing Group.

Kisdi, E. and Geritz, S. A. H. (1999). Adaptive Dynamics in Allele Space: Evolution of Genetic Polymorphism by Small Mutations in a Heterogeneous Environment. *Evolution*, 53(4):993–1008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1558-5646.1999.tb04515.x.

Kishi, M. J., Kashiwai, M., Ware, D. M., Megrey, B. A., Eslinger, D. L., Werner, F. E., Noguchi-Aita, M., Azumaya, T., Fujii, M., Hashimoto, S., Huang, D., Iizumi, H., Ishida, Y., Kang, S., Kantakov, G. A., Kim, H.-c., Komatsu, K., Navrotsky, V. V., Smith, S. L., Tadokoro, K., Tsuda, A., Yamamura, O., Yamanaka, Y., Yokouchi, K., Yoshie, N., Zhang, J., Zuenko, Y. I., and Zvalinsky, V. I. (2007). NEMURO—a lower trophic level model for the North Pacific marine ecosystem. *Ecological Modelling*, 202(1):12–25.

Kiørboe, T., Visser, A., and Andersen, K. H. (2018). A trait-based approach to ocean ecology. *ICES Journal of Marine Science*, 75(6):1849–1863. Publisher: Oxford Academic.

Konwar, K. M., Hanson, N. W., Bhatia, M. P., Kim, D., Wu, S.-J., Hahn, A. S., Morgan-Lang, C., Cheung, H. K., and Hallam, S. J. (2015). MetaPathways v2.5: quantitative functional, taxonomic and usability improvements. *Bioinformatics*, 31(20):3345–3347. Publisher: Oxford Academic.

Koonin, E. V., Makarova, K. S., and Aravind, L. (2001). Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. *Annual Review of Microbiology*, 55(1):709–742. _eprint: https://doi.org/10.1146/annurev.micro.55.1.709.

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., Fernandez-Guerra, A., Jeanthon, C., Rahav, E., Ullrich, M., Wichels, A., Gerdts, G., Polymenakou, P., Kotoulas, G., Siam, R., Abdallah, R. Z., Sonnenschein, E. C., Cariou, T., O'Gara, F., Jackson, S., Orlic, S., Steinke, M., Busch, J., Duarte, B., Caçador, I., Canning-Clode, J., Bobrova, O., Marteinsson, V., Reynisson, E., Loureiro, C. M., Luna, G. M., Quero, G. M., Löscher, C. R., Kremp, A., DeLorenzo, M. E., Øvreås, L., Tolman, J., LaRoche, J., Penna, A., Frischer, M., Davis, T., Katherine, B., Meyer, C. P., Ramos, S., Magalhães, C., Jude-Lemeilleur, F., Aguirre-Macedo, M. L., Wang, S., Poulton, N., Jones, S., Collin, R., Fuhrman, J. A., Conan, P., Alonso, C., Stambler, N., Goodwin, K., Yakimov, M. M., Baltar, F., Bodrossy, L., Kamp, J. V. D., Frampton, D. M., Ostrowski, M., Ruth, P. V., Malthouse, P., Claus, S., Deneudt, K., Mortelmans, J., Pitois, S., Wallom, D., Salter, I., Costa, R., Schroeder, D. C., Kandil, M. M., Amaral, V., Biancalana, F., Santana, R., Pedrotti, M. L., Yoshida, T., Ogata, H., Ingleton, T., Munnik, K., Rodriguez-Ezpeleta, N., Berteaux-Lecellier, V., Wecker, P., Cancio, I., Vaulot, D., Bienhold, C., Ghazal, H., Chaouni, B., Essayeh, S., Ettamimi, S., Zaid, E. H., Boukhatem, N., Bouali, A., Chahboune, R., Barrijal, S., Timinouni, M., Otmani, F. E., Bennani, M., Mea, M., Todorova, N., Karamfilov, V., Hoopen, P. t., Cochrane, G., L'Haridon, S., Bizsel, K. C., Vezzi, A., Lauro, F. M., Martin,

P., Jensen, R. M., Hinks, J., Gebbels, S., Rosselli, R., Pascale, F. D., Schiavon, R., Santos, A. d., Villar, E., Pesant, S., Cataletto, B., Malfatti, F., Edirisinghe, R., Silveira, J. A. H., Barbier, M., Turk, V., Tinta, T., Fuller, W. J., Salihoglu, I., Serakinci, N., Ergoren, M. C., Bresnan, E., Iriberri, J., Nyhus, P. A. F., Bente, E., Karlsen, H. E., Golyshin, P. N., Gasol, J. M., Moncheva, S., Dzhembekova, N., Johnson, Z., Sinigalliano, C. D., Gidley, M. L., Zingone, A., Danovaro, R., Tsiamis, G., Clark, M. S., Costa, A. C., Bour, M. E., Martins, A. M., Collins, R. E., Ducluzeau, A.-L., Martinez, J., Costello, M. J., Amaral-Zettler, L. A., Gilbert, J. A., Davies, N., Field, D., and Glöckner, F. O. (2015). The ocean sampling day consortium. *GigaScience*, 4(1). Publisher: Oxford Academic.

Kuhn, M. (2008). Building Predictive Models in *R* Using the **caret** Package. *Journal of Statistical Software*, 28(5).

Kuile, B. t. and Erez, J. (1984). In situ growth rate experiments on the symbiont-bearing foraminifera Amphistegina lobifera and Amphisorus hemprichii. *Journal of Foraminiferal Research*, 14(4):262–276.

Lamanna, C., Blonder, B., Violle, C., Kraft, N. J. B., Sandel, B., imova, I., Donoghue, J. C., Svenning, J.-C., McGill, B. J., Boyle, B., Buzzard, V., Dolins, S., Jorgensen, P. M., Marcuse-Kubitza, A., Morueta-Holme, N., Peet, R. K., Piel, W. H., Regetz, J., Schildhauer, M., Spencer, N., Thiers, B., Wiser, S. K., and Enquist, B. J. (2014). Functional trait space and the latitudinal diversity gradient. *Proceedings of the National Academy of Sciences*, 111(38):13745–13750.

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., and Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, 28(2):420–430. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.14920.

Lannes, R., Olsson-Francis, K., Lopez, P., and Bapteste, E. (2019). Carbon Fixation by Marine Ultrasmall Prokaryotes. *Genome Biology and Evolution*, 11(4):1166–1177. Publisher: Oxford Academic.

Le Gac, M., Metegnier, G., Chomérat, N., Malestroit, P., Quéré, J., Bouchez, O., Siano, R., Destombe, C., Guillou, L., and Chapelle, A. (2016). Evolutionary processes and cellular functions underlying divergence in Alexandrium minutum. *Molecular Ecology*, 25(20):5129–5143. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.13815.

Le Quéré, C., Andrew, R. M., Friedlingstein, P., Sitch, S., Hauck, J., Pongratz, J., Pickers, P. A., Korsbakken, J. I., Peters, G. P., Canadell, J. G., Arneth, A., Arora, V. K., Barbero, L., Bastos, A., Bopp, L., Chevallier, F., Chini, L. P., Ciais, P., Doney, S. C., Gkritzalis, T., Goll, D. S., Harris, I., Haverd, V., Hoffman, F. M., Hoppema, M., Houghton, R. A., Hurtt, G., Ilyina, T., Jain, A. K., Johannessen, T., Jones, C. D., Kato, E., Keeling, R. F., Goldewijk, K. K., Landschützer, P., Lefèvre, N., Lienert, S., Liu, Z., Lombardozzi, D., Metzl, N., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-i., Neill, C., Olsen, A., Ono, T., Patra, P., Peregon, A., Peters, W., Peylin, P., Pfeil, B., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rocher, M., Rödenbeck, C., Schuster, U., Schwinger, J., Séférian, R., Skjelvan, I., Steinhoff, T., Sutton, A., Tans, P. P.,

Tian, H., Tilbrook, B., Tubiello, F. N., Laan-Luijkx, I. T. v. d., Werf, G. R. v. d., Viovy, N., Walker, A. P., Wiltshire, A. J., Wright, R., Zaehle, S., and Zheng, B. (2018). Global Carbon Budget 2018. *Earth System Science Data*, 10(4):2141–2194. Publisher: Copernicus GmbH.

Le Quéré, C., Harrison, S. P., Colin Prentice, I., Buitenhuis, E. T., Aumont, O., Bopp, L., Claustre, H., Cotrim Da Cunha, L., Geider, R., Giraud, X., and others (2005). Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology*, 11(11):2016–2040.

Lee, K. S., Palatinszky, M., Pereira, F. C., Nguyen, J., Fernandez, V. I., Mueller, A. J., Menolascina, F., Daims, H., Berry, D., Wagner, M., and Stocker, R. (2019). An automated Raman-based platform for the sorting of live cells by functional properties. *Nature Microbiology*, 4(6):1035–1048. Number: 6 Publisher: Nature Publishing Group.

Legendre, L., Rivkin, R. B., Weinbauer, M. G., Guidi, L., and Uitz, J. (2015). The microbial carbon pump concept: Potential biogeochemical significance in the globally changing ocean. *Progress in Oceanography*, 134:432–450.

Legendre, P. and Legendre, L. F. J. (1998). *Numerical Ecology*. Elsevier Science. Google-Books-ID: KBoHuoNRO5MC.

Leles, S. G., Mitra, A., Flynn, K. J., Stoecker, D. K., Hansen, P. J., Calbet, A., McManus, G. B., Sanders, R. W., Caron, D. A., Not, F., Hallegraeff, G. M., Pitta, P., Raven, J. A., Johnson, M. D., Glibert, P. M., and Våge, S. (2017). Oceanic protists with different forms of acquired phototrophy display contrasting biogeographies and abundance. *Proceedings of the Royal Society B: Biological Sciences*, 284(1860):20170664.

Leles, S. G., Mitra, A., Flynn, K. J., Tillmann, U., Stoecker, D., Jeong, H. J., Burkholder, J., Hansen, P. J., Caron, D. A., Glibert, P. M., Hallegraeff, G., Raven, J. A., Sanders, R. W., and Zubkov, M. (2019). Sampling bias misrepresents the biogeographical significance of constitutive mixotrophs across global oceans. *Global Ecology and Biogeography*, 28(4):418–428. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/geb.12853.

Leles, S. G., Valentin, J. L., and Figueiredo, G. M. (2016). Evaluation of the complexity and performance of marine planktonic trophic models. *Anais da Academia Brasileira de Ciências*, 88(3):1971–1991.

Levine, N. M., Varaljay, V. A., Toole, D. A., Dacey, J. W. H., Doney, S. C., and Moran, M. A. (2012). Environmental, biochemical and genetic drivers of DMSP degradation and DMS production in the Sargasso Sea. *Environmental Microbiology*, 14(5):1210–1223. _eprint: https://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1462-2920.2012.02700.x.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*. arXiv: 1303.3997.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. 2:5.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J. C., Roux, S., Vincent, F., and others (2015). Determinants of community structure in the global plankton interactome. *Science*, 348(6237):1262073.

Lindeque, P. K., Parry, H. E., Harmer, R. A., Somerfield, P. J., and Atkinson, A. (2013). Next Generation Sequencing Reveals the Hidden Diversity of Zooplankton Assemblages. *PLOS ONE*, 8(11):e81327.

Litchman, E., de Tezanos Pinto, P., Edwards, K. F., Klausmeier, C. A., Kremer, C. T., and Thomas, M. K. (2015a). Global biogeochemical impacts of phytoplankton: a trait-based perspective. *Journal of Ecology*, 103(6):1384–1396.

Litchman, E., Edwards, K. F., and Klausmeier, C. A. (2015b). Microbial resource utilization traits and trade-offs: implications for community structure, functioning, and biogeochemical impacts at present and in the future. *Frontiers in Microbiology*, 6. Publisher: Frontiers.

Litchman, E. and Klausmeier, C. A. (2008). Trait-Based Community Ecology of Phytoplankton. *Annual Review of Ecology, Evolution, and Systematics*, 39(1):615–639. _eprint: https://doi.org/10.1146/annurev.ecolsys.39.110707.173549.

Litchman, E., Klausmeier, C. A., Schofield, O. M., and Falkowski, P. G. (2007). The role of functional traits and trade-offs in structuring phytoplankton communities: scaling from cellular to ecosystem level. *Ecology Letters*, 10(12):1170–1181. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1461-0248.2007.01117.x.

Litchman, E., Ohman, M. D., and Kiørboe, T. (2013). Trait-based approaches to zooplankton communities. *Journal of Plankton Research*, 35(3):473–484. Publisher: Oxford Academic.

Liu, H., Aris-Brosou, S., Probert, I., and de Vargas, C. (2010). A Time line of the Environmental Genetics of the Haptophytes. *Molecular Biology and Evolution*, 27(1):161–176.

Liu, Z., Campbell, V., Heidelberg, K. B., and Caron, D. A. (2016). Gene expression characterizes different nutritional strategies among three mixotrophic protists. *FEMS Microbiology Ecology*, 92(7).

Lobb, B., Kurtz, D. A., Moreno-Hagelsieb, G., and Doxey, A. C. (2015). Remote homology and the functions of metagenomic dark matter. *Frontiers in Genetics*, 6. Publisher: Frontiers.

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Gobet, A., Kooistra, W. H. C. F., Mahé, F., Not, F., Ogata, H., Pawlowski, J., Pernice, M. C., Romac, S., Shalchian-Tabrizi, K., Simon, N., Stoeck, T., Santini, S., Siano, R., Wincker, P., Zingone, A., Richards, T. A., de Vargas, C., and Massana, R. (2014). Patterns of Rare and Abundant Marine Microbial Eukaryotes. *Current Biology*, 24(8):813–821. Publisher: Elsevier.

Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemmann, L., Sosik, H. M., Schulz, J., Romagnan, J.-B., Picheral, M., Pearlman, J., Ohman, M. D., Niehoff, B., Möller, K. O.,

Miloslavich, P., Lara-Lpez, A., Kudela, R., Lopes, R. M., Kiko, R., Karp-Boss, L., Jaffe, J. S., Iversen, M. H., Irisson, J.-O., Fennel, K., Hauss, H., Guidi, L., Gorsky, G., Giering, S. L. C., Gaube, P., Gallager, S., Dubelaar, G., Cowen, R. K., Carlotti, F., Briseño-Avena, C., Berline, L., Benoit-Bird, K., Bax, N., Batten, S., Ayata, S. D., Artigas, L. F., and Appeltans, W. (2019). Globally Consistent Quantitative Observations of Planktonic Ecosystems. *Frontiers in Marine Science*, 6. Publisher: Frontiers.

Longhurst, A., Sathyendranath, S., Platt, T., and Caverhill, C. (1995). An estimate of global primary production in the ocean from satellite radiometer data. *Journal of Plankton Research*, 17(6):1245–1271. Publisher: Oxford Academic.

Longhurst, A. R. (1998). *Ecological Geography of the Sea*. Academic Press. Google-Books-ID: MFHK18F5aCsC.

Lopez, P., Halary, S., and Bapteste, E. (2015). Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life. *Biology Direct*, 10.

Louca, S., Hawley, A. K., Katsev, S., Torres-Beltran, M., Bhatia, M. P., Kheirandish, S., Michiels, C. C., Capelle, D., Lavik, G., Doebeli, M., Crowe, S. A., and Hallam, S. J. (2016a). Integrating biogeochemistry with multiomic sequence information in a model oxygen minimum zone. *Proceedings of the National Academy of Sciences*, 113(40):E5925–E5933.

Louca, S., Jacques, S. M. S., Pires, A. P. F., Leal, J. S., Srivastava, D. S., Parfrey, L. W., Farjalla, V. F., and Doebeli, M. (2016b). High taxonomic variability despite stable functional structure across microbial communities. *Nature Ecology & Evolution*, 1(1):1–12.

Louca, S., Wegener Parfrey, L., and Doebeli, M. (2016c). Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277.

Lowe, C. D., Martin, L. E., Montagnes, D. J. S., and Watts, P. C. (2012). A legacy of contrasting spatial genetic structure on either side of the Atlantic-Mediterranean transition zone in a marine protist. *Proceedings of the National Academy of Sciences*, 109(51):20998–21003.

Lowe, T. M. and Chan, P. P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research*, 44(W1):W54–W57. Publisher: Oxford Academic.

Luo, J. Y., Irisson, J.-O., Graham, B., Guigand, C., Sarafraz, A., Mader, C., and Cowen, R. K. (2018). Automated plankton image analysis using convolutional neural networks. *Limnology and Oceanography: Methods*, 16(12):814–827. _eprint: https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.1002/lom3.10285.

Lv, J., Zhao, F., Feng, J., Liu, Q., Nan, F., Liu, X., and Xie, S. (2019). Transcriptomic analysis reveals the mechanism on the response of Chlorococcum sp. GD to glucose concentration in mixotrophic cultivation. *Bioresource Technology*, 288:121568.

Lynch, M. D. J. and Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology*, 13(4):217–229.

Lévy, M., Jahn, O., Dutkiewicz, S., and Follows, M. J. (2014). Phytoplankton diversity and community structure affected by oceanic dispersal and mesoscale turbulence: Dispersal Impact on Plankton Diversity. *Limnology and Oceanography: Fluids and Environments*, 4(1):67–84.

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

Mangot, J.-F., Logares, R., Sánchez, P., Latorre, F., Seeleuthner, Y., Mondy, S., Sieracki, M. E., Jaillon, O., Wincker, P., Vargas, C. d., and Massana, R. (2017). Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Scientific Reports*, 7(1):41498. Number: 1 Publisher: Nature Publishing Group.

Marchetti, A., Schruth, D. M., Durkin, C. A., Parker, M. S., Kodner, R. B., Berthiaume, C. T., Morales, R., Allen, A. E., and Armbrust, E. V. (2012). Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences*, 109(6):E317–E325. Publisher: National Academy of Sciences Section: PNAS Plus.

Margalef, R. (1978). Life-forms of phytoplankton as survival alternatives in an unstable environment. *Oceanologica acta*, 1(4):493–509. Publisher: Gauthier-Villars.

Martiny, A. C., Vrugt, J. A., and Lomas, M. W. (2014). Concentrations and ratios of particulate organic carbon, nitrogen, and phosphorus in the global ocean. *Scientific Data*, 1(1):1–7. Number: 1 Publisher: Nature Publishing Group.

Martínez-Pérez, C., Mohr, W., Löscher, C. R., Dekaezemacker, J., Littmann, S., Yilmaz, P., Lehnen, N., Fuchs, B. M., Lavik, G., Schmitz, R. A., LaRoche, J., and Kuypers, M. M. M. (2016). The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nature Microbiology*, 1(11):1–7. Number: 11 Publisher: Nature Publishing Group.

Mason, N. W. H., Mouillot, D., Lee, W. G., and Wilson, J. B. (2005). Functional richness, functional evenness and functional divergence: the primary components of functional diversity. *Oikos*, 111(1):112–118. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0030-1299.2005.13886.x.

McGill, B. J., Enquist, B. J., Weiher, E., and Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution*, 21(4):178–185.

McKie-Krisberg, Z. M., Sanders, R. W., and Gast, R. J. (2018). Evaluation of Mixotrophy-Associated Gene Expression in Two Species of Polar Marine Algae. *Frontiers in Marine Science*, 5. Publisher: Frontiers.

Members of the Complex Trait Consortium (2003). The nature and identification of quantitative trait loci: a community's view. *Nature Reviews Genetics*, 4(11):911–916. Number: 11 Publisher: Nature Publishing Group.

Meng, A., Corre, E., Peterlongo, P., Marchet, C., Alberti, A., Silva, C. D., Wincker, P., Probert, I., Suzuki, N., Crom, S. L., Bittner, L., and Not, F. (2017). A transcriptomic approach to study marine plankton holobionts.

Meng, A., Corre, E., Probert, I., Gutierrez-Rodriguez, A., Siano, R., Annamale, A., Alberti, A., Silva, C. D., Wincker, P., Crom, S. L., Not, F., and Bittner, L. (2018). Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network. *Molecular Ecology*, 27(10):2365–2380.

Mevarech, M., Rice, D., and Haselkorn, R. (1980). Nucleotide sequence of a cyanobacterial nifH gene coding for nitrogenase reductase. *Proceedings of the National Academy of Sciences*, 77(11):6476–6480.

Michaels, A. F., Caron, D. A., Swanberg, N. R., Howse, F. A., and Michaels, C. M. (1995). Planktonic sarcodines (Acantharia, Radiolaria, Foraminifera) in surface waters near Bermuda: abundance, biomass and vertical flux. *Journal of Plankton Research*, 17(1):131–163. Publisher: Oxford Academic.

Miki, T., Yokokawa, T., Nagata, T., and Yamamura, N. (2008). Immigration of prokaryotes to local environments enhances remineralization efficiency of sinking particles: a metacommunity model. *Marine Ecology Progress Series*, 366:1–14.

Millette, N. C., Grosse, J., Johnson, W. M., Jungbluth, M. J., and Suter, E. A. (2018). Hidden in plain sight: The importance of cryptic interactions in marine plankton. *Limnology and Oceanography Letters*, 3(4):341–356.

Mitra, A., Flynn, K. J., Burkholder, J. M., Berge, T., Calbet, A., Raven, J. A., Granéli, E., Glibert, P. M., Hansen, P. J., Stoecker, D. K., Thingstad, F., Tillmann, U., Våge, S., Wilken, S., and Zubkov, M. V. (2014). The role of mixotrophic protists in the biological carbon pump. *Biogeosciences*, 11(4):995–1005.

Mitra, A., Flynn, K. J., Tillmann, U., Raven, J. A., Caron, D., Stoecker, D. K., Not, F., Hansen, P. J., Hallegraeff, G., Sanders, R., Wilken, S., McManus, G., Johnson, M., Pitta, P., Våge, S., Berge, T., Calbet, A., Thingstad, F., Jeong, H. J., Burkholder, J., Glibert, P. M., Granéli, E., and Lundgren, V. (2016). Defining Planktonic Protist Functional Groups on Mechanisms for Energy and Nutrient Acquisition: Incorporation of Diverse Mixotrophic Strategies. *Protist*, 167(2):106–120.

Mock, T., Daines, S. J., Geider, R., Collins, S., Metodiev, M., Millar, A. J., Moulton, V., and Lenton, T. M. (2016). Bridging the gap between omics and earth system science to better understand how environmental change impacts marine microbes. *Global Change Biology*, 22(1):61–75.

Mock, T., Otillar, R. P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B. J., Allen, A. E., Dupont, C. L., Frickenhaus, S., Maumus, F., Veluchamy, A., Wu, T., Barry, K. W., Falciatore, A., Ferrante, M. I., Fortunato, A. E., Glöckner, G., Gruber, A., Hipkin, R., Janech, M. G., Kroth, P. G., Leese, F., Lindquist, E. A., Lyon, B. R.,

Martin, J., Mayer, C., Parker, M., Quesneville, H., Raymond, J. A., Uhlig, C., Valas, R. E., Valentin, K. U., Worden, A. Z., Armbrust, E. V., Clark, M. D., Bowler, C., Green, B. R., Moulton, V., van Oosterhout, C., and Grigoriev, I. V. (2017). Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. *Nature*, 541(7638):536–540. Number: 7638 Publisher: Nature Publishing Group.

Mock, T., Samanta, M. P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K., Durkin, C., BonDurant, S. S., Richmond, K., Rodesch, M., Kallas, T., Huttlin, E. L., Cerrina, F., Sussman, M. R., and Armbrust, E. V. (2008). Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. *Proceedings of the National Academy of Sciences*, 105(5):1579–1584.

Monferrer, N. L., Boltovskoy, D., Tréguer, P., Sandin, M. M., Not, F., and Leynaert, A. (2020). Estimating Biogenic Silica Production of Rhizaria in the Global Ocean. *Global Biogeochemical Cycles*, 34(3):e2019GB006286. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019GB006286.

Moore, J. K., Doney, S. C., Glover, D. M., and Fung, I. Y. (2001a). Iron cycling and nutrient-limitation patterns in surface waters of the World Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography*, 49(1):463–507.

Moore, J. K., Doney, S. C., Kleypas, J. A., Glover, D. M., and Fung, I. Y. (2001b). An intermediate complexity marine ecosystem model for the global domain. *Deep Sea Research Part II: Topical Studies in Oceanography*, 49(1):403–462.

Moorthi, S., Caron, D. A., Gast, R. J., and Sanders, R. W. (2009). Mixotrophy: a widespread and important ecological strategy for planktonic and sea-ice nanoflagellates in the Ross Sea, Antarctica. *Aquatic Microbial Ecology*, 54(3):269–277.

Moran, M. A., Reisch, C. R., Kiene, R. P., and Whitman, W. B. (2012). Genomic Insights into Bacterial DMSP Transformations. *Annual Review of Marine Science*, 4(1):523–542. _eprint: https://doi.org/10.1146/annurev-marine-120710-100827.

Moran, M. A., Satinsky, B., Gifford, S. M., Luo, H., Rivers, A., Chan, L.-K., Meng, J., Durham, B. P., Shen, C., Varaljay, V. A., Smith, C. B., Yager, P. L., and Hopkinson, B. M. (2013). Sizing up metatranscriptomics. *The ISME Journal*, 7(2):237–243.

Morel, A., Huot, Y., Gentili, B., Werdell, P. J., Hooker, S. B., and Franz, B. A. (2007). 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 Sensing of Environment*, 111(1):69–88.

Naeem, S., Duffy, J. E., and Zavaleta, E. (2012). The Functions of Biological Diversity in an Age of Extinction. *Science*, 336(6087):1401–1406. Publisher: American Association for the Advancement of Science Section: Review.

Nalbantoglu, O. U., Way, S. F., Hinrichs, S. H., and Sayood, K. (2011). RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformatics*, 12:41.

Needham, D. M., Poirier, C., Hehenberger, E., Jiménez, V., Swalwell, J. E., Santoro, A. E., and Worden, A. Z. (2019). Targeted metagenomic recovery of four divergent viruses reveals shared and distinctive characteristics of giant viruses of marine eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1786):20190086. Publisher: Royal Society.

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D. R., Gautier, L., Pedersen, A. G., Le Chatelier, E., Pelletier, E., Bonde, I., Nielsen, T., Manichanh, C., Arumugam, M., Batto, J.-M., Quintanilha dos Santos, M. B., Blom, N., Borruel, N., Burgdorf, K. S., Boumezbeur, F., Casellas, F., Doré, J., Dworzynski, P., Guarner, F., Hansen, T., Hildebrand, F., Kaas, R. S., Kennedy, S., Kristiansen, K., Kultima, J. R., Léonard, P., Levenez, F., Lund, O., Moumen, B., Le Paslier, D., Pons, N., Pedersen, O., Prifti, E., Qin, J., Raes, J., Sørensen, S., Tap, J., Tims, S., Ussery, D. W., Yamada, T., Renault, P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S., and Ehrlich, S. D. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8):822–828. Number: 8 Publisher: Nature Publishing Group.

Niklas, J., Schneider, K., and Heinzle, E. (2010). Metabolic flux analysis in eukaryotes. *Current Opinion in Biotechnology*, 21(1):63–69.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2017). *vegan: Community Ecology Package*.

Onuma, R., Hirooka, S., Kanesaki, Y., Fujiwara, T., Yoshikawa, H., and Miyagishima, S.-y. (2020). Changes in the transcriptome, ploidy, and optimal light intensity of a cryptomonad upon integration into a kleptoplastic dinoflagellate. *The ISME Journal*, 14(10):2407–2423. Number: 10 Publisher: Nature Publishing Group.

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

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., and Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2(11):1533–1542.

Parr, C. S., Schulz, K. S., Hammock, J., Wilson, N., Leary, P., Rice, J., and Corrigan Jr., R. J. (2016). TraitBank: Practical semantics for organism attribute data. *Semantic Web*, 7(6):577–588. Publisher: IOS Press.

Partensky, F., Hess, W. R., and Vaulot, D. (1999). Prochlorococcus, a Marine Photosynthetic Prokaryote of Global Significance. *Microbiology and Molecular Biology Reviews*, 63(1):106–127.

Patarnello, T., Volckaert, F. a. M. J., and Castilho, R. (2007). Pillars of Hercules: is the Atlantic–Mediterranean transition a phylogeographical break? *Molecular Ecology*, 16(21):4426–4444.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419.

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S. S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A. M., Gile, G. H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P. J., Kostka, M., Kudryavtsev, A., Lara, E., Lukeš, J., Mann, D. G., Mitchell, E. A. D., Nitsche, F., Romeralo, M., Saunders, G. W., Simpson, A. G. B., Smirnov, A. V., Spouge, J. L., Stern, R. F., Stoeck, T., Zimmermann, J., Schindel, D., and Vargas, C. d. (2012). CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *PLOS Biology*, 10(11):e1001419.

Peltonen, L., Palotie, A., and Lange, K. (2000). Use of population isolates for mapping complex traits. *Nature Reviews Genetics*, 1(3):182–190. Number: 3 Publisher: Nature Publishing Group.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Bescot, N. L., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., Dimier, C., and Searson, S. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, 2:150023.

Planes, S., Allemand, D., Agostini, S., Banaigs, B., Boissin, E., Boss, E., Bourdin, G., Bowler, C., Douville, E., Flores, J. M., Forcioli, D., Furla, P., Galand, P. E., Ghiglione, J.-F., Gilson, E., Lombard, F., Moulin, C., Pesant, S., Poulain, J., Reynaud, S., Romac, S., Sullivan, M. B., Sunagawa, S., Thomas, O. P., Troublé, R., Vargas, C. d., Thurber, R. V., Voolstra, C. R., Wincker, P., Zoccola, D., and Consortium, t. T. P. (2019). The Tara Pacific expedition—A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean. *PLOS Biology*, 17(9):e3000483. Publisher: Public Library of Science.

Pousti, M., Zarabadi, M. P., Amirdehi, M. A., Paquet-Mercier, F., and Greener, J. (2018). Microfluidic bioanalytical flow cells for biofilm studies: a review. *Analyst*, 144(1):68–86. Publisher: The Royal Society of Chemistry.

Probert, I., Siano, R., Poirier, C., Decelle, J., Biard, T., Tuji, A., Suzuki, N., Not, F., and Lane, C. (2014). Brandtodinium gen. nov. and B. nutricula comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. *Journal of Phycology*, 50(2):388–399.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596. Publisher: Oxford Academic.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramond, P., Siano, R., and Sourisseau, M. (2018). Functional traits of marine protists. type: dataset.

Ramond, P., Sourisseau, M., Simon, N., Romac, S., Schmitt, S., Rigaut-Jalabert, F., Henry, N., Vargas, C. d., and Siano, R. (2019). Coupling between taxonomic and functional diversity in protistan coastal communities. *Environmental Microbiology*, 21(2):730–749. _eprint: https://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.14537.

Rappé, M. S. and Giovannoni, S. J. (2003). The Uncultured Microbial Majority. *Annual Review of Microbiology*, 57(1):369–394. _eprint: https://doi.org/10.1146/annurev.micro.57.030502.090759.

Redfield, A. (1934). James Johnstone memorial volume. *University of Liverpool Press, Liverpool UK*, pages 176–192.

Reed, D. C., Algar, C. K., Huber, J. A., and Dick, G. J. (2014). Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proceedings of the National Academy of Sciences*, 111(5):1879–1884.

Reimers, A.-M., Knoop, H., Bockmayr, A., and Steuer, R. (2017). Cellular trade-offs and optimal resource allocation during cyanobacterial diurnal growth. *Proceedings of the National Academy of Sciences*, 114(31):E6457–E6465. Publisher: National Academy of Sciences Section: PNAS Plus.

Reu, B., Proulx, R., Bohn, K., Dyke, J. G., Kleidon, A., Pavlick, R., and Schmidtlein, S. (2011). The role of climate and plant functional trade-offs in shaping global biome and biodiversity patterns. *Global Ecology and Biogeography*, 20(4):570–581. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1466-8238.2010.00621.x.

Reuter, J., Spacek, D. V., and Snyder, M. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4):586–597.

Rhodes, L. and Burke, B. (1996). Morphology and growth characteristics of *Chrysochromulina* species (Haptophyceae = Prymnesiophyceae) isolated from New Zealand coastal waters. *New Zealand Journal of Marine and Freshwater Research*, 30(1):91–103.

Richter, D. J., Watteaux, R., Vannier, T., Leconte, J., Frémont, P., Reygondeau, G., Maillet, N., Henry, N., Benoit, G., Fernàndez-Guerra, A., Suweis, S., Narci, R., Berney, C., Eveillard, D., Gavory, F., Guidi, L., Labadie, K., Mahieu, E., Poulain, J., Romac, S., Roux, S., Dimier, C., Kandels, S., Picheral, M., Searson, S., Coordinators, T. O., Pesant, S., Aury, J.-M., Brum, J. R., Lemaitre, C., Pelletier, E., Bork, P., Sunagawa, S., Karp-Boss, L., Bowler, C., Sullivan, M. B., Karsenti, E., Mariadassou, M., Probert, I., Peterlongo, P., Wincker, P., Vargas, C. d., d'Alcalà, M. R., Iudicone, D., Jaillon, O., and Coordinators, T. O. (2019). Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *bioRxiv*, page 867739.

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). Metagenomics: Genomic Analysis of Microbial Communities. *Annual Review of Genetics*, 38(1):525–552. _eprint: https://doi.org/10.1146/annurev.genet.38.072902.091216.

Riley, G. A. (1946). Factors controlling phytoplankton populations on Georges Bank. *J. mar. Res*, 6(1):54–73.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P., and Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437. Number: 7459 Publisher: Nature Publishing Group.

Rizzolo, K., Cohen, S. E., Weitz, A. C., López Muñoz, M. M., Hendrich, M. P., Drennan, C. L., and Elliott, S. J. (2019). A widely distributed diheme enzyme from Burkholderia that displays an atypically stable bis -Fe(IV) state. *Nature Communications*, 10(1):1–10.

Rombouts, I., Beaugrand, G., Ibañez, F., Gasparini, S., Chiba, S., and Legendre, L. (2009). Global latitudinal variations in marine copepod diversity and environmental factors. *Proceedings of the Royal Society B: Biological Sciences*, 276(1670):3053–3062. Publisher: Royal Society.

Routtu, J., Hall, M. D., Albere, B., Beisel, C., Bergeron, R. D., Chaturvedi, A., Choi, J.-H., Colbourne, J., De Meester, L., Stephens, M. T., Stelzer, C.-P., Solorzano, E., Thomas, W. K., Pfrender, M. E., and Ebert, D. (2014). An SNP-based second-generation genetic map of Daphnia magna and its application to QTL analysis of phenotypic traits. *BMC Genomics*, 15(1):1033.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y.-H., Falcón, L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R., Strausberg, R. L., Nealson, K., Friedman, R., Frazier, M., and Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology*, 5(3).

Ruvkun, G. B. and Ausubel, F. M. (1980). Interspecies homology of nitrogenase genes. *Proceedings of the National Academy of Sciences*, 77(1):191–195.

Safi, K. A. and Hall, J. A. (1999). Mixotrophic and heterotrophic nanoflagellate grazing in the convergence zone east of New Zealand. *Aquatic Microbial Ecology*, 20(1):83–93.

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., Field, C. M., Coelho, L. P., Cruaud, C., Engelen, S., Gregory, A. C., Labadie, K., Marec, C., Pelletier, E., Royo-Llonch, M., Roux, S., Sánchez, P., Uehara, H., Zayed, A. A., Zeller, G., Carmichael, M., Dimier, C., Ferland, J., Kandels, S., Picheral, M., Pisarev, S., Poulain, J., Acinas, S. G., Babin, M.,

Bork, P., Boss, E., Bowler, C., Cochrane, G., Vargas, C. d., Follows, M., Gorsky, G., Grimsley, N., Guidi, L., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Karp-Boss, L., Karsenti, E., Not, F., Ogata, H., Pesant, S., Poulton, N., Raes, J., Sardet, C., Speich, S., Stemmann, L., Sullivan, M. B., Sunagawa, S., Wincker, P., Acinas, S. G., Babin, M., Bork, P., Bowler, C., Vargas, C. d., Guidi, L., Hingamp, P., Iudicone, D., Karp-Boss, L., Karsenti, E., Ogata, H., Pesant, S., Speich, S., Sullivan, M. B., Wincker, P., and Sunagawa, S. (2019). Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell*, 179(5):1068–1083.e21.

Sanders, R. W. and Gast, R. J. (2012). Bacterivory by phototrophic picoplankton and nanoplankton in Arctic waters. *FEMS Microbiology Ecology*, 82(2):242–253.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467. Publisher: National Academy of Sciences Section: Biological Sciences: Biochemistry.

Santoferrara, L. F. (2019). Current practice in plankton metabarcoding: optimization and error management. *Journal of Plankton Research*, 41(5):571–582. Publisher: Oxford Academic.

Santoro, A. E., Richter, R. A., and Dupont, C. L. (2019). Planktonic Marine Archaea. *Annual Review of Marine Science*, 11(1):131–158. _eprint: https://doi.org/10.1146/annurev-marine-121916-063141.

Sarmiento, J. L., Slater, R. D., Fasham, M. J. R., Ducklow, H. W., Toggweiler, J. R., and Evans, G. T. (1993). A seasonal three-dimensional ecosystem model of nitrogen cycling in the North Atlantic Euphotic Zone. *Global Biogeochemical Cycles*, 7(2):417–450. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/93GB00375.

Sauterey, B., Ward, B., Rault, J., Bowler, C., and Claessen, D. (2017). The Implications of Eco-Evolutionary Processes for the Emergence of Marine Plankton Community Biogeography. *The American Naturalist*, 190(1):116–130. Publisher: The University of Chicago Press.

Sauterey, B., Ward, B. A., Follows, M. J., Bowler, C., and Claessen, D. (2015). When everything is not everywhere but species evolve: an alternative method to model adaptive properties of marine ecosystems. *Journal of Plankton Research*, 37(1):28–47. Publisher: Oxford Academic.

Schneider, F. D., Fichtmueller, D., Gossner, M. M., Güntsch, A., Jochum, M., König-Ries, B., Provost, G. L., Manning, P., Ostrowski, A., Penone, C., and Simons, N. K. (2019). Towards an ecological trait-data standard. *Methods in Ecology and Evolution*, 10(12):2006–2019. _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13288.

Schoemann, V., Becquevort, S., Stefels, J., Rousseau, V., and Lancelot, C. (2005). Phaeocystis blooms in the global ocean and their controlling mechanisms: a review. *Journal of Sea Research*, 53(1):43–66.

Schoener, D. M. and McManus, G. B. (2012). Plastid retention, use, and replacement in a kleptoplastidic ciliate. *Aquatic Microbial Ecology*, 67(3):177–187.

Selosse, M.-A., Charpin, M., Not, F., and Jeyasingh, P. (2017). Mixotrophy everywhere on land and in water: the grand écart hypothesis. *Ecology Letters*, 20(2):246–263.

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., Rosa, S. L. L., Solden, L. M., Liu, P., Narrowe, A. B., Rodríguez-Ramos, J., Bolduc, B., Gazitua, M. C., Daly, R. A., Smith, G. J., Vik, D. R., Pope, P. B., Sullivan, M. B., Roux, S., and Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. *bioRxiv*, page 2020.06.29.177501. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Sharma, S. and Steuer, R. (2019). Modelling microbial communities using biochemical resource allocation analysis. *Journal of The Royal Society Interface*, 16(160):20190474. Publisher: Royal Society.

Shaw, G. E. (1983). Bio-controlled thermostasis involving the sulfur cycle. *Climatic Change*, 5(3):297–303.

Shin, Y., Lee, H., Lee, Y.-J., Seo, D. K., Jeong, B., Hong, S., Kim, J., Kim, T., Lee, J.-K., and Heo, T.-Y. (2019). The Prediction of Diatom Abundance by Comparison of Various Machine Learning Methods. ISSN: 1024-123X Library Catalog: www.hindawi.com Pages: e5749746 Publisher: Hindawi Volume: 2019.

Shuter, B. (1979). A model of physiological adaptation in unicellular algae. *Journal of Theoretical Biology*, 78(4):519–552.

Siegenthaler, U. and Sarmiento, J. L. (1993). Atmospheric carbon dioxide and the ocean. *Nature*, 365(6442):119–125. Number: 6442 Publisher: Nature Publishing Group.

Simó, R. (2001). Production of atmospheric sulfur by oceanic plankton: biogeochemical, ecological and evolutionary links. *Trends in Ecology & Evolution*, 16(6):287–294.

Simó, R., Archer, S. D., Pedrós-Alió, C., Gilpin, L., and Stelfox-Widdicombe, C. E. (2002). Coupled dynamics of dimethylsulfoniopropionate and dimethylsulfide cycling and the microbial food web in surface waters of the North Atlantic. *Limnology and Oceanography*, 47(1):53–61. _eprint: https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.4319/lo.2002.47.1.0053.

Singh, J., Behal, A., Singla, N., Joshi, A., Birbian, N., Singh, S., Bali, V., and Batra, N. (2009). Metagenomics: Concept, methodology, ecological inference and recent advances. *Biotechnology Journal*, 4(4):480–494. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/biot.200800201.

Sinha, B., Buitenhuis, E. T., Quéré, C. L., and Anderson, T. R. (2010). Comparison of the emergent behavior of a complex ecosystem model in two ocean general circulation models. *Progress in Oceanography*, 84(3):204–224.

Smith, T. and Huston, M. (1990). A theory of the spatial and temporal dynamics of plant communities. In Grabherr, G., Mucina, L., Dale, M. B., and Ter Braak, C. J. F., editors, *Progress in theoretical vegetation science*, Advances in vegetation science, pages 49–69. Springer Netherlands, Dordrecht.

Smith, T. M., Shugart, H. H., Woodward, F. I., and Burton, P. J. (1993). Plant Functional Types. In Solomon, A. M. and Shugart, H. H., editors, *Vegetation Dynamics & Global Change*, pages 272–292. Springer US, Boston, MA.

Snow, G. (2016). *TeachingDemos: Demonstrations for Teaching and Learning*.

Sonah, H., O'Donoughue, L., Cober, E., Rajcan, I., and Belzile, F. (2015). Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnology Journal*, 13(2):211–221. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pbi.12249.

Song, H.-S., Cannon, W. R., Beliaev, A. S., and Konopka, A. (2014). Mathematical Modeling of Microbial Community Dynamics: A Methodological Review. *Processes*, 2(4):711–752. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Sonnewald, M., Dutkiewicz, S., Hill, C., and Forget, G. (2020). Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Science Advances*, 6(22):eaay4740. Publisher: American Association for the Advancement of Science Section: Research Article.

Spero, H. J. and Parker, S. L. (1985). Photosynthesis in the symbiotic planktonic foraminifer Orbulina universa, and its potential contribution to oceanic primary productivity. *Journal of Foraminiferal Research*, 15(4):273–281.

Stec, K. F., Caputi, L., Buttigieg, P. L., D'Alelio, D., Ibarbalz, F. M., Sullivan, M. B., Chaffron, S., Bowler, C., Ribera d'Alcalà, M., and Iudicone, D. (2017). Modelling plankton ecosystems in the meta-omics era. Are we ready? *Marine Genomics*, 32:1–17.

Steele, J. H. (1958). *Plant production in the northern North Sea*. HM Stationery Office.

Stephens, T. G., Ragan, M. A., Bhattacharya, D., and Chan, C. X. (2018). Core genes in diverse dinoflagellate lineages include a wealth of conserved dark genes with unknown functions. *Scientific Reports*, 8(1):17175. Number: 1 Publisher: Nature Publishing Group.

Steuer, R., Knoop, H., and Machne, R. (2012). Modelling cyanobacteria: from metabolism to integrative models of phototrophic growth. *Journal of Experimental Botany*, 63(6):2259–2274.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H.-W., and Richards, T. A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, 19:21–31.

Stoecker, D. K., Hansen, P. J., Caron, D. A., and Mitra, A. (2017). Mixotrophy in the Marine Plankton. *Annual Review of Marine Science*, 9(1):311–335.

Stoecker, D. K., Johnson, M. D., Vargas, C. d., and Not, F. (2009). Acquired phototrophy in aquatic protists. *Aquatic Microbial Ecology*, 57(3):279–310.

Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., and Stahl, D. A. (2007). Metabolic modeling of a mutualistic microbial community. *Molecular Systems Biology*, 3(1):92. Publisher: John Wiley & Sons, Ltd.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans coordinators, Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., Bork, P., Boss, E., Bowler, C., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E. G., Sardet, C., Sieracki, M., and Velayoudon, D. (2015). Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359–1261359.

Swanberg, N. R. (1983). The trophic role of colonial Radiolaria in oligotrophic oceanic environments1,2. *Limnology and Oceanography*, 28(4):655–666. _eprint: https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.4319/lo.1983.28.4.0655.

Tang, W. and Cassar, N. (2019). Data-Driven Modeling of the Distribution of Diazotrophs in the Global Ocean. *Geophysical Research Letters*, 46(21):12258–12269. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019GL084376.

Tang, W., Li, Z., and Cassar, N. (2019). Machine Learning Estimates of Global Marine Nitrogen Fixation. *Journal of Geophysical Research: Biogeosciences*, 124(3):717–730.

Tilman, D., Knops, J., Wedin, D., Reich, P., Ritchie, M., and Siemann, E. (1997). The Influence of Functional Diversity and Composition on Ecosystem Processes. *Science*, 277(5330):1300–1302. Publisher: American Association for the Advancement of Science Section: Report.

Todd, J. D., Rogers, R., Li, Y. G., Wexler, M., Bond, P. L., Sun, L., Curson, A. R. J., Malin, G., Steinke, M., and Johnston, A. W. B. (2007). Structural and Regulatory Genes Required to Make the Gas Dimethyl Sulfide in Bacteria. *Science*, 315(5812):666–669. Publisher: American Association for the Advancement of Science Section: Report.

Toseland, A., Daines, S. J., Clark, J. R., Kirkham, A., Strauss, J., Uhlig, C., Lenton, T. M., Valentin, K., Pearson, G. A., Moulton, V., and Mock, T. (2013). The impact of temperature on marine phytoplankton resource allocation and metabolism. *Nature Climate Change*, 3(11):979–984.

Tringe, S. G. and Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, 6(11):805–814.

Tully, B. J., Graham, E. D., and Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, 5:170203.

Tyler, E. H. M., Somerfield, P. J., Berghe, E. V., Bremner, J., Jackson, E., Langmead, O., Palomares, M. L. D., and Webb, T. J. (2012). Extensive gaps and biases in our knowledge of a well-known fauna: implications for integrating biological traits into macroecology. *Global Ecology and Biogeography*, 21(9):922–934. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1466-8238.2011.00726.x.

Tyrrell, T. (1999). The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature*, 400(6744):525–531. Number: 6744 Publisher: Nature Publishing Group.

Unrein, F., Gasol, J. M., and Massana, R. (2010). Dinobryon faculiferum (Chrysophyta) in coastal Mediterranean seawater: presence and grazing impact on bacteria. *Journal of Plankton Research*, 32(4):559–564.

Vanni, C., Schechter, M., Acinas, S., Barberán, A., Buttigieg, P. L., Casamayor, E. O., Delmont, T. O., Duarte, C. M., Eren, A. M., Finn, R., Mitchell, A., Sanchez, P., Siren, K., Steinegger, M., Glöckner, F. O., and Fernandez-Guerra, A. (2020). Light into the darkness: Unifying the known and unknown coding sequence space in microbiome analyses. *bioRxiv*, page 2020.06.30.180448. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., and Smith, H. O. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74.

Venturelli, O. S., Carr, A. V., Fisher, G., Hsu, R. H., Lau, R., Bowen, B. P., Hromada, S., Northen, T., and Arkin, A. P. (2018). Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular Systems Biology*, 14(6):e8157. Publisher: John Wiley & Sons, Ltd.

Violle, C., Enquist, B. J., McGill, B. J., Jiang, L., Albert, C. H., Hulshof, C., Jung, V., and Messier, J. (2012). The return of the variance: intraspecific variability in community ecology. *Trends in Ecology & Evolution*, 27(4):244–252.

Violle, C., Navas, M.-L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., and Garnier, E. (2007). Let the concept of trait be functional! *Oikos*, 116(5):882–892. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0030-1299.2007.15559.x.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, 101(1):5–22.

Volk, T. and Hoffert, M. I. (1985). Ocean Carbon Pumps: Analysis of Relative Strengths and Efficiencies in Ocean-Driven Atmospheric CO2 Changes. In *The Carbon Cycle and Atmospheric CO2: Natural Variations Archean to Present*, pages 99–110. American Geophysical Union (AGU). _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/GM032p0099.

Vorobev, A., Dupouy, M., Carradec, Q., Delmont, T. O., Annamalé, A., Wincker, P., and Pelletier, E. (2019). Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *bioRxiv*, page 812974.

Vorobev, A., Dupouy, M., Carradec, Q., Delmont, T. O., Annamalé, A., Wincker, P., and Pelletier, E. (2020). Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Research*. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Wang, W.-L., Moore, J. K., Martiny, A. C., and Primeau, F. W. (2019a). Convergent estimates of marine nitrogen fixation. *Nature*, 566(7743):205–211. Number: 7743 Publisher: Nature Publishing Group.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A. G., Park, B.-S., Weisshaar, B., Liu, B., Li, B., Liu, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G. J., Bonnema, G., Tang, H., Wang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Wang, H., Jin, H., Parkin, I. A. P., Batley, J., Kim, J.-S., Just, J., Li, J., Xu, J., Deng, J., Kim, J. A., Li, J., Yu, J., Meng, J., Wang, J., Min, J., Poulain, J., Wang, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M. G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P. J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Zhang, S., Huang, S., Sato, S., Sun, S., Kwon, S.-J., Choi, S.-R., Lee, T.-H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Y., Wang, Z., Li, Z., Wang, Z., Xiong, Z., and Zhang, Z. (2011). The genome of the mesopolyploid crop species Brassica rapa. *Nature Genetics*, 43(10):1035–1039. Number: 10 Publisher: Nature Publishing Group.

Wang, X., Xia, K., Yang, X., and Tang, C. (2019b). Growth strategy of microbes on mixed carbon sources. *Nature Communications*, 10(1):1279. Number: 1 Publisher: Nature Publishing Group.

Ward, B. A., Collins, S., Dutkiewicz, S., Gibbs, S., Bown, P., Ridgwell, A., Sauterey, B., Wilson, J. D., and Oschlies, A. (2019). Considering the Role of Adaptive Evolution in Models of the Ocean and Climate System. *Journal of Advances in Modeling Earth Systems*, 11(11):3343–3361. _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001452.

Ward, B. A., Dutkiewicz, S., Barton, A. D., and Follows, M. J. (2011). Biophysical Aspects of Resource Acquisition and Competition in Algal Mixotrophs. *The American Naturalist*, 178(1):98–112.

Ward, B. A., Dutkiewicz, S., Jahn, O., and Follows, M. J. (2012). A size-structured food-

web model for the global ocean. *Limnology and Oceanography*, 57(6):1877–1891. _eprint: https://aslopubs.onlinelibrary.wiley.com/doi/pdf/10.4319/lo.2012.57.6.1877.

Ward, B. A. and Follows, M. J. (2016). Marine mixotrophy increases trophic transfer efficiency, mean organism size, and vertical carbon flux. *Proceedings of the National Academy of Sciences*, 113(11):2958–2963.

Watson, A. K., Lannes, R., Pathmanathan, J. S., Méheust, R., Karkar, S., Colson, P., Corel, E., Lopez, P., and Bapteste, E. (2019). The Methodology Behind Network Thinking: Graphs to Analyze Microbial Complexity and Evolution. In Anisimova, M., editor, *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, pages 271–308. Springer, New York, NY.

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wit, R. D. and Bouvier, T. (2006). 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology*, 8(4):755–758.

Worden, A. Z., Follows, M. J., Giovannoni, S. J., Wilken, S., Zimmerman, A. E., and Keeling, P. J. (2015). Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science*, 347(6223). Publisher: American Association for the Advancement of Science Section: Review.

Wu, J., Sunda, W., Boyle, E. A., and Karl, D. M. (2000). Phosphate Depletion in the Western North Atlantic Ocean. *Science*, 289(5480):759–762. Publisher: American Association for the Advancement of Science Section: Report.

Wu, S., Xiong, J., and Yu, Y. (2015). Taxonomic Resolutions Based on 18S rRNA Genes: A Case Study of Subclass Copepoda. *PLOS ONE*, 10(6):e0131498.

Wyman, S. K., Avila-Herrera, A., Nayfach, S., and Pollard, K. S. (2018). A most wanted list of conserved microbial protein families with no known domains. *PLOS ONE*, 13(10):e0205749. Publisher: Public Library of Science.

Zak, L. J., Gaustad, A. H., Bolarin, A., Broekhuijse, M. L. W. J., Walling, G. A., and Knol, E. F. (2017). Genetic control of complex traits, with a focus on reproduction in pigs. *Molecular Reproduction and Development*, 84(9):1004–1011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrd.22875.

Zehr, J. P. (2011a). Nitrogen fixation by marine cyanobacteria. *Trends in Microbiology*, 19(4):162–173.

Zehr, J. P. (2011b). Nitrogen fixation by marine cyanobacteria. *Trends in Microbiology*, 19(4):162–173.

Zehr, J. P., Jenkins, B. D., Short, S. M., and Steward, G. F. (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environmental microbiology*, 5(7):539–554.

Zehr, J. P. and Kudela, R. M. (2011). Nitrogen Cycle of the Open Ocean: From Genes to Ecosystems. *Annual Review of Marine Science*, 3(1):197–225. _eprint: https://doi.org/10.1146/annurev-marine-120709-142819.

Zehr, J. P., Mellon, M. T., and Zani, S. (1998). New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (nifH) genes. *Applied and environmental microbiology*, 64(9):3444–3450.

Appendices

Appendix A

Co-authored manuscript: Martini et al. 2020

A.1 Functional trait-based approaches as a common framework for aquatic ecologists

Authors: Séverine Martini[1,2,#], Floriane Larras[3,*], Aurélien Boyé[4,*], Emile Faure[1,5,*], Nicole Aberle[6], Philippe Archambault[7], Lise Bacouillard[8], Beatrix E Beisner[9], Lucie Bittner[5], Emmanuel Castella[8], Michael Danger[11,12], Olivier Gauthier[4], Lee Karp-Boss[13], Fabien Lombard[1,12], Frédéric Maps[7], Lars Stemmann[1], Eric Thiébaut[8], Philippe Usseglio-Polatera[11], Meike Vogt[14], Martin Laviale[11,**], Sakina-Dorothée Ayata[1,5,#,**]

1- Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche, LOV, 06230 Villefranche-sur-mer, France

2- Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110, 13288 Marseille, France

3- Helmholtz Center for Environmental Research, Leipzig, Germany

4- Laboratoire des Sciences de l'Environnement Marin (LEMAR) UMR 6539 CNRS UBO IRD IFREMER, Institut Universitaire Européen de la Mer, Université de Bretagne Occidentale (UBO), Plouzané, France

5- Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'histoire naturelle, CNRS, Sorbonne Université, EPHE, CP 50, 57 rue Cuvier, 75005 Paris, France

6- Norwegian University of Science and Technology (NTNU), Trondhjem Biological Station, Trondheim, Norway

7- Québec-Océan and Unité Mixte Internationale Takuvik Ulaval-CNRS, Département de Biologie, Université Laval, Québec, Canada

8- Sorbonne Université, CNRS, Station Biologique de Roscoff, Laboratoire Adaptation et Diversité en Milieu Marin, Place Georges Teissier, 29680 Roscoff, France

9- Department of Biological Sciences, University of Québec at Montréal, Montréal, Québec, Canada

10- Department F.-A. Forel for Environmental and Aquatic Sciences, Earth and Environmental Science Section and Institute for Environmental Sciences, University of Geneva, Geneva, Switzerland

11- Université de Lorraine Lorraine, CNRS, LIEC, F-57000 Metz, France

12- Institut Universitaire de France Paris, Île-de-France, France

13- School of Marine Sciences, University of Maine, Orono, ME USA

14- Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, Switzerland

#- Corresponding authors

*- Contributed equally

**- Contributed equally

Abstract

Aquatic ecologists face challenges in identifying the general rules of the functioning of ecosystems. A common framework, including freshwater, marine, benthic, and pelagic ecologists is needed to bridge communication gaps and foster knowledge sharing. This framework should transcend local specificities and taxonomy in order to provide a common ground and shareable tools to address common scientific challenges. Here, we advocate the use of functional trait-based approaches (FTBAs) for aquatic ecologists, and propose concrete paths to go forward. Firstly, we propose to unify existing definitions in FTBAs to adopt a common language. Secondly, we list the numerous databases referencing functional traits for aquatic organisms. Thirdly, we present a synthesis on traditional as well as recent promising methods for the study of aquatic functional traits, including imaging and genomics. Finally, we conclude with a highlight on scientific challenges and promising venues for which FTBAs should foster opportunities for future research. By offering practical tools, our framework provides a clear path forward to the adoption of trait-based approaches in aquatic ecology.

Keywords: Functional trait; Marine; Freshwater; Trait-based approaches; Databases: Imaging; Omics; Aquatic ecology; Limnology; Oceanography

A.1.1 Introduction

The aquatic realm encompasses very diverse environments from freshwater ponds, lakes, and rivers to estuaries, salt marshes, mangroves, coasts, continental shelves, deep-seas, marginal seas, and open ocean areas. It plays a major role in the Earth's climate system and supplies important ecosystem services for human populations (Grizzetti et al. 2016). Yet, different aquatic ecosystems are still studied by distinct scientific communities that have limited interactions with each other, as illustrated by the tendency to train graduate students independently, to publish in different journals and to attend distinct conferences (with a few exceptions, such as the Association for the Sciences of Limnology and Oceanography (ASLO) and its conferences and journals, including Limnology and Oceanography). Freshwater and marine ecosystems even belong to different Sustainable Development Goals for the United Nations, with one dedicated to the marine environment (#14: Life below water), and another to terrestrial systems including freshwater (#15: Life on land) (United Nations 2015).

Ecology seeks to understand interactions between organisms and the environment, as well as to identify general rules that elucidate the functioning of ecosystems, to ultimately improve our ability

to predict ecosystem changes (Loreau 2010). In both freshwater and marine environments, and for both pelagic and benthic habitats, the crucial questions remain the same (Heino et al. 2015): (1) What are the processes that control the structure and functioning of aquatic ecosystems? (2) What ecological patterns emerge at various spatio-temporal scales, and what are their key drivers? (3) How will aquatic organisms respond to increasing anthropogenic pressures? Some efforts have been made to integrate aquatic ecology for planktonic (Margalef 1978; Hecky and Kilham 1988; Leibold and Norberg 2004; Litchman and Klausmeier 2008) and benthic (Mermillod-Blondin and Rosenberg 2006) studies. Despite recent efforts to bring the communities together (*e.g.* the AQUASHIFT and DynaTrait projects priority programmes of the German Research Foundation, or such as the bi-annual Trait workshop https://www.traitspace.com/ including limnologists, benthic ecologists, terrestrial ecologists), a unified framework for addressing ecological questions in pelagic and benthic habitats of both environments has been slow to develop. A recent review highlights the potential of trait-based ecology for studying aquatic ecosystems and the need for collaborative approaches among aquatic ecologists was emphasized (Kremer et al. 2017). In addition to bridging the gap between freshwater and marine studies, there is a crucial need to integrate planktonic and benthic studies, especially because of the strong coupling between these two habitats (Griffiths et al. 2017). The present synthesis proposes a practical framework to address these needs.

Indeed, trait-based approaches, defined in ecological research as any method that focuses on individual properties of organisms (so-called traits) rather than species, could provide this common framework (McGill et al. 2006; Kremer et al. 2017). These approaches emerged from terrestrial ecology when attributes at the individual level, initially used to describe ecosystem function based on elements common to multiple species, were considered to gather individuals into functional groups (*i.e.*, "plant functional types") based on their physical, phylogenetic and phenological characteristics, rather than on their taxonomy (*e.g.* species). Trait-based models of aquatic ecosystems can be traced back to the pioneering work of Riley in the 1940's (Riley 1946), who modelled the phytoplankton bloom dynamics in the North Atlantic focusing on the main physiological and biological characteristics of phytoplankton as a group. Since earlier attempts to classify phytoplankton by "life-forms" (Sournia 1982; Reynolds 1988), a similar approach was applied to identify functional groups for freshwater benthic macrofauna (Usseglio-Polatera et al. 2000, 2001), marine benthos (Rigolet et al. 2014), benthic algae (Tapolczai et al. 2016), submerged plants (Willby et al. 2000; Lukács et al. 2019), or marine zooplankton (Benedetti et al. 2016). The underlying assumption is that functional grouping would make it easier to link community ecology to biogeochemical processes and biodiversity to ecosystem functioning (Naeem and Wright 2003). Through the study of functional diversity and functional traits, these approaches allow for the quantitative assessment of community or ecosystem resistance or resilience to changes through functional redundancy (Lavorel and Garnier 2002; McGill et al. 2006) which could potentially enhance generality and predictability in future projections of ecosystem function and service provision than the species-centred or taxonomic approaches (Levine 2016).

In aquatic ecology alone, more than 2,476 articles were published between 1991 and 2018 using the terms "functional trait" or "trait-based" (see Supplementary Information). The percentage of

those publications relative to the total ones published in freshwater and marine ecology (using those terms as keywords in Web of Science) has increased over time. This emergent and still increasing area of research in aquatic ecology has been the topic of several recent reviews, which summarize the state of the knowledge with regard to specific taxonomic or trophic groups, or traits (Litchman and Klausmeier 2008; Litchman et al. 2013; Nock et al. 2016; Meunier et al. 2017; Hébert et al. 2017; Kremer et al. 2017; Beauchard et al. 2017; Degen et al. 2018; Kiørboe et al. 2018). Previous studies focused either on one species (Pardo and Johnson 2005), on one taxonomic group of organisms (*e.g.* crustaceans in Hébert et al. 2016, 2017), on one compartment of the ecosystem (*e.g.* pelagic primary producers in Litchman and Klausmeier (2008); benthic primary producers in Tapolczai et al. (2016); zooplankton in Litchman et al. (2013) and Hébert et al. (2017); stream fish in Frimpong and Angermeier (2010)), on a particular ecosystem (*e.g.* oceans in Barton et al. (2013) and Kiørboe et al. (2018) marine benthos in Degen et al. (2018); running water benthos in Statzner and Bêche (2010)) or even on a single type of trait (*e.g.* size in Andersen et al. (2016) or stoichiometric traits in Meunier et al. (2017)). A network analysis of key words associated with the aquatic trait-based literature highlights differences between studies, both in the terminology used to characterize traits and in the application of trait-based approaches in studies of freshwater and marine systems (Figure A.4).

The goal of this review is to facilitate exchanges of FTBAs and their products across different aquatic fields. To do so, we propose: 1) A table compiling the main definitions of traits that are commonly used in trait-based studies, in addition to recommendations for using a common and unambiguous vocabulary, 2) A unified typology of 40 aquatic functional traits that could be used in multicompartment studies (including several biological compartments, or different habitats *e.g.* sediment and water), 3) A summary of existing databases that contain information on functional traits, 4) A review of traditional and emerging methods for estimating and using traits of aquatic organisms, and 5) The main challenges that aquatic ecologists can now address using FTBAs and that should inspire future studies.

A.1.2 Trait definition and aquatic trait description

The term "trait" depicts specific attributes of an individual that are both inherent and characteristic to its nature. However, as highlighted by our literature survey (Supplementary Figure A.4 and A.5, see also Supplementary Information), this term is used in multiple contexts to describe a diverse set of attributes such as: "physiological traits", "functional traits", "life history traits", "biological traits", "ecological traits", "response traits", "effect traits", "behavioral traits", etc (see Table A.1). To avoid misunderstandings, clear definitions of these concepts are needed (Violle et al. 2007).

A.1.2.1 Adopting common definitions for aquatic FTBAs

Trait definitions vary between scientific communities, from the individual organism (*e.g.* life-history traits) to the population (*e.g.* demographic traits), community (*e.g.* response traits) and the ecosystem scale (*e.g.* effect traits; Hébert et al. 2017). Traits can also be directly measured

Term	Recommended definitions	References	Examples
Trait	Any morphological, physiological or phenological feature measurable at the individual level, from the cell to the whole-organism level, without reference to the environment or any other level of organization.	(Violle et al. 2007)	
Functional trait	Any trait that impacts fitness indirectly via its effects on growth, reproduction and survival.	(Violle et al. 2007)	
Realized trait	Trait actually measured *in situ* or in the laboratory	(Reu et al. 2011)	
Potential trait	Trait described from the literature, usually at the species level, and ideally covering a large variety of environmental conditions.	(Reu et al. 2011)	
Life history traits	Traits referring to life history	(Litchman and Klausmeier 2008)	Type of reproduction (sexual versus asexual) or the ability to form resting stages.
	Traits that are relevant at the individual level	(Hébert et al. 2017)	Fitness-related traits.
Morphological trait	Traits related to the morphology of organisms	(Litchman and Klausmeier 2008)	Cell size, cell shape.
Physiological trait	Traits related to the physiology of organisms	(Litchman and Klausmeier 2008)	Nutrient acquisition, response to light.
Behavioral trait	Traits related to the behavior of organisms	(Litchman and Klausmeier 2008)	Motility.

Table A.1 - Main definitions related to traits in aquatic trait-based studies.

in situ (*e.g.* realized traits) or inferred from the literature (*e.g.* potential traits). Realized traits are ultimately one of the sources for potential traits found in databases (see section A.1.3.1). To establish a unified framework and avoid subjectivity in these definitions, we recommend the use of the definitions that focus on the individual level only. These are the ones proposed by Violle et al. (2007), by Litchman and Klausmeier (2008) and by Reu et al. (2011) are summarized in Table A.1.

Since it is the diversity of organismal functions that structures communities and eventually ecosystems, trait-based approaches should rather refer to "functional traits" (sensu Violle et al. 2007: Any trait that impacts fitness indirectly via its effects on growth, reproduction and survival) than to "traits" and should in fact be called functional trait-based approaches. Functional traits have been further divided into four types: life history traits, morphological traits, physiological traits, and behavioral/mobility traits (Litchman and Klausmeier 2008; Litchman et al. 2013, 2015; Desrosiers et al. 2019). The term "ecological traits" has also been used in the context of "functional traits" to describe the environmental preference of the organisms, especially for benthic ones (*e.g.* Desrosiers et al. 2019). Where "ecological traits" refer to ecological or environmental preferences of organisms, they should rather be called physiological traits (*e.g.* salinity preference/tolerance) or behavioral traits (*e.g.* relationship with the substrate). In contrast, ecological traits referring to taxonomic information, sampling location or habitat features (*e.g.* depth, substratum type) should neither be considered as functional traits nor as traits.

A.1.2.2 Functional traits as a common framework beyond taxonomy to transcend ecosystems

Functional traits provide a "common currency across biological organizational levels and taxonomic groups" (Violle et al. 2014), beyond taxonomic variation and geographic or ecosystemic

peculiarities. Firstly, functional trait-based ecology describes emergent properties related to ecosystem functioning, without necessarily having to explicitly identify the organisms at a given taxonomic level. Secondly, FTBAs in aquatic ecology can account for a continuous degree of plasticity in the trait expressed (Chevenet et al. 1994), thus allowing for a better quantification of intra-specific variability (see also A.1.4.1). Moreover, phenotypic plasticity can result in substantial intra-specific variation (Des Roches et al. 2018), with clonal differences in plasticity. For instance, many aquatic species can exhibit a high degree of morphological plasticity in response to different environmental cues. Zooplankters such as Daphnia can form elongated carapaces (*e.g.* longer tailspines or helmets) in response to strong predation (O'Brien et al. 1979; Lüning 1992; Swaffar and O'Brien 1996) while the freshwater phytoplankters Desmodesmus can increase the size of their colonies to avoid mortality from numerous grazers (Lürling 2003). These are examples among a vast amount of abilities for phenotypic plasticity that can in and of themselves be seen as functional trait of the organisms that possess this flexibility (Barnett et al. 2007; Weithoff and Beisner 2019). Intra-specific variability can be substantial in aquatic organisms (*e.g.* Sanford and Kelly 2011), and can impact community and ecosystem dynamics similarly to inter-specific trait variability (Des Roches et al. 2018; Raffard et al. 2019). Within the climate context, understanding the drivers and link between intra- and inter-specific trait variability is another argument for the use of FTBAs instead of species-centered approaches (Violle et al. 2012).

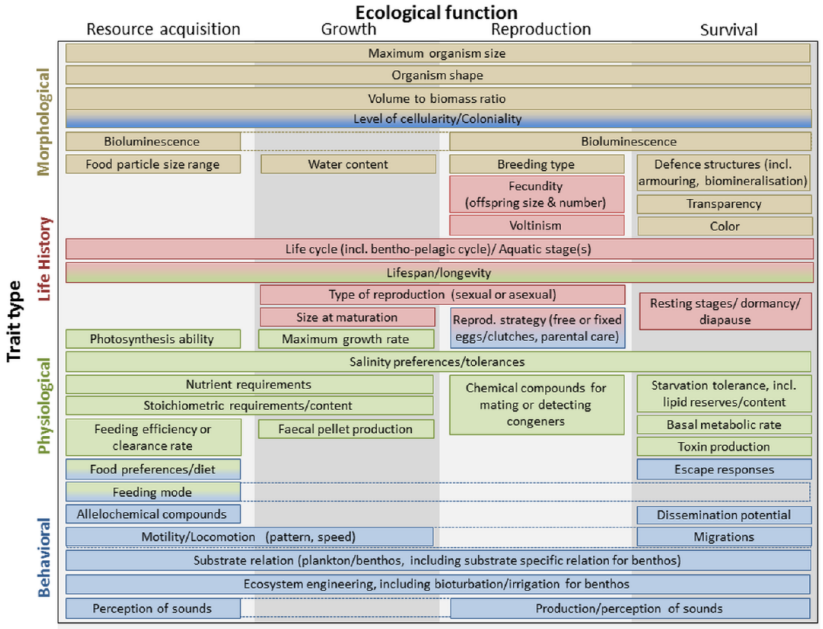

Figure A.1 - Unified typology of aquatic functional traits that could be used in multicompartmental studies. This typology focuses on the key functional traits that transcend taxonomic peculiarities of the different aquatic ecosystems. Traits are classified by type and ecological function (as in Litchman and Klausmeier, 2008) and most of them are quantitative. The dashed lines are a representation for similar traits crossing multiple ecological functions that are not close. A mental map providing a network visualisation of this figure is available online, with each trait node linking towards associated research articles (`http://doi.org/10.5281/zenodo.3635898`).

To go further towards a unified framework, we propose a common typology of functional traits for aquatic organisms (Figure A.1). It not only follows what was previously proposed for phy-

toplankton (Litchman and Klausmeier 2008) and zooplankton (Litchman et al. 2013; Brun et al. 2016) but now incorporates new elements proposed for marine benthic ecosystems (Degen et al. 2018). Moreover, in the typologies proposed by Litchman and colleagues, functional traits are classified in one of four types (morphological, physiological, behavioral and life-history) and associated to three main ecological functions (resource acquisition, reproduction and predator avoidance for phytoplankton; feeding, growth/reproduction, and survival for zooplankton). Here we propose to separate growth and reproduction into two distinct columns (Figure A.1). Compared to earlier typologies, ours identifies key functional traits that can be used for multicompartmental studies because they transcend the taxonomic specificities of the different aquatic ecosystems (Salguero-Gómez et al. 2018). For instance it includes some traits that have been disregarded so far in studies focusing on only one compartment. These traits are: water content, color, breeding type, life-cycle, life span, diapause, reproduction strategy, salinity preference/tolerance, chemical compounds for mating or detecting congeners, diet/food preference, allochemical compounds, dissemination potential, substrate relation (plankton/benthos, including substrate specific relation for benthos), ecosystem engineering, including bioturbation/irrigation for benthos, and finally perception/production of sounds. Most of the 40 functional traits presented in this typology can be estimated quantitatively (Costello et al. 2015), making them good candidates for comparative studies. In addition, a dynamic representation of this typology is proposed as an online mental map (http://doi.org/10.5281/zenodo.3635898) which links to associated research articles. This mental map is not only a different way to represent the functional traits proposed in Figure A.1, but it also provides a dynamic visual representation. It can serve as a pedagogical tool for teaching purposes and as a basis to identify trade-offs between related traits. Further work could initiate a globally shared ontology for aquatic traits, for instance as part of Open Biological and Biomedical Ontology (OBO) Foundry (http://obofoundry.org/).

A.1.2.3 Estimating functional diversity from functional traits

Traits are useful tools to quantify not only the functional biogeography of a system or organism, but also the diversity of a system, its functional redundancy, and/or its likely resilience to perturbations. Those traits that have been measured at the individual level, or estimated for each species of a given community, can be used to estimate trait-based Shannon diversity (Usseglio-Polatera et al. 2000) or Rao's quadratic entropy indices (Rao 1982). Functional diversity (FD) and its various dimensions, such as functional richness, functional divergence, or functional evenness (Mason et al. 2005; Ricotta 2005) can further be quantified, either using dendrogram-based metrics (*e.g.* Petchey and Gaston 2007; Mouchet et al. 2008), or from the definition of a functional space (*e.g.* Villéger et al. 2008; Laliberté and Legendre 2010) Several indices taking explicitly into account intraspecific trait variability were also proposed (*e.g.* Bello et al. 2011; Carmona et al. 2016). Functional beta diversity can be estimated too, including through the more classical Biological Trait Analysis (BTA) (*e.g.* Bremner et al. 2006; Beauchard et al. 2017). To aid ecologists in finding their way among the many functional diversity metrics, several guides were published about their definition and use (Schleuter et al. 2010; Mason et al. 2013; Mouillot et al. 2013; Carmona et

al. 2016; Schmera et al. 2017; Legras et al. 2018). Many of these indices are sensitive to the number and the type of traits that are considered (*e.g.* Legras et al. 2019), as well as to the species richness of the communities, meaning that the comparison of sites with different richness levels would require using comparable indices that are unbiased by species richness and trait selection.

A.1.3 Estimating and using traits: tools and limits for studying functional traits

Several observational methods, both used in situ as well as in vitro, allow for the quantification or identification of functional traits; but they are predominantly used in either oceanographical or limnological applications, not both. Currently available methods to measure or estimate traits include classical trait measurements (laboratory and field), imaging and acoustic techniques, as well as molecular sequencing (-omics). These methods will be described in the following sections and opportunities for sharing between scientific communities will be outlined.

A.1.3.1 Empirical studies of traits as a source for trait databases

The investigation of functional traits has been largely based on empirical studies. Such studies rely on three complementary approaches that can be described by: 1) measurements of traits in situ, 2) measurements of traits under controlled laboratory conditions, and 3) metadata analyses of databases and literature (Figure A.2A). The metadata approach has been undoubtedly the most developed across aquatic ecosystems (Degen et al. 2018; Kiørboe et al. 2018) and the literature has been the basis of a number of reviews describing functional traits. For example, in freshwater ecology, Kolkwitz and Marsson (1909) pioneered a compilation of types of organisms in relation to the presence of various pollution levels. In marine ecosystems, metadata compilations allowed mapping of key traits of marine copepods at a global scale and evaluation of their relationships with environmental conditions (Brun et al. 2016b; Benedetti et al. 2018). One effective way to merge functional traits with taxa, based on a variety of sources and literature, is the fuzzy coding procedure (*e.g.* Chevenet et al. 1994). In functional trait-based approaches, the fuzzy coding uses positive scores to describe the affinity of a species for the different categories of a given trait, *e.g.* using “0”, “1”, “2” and “3” for species exhibiting respectively “no”, “weak”, “moderate” and “strong” link with a given trait category (Chevenet et al. 1994; Usseglio-Polatera et al. 2000). When a trait applies to a subset of the different stages of the species life cycle (egg, larva, pupa, and adult), the relative duration of each stage is considered in assigning appropriate scores to the different categories of this trait. To standardize the description of species attributes, trait category scores are converted into a relative abundance distribution so that the sum of the trait category scores for an individual trait and a given taxon equals one. This technique of coding is robust enough to compensate for different types and levels of information available for different taxa.

Table A.2 - Online databases documenting functional traits of aquatic organisms. Databases without a primary focus on traits, but that also provide trait information, are included. This list is available at `https://github.com/severine13/FonctionalTrait_databases`.

Name of the database	Taxonomic groups of interest and habitats	Reference	Brief description	Web link
Traitbank - Encyclopedia of Life	All taxa across the tree of life, including marine and freshwater organisms	(Parr et al. 2014)	Provides traits, measurements, interactions and other facts. Actively growing resource covering all ecosystems (not restricted to aquatic ecosystems).	http://eol.org/info/516
Bromeliad invertebrate traits	Aquatic invertebrates in bromeliads from South America	(Céréghino et al. 2018)	12 functional traits of 852 taxa	https://knb.ecoinformatics.org/#view/doi:10.5063/F1VD6WMF
South-East Australian freshwater macroinvertebrate traits	Freshwater macroinvertebrates from South-East Australia	(Schäfer et al. 2011)	9 traits, described at the family level for 172 taxa	Supplementary information to the article
EPA Freshwater Biological Traits Database	Freshwater macroinvertebrates from North America rivers and streams	(U.S. EPA. 2012)	Includes functional traits (e.g. life history, mobility, morphology traits) but also ecological and habitat information for 3,857 North American taxa.	https://www.epa.gov/risk/freshwater-biological-traits-database-traits
Biological Traits Information Catalogue (BIOTIC)	Benthic marine macrofauna and macroalgae	(MARLIN 2006)	Includes 40 biological trait categories.	http://www.marlin.ac.uk/biotic
EMODnet Biology database	European seaweeds	(Robuchon et al. 2015)	Functional traits (morphology, life history, ecophysiology) and ecological information (incl. biogeography) for the 1800 seaweed species listed in Europe.	Ongoing work
Functional traits of marine macrophytes	European marine macrophytes, including seaweeds	(Jänes et al. 2017)	Functional traits (morphology, ecophysiology) and ecological information for 68 species.	https://www.datadryad.org/resource/doi:10.5061/dryad.964pf/1
POLYTRAITS	Marine polychaetes	(Faulwetter et al. 2014)	47 traits describing morphological, behavioural, physiological, life-history characteristics, as well as the environmental preferences, for a total of 27198 trait records for 952 species.	http://polytraits.lifewatchgreece.eu/
The Arctic Traits Database	Marine organisms from the Arctic	(Degen and Faulwetter 2019)	Traits for 478 species-level taxa.	https://www.univie.ac.at/arctictraits/team
WoRMS Marine Species Traits portal	Marine species	(WoRMS Editorial Board 2019)	Provides 10 traits that have been prioritized within EMODnet Biology, as part of the World Register of Marine Species (WoRMS).	http://www.marinespecies.org/traits/index.php
Functional traits of marine protists	Marine protists, including fungi.	(Ramond et al. 2018)	Provides 30 functional traits for 2,007 taxonomic references associated to V4 18S rDNA sequences.	https://doi.org/10.17882/51662
COPEPEDIA/ COPEPOD	Marine plankton	(O'Brien 2014)	Database of plankton taxa distribution maps, photographs, biometric traits, and genetic markers.	https://www.st.nmfs.noaa.gov/copepod/documentation/contact-us.html
Trait database for marine copepods	Marine pelagic copepods	(Brun 2017)	Trait databases providing 9,306 records for 14 functional traits of about 2,600 species.	https://doi.pangaea.de/10.1594/PANGAEA.862968

Mediterranean copepods' functional traits	Marine copepods present in the Mediterranean Sea	(Benedetti 2015; Benedetti et al. 2016)	Seven functional traits for 191 species.	https://doi.org/10.1594/PANGAEA.854331
Freshwater Ecology	European freshwater organisms belonging to fish, macro-invertebrates, macrophytes, diatoms and phytoplankton	(Schmidt-Kloiber and Hering 2015)	Covers environmental preferences, distribution patterns, and functional traits for 20,000 taxa.	https://www.freshwaterecology.info/
Phytoplankton of temperate lakes	Phytoplankton of temperate lakes	(Rimet and Druart 2018)	Database of morphological and physiological traits of more than 1,200 taxa.	https://zenodo.org/record/1164834#.XRNrPXvgrOR
FishBase	Fish	(Froese and Pauly 2019; Beukhof et al. 2019)	Provides information on 34,100 species, including traits related to trophic ecology and life history.	www.fishbase.org https://doi.org/10.1594/PANGAEA.900866
The Coral Trait Database	Coral species from the global oceans	(Madin et al. 2016)	Includes 68,494 coral observations with 106,462 trait entries of 158 traits for 1,548 coral species.	https://coraltraits.org/
FishTraits	Freshwater fishes of the United States.	(Frimpong and Angermeier 2010)	More than 100 traits are informed for 809 fish species of the USA, including 731 native and 78 exotic species.	http://www.fishtraits.info/
ECOTAXA	Marine planktonic eukaryotes and prokaryotes (Viruses in prep.)	(Picheral et al. 2017)	50 morphological features including size, shape or opacity.	http://ecotaxa.obs-vlfr.fr/ http://ecotaxa.sb-roscoff.fr
Protist Ribosomal Reference database (PR2)	Protists	(Guillou et al. 2013)	Sequence database for which the inclusion of functional traits is under development.	https://github.com/pr2database/pr2database
Eukaryotic Reference Database (EukRef)	A wide range of eukaryotic organisms across the tree of life	(del Campo et al. 2018)	Collaborative annotation initiative for referencing 18S rRNA sequences, for which the inclusion of functional traits is under development.	https://eukref.org
The Kyoto Encyclopedia of Genes and Genomes (KEGG)	A wide range of organisms across the tree of life	(Kanehisa and Goto 2000)	Collection of databases on genomes and biological pathways that provides molecular-level information on gene functions, which could inform on potential functional traits.	https://www.genome.jp/kegg/

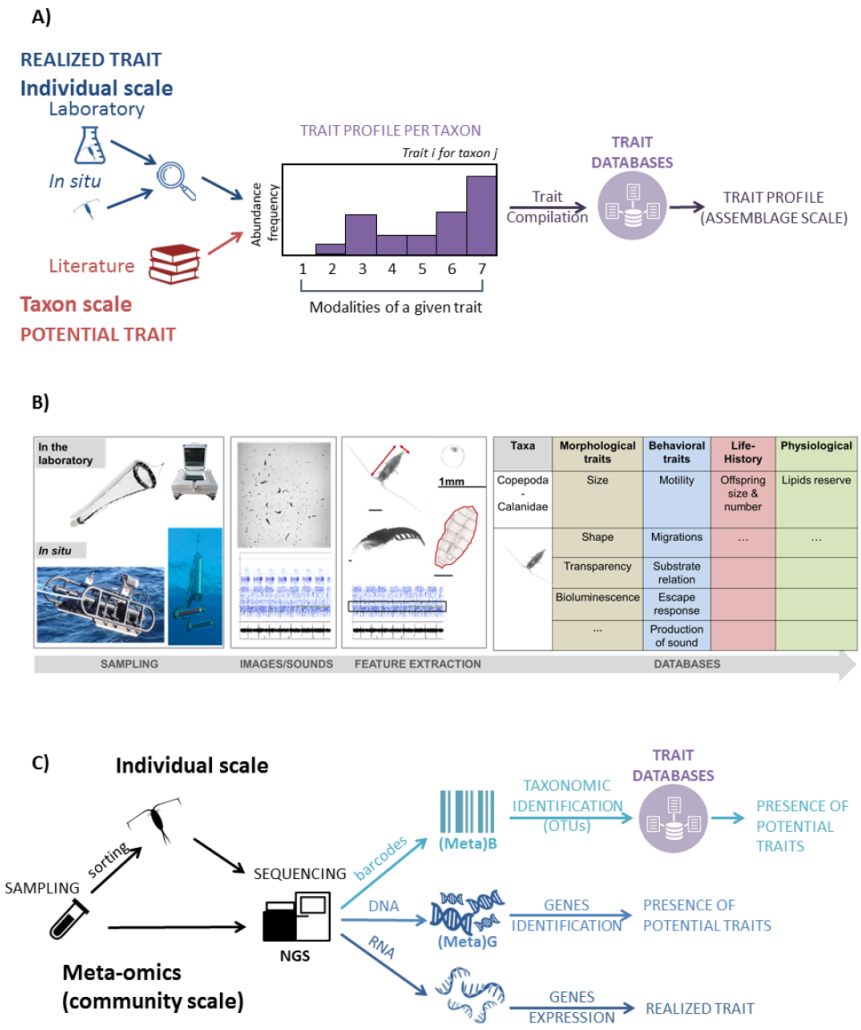

Figure A.2 - Main methods to study traits. A: Use of empirical studies to measure realized traits (in laboratory or in situ) or to estimate potential traits from the literature. The abundance frequency the modalities a given trait can be used to code trait profile by taxon using fuzzy coding and thereby inform trait databases. B: Use of imaging and acoustic techniques to identify or measure functional traits of aquatic organisms from sampling, images/sound recording, features extraction to databases (ZooVis picture has been kindly provided by H. Bi). C: Use of sequencing techniques to identify or measure functional traits. Sequencing can be done at the community scale (meta-omics) or at the individual scale after manual or automatic sorting. (Meta)B: (Meta)Barcoding, (Meta)G: (Meta)Genomics, (Meta)T: (Meta)Transcriptomic.

In recent years, numerous open access databases recording functional traits have been developed to document traits included in existing databases (Table A.2). This diversity of databases gathers trait information not only for widely studied traits (*e.g.* body size or feeding strategy), but also for less common traits or for those that are more difficult to measure (*e.g.* age at first reproduction, migration mode, or nutrient affinities). Some large trait databases were published online and open access (*e.g.* Herring 1987; Barnett et al. 2007; Benedetti 2015; Hébert et al. 2016; Degen and Faulwetter 2019), thus allowing for follow-up studies that compare and merge trait data across taxa, species and environments. In some instances (*e.g.* freshwater invertebrates) published databases rapidly became foundational for environmental assessment procedures (*e.g.* Mondy et al. 2012; Mondy and Usseglio-Polatera 2013; Larras et al. 2017). The main caveat of these FTBAs

is that only a limited number of species and/or traits have been reported so far, thus not yet allowing for a generalisation of findings across taxa, the definition of fitness landscapes, and/or the characterisation of ecological niches or responses to environmental change. Indeed, these databases often focus on the dominant and most easily sampled or cultivable species. Moreover, metadata associated with trait measurement methods are usually lacking. Until now, shortfalls in the knowledge of many aquatic taxa (Troudet et al. 2017) restrict the application of trait databases at the community scale and remain a limiting factor for the integration of FTBAs into macroecology (Tyler et al. 2012; Borgy et al. 2017). However, the main limit so far to provide and share trait data remains the lack of an ecological standard for data (Schneider et al. 2018). Attempts to increase unification are currently emerging on various fronts such as the terminology of traits (*e.g.* Schmera et al. 2015 for stream ecology), the cross-taxa compatibility of functional traits (*e.g.* Weiss and Ray 2019 for plants and animals) or the actual measurements of such traits (*e.g.* Moretti et al. 2017 for terrestrial invertebrates). Large efforts are still needed to combine and integrate all these various trait databases (Degen et al. 2018), but applying Open Science principles should accelerate trait-based science (see for example the Open Traits Network initiative, Gallagher et al. 2019). Such databases are already numerous, large-sized and of increasing complexity. Therefore, their manipulation requires strong computational abilities (Durden et al. 2017). As a result, aquatic research is evolving into a more biostatistical- and bioinformatical-based field, enabling the extraction of large-scale information on traits and putting to full use taxonomic surveys recorded over time. Despite this, naturalist taxonomic knowledge per se remains critical and future challenges in ecology will undoubtedly benefit from a combination of modern functional trait-based approaches and a modern integrative taxonomic knowledge.

The traits documented by these databases originate from direct measurements of realized traits in the laboratory or in situ (Figure A.2A). Laboratory experiments allow for the quantification of functional traits of model species within a large range of controlled environmental conditions. They provide a well-constrained system, both in physical variables and species content, to measure functional traits at the individual level. However, they are often limited to a few cultured species that do not necessarily reflect the actual functional diversity and complexity of whole ecosystems, as should FTBAs do. One of the few examples of lab-measured traits tested the existence of trade-offs across many phytoplankton species between maximum growth rate, competitive ability for phosphorus acquisition, and ability to store phosphorus (Edwards et al. 2013a).

In recent years, innovative instruments and tools have become available to measure in situ new functional traits. They include imaging and genomics tools that have the potential to provide a comprehensive picture of aquatic ecosystem composition, structure and function. Their implementation should greatly help advance the use of functional traits in aquatic studies.

A.1.3.2 Imaging and acoustic techniques

Imaging systems are best suited for the quantification of morphological traits, such as size, transparency, bioluminescence or shape (Forest et al. 2012; Barton et al. 2013; Fontana et al. 2014; Andersen et al. 2016), but also for the estimation of some behavioral (*e.g.* motility or substrate re-

lationships), life-history or physiological traits (Table A.1; Figure A.2B,Schmid et al. 2018; Ohman 2019). Imagery has been used as a tool in marine science since the 1950's and a variety of imaging systems have been successfully developed to record individual characteristics (see imaging and acoustic instruments listed in Table A.3; *e.g.* Lombard et al. 2019). Over the last 15 years, novel imaging techniques have allowed for rapid and less-intrusive visual observation of organisms' traits from pico- to macro-scales (*e.g.* (Culverhouse et al. 2006; Stemmann et al. 2008; Sieracki et al. 2010; Biard et al. 2016). To date, imaging tools have mostly been used by marine ecologists (Table A.3), in both benthic and pelagic ecosystems, with only a few implementations in freshwater environments (*e.g.* (Althaus et al. 2015; González-Rivero et al. 2016). This is mainly due to the large amount of particles, the higher turbidity and the relatively smaller size of the crustacean zooplankton in freshwater ecosystems. Benthic imaging tools include baited, unbaited, towed, autonomous- and diver-operated systems (Matabos et al. 2014; de Juan et al. 2015; Mérillet et al. 2018), while pelagic ones are mainly in situ or bench-top systems. Since the turbidity and obstacles in benthic, coastal or river ecosystems strongly modify optical characteristics, systems with external light are more commonly used to efficiently capture morphological traits of aquatic organisms.

Taxonomic groups	Instrument	**References**	**Applications**
Protists	FlowCam	(Sieracki et al. 1998)	Marine microplankton, abundance, size
	Imaging FlowCytobot (IFCB)	(Olson and Sosik 2007)	Marine coastal, nano- and microplankton, quantification, particle profile (morphology)
	Cytobuoy	(Dubelaar et al. 1999; Dubelaar and Gerritzen 2000)	Freshwater and marine coastal, phytoplankton biomass, particle profile (morphology)
Large protists and meso-plancton	Zooscan	(Gorsky et al. 2010)	Marine, shelf, coastal, pelagic plankton, morphological features
	Zooglider	(Ohman et al. 2019)	Imaging and acoustics, marine, shelf, coastal, pelagic plankton, morphological features
Macro-organisms and fish	Laser Optical Plankton Counter (LOPC)	(Finlay et al. 2007)	Freshwater and marine, zooplankton size, biomass, abundance
	ZOOplankton Visualization and Imaging System (ZOOVIS)	(Bi et al. 2012)	Marine pelagic, zooplankton, size
	Underwater Video Profiler (UVP)	(Picheral et al. 2010)	Marine, shelf, coastal, pelagic plankton, morphological features
	Lightframe On-sight Keyspecies Investigation (LOKI)	(Schulz et al. 2010; Schmid et al. 2016)	Marine zooplankton, species, stages, morphological features
	In situ Ichthyoplankton Imaging System (ISIIS)	(Cowen and Guigand 2008)	Marine, ichtyoplankton, meso-zooplankton, abundances, species
	Hydrophone	(Coquereau et al. 2016; Desjonquères 2016)	Marine, freshwater, benthic

Table A.3 - Examples of instruments for imaging and acoustic assessment, used for trait description and quantification in aquatic ecosystems.

A major advantage of imaging systems is their variable degree of invasiveness during observation. Imaging systems can analyse discrete measurement of water samples (living or fixed samples), but they can also acquire in situ continuous records on living organisms. For instance, imaging techniques applied to marine plankton revealed that the abundance of the most fragile organisms (such as gelatinous zooplankton, Rhizaria, etc.) has been underestimated for a long time using traditional observation techniques (*e.g.* Biard et al. 2016), as they tend to break when collected using plankton nets (Stemmann et al. 2008). The use of in situ imaging systems also provides information on poorly studied traits, such as transparency and water content of gelatinous organisms. For benthic systems, imaging techniques provide non-intrusive and non-destructive methods that can be valuable to assess endangered habitats and/or marine protected areas and to collect information on the distribution of large over-dispersed epifaunal species inadequately sampled by traditional gears like grabs (*e.g.* Althaus et al. 2015).

In addition to classical imaging, acoustic methods (passive and active) are also tools of increasing importance to quantify particular functional traits. Acoustic Doppler current profilers (ADCPs) have been successfully used in lakes to capture diel migration behavior in larger planktonic species such as the insect larval predators of zooplankton (*e.g.* Chaoborus; Lorke et al. 2004). Hydrophone recordings can be used to record sound emissions by the organisms themselves. The sounds produced by freshwater organisms represent a highly overlooked trait and such trait recordings might provide relevant non-invasive tools to monitor the complexity and changes in aquatic communities. In a literature survey, Desjonquères (2016) showed that at least 271 freshwater species amongst French aquatic fauna (89% insects, but also fish and crustaceans) produce sounds. Using continuous underwater recordings with hydrophones, it was shown that the acoustic diversity of ponds and floodplain water bodies reflects the taxonomic diversity of aquatic communities (Desjonquères et al. 2018). Similarly, sound production by benthic invertebrates in the bay of Brest (France) was used to describe the soundscape and assess the ecological status of maerl beds (Coquereau et al. 2016).

One of the main caveats of imaging methods for FTBAs is that imaging tools have a low resolution below a certain size (most of these tools are of limited accuracy below a size limit of 200 µm for zooplankton, and 30-40 µm for phytoplankton, see Table A.3), and may not allow for a reliable analysis of smaller size fractions, often associated with detrital matter or particles with a lack of discernible morphological differences. This limit is especially true for organisms without hard structures such as naked dinoflagellates or aloricate ciliates. However, imaging and acoustic methods generate high frequency and automated datasets at large spatial scales, with some of them recorded by inter-calibrated instruments, which allow for their comparison and combination in space and time (*e.g.* UVP for marine plankton; Table A.3). These data are also suitable for the validation of trait-based marine ecosystem models (Kiørboe et al. 2018) and new ecological questions have been addressed by combining both recent imaging techniques and FTBAs (Schmid et al. 2018). New opportunities using imaging and acoustics include the evaluation of feeding behaviors and network associations (Choy et al. 2017), filtration rates and carbon fluxes (Katija et al. 2017) and migration patterns of zooplankton (Benoit-Bird and Lawson 2016).

Because the number of images stored on acoustic and imaging systems is limited, and even short deployment times lead to considerable data volumes, the development of artificial intelligence (AI) techniques such as machine learning, deep learning recognition and classification has been a crucial tipping point in the extraction of traits from these large datasets (Villon et al. 2016; Maps et al. 2019). Bigger storage capacity, standardized learning sets for machine learning combined with the automatized pre-processing of data directly in autonomous sampling instruments are already under development and will be an asset for the future of functional traits quantification by imaging.

A.1.3.3 Omics techniques for FTBAs

Another opportunity for automatic measurements of functional traits has emerged from the recent rise of high-throughput sequencing techniques (HTS, also called NGS, for Next Generation Sequencing, or "-omics" in the broader sense). These techniques provide fast and relatively cheap nucleic acid sequencing and have opened new perspectives for investigating the structure and functioning of aquatic communities, both in marine (Raes et al. 2011; Sunagawa et al. 2015; Mock et al. 2016) and freshwater systems (Chonova et al. 2019). Methods based on DNA or RNA sequencing can be used for large-scale studies of environmental samples, investigating water samples in which any nucleic acid that is present can theoretically be retrieved.

For FTBAs, the identification of targeted DNA sequences (or metabarcoding; Bucklin et al. 2011; Valentini et al. 2016) can be used as a first step for fast and automatic taxon recognition, prior to the attribution of traits to the respective taxa using trait databases (Figure A.2C; Table A.2). This was recently done to describe the biogeography of mixotrophic traits of marine protists at global scale (Faure et al. 2019), or to estimate the functional diversity of coastal protist communities (Ramond et al. 2019). In freshwater systems, metabarcoding of benthic diatoms was used to assess the water quality status of rivers (Vasselon et al. 2017) and metabarcoding was combined with text-mining or phylogenetic inference of ecological profiles and traits for biomonitoring (Keck et al. 2018; Compson et al. 2018). Yet, metabarcoding is inherently biased in multiple ways, such as its lack of quantitative link between the number of copies of barcodes (targeted DNA sequences) and the biomass or abundance distribution of organisms, the risk of gene amplification from dead material (not currently influencing ecosystem function), or the use of universal barcodes that may not be adapted to distinguish taxa for all lineages (*e.g.* Deiner et al. 2017). However, the main obstacle to using metabarcoding data for FTBAs is the low number of taxa for which barcodes have been documented (in addition to the low number of taxa for which trait information is available). This limitation precludes a full assessment of ecosystem structure from metabarcoding (*e.g.* de Vargas et al. 2015; Le Bescot et al. 2016). Thus, a strong effort remains to be made to supplement existing genomic databases with more taxonomically-referenced sequences and trait information to allow the metabarcoding-based monitoring of aquatic functional traits (*e.g.* Ramond et al. 2018; PR2: Guillou et al. 2013; EukRef: del Campo et al. 2018; Diat.barcode: Rimet et al. 2019).

Beyond metabarcoding, -omics approaches are of particular interest to identify or measure functional traits linked to metabolic pathways (*e.g.* photosynthesis, nitrification, diazotrophy, cal-

cification, etc.), using either (meta-)genomic or (meta-)transcriptomic approaches (Figure A.2C). When combined with databases like KEGG (Kanehisa and Goto 2000), which includes the genes (for genomics/transcriptomics), proteins (for proteomics) and metabolites (for metabolomics) implied in a specific pathway, -omics approaches open up the possibility of monitoring functional traits (defined at the individual level) across different levels of biological organisation (from organisms to communities). For example, approaches that report the expression level of genes, proteins and metabolites are increasingly used in ecotoxicology to assess functional traits (*e.g.* photosynthesis, chemical degradation) in response to stressor(s) via targeted approaches (*e.g.* q-PCR on pre-identified candidate genes, Pesce et al. 2013; Moisset et al. 2015). Although it is still very challenging to relate -omics data to functional traits (Stec et al. 2017), the identification of certain genes coding for particular metabolic or physiological traits (*e.g.* iron uptake, nitrogen fixation) may help to directly link ecosystem structure to ecosystem functions (Mock et al. 2016), while taking into account the majority of organisms that in fact cannot be classified based on their morphological characteristics (*e.g.* picophytoplankton), and/or cannot be captured by imaging methods due to their small size or behavior. For instance, using metagenomic data, Farrell et al. (2018) created a machine-learning algorithm that can predict values of 65 phenotypic traits with more than 90% accuracy, thus allowing the investigation of the functional profiles of 660 uncultured marine prokaryotes based only on their metagenomically-assembled genomes or MAGs (*i.e.* genomes putatively reconstructed from metagenomics data). This very promising method cannot yet be applied to eukaryotes, as relating genes to potential traits in eukaryotes remains much more challenging than for prokaryotes (Sunagawa et al. 2015; Salazar et al. 2019). However, transcriptomics techniques were successfully used to estimate putative traits for marine protists using sequence similarity network-based approaches (toxicity and symbiosis for dinoflagellates; Meng et al. 2018). For pluricellular organisms, many challenges remain for the application of such methods in FTBAs, especially because of the large size of their genomes and because reference genomes are lacking (hence, the function of their DNA or RNA remains unknown). Yet, the use of transcriptomics approaches seem promising for these organisms (*e.g.* Lenz et al. 2014; Blanco-Bercial and Maas 2018).

Substantial progress remains to be made before aquatic ecologists can fully exploit -omics information using a FTBA. This includes the design of new methods to estimate the quantitative aspects of -omics information, but also to decipher the large quantity of sequences that cannot be assigned to any taxon in an environmental sample, and to circumvent the low proportion of genomic functional annotation (especially for eukaryotes). However, ongoing and future -omics studies may allow skipping taxonomic assignation and even the identification of gene functions as an intermediary between ecosystem composition and function. Such studies would fully contribute to FTBAs of aquatic ecosystems by targeting the -omics signature of relevant functional traits (Mock et al. 2016; Stec et al. 2017).

Another application would be the use of -omics data to develop a new generation of trait-based models (Mock et al. 2016; Stec et al. 2017; Coles et al. 2017). Metatranscriptomic data could be used to identify physiological traits of phytoplankton, combined with a mechanistic model of the phytoplankton cell, and used to construct a trait-based global marine ecosystem model (Mock

et al. 2016). Emergent communities of marine microbes (from bacteria to phytoplankton) have already been predicted by directly simulating their metagenomes and metatranscriptomes (Coles et al. 2017). In summary, the idea of improving ecosystem models using -omics is not new (Hood et al. 2006), but FTBAs could constitute the common framework needed for next-generation ecosystem modellers, observers, molecular biologists, and ecologists working in limnology and oceanography. This would advance our ecological understanding of aquatic ecosystems and the links between ecosystem structure, function and ecosystem services or bioindicators relevant for ecosystem monitoring and management.

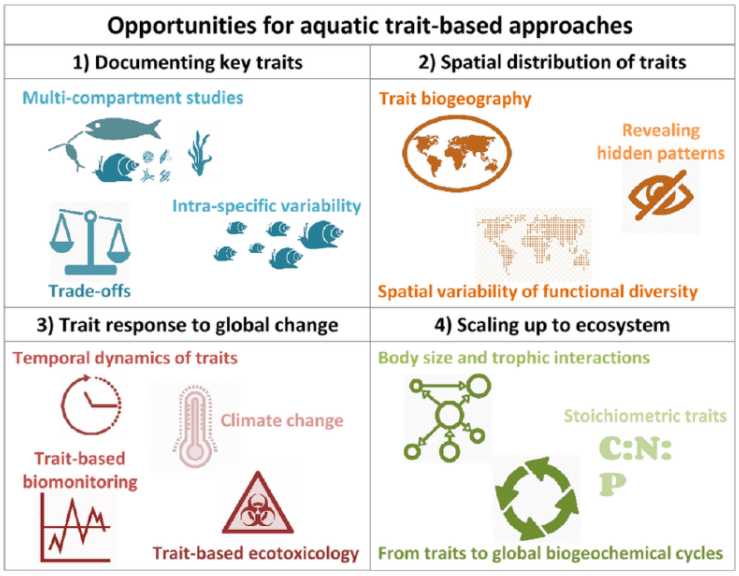

Figure A.3 - Main opportunities for trait-based approaches in aquatic ecology. These four opportunities are described in section A.1.4.

Using either empirical studies, imaging/acoustics, or -omics, both quantitative and qualitative traits can be estimated. One advantage of qualitative traits is that they do not have to be measured using the same instruments, and can be more easily described across compartments and realms. For quantitative traits, metrics and indices, one challenge is to be able to compare trait-based functional diversity among studies. In the next section we will focus on such traits that can be shared between ecological compartments and we will describe new opportunities in aquatic ecology to highlight spatio-temporal patterns, study anthropogenic impacts and better describe trophic interactions between plankton species (Figure A.3).

A.1.4 Future opportunities for aquatic FTBAs

A.1.4.1 Going further towards a trait-based aquatic ecology by identifying key traits

Documenting key traits in multi-compartment studies

Given that the main power of FTBAs is to transcend both taxonomy and realms, trait-based ecological studies could result in a common set of ecological rules and theoretical principles

that could be applied to multiple systems (*e.g.* benthos and pelagos, including plankton and fish). Following our framework, aquatic functional traits could be described at various spatio-temporal scales for both benthic and planktonic organisms, for instance taking benthic-pelagic coupling into account. To do so, we recommend a closer collaboration among aquatic ecologists, including process-oriented projects and comparative studies of freshwater and marine systems, focusing on the aquatic functional traits that we have identified and on their links to ecosystem functions (Figure A.1). In addition to morphological traits that are relatively easy to identify and to measure (such as size, shape, cellularity, defences and colour), priority traits to be investigated should also include: 1) life history traits such as voltinism (number of generation or breeding per year), life cycle, life span, type of reproduction, and reproduction strategy, 2) physiological traits such as photosynthesis ability, diet, feeding mode, salinity preference, and trophic regime, and 3) behavioral traits like motility, dispersal potential, and substrate relation. Indeed, among all the traits identified in Figure A.1, these traits are the most easily identified for any aquatic organisms, including both uni- and pluricellular organisms, and cover all ecological functions and all types of traits (Figure A.3.1). One recent example, that could lead future novelty in trait-based studies, is the use of morphological traits estimated from multiple images. Statistically-defined multidimensional morphological space can be synthesized from many individual images to generate a suite of interpretable continuous traits. Looking at the spatial distribution across the Arctic ecosystem of key traits, including body size, opacity, or appendage visibility, revealed meaningful information of copepods distribution and ecology in relation to ice-coverage (Vilgrain et al. under review). Such statistical approaches using these continuous traits can easily be applied to multi-compartment studies, (for example using transparency to describe gelatinous ecological patterns).

Documenting the trade-offs between key traits

Multi-compartmental studies that aggregate effects across species and trophic levels, hence taking into account the network structure of a community or the food-web structure of an ecosystem, would also enable a better understanding and quantification of the trade-offs occurring between two or multiple traits. Trade-offs, which result from the inherent metabolic, energetic or behavioral costs associated with each expressed trait, provide the fundamental basis to understand species coexistence and the trait composition of communities (Ehrlich et al. 2017). In particular, the competition-colonization trade-off is a major mechanism for biodiversity maintenance (Tilman 1994; Muthukrishnan et al. 2020; Ehrlich et al. 2020). Strong competitors able to exclude other species in any given habitat are often slow dispersers. In contrast, poor competitors are often strong colonizers, able to easily disperse and find unoccupied niches. A trade-off between resource acquisition and survival (or predation vulnerability) was reported for zooplankton: organisms that feed using feeding currents, increase their risk of being detected by predators that are sensitive to flow disturbances (Kiørboe and Hirst 2014). Unexpected trade-offs can often explain the relative mismatch between expected and observed individual traits in aquatic communities along gradients of anthropogenic pressure, complexifying the trait-based diagnostic of water bodies (Resh et al. 1994; Mondy and Usseglio-Polatera 2014; Desrosiers et al. 2019). Indeed, the success of a species in adverse conditions might be due to a particularly effective adaptation without the need

for further adaptive traits. Moreover, investing in a given adaptation can leave fewer resources available for the investment in another adaptation. Species of various lineages (*e.g.* different phyla in invertebrate assemblages) may also solve the same ecological constraint with different adaptations.

Trade-offs have been globally recognised as a central component of trait-based approaches in aquatic ecology (Resh et al. 1994; Kremer et al. 2017) especially in plankton ecology (Litchman et al. 2007, 2013, 2015; Litchman and Klausmeier 2008; Hébert et al. 2017; Kiørboe et al. 2018; Ehrlich et al. 2020). In benthic studies, there has been a clear lack of work that considers simultaneously several traits relative to what has been done in studies on marine plankton (*e.g.* Litchman et al. 2013) and in freshwater ecology (Verberk et al. 2008). As a case in point, the term trade-off is not mentioned in the recent review on benthic traits by Degen et al. (2018). More studies are needed to explore trade-offs among traits across compartments and realms in order to identify the rules governing the links between traits, trade-offs, community structure and function. To accomplish this, researchers will have to put effort on measuring multiple traits, focusing on those related to resource acquisition, growth, storage and predation avoidance (*i.e.* directly related to fitness) on a variety of taxa within the same habitat. Comparison of how such relationships that trade-off (*i.e.* negatively related) change under different abiotic or biotic conditions will allow determination of how flexible such trade-offs are as ecological conditions change. Recently, the shape of the trade-off curve, representing the boundary of the set of feasible trait combinations, has been described as explaining traits of co-existing species and changes in trait values along environmental gradients (Ehrlich et al. 2017, 2020). Convex trade-offs would facilitate the coexistence of specialized species with extreme trait values while concave trade-offs would promote species with intermediate trait values.

To further explore trait relationships, aquatic ecologists may be inspired by what has been done in terrestrial plant ecology: the identification of so-called trait syndromes. Trait syndromes are relationships between traits that are defined by fundamental trade-offs amongst taxa that determine their ecological roles in ecosystems. The classic example in plant ecology is the "leaf economics spectrum" that characterizes taxa according to the speed at which they are able to take up nutrients and invest in leaf biomass (Wright et al. 2004). In this vein, some work was done with aquatic organisms by considering trade-offs amongst lotic insects (*e.g.* Poff et al. 2006), fishes (*e.g.* Winemiller et al. 2015) and phytoplankton (*e.g.* Edwards et al. 2013a). By considering trait syndromes, FTBAs are likely to better predict competitive outcomes as well as distributions of traits across environmental gradients. We thus encourage the aquatic ecology community to engage with the vast array of accessible trait databases provided in Table A.2 and to take the next steps to characterize trait syndromes across the different groups of aquatic organisms.

Documenting the variability of key traits

Finally, more attention should be given to document the variability of all key traits at all organisational levels, *i.e.* at the community scale, between individuals in a given population (*i.e.* intra-specific variability; Raffard et al. 2019), but also for one individual throughout its lifespan (*i.e.* ontogenic variability; *e.g.* Zhao et al. 2014). Indeed, with the exception of a few studies (*e.g.*

Maps et al. 2014b; Banas and Campbell 2016), intra-specific variability of traits is rarely taken into account, mainly because of a lack of empirical information on this variability. For example, the ability to engage fully in autotrophy or to add in heterotrophic feeding is a characteristic of mixotrophic phytoplankton taxa. By characterizing the conditions under which one or the other condition is utilized by a taxon, we can begin to characterize intra-specific variability. Therefore, the question of the scale of variation of functional traits, both at community and population scales, and its impact on ecosystem structure and functioning should be further explored, especially with the use of new methodological development to measure traits (see section A.1.3). Trait-based models could also be used (see review on trait-based modeling in Kiørboe et al. 2018) to quantify the impact of environmental changes on the intra- and inter-specific variability of functional traits (*e.g.* lipid content and size of copepods, Renaud et al. 2018), and to assess the variation of peculiar traits along environmental gradients (Edwards et al. 2012).

Identifying key traits common in limnology and oceanography and their trade-offs, syndromes, and variability, will allow aquatic ecologists to better address central ecological questions, including understanding: 1) the spatial patterns of functional diversity and its drivers, 2) the effects of environmental and anthropogenic pressures on ecosystem structure and functioning, and 3) the interactions among organisms and associated food web organisation and dynamics. For each of these main opportunities, we will briefly describe what has been done to date and then identify potential ways to advance the field of aquatic ecology using FTBAs.

A.1.4.2 New opportunities emerging from the study of the spatial distribution of aquatic traits

The description of aquatic trait biogeography

To date, trait biogeography has been studied for a few compartments in marine ecosystems, such as marine plankton (Barton et al. 2013) including: bacterioplankton (Brown et al. 2014), zooplankton (Prowe et al. 2019), copepods (Brun et al. 2016b; Record et al. 2018), pelagic diatoms (Fragoso et al. 2018), estuarine fish (Henriques et al. 2017) and reef fish (*e.g.* Stuart-Smith et al. 2013). Large-scale studies of the trait biogeography of freshwater organisms are more rare (*e.g.* for amphibians see Trakimas et al. 2016). Aquatic trait biogeography studies covering multi-compartments, including plankton, fish and benthos, remain scarce and usually focus on one realm (*e.g.* marine organisms in Pecuchet et al. 2018). Similarly, aquatic trait biogeography studies covering different environments (marine and freshwater) are few and usually target only one compartment (*e.g.* phytoplankton in Thomas et al. 2016) .

Based on the biogeography of some key traits (*e.g.* size, feeding strategy), aquatic ecologists can now relate functional traits to environmental conditions and identify general rules governing trait diversity distribution. For instance, the description of key traits of marine copepods (body size, offspring size and myelination) has highlighted latitudinal global patterns in trait biogeography. These patterns are in agreement with the temperature-size rule and have unveiled relationships between these traits and environmental conditions, such as water column transparency, but also

biotic conditions, such as chlorophyll seasonality or phytoplankton size (Brun et al. 2016a). More recently, the study of taxonomic and functional diversity of fish communities between two different regions (Caribbean and Great Barrier Reef) and among three habitats (coral reef, seagrass, and mangrove) revealed that traits and functional groups varied among habitats, whereas taxonomic composition varied between regions (Hemingson and Bellwood 2018). Similar relationships should now be tested across ecosystems, geographical regions and trophic levels to verify whether these findings can be generalized to other aquatic organisms/ecosystems (Figure A.3.2). The trait databases now available for many groups of aquatic organisms (see Table A.2) should provide relevant information to explore this direction.

Using traits for revealing hidden community assembly rules at various spatial scales

Based on the spatial description of functional traits, hypotheses underlying community assembly rules can also be tested and community composition can be predicted (Cadotte et al. 2015). For example, the description of physiological and behavioral traits of dragonfly larvae in various lakes recently suggested that traits can drive species distribution and community assembly, through the direct impact of physiological and behavioral traits (activity rate and burst swimming speed) on foraging and predator avoidance behavior (Start et al. 2018). The traits considered in this study were driven by two biomolecules, the expression of which could predict more than 80% of the variation in dragonfly community structure across lakes, and which were involved in the interactions between the dragonfly larvae and their fish predators. Measurements made by new observational methods such as metabolomics, transcriptomics (see section A.1.3.3) or in situ imaging (see section A.1.3.2) would nicely complement presence-absence data by providing indication of the physiological state (*e.g.* healthy or stressed) of the individuals and hence help teasing apart the ideal and realized niche of organisms.

From trait biogeography to spatial variation in functional diversity

Traits that are shared among compartments could also be used to describe the spatial variability of functional diversity (Petchey and Gaston 2006). Among the metrics that were proposed to measure functional diversity and its different dimensions (see section A.1.2.3), aquatic ecologists have to adopt common metrics for comparative studies. Based on these common metrics, the spatial variation of the functional diversity of aquatic communities could be estimated across environments and in multi-compartment studies. For example, the functional diversity of macrophytes was described along a water depth gradient in a freshwater lake (Fu et al. 2014): future studies could cover similar environmental gradients in both freshwater and marine environments (*e.g.* rivers, estuaries, coasts, islands, etc.) and also include other organisms and higher trophic levels, both benthic and pelagic, to test whether the resulting spatial patterns of functional diversity can be generalized. The spatial distribution of traits and functional diversity could also be used to identify functional diversity hotspots and propose protected areas for a trait-based conservation. The diversity of functional traits is indeed correlated to both taxonomic diversity (*e.g.* Petchey and Gaston 2006) and the provision of ecosystem services. Conservation programs usually aim to protect both. Trait-based conservation could then rely on the rarity of species traits (or functional rarity) to identify conservation priorities (*e.g.* for coral reef fish in Grenié et al. 2018).

In addition to studying spatial patterns, traits can be used to study the temporal variation of functional diversity and how aquatic organisms respond to increasing global changes from anthropogenic pressures in the context of biomonitoring.

A.1.4.3 Trait response to global changes

Temporal dynamics of traits and their response to climate change

FTBAs can be used to estimate the temporal response of aquatic organisms and ecosystems to environmental forcing (Figure A.3.3). For example, functional traits have been shown to explain community structure and seasonal dynamics of marine phytoplankton (Edwards et al. 2013b). It is also possible to combine classical data sets, and especially time-series of species abundance, with trait databases described at the species level (see section A.1.3.1) to apply a FTBA to in situ observations and/or monitoring datasets previously collected. In such reanalyses, key traits could be targeted (see section A.1.4.1) to compare their temporal changes, identify tipping points, and reveal trade-offs among traits. Hence, FTBAs offer novel perspectives for a posteriori (re)analysis of historical or long term monitoring data for the study of climate change and its impact on communities and ecosystem functioning (Pomerleau et al. 2015; Abonyi et al. 2018; Floury et al. 2018).

Marine ecologists have long since used FTBAs to study the impact of climate change on aquatic ecosystems (c.f. purple cluster in Supplementary Figure A.4). For example, numerous marine studies explored the response of individual size to climate change (*e.g.* Schmidt et al. 2006; Genner et al. 2010; Finkel et al. 2010), showing that ocean warming is likely to cause a shift towards a larger contribution of smaller organisms to total biomass. Freshwater ecology could benefit from this experience, but currently, two main challenges can be pointed out for both freshwater and marine systems: the identification of links between functional traits and climate-change related variables (*e.g.* acidification and temperature increase in oceans, rivers and lakes; increase of freshwater shortage/scarcity in small streams) but also the deconvolution of the effects of multiple stressors on marine ecosystems (Mouillot et al. 2013). The joint pressure of multiple simultaneous stressors makes the identification of relationships between stressors and functional traits even more complicated, since interactions (*e.g.* synergism, antagonism, additivity, or inhibition) need to be taken into account. Under such conditions, monitoring functional traits of various types (Figure A.1) may prove useful to disentangle these complex interactions.

Impact of climate change on functional diversity

The study of functional diversity may also reveal functional redundancies at the community scale, which may have implications for ecosystem responses to climate change. As a consequence, because functional groups gather together individuals belonging to different species, the loss of a given species with a particular function does not necessarily mean that such function will be lost at higher ecological scale. Indeed, some ecosystems were shown to be insensitive to species loss because multiple species share similar functional roles (mixotrophy, nutrient uptake or requirements), or some species only make a small contribution to the ecosystem processes (Hooper et al.

2005). Recently, it was suggested that climate change may have minor impacts on marine zooplankton functional diversity, due to strong functional redundancy (Benedetti et al. 2019). Conversely, climate change may have contrasting impacts on stream fishes (Buisson and Grenouillet 2009) or decrease their functional diversity (Buisson et al. 2013). By combining climate change scenario modeling with species distribution modeling and functional trait databases, the impact of climate change on the functional diversity of aquatic ecosystems can be assessed at broader scales and across biological compartments and ecosystem boundaries. For aquatic insects, such a combined modeling-FTBA study revealed the spatial patterns of vulnerability to climate change, which also opens opportunities for biomonitoring (Conti et al. 2014). However, limitations remain in the use of trait-based approaches for the assessment of the effects of multiple stressors in the context of climate change, as emphasized recently by (Hamilton et al. 2019) for freshwater invertebrates. These authors pointed out the need to better account for trait redundancy, to better define the appropriate spatial scales for trait applicability and to progress towards the quantification of categorical traits.

Trait-based biomonitoring

Traditionally, the ecological health or "good environmental status" of aquatic ecosystems has been assessed in terms of species composition or relative abundance/biomass of specific indicators, initially within the context of the European Water Framework Directive (WFD-2000/60/CE). However, trait-based approaches offer new opportunities for the monitoring of aquatic ecosystems (Culp et al. 2011), since they can provide new tools that transcend taxonomical denomination, directly related to ecological functions, and exploit the traits available in open databases (Usseglio-Polatera et al. 2000; Baird et al. 2011, see also Table A.2). To date, trait-based biomonitoring has been mainly applied to freshwater ecosystems (cf. the corresponding cluster in Supplementary Figure A.4). Indeed, the links between traits of organisms and natural environmental variables (*e.g.* pH, flow velocity) or even anthropogenic pressures (*e.g.* nutrient or organic matter contamination) have been explored for decades by freshwater ecologists. More specifically, biomonitoring studies put a strong emphasis on the definition and the attribution of traits to taxa such as freshwater benthic macroinvertebrates (Usseglio-Polatera et al. 2000; Menezes et al. 2010), benthic diatoms (Van Dam et al. 1994; Passy 2007) and phytoplankton (Reynolds et al. 2002). As a consequence, the last versions of several biological indices for stream monitoring are mainly based on functional traits (*e.g.* I2M2 in Mondy et al. 2012, BDI in Coste et al. 2009). Within the context of lake monitoring, FTBAs mainly investigated the abundance and the seasonal variability of phytoplanktonic functional groups, as they are known to respond to nutrient concentrations (St-Gelais et al. 2017; Huang et al. 2018). Furthermore, the traits of macroinvertebrates (*e.g.* reproduction mode, size) and diatoms (*e.g.* auto-ecological guilds, life form) are now used in ecotoxicological and ecological models to identify the probability that chemical and/or land use related pressures impair natural communities (Mondy and Usseglio-Polatera 2013; Larras et al. 2017) even in multiple stressor scenarios. Combined with statistical modeling, traits also allow deriving stressor-specific models to assess environmental quality (*e.g.* focusing on invertebrates inhabiting large rivers, Desrosiers et al. 2019).

In comparison to routine monitoring activities in freshwater systems, the use of trait-based monitoring of marine ecosystem is still in its infancy, yet under active development for coastal environments, especially through the implementation of the European Marine Strategy Framework Directive (MSFD-2008/56/EC). FTBAs were proposed to monitor the effects of human activities on benthic communities (*e.g.* Xu et al. 2018), such as bottom trawling and dredging (Tillin et al. 2006), aggregate dredging (Bolam et al. 2016) or pollution (Oug et al. 2012). These approaches can also be used to estimate the success of management strategies, and to predict the effects of future disturbances (including climate change) for marine benthos, by defining critical limits beyond which ecosystem functioning is altered (Bremner 2008). However, functional traits are not yet included in biological indicators and institutional monitoring programs of marine ecosystems, in contrast to what is included in freshwater monitoring efforts. Assessments of functional diversity could inform different MSFD indicators (such as 'biological diversity', 'habitat condition', and 'ecosystem structure'). To our knowledge, trait-based monitoring on marine pelagic ecosystems does not exist. Similar efforts should be extended to open ocean monitoring, for example by incorporating trait data in the reanalysis of long term observations such as the Continuous Plankton Recorder (CPR) time-series in the North Atlantic existing since the 1930s (Richardson et al. 2006).

Both the European Water Framework Directive and the European Marine Strategy Framework Directive require the estimation of the biological status of aquatic ecosystems from the evaluation of each compartment (benthic diatoms, macrophytes and macroalgae, benthic macroinvertebrates, phytoplankton, zooplankton, and fish) independently (*e.g.* Birk et al. 2012). Universal and standardized trait-based indices for biomonitoring should now cover all compartments (Borja et al. 2010). To this end, freshwater ecologists, who have a greater experience in multi-compartment monitoring (Lainé et al. 2014), could inspire marine ecologists, who are more used to focus on one compartment only (*e.g.* benthos, plankton, or fish).

Trait-based ecotoxicology

Besides the policy frameworks, FTBAs can also be used in ecotoxicology to highlight the impact of various stressors (*e.g.* organic synthetics products) on aquatic ecosystems (Baird and Van den Brink 2007). In fact, trait-based ecological risk assessments have been proposed as the new frontier in ecotoxicology (Baird et al. 2008; Rubach et al. 2011). In freshwater systems, diatom traits such as life form (*e.g.* colonial, solitary) or affinities to water quality have already been linked to pesticides contamination (Roubeix et al. 2011). The deformation of their silicified exoskeleton (teratology) has also been considered as a morphological trait that can inform on organisms exposure to heavy metals or pesticides (Lavoie et al. 2017). Similar studies have reported the response of freshwater benthic macroinvertebrate traits to environmental stressors (*e.g.* Statzner and Bêche 2010). For example, Peter et al. (2018) demonstrated that functional traits such as the feeding mode of zooplankton can be used as indicators for the level of metal pollution in freshwater invertebrates at the community level. For marine ecosystems, trait-based ecological risk assessments remain scarce (*e.g.* Neuparth et al. 2002 for marine amphipods). More recently, -omics techniques offer new ways for estimating physiological traits related to pollutant catabolism, for example, by detecting the activity of particular genes (*e.g.* mercury methylating genes in the ocean, Villar et

al. 2019). The recent development of appropriate statistical tools will help to integrate omics data within the framework for ecological risk assessment (Larras et al. 2018). Similarly, imaging could allow to automatically identify changes in morphological traits as a response to environmental stressors (*e.g.* Maps et al. 2019). Altogether, the high-throughput acquisition of -omics data and images could allow the detection of new environmental stressors (*e.g.* Bowler et al. 2009; Reid and Whitehead 2016). Such state-of-the-art tools can contribute to the development of universal multi-compartment indices, that could provide estimates automatically and in almost real time. Ultimately, this could expand biomonitoring approaches beyond traditional taxonomically based assessments.

A.1.4.4 Scaling up from functional traits to community structure and ecosystem functions

Finally, FTBAs could be used to explore trophic interactions and food webs (Reiss et al. 2009). Indeed, several traits directly reflect trophic interactions (*e.g.* diet, size, stoichiometric traits) and can be used to better understand food web structure and dynamics (Figure A.3.4). However, scaling-up from individual traits to populations, communities, and ecosystems requires taking trait variation at multiple intermediate organisation scales into account (*e.g.* population, meta-population, and community scales; Gibert et al. 2015). Rather than considering a collection of traits independently, one approach is to analyse how these traits influence or reveal the biotic interactions and trophic structure of aquatic communities. To do so, the numerous traits that are directly related to the way consumers interact with their prey (*e.g.* diet, feeding modes, motility, and perception of sounds) or the way prey interact with their predators (*e.g.* toxin production, bioluminescence, migration) are emphasized in the following subparts.

Body size as a major functional trait driving trophic interactions

The functional trait of body/cell size plays a particularly important role and is often referred to as “a master trait”. Size influences most of the ecological, physiological and behavioral functions of organisms due to metabolic laws, underpinning trophic position and interactions that are especially influenced by relative prey and predator sizes (Weitz and Levin 2006; Conley et al. 2018). Size or morphological characteristics can potentially be measured directly using imaging methods (see section A.1.3.2), and could be used to infer trophic relationships. Predator traits (*i.e.*, body/cell size and motility type) may also be responsible for the body-size architecture of natural food webs in freshwater, marine and terrestrial ecosystems (Brose et al. 2019). At large spatial scales, body size and prey selection were shown to be modified by climate change and therefore to strongly impact food webs and ecosystem functions in return (Hoegh-Guldberg and Bruno 2010; Sheridan and Bickford 2011). For example, ocean warming was associated with a reduction in copepod body size, which may impact upper trophic levels and ultimately fisheries (Beaugrand et al. 2010, but see also Renaud et al. 2018). More general laws between size, trophic interactions and environmental variables could be tested in future trait-based studies, especially by taking advantage of automatic morphological measurements (including but not restricted to size) through imaging. More specifically, direct observations of predator-prey interactions and

associated traits could be performed by combining imaging (Choy et al. 2017; Ohman 2019) with gut content and/or faeces analysis based on taxonomic and/or -omics description, such tools being complementary and sometimes even more informative than the stable isotope methods that have been traditionally used so far (*e.g.* Majdi et al. 2018).

Including stoichiometric traits to study trophic interactions

In addition to body size, stoichiometric traits are highly promising for integrating FTBAs into food web models and to bridge the gap between community structure and ecosystem functioning (Meunier et al. 2017). Because all organisms are composed of the same major elements (*e.g.* C, N, and P), their balance not only reflects nutrient cycling in the ecosystem but also food web topologies. Quantifying stoichiometric traits across taxonomic and trophic groups allows the depiction of trophic interactions. In food web approaches, trophic position is associated with significant changes in C:N:P ratios, as well as altered isotope ratios due to selective uptake. As an example, heterotrophs are generally relatively less rich in carbon than autotrophs (Hessen et al. 2004; Persson et al. 2010). However, while stoichiometric composition and variation have been quantified for some species in different taxonomic groups (*e.g.* some plants, marine bacteria or plankton), there is still a lack of knowledge of the C:N:P ratios and their variations for numerous taxa, including higher-level consumers (*e.g.* Frost et al. 2002, 2006). Stoichiometric gradients may also inform on some specific traits such as growth rate, food preferences, nutrient acquisition and on some life history traits type such as fecundity, or even genome and cell size (see review in Carnicer et al. 2015). Indeed, stoichiometric ratios have the advantage of being directly related to organismal growth rates, which are central life history traits. The "growth rate hypothesis" demonstrates that rapidly growing organisms commonly have low biomass C:P and N:P ratios. This observation is explained by a high demand for P-rich ribosomal RNA, but also by the shorter lifespan of faster growing organisms, which prevents large investments into reserve structures (Elser et al. 1996, 2003). Consumers stoichiometry, in addition to metabolic characteristics, also gives important information on consumers driven nutrient recycling (Allen and Gillooly 2009). Better documenting the stoichiometric ratios of aquatic organisms in existing trait databases would help to identify their drivers and thus improve our understanding of the impact of stoichiometric traits on food web dynamics and ecosystem functioning.

From aquatic functional traits to global biogeochemical cycles

Finally, studying aquatic food webs following a FTBA should improve predictions of nutrient and carbon fluxes at the ecosystem scale (Vanni and McIntyre 2016). For example, trait-based models of food webs could be constructed to infer trophic interactions influencing ecosystem stocks and fluxes (Woodward et al. 2005). In addition to size and stoichiometry, several other functional traits could be taken into account in these models, such as predator foraging and prey vulnerability traits (Boukal 2014). To do so, one promising pathway is to increase the exploitation of trait databases. For example, global datasets of marine plankton abundances and biomass were recently coupled with a trait-based model used to predict dominant feeding strategies in pelagic ambush predators and to estimate the effects of these feeding traits on energy and biomass transfer efficiency (Prowe et al. 2019). For fish, diets and trophic strategies can be predicted from their functional traits

(Albouy et al. 2011). This approach could be extended to other aquatic organisms. Scaling-up from individual traits to food web dynamics should ultimately contribute to better understand the response of aquatic ecosystems to environmental changes in terms of biogeochemical cycling, ultimately improving long-term prediction of ecosystem dynamics and feedback mechanisms to climate.

A.1.5 Conclusions

The main goal of FTBAs is to improve our understanding of the links between community structure, ecosystem function and ecosystem service provision. The main advantages of such approaches come from the definition of traits at the individual level. Indeed, this allows for the direct measurement of the functional traits of any organism without an additional step of taxonomic assignment that may be time-consuming. This can also provide access to universal ecological rules (transcending trophic levels and ecosystems). On the contrary, FTBAs would not be adapted to study population dynamics that require taxonomic description at the species level, nor to directly estimate bulk properties of the communities (which would require summing of individual-level information). For these reasons, the description and quantification of functional traits provide a common basis across diverse ecological fields, from ecophysiology to community and ecosystem ecology, via population and evolutionary biology. Yet, distinct questions and methods are often specific to each identified habitat (*i.e.* benthic and pelagic) or even to each biological compartment (i.e. invertebrates and diatoms for the freshwater benthic habitat). Here, we proposed functional trait-based pathways across multiple ecological components. As a first step, we: i) homogenized the terminology used in FTBAs and provided a common typology for aquatic functional traits that can be used across various aquatic systems and for multi-compartment studies, ii) listed the currently available databases dedicated to (aquatic) functional traits, iii) described classical and emerging methods for estimating traits of marine and freshwater organisms, and iv) highlighted some key traits that could be used for multi-compartment and trans-ecosystem studies. Establishing such a common ground among aquatic ecologists is required to further encourage and stimulate collaborative research across disciplines. The next step would be to create a common ontology dedicated to FTBAs, such as the Open Traits Network initiative (Gallagher et al. 2019), in order to improve the sharing of trait information in databases.

The recent methodologies we described offer new opportunities to study traits at various scales, from -omic sequences to whole-ecosystem approaches and biogeochemical cycles. Imaging, -omics and modeling tools are amongst the most promising emerging approaches to work with traits across the tree of life. We propose extending discussions within aquatic ecologists, including freshwater, marine, benthic and pelagic fields, to better share expertise in these tools, thereby improving our knowledge on potential and realized functional traits. With these methodologies, FTBAs provide promising foundations for the development of integrated frameworks that combine ecological theories with empirical knowledge across scales.

References

Abonyi, A., Z. Horváth, and R. Ptacnik. 2018. Functional richness outperforms taxonomic richness in predicting ecosystem functioning in natural phytoplankton communities. Freshw. Biol. 63: 178–186. doi:10.1111/fwb.13051

Albouy, C., F. Guilhaumon, S. Villéger, and others. 2011. Predicting trophic guild and diet overlap from functional traits: statistics, opportunities and limitations for marine ecology. Mar. Ecol. Prog. Ser. 436: 17–28. doi:10.3354/meps09240

Allen, A. P., and J. F. Gillooly. 2009. Towards an integration of ecological stoichiometry and the metabolic theory of ecology to better understand nutrient cycling. Ecol. Lett. 12: 369–384. doi:10.1111/j.1461-0248.2009.01302.x

Althaus, F., N. Hill, R. Ferrari, and others. 2015. A Standardised Vocabulary for Identifying Benthic Biota and Substrata from Underwater Imagery: The CATAMI Classification Scheme J. Hewitt [ed.]. PLOS ONE 10: e0141039. doi:10.1371/journal.pone.0141039

Andersen, K. H., N. S. Jacobsen, and K. D. Farnsworth. 2016. The theoretical foundations for size spectrum models of fish communities J. Baum [ed.]. Can. J. Fish. Aquat. Sci. 73: 575–588. doi:10.1139/cjfas-2015-0230

Baird, D. J., C. J. O. Baker, R. B. Brua, M. Hajibabaei, K. McNicol, T. J. Pascoe, and D. de Zwart. 2011. Toward a knowledge infrastructure for traits-based ecological risk assessment. Integr. Environ. Assess. Manag. 7: 209–215. doi:10.1002/ieam.129

Baird, D. J., M. N. Rubach, and P. J. V. den Brinkt. 2008. Trait-based ecological risk assessment (TERA): The new frontier? Integr. Environ. Assess. Manag. 4: 2–3. doi:10.1897/IEAM_2007-063.1

Baird, D. J., and P. J. Van den Brink. 2007. Using biological traits to predict species sensitivity to toxic substances. Ecotoxicol. Environ. Saf. 67: 296–301. doi:10.1016/j.ecoenv.2006.07.001

Banas, N. S., and R. G. Campbell. 2016. Traits controlling body size in copepods: separating general constraints from species-specific strategies. Mar. Ecol. Prog. Ser. 558: 21–33. doi:10.3354/meps11873

Barnett, A. J., K. Finlay, and B. E. Beisner. 2007. Functional diversity of crustacean zooplankton communities: towards a trait-based classification. Freshw. Biol. 52: 796–813. doi:10.1111/j.1365-2427.2007.01733.x

Barton, A. D., A. J. Pershing, E. Litchman, N. R. Record, K. F. Edwards, Z. V. Finkel, T. Kiørboe, and B. A. Ward. 2013. The biogeography of marine plankton traits. Ecol. Lett. 16: 522–534. doi:10.1111/ele.12063

Beauchard, O., H. Veríssimo, A. M. Queirós, and P. M. J. Herman. 2017. The use of multiple biological traits in marine community ecology and its potential in ecological indicator development. Ecol. Indic. 76: 81–96. doi:10.1016/j.ecolind.2017.01.011

Beaugrand, G., M. Edwards, and L. Legendre. 2010. Marine biodiversity, ecosystem functioning, and carbon cycles. Proc. Natl. Acad. Sci. 107: 10120–10124. doi:10.1073/pnas.0913855107

Bello, F. de, S. Lavorel, C. H. Albert, W. Thuiller, K. Grigulis, J. Dolezal, Š. Janeček, and J. Lepš. 2011. Quantifying the relevance of intraspecific trait variability for functional diversity. Methods Ecol. Evol. 2: 163–174. doi:10.1111/j.2041-210X.2010.00071.x

Benedetti, F. 2015. Mediterranean copepods' functional traits. PANGAEA. doi:https://doi.org/10.1594/PANGAEA.854331

Benedetti, F., S. Gasparini, and S.-D. Ayata. 2016. Identifying copepod functional groups from species functional traits. J. Plankton Res. 38: 159–166. doi:10.1093/plankt/fbv096

Benedetti, F., L. Jalabert, M. Sourisseau, and others. 2019. The Seasonal and Inter-Annual Fluctuations of Plankton Abundance and Community Structure in a North Atlantic Marine Protected Area. Front. Mar. Sci. 6. doi:10.3389/fmars.2019.00214

Benedetti, F., M. Vogt, D. Righetti, F. Guilhaumon, and S.-D. Ayata. 2018. Do functional groups of planktonic copepods differ in their ecological niches? J. Biogeogr. 45: 604–616. doi:10.1111/jbi.13166

Benoit-Bird, K. J., and G. L. Lawson. 2016. Ecological Insights from Pelagic Habitats Acquired Using Active Acoustic Techniques. Annu. Rev. Mar. Sci. 8: 463–490. doi:10.1146/annurev-marine-122414-034001

Beukhof, E., T. S. Dencker, M. L. D. Palomares, and A. Maureaud. 2019. A trait collection of marine fish species from North Atlantic and Northeast Pacific continental shelf seas.doi:https://doi.org/10.1594/PANGAEA.900866

Bi, H., S. Cook, H. Yu, M. C. Benfield, and E. D. Houde. 2012. Deployment of an imaging system to investigate fine-scale spatial distribution of early life stages of the ctenophore Mnemiopsis leidyi in Chesapeake Bay. J. Plankton Res. 35: 270–280.

Biard, T., L. Stemmann, M. Picheral, and others. 2016. In situ imaging reveals the biomass of giant protists in the global ocean. Nature 532: 504–507. doi:10.1038/nature17652

Birk, S., W. Bonne, A. Borja, and others. 2012. Three hundred ways to assess Europe's surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. Ecol. Indic. 18: 31–41. doi:10.1016/j.ecolind.2011.10.009

Blanco-Bercial, L., and A. E. Maas. 2018. A transcriptomic resource for the northern krill Meganyctiphanes norvegica based on a short-term temperature exposure experiment. Mar. Genomics 38: 25–32. doi:10.1016/j.margen.2017.05.013

Bolam, S. G., P. S. O. McIlwaine, and C. Garcia. 2016. Application of biological traits to further our understanding of the impacts of dredged material disposal on benthic assemblages. Mar. Pollut. Bull. 105: 180–192. doi:10.1016/j.marpolbul.2016.02.031

Borgy, B., C. Violle, P. Choler, and others. 2017. Sensitivity of community-level trait–environment relationships to data representativeness: A test for functional biogeography. Glob. Ecol. Biogeogr. 26: 729–739. doi:10.1111/geb.12573

Borja, Á., M. Elliott, J. Carstensen, A.-S. Heiskanen, and W. van de Bund. 2010. Marine manage-

ment – Towards an integrated implementation of the European Marine Strategy Framework and the Water Framework Directives. Mar. Pollut. Bull. 60: 2175–2186. doi:10.1016/j.marpolbul.2010.09.026

Boukal, D. S. 2014. Trait- and size-based descriptions of trophic links in freshwater food webs: current status and perspectives. J. Limnol. doi:10.4081/jlimnol.2014.826

Bowler, C., D. M. Karl, and R. R. Colwell. 2009. Microbial oceanography in a sea of opportunity. Nature 459: 180–184. doi:10.1038/nature08056

Bremner, J. 2008. Species' traits and ecological functioning in marine conservation and management. J. Exp. Mar. Biol. Ecol. 366: 37–47. doi:10.1016/j.jembe.2008.07.007

Bremner, J., S. I. Rogers, and C. L. J. Frid. 2006. Methods for describing ecological functioning of marine benthic assemblages using biological traits analysis (BTA). Ecol. Indic. 6: 609–622. doi:10.1016/j.ecolind.2005.08.026

Brose, U., P. Archambault, A. D. Barnes, and others. 2019. Predator traits determine food-web architecture across ecosystems. Nat. Ecol. Evol. 3: 919–927. doi:10.1038/s41559-019-0899-x

Brown, M. V., M. Ostrowski, J. J. Grzymski, and F. M. Lauro. 2014. A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades. Mar. Genomics 15: 17–28. doi:10.1016/j.margen.2014.03.002

Brun, P., T. Kiørboe, P. Licandro, and M. R. Payne. 2016a. The predictive skill of species distribution models for plankton in a changing climate. Glob. Change Biol. 22: 3170–3181. doi:10.1111/gcb.13274

Brun, P., M. R. Payne, and T. Kiørboe. 2016b. Trait biogeography of marine copepods – an analysis across scales. Ecol. Lett. 19: 1403–1413. doi:10.1111/ele.12688

Brun, P., M. R. Payne, and T. Kiørboe. 2017. A trait database for marine copepods. Earth Syst. Sci. Data Discuss. 9: 99–113. doi:10.5194/essd-9-99-2017

Bucklin, A., D. Steinke, and L. Blanco-Bercial. 2011. DNA Barcoding of Marine Metazoa. Annu. Rev. Mar. Sci. 3: 471–508. doi:10.1146/annurev-marine-120308-080950

Buisson, L., and G. Grenouillet. 2009. Contrasted impacts of climate change on stream fish assemblages along an environmental gradient. Divers. Distrib. 15: 613–626. doi:10.1111/j.1472-4642.2009.00565.x

Buisson, L., G. Grenouillet, S. Villéger, J. Canal, and P. Laffaille. 2013. Toward a loss of functional diversity in stream fish assemblages under climate change. Glob. Change Biol. 19: 387–400. doi:10.1111/gcb.12056

Cadotte, M. W., C. A. Arnillas, S. W. Livingstone, and S.-L. E. Yasui. 2015. Predicting communities from functional traits. Trends Ecol. Evol. 30: 510–511. doi:10.1016/j.tree.2015.07.001

del Campo, J., M. Kolisko, V. Boscaro, and others. 2018. EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. PLOS Biol. 16: e2005849. doi:10.1371/journal.pbio.2005849

Carmona, C. P., F. de Bello, N. W. H. Mason, and J. Lepš. 2016. Traits Without Borders: Integrating Functional Diversity Across Scales. Trends Ecol. Evol. 31: 382–394. doi:10.1016/j.tree.2016.02.003

Carnicer, J., J. Sardans, C. Stefanescu, A. Ubach, M. Bartrons, D. Asensio, and J. Peñuelas. 2015. Global biodiversity, stoichiometry and ecosystem function responses to human-induced C–N–P imbalances. J. Plant Physiol. 172: 82–91. doi:10.1016/j.jplph.2014.07.022

Céréghino, R., V. D. Pillar, D. S. Srivastava, and others. 2018. Constraints on the functional trait space of aquatic invertebrates in bromeliads. Funct. Ecol. 32: 2435–2447. doi:10.1111/1365-2435.13141

Chapman, A. S. A., S. E. Beaulieu, A. Colaço, and others. 2019. sFDvent: A global trait database for deep-sea hydrothermal-vent fauna. Glob. Ecol. Biogeogr. 28: 1538–1551. doi:10.1111/geb.12975

Chevenet, Fran., S. Doleadec, and D. Chessel. 1994. A fuzzy coding approach for the analysis of long-term ecological data. Freshw. Biol. 31: 295–309. doi:10.1111/j.1365-2427.1994.tb01742.x

Chonova, T., R. Kurmayer, F. Rimet, J. Labanowski, V. Vasselon, F. Keck, P. Illmer, and A. Bouchez. 2019. Benthic Diatom Communities in an Alpine River Impacted by Waste Water Treatment Effluents as Revealed Using DNA Metabarcoding. Front. Microbiol. 10. doi:10.3389/fmicb.2019.00653

Choy, C. A., Haddock, S. H. D., and B. H. Robison. 2017. Deep pelagic food web structure as revealed by in situ feeding observations. Proc. R. Soc. B Biol. Sci. 284: 20172116. doi:10.1098/rspb.2017.2116

Coles, V. J., M. R. Stukel, M. T. Brooks, and others. 2017. Ocean biogeochemistry modeled with emergent trait-based genomics. Science 358: 1149–1154. doi:10.1126/science.aan5712

Compson, Z. G., W. A. Monk, C. J. Curry, and others. 2018. Chapter Two - Linking DNA Metabarcoding and Text Mining to Create Network-Based Biomonitoring Tools: A Case Study on Boreal Wetland Macroinvertebrate Communities, p. 33–74. In D.A. Bohan, A.J. Dumbrell, G. Woodward, and M. Jackson [eds.], Advances in Ecological Research. Academic Press.

Conley, K. R., F. Lombard, and K. R. Sutherland. 2018. Mammoth grazers on the ocean's minuteness: a review of selective feeding using mucous meshes. Proc. Biol. Sci. 285. doi:10.1098/rspb.2018.0056

Conti, L., A. Schmidt-Kloiber, G. Grenouillet, and W. Graf. 2014. A trait-based approach to assess the vulnerability of European aquatic insects to climate change. Hydrobiologia 721: 297–315. doi:10.1007/s10750-013-1690-7

Coquereau, L., J. Grall, L. Chauvaud, C. Gervaise, J. Clavier, A. Jolivet, and L. Di Iorio. 2016. Sound production and associated behaviours of benthic invertebrates from a coastal habitat in the north-east Atlantic. Mar. Biol. 163: 127. doi:10.1007/s00227-016-2902-2

Coste, M., S. Boutry, J. Tison-Rosebery, and F. Delmas. 2009. Improvements of the Biological

Diatom Index (BDI): Description and efficiency of the new version (BDI-2006). Ecol. Indic. 9: 621–650. doi:10.1016/j.ecolind.2008.06.003

Costello, M. J., S. Claus, S. Dekeyzer, L. Vandepitte, É. Ó. Tuama, D. Lear, and H. Tyler-Walters. 2015. Biological and ecological traits of marine species. PeerJ 3: e1201. doi:10.7717/peerj.1201

Cowen, R. K., and C. M. Guigand. 2008. In situ ichthyoplankton imaging system (ISIIS): system design and preliminary results. Limnol. Oceanogr. Methods 6: 126–132. doi:10.4319/lom.2008.6.126

Culp, J. M., D. G. Armanini, M. J. Dunbar, J. M. Orlofske, N. L. Poff, A. I. Pollard, A. G. Yates, and G. C. Hose. 2011. Incorporating traits in aquatic biomonitoring to enhance causal diagnosis and prediction. Integr. Environ. Assess. Manag. 7: 187–197. doi:10.1002/ieam.128

Culverhouse, P. F., R. Williams, M. Benfield, P. R. Flood, A. F. Sell, M. G. Mazzocchi, I. Buttino, and M. Sieracki. 2006. Automatic image analysis of plankton: future perspectives. Mar. Ecol. Prog. Ser. 312: 297–309. doi:10.3354/meps312297

Degen, R., M. Aune, B. A. Bluhm, and others. 2018. Trait-based approaches in rapidly changing ecosystems: A roadmap to the future polar oceans. Ecol. Indic. 91: 722–736. doi:https://doi.org/10.1016/j.ecolind.2018.04.050>

Degen, R., and S. Faulwetter. 2019. The Arctic Traits Database – a repository of Arctic benthic invertebrate traits. Earth Syst. Sci. Data 11: 301–322. doi:https://doi.org/10.5194/essd-11-301-2019

Deiner, K., H. M. Bik, E. Mächler, and others. 2017. Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Mol. Ecol. 26: 5872–5895. doi:10.1111/mec.14350

Des Roches, S., D. M. Post, N. E. Turley, J. K. Bailey, A. P. Hendry, M. T. Kinnison, J. A. Schweitzer, and E. P. Palkovacs. 2018. The ecological importance of intraspecific variation. Nat. Ecol. Evol. 2: 57–64. doi:10.1038/s41559-017-0402-5

Desjonquères, C. 2016. Ecologie et diversité acoustique des milieux aquatiques: exploration en milieux tempérés. thesis. Paris, Muséum national d'histoire naturelle.

Desjonquères, C., F. Rybak, E. Castella, D. Llusia, and J. Sueur. 2018. Acoustic communities reflects lateral hydrological connectivity in riverine floodplain similarly to macroinvertebrate communities. Sci. Rep. 8: 14387. doi:10.1038/s41598-018-31798-4

Desrosiers, M., P. Usseglio-Polatera, V. Archaimbault, F. Larras, G. Méthot, and B. Pinel-Alloul. 2019. Assessing anthropogenic pressure in the St. Lawrence River using traits of benthic macroinvertebrates. Sci. Total Environ. 649: 233–246. doi:10.1016/j.scitotenv.2018.08.267

Dubelaar, G. B. J., and P. L. Gerritzen. 2000. CytoBuoy: a step forward towards using flow cytometry in operational oceanography. Sci. Mar. 64: 255–265. doi:10.3989/scimar.2000.64n2255

Dubelaar, G. B. J., P. L. Gerritzen, A. E. R. Beeker, R. R. Jonker, and K. Tangen. 1999. Design and first results of CytoBuoy: A wireless flow cytometer for in situ analysis of marine and fresh waters.

Cytometry 37: 247–254. doi:10.1002/(SICI)1097-0320(19991201)37:4<247::AID-CYTO1>3.0.CO;2-9

Durden, J. M., J. Y. Luo, H. Alexander, A. M. Flanagan, and L. Grossmann. 2017. Integrating "Big Data" into Aquatic Ecology: Challenges and Opportunities. Limnol. Oceanogr. Bull. 26: 101–108. doi:10.1002/lob.10213

Edwards, K. F., C. A. Klausmeier, and E. Litchman. 2013a. A Three-Way Trade-Off Maintains Functional Diversity under Variable Resource Supply. Am. Nat. 182: 786–800. doi:10.1086/673532

Edwards, K. F., E. Litchman, and C. A. Klausmeier. 2013b. Functional traits explain phytoplankton community structure and seasonal dynamics in a marine ecosystem. Ecol. Lett. 16: 56–63. doi:10.1111/ele.12012

Edwards, K. F., M. K. Thomas, C. A. Klausmeier, and E. Litchman. 2012. Allometric scaling and taxonomic variation in nutrient utilization traits and maximum growth rate of phytoplankton. Limnol. Oceanogr. 57: 554–566. doi:10.4319/lo.2012.57.2.0554

Ehrlich, E., L. Becks, and U. Gaedke. 2017. Trait–fitness relationships determine how trade-off shapes affect species coexistence. Ecology 98: 3188–3198. doi:10.1002/ecy.2047

Ehrlich, E., N. J. Kath, and U. Gaedke. 2020. The shape of a defense-growth trade-off governs seasonal trait dynamics in natural phytoplankton. ISME J. 14: 1451–1462. doi:10.1038/s41396-020-0619-1

Elser, J. J., K. Acharya, M. Kyle, and others. 2003. Growth rate–stoichiometry couplings in diverse biota. Ecol. Lett. 6: 936–943. doi:10.1046/j.1461-0248.2003.00518.x

Elser, J. J., D. R. Dobberfuhl, N. A. MacKay, and J. H. Schampel. 1996. Organism Size, Life History, and N:P Stoichiometry. BioScience 46: 674–684. doi:10.2307/1312897

Farrell, F., O. S. Soyer, and C. Quince. 2018. Machine learning based prediction of functional capabilities in metagenomically assembled microbial genomes. bioRxiv 307157. doi:10.1101/307157

Faulwetter, S., V. Markantonatou, C. Pavloudi, and others. 2014. Polytraits: A database on biological traits of marine polychaetes. Biodivers. Data J. e1024. doi:10.3897/BDJ.2.e1024

Faure, E., F. Not, A.-S. Benoiston, K. Labadie, L. Bittner, and S.-D. Ayata. 2019. Mixotrophic protists display contrasted biogeographies in the global ocean. ISME J. 13: 1072. doi:10.1038/s41396-018-0340-5

Finkel, Z. V., J. Beardall, K. J. Flynn, A. Quigg, T. A. V. Rees, and J. A. Raven. 2010. Phytoplankton in a changing world: cell size and elemental stoichiometry. J. Plankton Res. 32: 119–137. doi:10.1093/plankt/fbp098

Finlay, K., B. E. Beisner, and A. J. D. Barnett. 2007. The use of the Laser Optical Plankton Counter to measure zooplankton size, abundance, and biomass in small freshwater lakes. Limnol. Oceanogr. Methods 5: 41–49. doi:10.4319/lom.2007.5.41

Floury, M., Y. Souchon, and K. V. Looy. 2018. Climatic and trophic processes drive long-term

changes in functional diversity of freshwater invertebrate communities. Ecography 41: 209–218. doi:10.1111/ecog.02701

Fontana, S., J. Jokela, and F. Pomati. 2014. Opportunities and challenges in deriving phytoplankton diversity measures from individual trait-based data obtained by scanning flow-cytometry. Front. Microbiol. 5. doi:10.3389/fmicb.2014.00324

Forest, A., L. Stemmann, M. Picheral, L. Burdorf, D. Robert, L. Fortier, and M. Babin. 2012. Size distribution of particles and zooplankton across the shelf-basin system in southeast Beaufort Sea: combined results from an Underwater Vision Profiler and vertical net tows. Biogeosciences 9: 1301–1320. doi:10.5194/bg-9-1301-2012

Fragoso, G. M., A. J. Poulton, I. M. Yashayaev, E. J. H. Head, G. Johnsen, and D. A. Purdie. 2018. Diatom Biogeography From the Labrador Sea Revealed Through a Trait-Based Approach. Front. Mar. Sci. 5.

Frimpong, E., and P. Angermeier. 2010. Trait-based approaches in the analysis of stream fish communities. 73: 109–136.

Froese, R., and D. Pauly. 2019. FishBase. World Wide Web electronic publication.

Frost, P. C., J. P. Benstead, W. F. Cross, H. Hillebrand, J. H. Larson, M. A. Xenopoulos, and T. Yoshida. 2006. Threshold elemental ratios of carbon and phosphorus in aquatic consumers. Ecol. Lett. 9: 774–779. doi:10.1111/j.1461-0248.2006.00919.x

Frost, P. C., R. S. Stelzer, G. A. Lamberti, and J. Elser. 2002. Ecological stoichiometry of trophic interactions in the benthos: Understanding the role of C:N:P ratios in lentic and lotic habitats. J. North Am. Benthol. Soc. 21: 515–528.

Fu, H., J. Zhong, G. Yuan, L. Ni, P. Xie, and T. Cao. 2014. Functional traits composition predict macrophytes community productivity along a water depth gradient in a freshwater lake. Ecol. Evol. 4: 1516–1523. doi:10.1002/ece3.1022

Gallagher, R., D. S. Falster, B. Maitner, and others. 2019. The Open Traits Network: Using Open Science principles to accelerate trait-based science across the Tree of Life. preprint EcoEvoRxiv.

Genner, M. J., D. W. Sims, A. J. Southward, and others. 2010. Body size-dependent responses of a marine fish assemblage to climate change and fishing over a century-long scale. Glob. Change Biol. 16: 517–527. doi:10.1111/j.1365-2486.2009.02027.x

Gibert, J. P., A. I. Dell, J. P. DeLong, and S. Pawar. 2015. Chapter One - Scaling-up Trait Variation from Individuals to Ecosystems, p. 1–17. In S. Pawar, G. Woodward, and A.I. Dell [eds.], Advances in Ecological Research. Academic Press.

González-Rivero, M., O. Beijbom, A. Rodriguez-Ramirez, and others. 2016. Scaling up Ecological Measurements of Coral Reefs Using Semi-Automated Field Image Collection and Analysis. Remote Sens. 8: 30. doi:10.3390/rs8010030

Gorsky, G., M. D. Ohman, M. Picheral, and others. 2010. Digital zooplankton image analysis using the ZooScan integrated system. J. Plankton Res. 32: 285–303. doi:10.1093/plankt/fbp124

Grenié, M., D. Mouillot, S. Villéger, P. Denelle, C. M. Tucker, F. Munoz, and C. Violle. 2018. Functional rarity of coral reef fishes at the global scale: Hotspots and challenges for conservation. Biol. Conserv. 226: 288–299. doi:10.1016/j.biocon.2018.08.011

Griffiths, J. R., M. Kadin, F. J. A. Nascimento, and others. 2017. The importance of benthic–pelagic coupling for marine ecosystem functioning in a changing world. Glob. Change Biol. 23: 2179–2196. doi:10.1111/gcb.13642

Grizzetti, B., D. Lanzanova, C. Liquete, A. Reynaud, and A. C. Cardoso. 2016. Assessing water ecosystem services for water resource management. Environ. Sci. Policy 61: 194–203. doi:10.1016/j.envsci.2016.04.008

Guillou, L., D. Bachar, S. Audic, and others. 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41: D597–D604. doi:10.1093/nar/gks1160

Hamilton, A. T., R. B. Schäfer, M. I. Pyne, and others. 2019. Limitations of trait-based approaches for stressor assessment: The case of freshwater invertebrates and climate drivers. Glob. Change Biol. doi:10.1111/gcb.14846

Hébert, M.-P., B. E. Beisner, and R. Maranger. 2016. A meta-analysis of zooplankton functional traits influencing ecosystem function. Ecology 97: 1069–1080. doi:10.1890/15-1084.1

Hébert, M.-P., B. E. Beisner, and R. Maranger. 2017. Linking zooplankton communities to ecosystem functioning: toward an effect-trait framework. J. Plankton Res. 39: 3–12. doi:10.1093/plankt/fbw068

Hecky, R. E., and P. Kilham. 1988. Nutrient limitation of phytoplankton in freshwater and marine environments: A review of recent evidence on the effects of enrichment. Limnol. Oceanogr. 33: 796–822. doi:10.4319/lo.1988.33.4part2.0796

Heino, J., J. Soininen, J. Alahuhta, J. Lappalainen, and R. Virtanen. 2015. A comparative analysis of metacommunity types in the freshwater realm. Ecol. Evol. 5: 1525–1537. doi:10.1002/ece3.1460

Hemingson, C. R., and D. R. Bellwood. 2018. Biogeographic patterns in major marine realms: function not taxonomy unites fish assemblages in reef, seagrass and mangrove systems. Ecography 41: 174–182. doi:10.1111/ecog.03010

Henriques, S., F. Guilhaumon, S. Villéger, S. Amoroso, S. França, S. Pasquaud, H. N. Cabral, and R. P. Vasconcelos. 2017. Biogeographical region and environmental conditions drive functional traits of estuarine fish assemblages worldwide. Fish Fish. 18: 752–771. doi:10.1111/faf.12203

Herring, P. J. 1987. Systematic distribution of bioluminescence in living organisms. J. Biolumin. Chemilumin. 1: 147–163. doi:10.1002/bio.1170010303

Hessen, D. O., G. I. Ågren, T. R. Anderson, J. J. Elser, and P. C. de Ruiter. 2004. Carbon Sequestration in Ecosystems: The Role of Stoichiometry. Ecology 85: 1179–1192. doi:10.1890/02-0251

Hoegh-Guldberg, O., and J. F. Bruno. 2010. The Impact of Climate Change on the World's Marine Ecosystems. Science 328: 1523–1528. doi:10.1126/science.1189930

Hood, R. R., E. A. Laws, R. A. Armstrong, and others. 2006. Pelagic functional group modeling: Progress, challenges and prospects. Deep Sea Res. Part II Top. Stud. Oceanogr. 53: 459–512. doi:10.1016/j.dsr2.2006.01.025

Hooper, D. U., F. S. Chapin, J. J. Ewel, and others. 2016. Effects of Biodiversity on Ecosystem Functioning: A Consensus of Current Knowledge. Ecol. Monogr. 3–35. doi:10.1890/04-0922@10.1002/(ISSN)1557-7015(CAT)VirtualIssue(VI)ECM

Huang, G., X. Wang, Y. Chen, L. Xu, and D. Xu. 2018. Seasonal succession of phytoplankton functional groups in a reservoir in central China.doi:info:doi/10.1127/fal/2018/1083

Jänes, H., J. Kotta, M. Pärnoja, T. P. Crowe, F. Rindi, and H. Orav-Kotta. 2017. Functional traits of marine macrophytes predict primary production E. Carrington [ed.]. Funct. Ecol. 31: 975–986. doi:10.1111/1365-2435.12798

de Juan, S., J. Hewitt, S. Thrush, and D. Freeman. 2015. Standardising the assessment of Functional Integrity in benthic ecosystems. J. Sea Res. 98: 33–41. doi:10.1016/j.seares.2014.06.001

Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28: 27–30. doi:10.1093/nar/28.1.27

Katija, K., C. A. Choy, R. E. Sherlock, A. D. Sherman, and B. H. Robison. 2017. From the surface to the seafloor: How giant larvaceans transport microplastics into the deep sea. Sci. Adv. 3: e1700715. doi:10.1126/sciadv.1700715

Keck, F., V. Vasselon, F. Rimet, A. Bouchez, and M. Kahlert. 2018. Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of operational taxonomic units' ecological profiles. Mol. Ecol. Resour. 18: 1299–1309. doi:10.1111/1755-0998.12919

Kiørboe, T., and A. G. Hirst. 2014. Shifts in Mass Scaling of Respiration, Feeding, and Growth Rates across Life-Form Transitions in Marine Pelagic Organisms. Am. Nat. 183: E118–E130. doi:10.1086/675241

Kiørboe, T., A. Visser, and K. H. Andersen. 2018. A trait-based approach to ocean ecology. ICES J. Mar. Sci. 75: 1849–1863. doi:10.1093/icesjms/fsy090 Kolkwitz, R., and M. Marsson. 1909. Ökologie der tierischen Saprobien. Beiträge zur Lehre von der biologischen Gewässerbeurteilung. Int. Rev. Gesamten Hydrobiol. Hydrogr. 2: 126–152. doi:10.1002/iroh.19090020108

Kremer, C. T., A. K. Williams, M. Finiguerra, A. A. Fong, A. Kellerman, S. F. Paver, B. B. Tolar, and B. J. Toscano. 2017. Realizing the potential of trait-based aquatic ecology: New tools and collaborative approaches: Challenges of trait-based aquatic ecology. Limnol. Oceanogr. 62: 253–271. doi:10.1002/lno.10392

Lainé, M., S. Morin, and J. Tison-Rosebery. 2014. A Multicompartment Approach - Diatoms, Macrophytes, Benthic Macroinvertebrates and Fish - To Assess the Impact of Toxic Industrial

Releases on a Small French River. PLoS ONE 9. doi:10.1371/journal.pone.0102358

Laliberté, E., and P. Legendre. 2010. A distance-based framework for measuring functional diversity from multiple traits. Ecology 91: 299–305. doi:10.1890/08-2244.1

Larras, F., E. Billoir, V. Baillard, and others. 2018. DRomics: A Turnkey Tool to Support the Use of the Dose–Response Framework for Omics Data in Ecological Risk Assessment. Environ. Sci. Technol. 52: 14461–14468. doi:10.1021/acs.est.8b04752

Larras, F., R. Coulaud, E. Gautreau, E. Billoir, J. Rosebery, and P. Usseglio-Polatera. 2017. Assessing anthropogenic pressures on streams: A random forest approach based on benthic diatom communities. Sci. Total Environ. 586: 1101–1112. doi:10.1016/j.scitotenv.2017.02.096

Lavoie, I., P. B. Hamilton, S. Morin, and others. 2017. Diatom teratologies as biomarkers of contamination: Are all deformities ecologically meaningful? Ecol. Indic. 82: 539–550. doi:10.1016/j.ecolind.2017.06.048

Lavorel, S., and E. Garnier. 2002. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. Funct. Ecol. 16: 545–556. doi:10.1046/j.1365-2435.2002.00664.x

Le Bescot, N., F. Mahé, S. Audic, and others. 2016. Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. Environ. Microbiol. 18: 609–626. doi:10.1111/1462-2920.13039

Legras, G., N. Loiseau, and J.-C. Gaertner. 2018. Functional richness: Overview of indices and underlying concepts. Acta Oecologica 87: 34–44. doi:10.1016/j.actao.2018.02.007

Legras, G., N. Loiseau, J.-C. Gaertner, J.-C. Poggiale, and N. Gaertner-Mazouni. 2019. Assessing functional diversity: the influence of the number of the functional traits. Theor. Ecol. doi:10.1007/s12080-019-00433-x

Leibold, M. A., and J. Norberg. 2004. Biodiversity in metacommunities: Plankton as complex adaptive systems? Limnol. Oceanogr. 49: 1278–1289. doi:10.4319/lo.2004.49.4_part_2.1278

Lenz, P. H., V. Roncalli, R. P. Hassett, L.-S. Wu, M. C. Cieslak, D. K. Hartline, and A. E. Christie. 2014. De Novo Assembly of a Transcriptome for Calanus finmarchicus (Crustacea, Copepoda) – The Dominant Zooplankter of the North Atlantic Ocean. PLOS ONE 9: e88589. doi:10.1371/journal.pone.0088589

Levine, J. M. 2016. A trail map for trait-based studies. Nature 529: 163–164. doi:10.1038/nature16862

Litchman, E., and C. A. Klausmeier. 2008. Trait-Based Community Ecology of Phytoplankton. Annu. Rev. Ecol. Evol. Syst. 39: 615–639. doi:10.1146/annurev.ecolsys.39.110707.173549

Litchman, E., C. A. Klausmeier, O. M. Schofield, and P. G. Falkowski. 2007. The role of functional traits and trade-offs in structuring phytoplankton communities: scaling from cellular to ecosystem level. Ecol. Lett. 10: 1170–1181. doi:10.1111/j.1461-0248.2007.01117.x

Litchman, E., M. D. Ohman, and T. Kiørboe. 2013. Trait-based approaches to zooplankton

communities. J. Plankton Res. 35: 473–484. doi:10.1093/plankt/fbt019

Litchman, E., P. de T. Pinto, K. F. Edwards, C. A. Klausmeier, C. T. Kremer, and M. K. Thomas. 2015. Global biogeochemical impacts of phytoplankton: a trait-based perspective. J. Ecol. 103: 1384–1396. doi:10.1111/1365-2745.12438

Lombard, F., E. Boss, A. M. Waite, and others. 2019. Globally Consistent Quantitative Observations of Planktonic Ecosystems. Front. Mar. Sci. 6. doi:10.3389/fmars.2019.00196

Loreau. 2010. Linking biodiversity and ecosystems: towards a unifying ecological theory. Philos. Trans. R. Soc. B Biol. Sci. 365: 49–60. doi:10.1098/rstb.2009.0155

Lorke, A., D. F. McGinnis, P. Spaak, and A. Wüest. 2004. Acoustic observations of zooplankton in lakes using a Doppler current profiler. Freshw. Biol. 49: 1280–1292. doi:10.1111/j.1365-2427.2004.01267.x

Lukács, B. A., A. E-Vojtkó, T. Erős, A. M. V, S. Szabó, and L. Götzenberger. 2019. Carbon forms, nutrients and water velocity filter hydrophyte and riverbank species differently: A trait-based study. J. Veg. Sci. 30: 471–484. doi:10.1111/jvs.12738

Lüning, J. 1992. Phenotypic plasticity of Daphnia pulex in the presence of invertebrate predators: morphological and life history responses. Oecologia 92: 383–390. doi:10.1007/BF00317464

Lürling, M. 2003. Phenotypic plasticity in the green algae Desmodesmus and Scenedesmus with special reference to the induction of defensive morphology. Ann. Limnol. - Int. J. Limnol. 39: 85–101. doi:10.1051/limn/2003014

Madin, J. S., K. D. Anderson, M. H. Andreasen, and others. 2016. The Coral Trait Database, a curated database of trait information for coral species from the global oceans. Sci. Data 3: 160017. doi:10.1038/sdata.2016.17

Majdi, N., N. Hette-Tronquart, E. Auclair, and others. 2018. There's no harm in having too much: A comprehensive toolbox of methods in trophic ecology. Food Webs 17: e00100. doi:10.1016/j.fooweb.2018.e00100

Maps, F., J.-O. Irisson, and S.-D. Ayata. 2019. Book of abstract of the ARTIFACTZ Workshop. Proceedings of the ARTIFACTZ Workshop artificial intelligence for characterizing plankton traits from images. 12pp.

Maps, F., N. R. Record, and A. J. Pershing. 2014. A metabolic approach to dormancy in pelagic copepods helps explaining inter- and intra-specific variability in life-history strategies. J. Plankton Res. 36: 18–30. doi:10.1093/plankt/fbt100

Margalef, R. 1978. Life-forms of phytoplankton as survival alternatives in an unstable environment. Oceanol. Acta I 493–509.

MARLIN. 2006. BIOTIC - Biological Traits Information Catalogue. Marine Life Information Network.

Mason, N. W. H., F. de Bello, D. Mouillot, S. Pavoine, and S. Dray. 2013. A guide for using functional diversity indices to reveal changes in assembly processes along ecological gradients. J.

Veg. Sci. 24: 794–806. doi:10.1111/jvs.12013

Mason, N. W. H., D. Mouillot, W. G. Lee, and J. B. Wilson. 2005. Functional richness, functional evenness and functional divergence: the primary components of functional diversity. Oikos 111: 112–118. doi:10.1111/j.0030-1299.2005.13886.x

Matabos, M., A. O. V. Bui, S. Mihály, J. Aguzzi, S. K. Juniper, and R. S. Ajayamohan. 2014. High-frequency study of epibenthic megafaunal community dynamics in Barkley Canyon: A multi-disciplinary approach using the NEPTUNE Canada network. J. Mar. Syst. 130: 56–68. doi:10.1016/j.jmarsys.2013.05.002

McGill, B. J., B. J. Enquist, E. Weiher, and M. Westoby. 2006. Rebuilding community ecology from functional traits. Trends Ecol. Evol. 21: 178–185. doi:10.1016/j.tree.2006.02.002

Menezes, S., D. J. Baird, and A. M. V. M. Soares. 2010. Beyond taxonomy: a review of macroinvertebrate trait-based community descriptors as tools for freshwater biomonitoring. J. Appl. Ecol. 47: 711–719. doi:10.1111/j.1365-2664.2010.01819.x

Meng, A., E. Corre, I. Probert, and others. 2018. Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network. Mol. Ecol. 27: 2365–2380. doi:10.1111/mec.14579

Mérillet, L., M. Mouchet, M. Robert, M. Salaün, L. Schuck, S. Vaz, and D. Kopp. 2018. Using underwater video to assess megabenthic community vulnerability to trawling in the Grande Vasière (Bay of Biscay). Environ. Conserv. 45: 163–172. doi:10.1017/S0376892917000480

Mermillod-Blondin, F., and R. Rosenberg. 2006. Ecosystem engineering: the impact of bioturbation on biogeochemical processes in marine and freshwater benthic habitats. Aquat. Sci. 68: 434–442. doi:10.1007/s00027-006-0858-x

Meunier, C. L., M. Boersma, R. El-Sabaawi, H. M. Halvorson, E. M. Herstoff, D. B. Van de Waal, R. J. Vogt, and E. Litchman. 2017. From Elements to Function: Toward Unifying Ecological Stoichiometry and Trait-Based Ecology. Front. Environ. Sci. 5. doi:10.3389/fenvs.2017.00018

Mock, T., S. J. Daines, R. Geider, S. Collins, M. Metodiev, A. J. Millar, V. Moulton, and T. M. Lenton. 2016. Bridging the gap between omics and earth system science to better understand how environmental change impacts marine microbes. Glob. Change Biol. 22: 61–75. doi:10.1111/gcb.12983

Moisset, S., S. K. Tiam, A. Feurtet-Mazel, S. Morin, F. Delmas, N. Mazzella, and P. Gonzalez. 2015. Genetic and physiological responses of three freshwater diatoms to realistic diuron exposures. Environ. Sci. Pollut. Res. 22: 4046–4055. doi:10.1007/s11356-014-3523-2

Mondy, C. P., and P. Usseglio-Polatera. 2013. Using conditional tree forests and life history traits to assess specific risks of stream degradation under multiple pressure scenario. Sci. Total Environ. 461–462: 750–760. doi:10.1016/j.scitotenv.2013.05.072

Mondy, C. P., and P. Usseglio-Polatera. 2014. Using fuzzy-coded traits to elucidate the non-random role of anthropogenic stress in the functional homogenisation of invertebrate assemblages.

Freshw. Biol. 59: 584–600. doi:10.1111/fwb.12289

Mondy, C. P., B. Villeneuve, V. Archaimbault, and P. Usseglio-Polatera. 2012. A new macroinvertebrate-based multimetric index (I2M2) to evaluate ecological quality of French wadeable streams fulfilling the WFD demands: A taxonomical and trait approach. Ecol. Indic. 18: 452–467. doi:10.1016/j.ecolind.2011.12.013

Moretti, M., A. T. C. Dias, F. de Bello, and others. 2017. Handbook of protocols for standardized measurement of terrestrial invertebrate functional traits. Funct. Ecol. 31: 558–567. doi:10.1111/1365-2435.12776

Mouchet, M., F. Guilhaumon, S. Villéger, N. W. H. Mason, J.-A. Tomasini, and D. Mouillot. 2008. Towards a consensus for calculating dendrogram-based functional diversity indices. Oikos 117: 794–800. doi:10.1111/j.0030-1299.2008.16594.x

Mouillot, D., N. A. J. Graham, S. Villéger, N. W. H. Mason, and D. R. Bellwood. 2013. A functional approach reveals community responses to disturbances. Trends Ecol. Evol. 28: 167–177. doi:10.1016/j.tree.2012.10.004

Muthukrishnan, R., L. L. Sullivan, A. K. Shaw, and J. D. Forester. 2020. Trait plasticity alters the range of possible coexistence conditions in a competition–colonisation trade-off. Ecol. Lett. 23: 791–799. doi:10.1111/ele.13477

Naeem, S., and J. P. Wright. 2003. Disentangling biodiversity effects on ecosystem functioning: deriving solutions to a seemingly insurmountable problem. Ecol. Lett. 6: 567–579. doi:10.1046/j.1461-0248.2003.00471.x

Neuparth, T., F. O. Costa, and M. H. Costa. 2002. Effects of Temperature and Salinity on Life History of the Marine Amphipod Gammarus locusta. Implications for Ecotoxicological Testing. Ecotoxicology 11: 61–73. doi:10.1023/A:1013797130740

Neury-Ormanni, J., J. Vedrenne, M. Wagner, G. Jan, and S. Morin. 2019. Micro-meiofauna morphofunctional traits linked to trophic activity. Hydrobiologia. doi:10.1007/s10750-019-04120-0

Nock, C. A., R. J. Vogt, and B. E. Beisner. 2016. Functional Traits, p. 1–8. In eLS. American Cancer Society.

O'Brien, T. D. 2014. COPEPOD: The Global Plankton Database. An overview of the 2014 database contents, processing methods, and access interface. U.S. Dep. Commerce, NOAA Tech. Memo. NMFS-F/ST-37,.

O'Brien, W. J., D. Kettle, and H. Riessen. 1979. Helmets and Invisible Armor: Structures Reducing Predation from Tactile and Visual Planktivores. Ecology 60: 287–294. doi:10.2307/1937657

Ohman, M. D. 2019. A sea of tentacles: optically discernible traits resolved from planktonic organisms in situ. ICES J. Mar. Sci. 76: 1959–1972. doi:10.1093/icesjms/fsz184

Ohman, M. D., R. E. Davis, J. T. Sherman, K. R. Grindley, B. M. Whitmore, C. F. Nickels, and J. S. Ellen. 2019. Zooglider: An autonomous vehicle for optical and acoustic sensing of zooplankton.

Limnol. Oceanogr. Methods 17: 69–86. doi:10.1002/lom3.10301

Olson, R. J., and H. M. Sosik. 2007. A submersible imaging-in-flow instrument to analyze nano-and microplankton: Imaging FlowCytobot. Limnol. Oceanogr. Methods 5: 195–203. doi:10.4319/lom.2007.5.195

Oug, E., A. Fleddum, B. Rygg, and F. Olsgard. 2012. Biological traits analyses in the study of pollution gradients and ecological functioning of marine soft bottom species assemblages in a fjord ecosystem. J. Exp. Mar. Biol. Ecol. 432–433: 94–105. doi:10.1016/j.jembe.2012.07.019

Pardo, L. M., and L. E. Johnson. 2005. Explaining variation in life-history traits: growth rate, size, and fecundity in a marine snail across an environmental gradient lacking predators. Mar. Ecol. Prog. Ser. 296: 229–239. doi:10.3354/meps296229

Parr, C. S., N. Wilson, P. Leary, and others. 2014. The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. Biodivers. Data J. 2: e1079. doi:10.3897/BDJ.2.e1079

Passy, S. I. 2007. Diatom ecological guilds display distinct and predictable behavior along nutrient and disturbance gradients in running waters. Aquat. Bot. 86: 171–178. doi:10.1016/j.aquabot.2006.09.018

Pecuchet, L., G. Reygondeau, W. W. L. Cheung, P. Licandro, P. D. van Denderen, M. R. Payne, and M. Lindegren. 2018. Spatial distribution of life-history traits and their response to environmental gradients across multiple marine taxa. Ecosphere 9: e02460. doi:10.1002/ecs2.2460

Persson, J., P. Fink, A. Goto, J. M. Hood, J. Jonas, and S. Kato. 2010. To be or not to be what you eat: regulation of stoichiometric homeostasis among autotrophs and heterotrophs. Oikos 119: 741–751. doi:10.1111/j.1600-0706.2009.18545.x

Pesce, S., J. Beguet, N. Rouard, M. Devers-Lamrani, and F. Martin-Laurent. 2013. Response of a diuron-degrading community to diuron exposure assessed by real-time quantitative PCR monitoring of phenylurea hydrolase A and B encoding genes. Appl. Microbiol. Biotechnol. 97: 1661–1668. doi:10.1007/s00253-012-4318-3

Petchey, O. L., and K. J. Gaston. 2006. Functional diversity: back to basics and looking forward. Ecol. Lett. 9: 741–758. doi:10.1111/j.1461-0248.2006.00924.x Petchey, O. L., and K. J. Gaston. 2007. Dendrograms and measuring functional diversity. Oikos 116: 1422–1426. doi:10.1111/j.0030-1299.2007.15894.x

Peter, D. H., S. Sardy, J. Diaz Rodriguez, E. Castella, and V. I. Slaveykova. 2018. Modeling whole body trace metal concentrations in aquatic invertebrate communities: A trait-based approach. Environ. Pollut. 233: 419–428. doi:10.1016/j.envpol.2017.10.044

Picheral, M., S. Colin, and J. O. Irisson. 2017. EcoTaxa, a tool for the taxonomic classification of images.

Picheral, M., L. Guidi, L. Stemmann, D. M. Karl, G. Iddaoud, and G. Gorsky. 2010. The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. Limnol. Oceanogr. Methods 8: 462–473. doi:10.4319/lom.2010.8.462

Poff, N. L., J. D. Olden, N. K. M. Vieira, D. S. Finn, M. P. Simmons, and B. C. Kondratieff. 2006. Functional trait niches of North American lotic insects: traits-based ecological applications in light of phylogenetic relationships. J. North Am. Benthol. Soc. 25: 730–755. doi:10.1899/0887-3593(2006)025[0730:FTNONA]2.0.CO;2

Pomerleau, C., A. R. Sastri, and B. E. Beisner. 2015. Evaluation of functional trait diversity for marine zooplankton communities in the Northeast subarctic Pacific Ocean. J. Plankton Res. 37: 712–726. doi:10.1093/plankt/fbv045

Prowe, A. E. F., A. W. Visser, K. H. Andersen, S. Chiba, and T. Kiørboe. 2019. Biogeography of zooplankton feeding strategy. Limnol. Oceanogr. 64: 661–678. doi:10.1002/lno.11067

Raes, J., I. Letunic, T. Yamada, L. J. Jensen, and P. Bork. 2011. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. Mol. Syst. Biol. 7: 473. doi:10.1038/msb.2011.6

Raffard, A., F. Santoul, J. Cucherousset, and S. Blanchet. 2019. The community and ecosystem consequences of intraspecific diversity: a meta-analysis. Biol. Rev. 94: 648–661. doi:10.1111/brv.12472

Ramond, P., R. Siano, and M. Sourisseau. 2018. Functional traits of marine protists.doi:10.17882/51662

Ramond, P., M. Sourisseau, N. Simon, and others. 2019. Coupling between taxonomic and functional diversity in protistan coastal communities: Functional diversity of marine protists. Environ. Microbiol. 21: 730–749. doi:10.1111/1462-2920.14537

Rao, C. R. 1982. Diversity and dissimilarity coefficients: A unified approach. Theor. Popul. Biol. 21: 24–43. doi:10.1016/0040-5809(82)90004-1

Record, N. R., R. Ji, F. Maps, Ø. Varpe, J. A. Runge, C. M. Petrik, and D. Johns. 2018. Copepod diapause and the biogeography of the marine lipidscape. J. Biogeogr. 45: 2238–2251. doi:10.1111/jbi.13414

Reid, N. M., and A. Whitehead. 2016. Functional genomics to assess biological responses to marine pollution at physiological and evolutionary timescales: toward a vision of predictive ecotoxicology. Brief. Funct. Genomics 15: 358–364. doi:10.1093/bfgp/elv060

Reiss, J., J. R. Bridle, J. M. Montoya, and G. Woodward. 2009. Emerging horizons in biodiversity and ecosystem functioning research. Trends Ecol. Evol. 24: 505–514. doi:10.1016/j.tree.2009.03.018

Renaud, P. E., M. Daase, N. S. Banas, and others. 2018. Pelagic food-webs in a changing Arctic: a trait-based perspective suggests a mode of resilience. ICES J. Mar. Sci. 75: 1871–1881. doi:10.1093/icesjms/fsy063

Resh, V. H., A. G. Hildrew, B. Statzner, and C. R. Townsend. 1994. Theoretical habitat templets, species traits, and species richness: a synthesis of long-term ecological research on the Upper Rhône River in the context of concurrently developed ecological theory. Freshw. Biol. 31:

539–554. doi:10.1111/j.1365-2427.1994.tb01756.x

Reu, B., R. Proulx, K. Bohn, J. G. Dyke, A. Kleidon, R. Pavlick, and S. Schmidtlein. 2011. The role of climate and plant functional trade-offs in shaping global biome and biodiversity patterns. Glob. Ecol. Biogeogr. 20: 570–581. doi:10.1111/j.1466-8238.2010.00621.x

Reynolds, C. S. 1988. Functional morphology and the adaptive strategies of freshwater phytoplankton. Growth Reprod. Strateg. Freshw. Phytoplankton 388–433.

Reynolds, C. S., V. Huszar, C. Kruk, L. Naselli-Flores, and S. Melo. 2002. Towards a functional classification of the freshwater phytoplankton. J. Plankton Res. 24: 417–428. doi:10.1093/plankt/24.5.417

Richardson, A. J., A. W. Walne, A. W. G. John, T. D. Jonas, J. A. Lindley, D. W. Sims, D. Stevens, and M. Witt. 2006. Using continuous plankton recorder data. Prog. Oceanogr. 68: 27–74. doi:10.1016/j.pocean.2005.09.011

Ricotta, C. 2005. A note on functional diversity measures. Basic Appl. Ecol. 6: 479–486. doi:10.1016/j.baae.2005.02.008

Rigolet, C., S. F. Dubois, and E. Thiébaut. 2014. Benthic control freaks: Effects of the tubiculous amphipod Haploops nirae on the specific diversity and functional structure of benthic communities. J. Sea Res. 85: 413–427. doi:10.1016/j.seares.2013.07.013

Riley, G. A. 1946. Factors controlling phytoplankton population on George's Bank. J Mar Res 6: 54–73. Rimet, F., and A. Bouchez. 2012. Life-forms, cell-sizes and ecological guilds of diatoms in European rivers. Knowl. Manag. Aquat. Ecosyst. 01. doi:10.1051/kmae/2012018

Rimet, F., and J.-C. Druart. 2018. A trait database for Phytoplankton of temperate lakes. Ann. Limnol. - Int. J. Limnol. 54: 18. doi:10.1051/limn/2018009

Rimet, F., E. Gusev, M. Kahlert, and others. 2019. Diat.barcode, an open-access barcode library for diatoms.doi:10.15454/TOMBYZ

Robuchon, M., S. Vranken, L. Vandepitte, S. Dekeyzer, R. Julliard, L. Le Gall, and O. De Clerck. 2015. Towards a seaweed trait database for European species. Proceedings of the 6th European Phycological Congress.

Roubeix, V., N. Mazzella, L. Schouler, V. Fauvelle, S. Morin, M. Coste, F. Delmas, and C. Margoum. 2011. Variations of periphytic diatom sensitivity to the herbicide diuron and relation to species distribution in a contamination gradient: implications for biomonitoring. J. Environ. Monit. 13: 1768–1774. doi:10.1039/C0EM00783H

Rubach, M. N., R. Ashauer, D. B. Buchwalter, H. D. Lange, M. Hamer, T. G. Preuss, K. Töpke, and S. J. Maund. 2011. Framework for traits-based assessment in ecotoxicology. Integr. Environ. Assess. Manag. 7: 172–186. doi:10.1002/ieam.105

Salazar, G., L. Paoli, A. Alberti, and others. 2019. Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. Cell 179: 1068-1083.e21. doi:10.1016/j.cell.2019.10.014

Salguero-Gómez, R., C. Violle, O. Gimenez, and D. Childs. 2018. Delivering the promises of trait-based approaches to the needs of demographic approaches, and vice versa. Funct. Ecol. 32: 1424–1435. doi:10.1111/1365-2435.13148

Sanford, E., and M. W. Kelly. 2011. Local Adaptation in Marine Invertebrates. Annu. Rev. Mar. Sci. 3: 509–535. doi:10.1146/annurev-marine-120709-142756

Schäfer, R. B., B. J. Kefford, L. Metzeling, and others. 2011. A trait database of stream invertebrates for the ecological risk assessment of single and combined effects of salinity and pesticides in South-East Australia. Sci. Total Environ. 409: 2055–2063. doi:10.1016/j.scitotenv.2011.01.053

Schleuter, D., M. Daufresne, F. Massol, and C. Argillier. 2010. A user's guide to functional diversity indices. Ecol. Monogr. 80: 469–484. doi:10.1890/08-2225.1

Schmera, D., J. Heino, J. Podani, T. Erős, and S. Dolédec. 2017. Functional diversity: a review of methodology and current knowledge in freshwater macroinvertebrate research. Hydrobiologia 787: 27–44. doi:10.1007/s10750-016-2974-5

Schmera, D., J. Podani, J. Heino, T. Erős, and N. L. Poff. 2015. A proposed unified terminology of species traits in stream ecology. Freshw. Sci. 34: 823–830. doi:10.1086/681623

Schmid, M. S., C. Aubry, J. Grigor, and L. Fortier. 2016. The LOKI underwater imaging system and an automatic identification model for the detection of zooplankton taxa in the Arctic Ocean. Methods Oceanogr. 15–16: 129–160. doi:10.1016/j.mio.2016.03.003

Schmid, M. S., F. Maps, and L. Fortier. 2018. Lipid load triggers migration to diapause in Arctic Calanus copepods—insights from underwater imaging. J. Plankton Res. 40: 311–325. doi:10.1093/plankt/fby012

Schmidt, D. N., D. Lazarus, J. R. Young, and M. Kucera. 2006. Biogeography and evolution of body size in marine plankton. Earth-Sci. Rev. 78: 239–266. doi:10.1016/j.earscirev.2006.05.004

Schmidt-Kloiber, A., and D. Hering. 2015. www.freshwaterecology.info – An online tool that unifies, standardises and codifies more than 20,000 European freshwater organisms and their ecological preferences. Ecol. Indic. 53: 271–282. doi:10.1016/j.ecolind.2015.02.007

Schneider, F. D., M. Jochum, G. Le Provost, and others. 2018. Towards an Ecological Trait-data Standard. bioRxiv. doi:info:doi:10.1101/328302

Schulz, J., K. Barz, P. Ayon, A. Lüdtke, O. Zielinski, D. Mengedoht, and H.-J. Hirche. 2010. Imaging of plankton specimens with the lightframe on-sight keyspecies investigation (LOKI) system. J. Eur. Opt. Soc. - Rapid Publ. 5. doi:10.2971/jeos.2010.10017s

Sheridan, J. A., and D. Bickford. 2011. Shrinking body size as an ecological response to climate change. Nat. Clim. Change 1: 401–406. doi:10.1038/nclimate1259

Sieracki, C. K., M. E. Sieracki, and C. S. Yentsch. 1998. An imaging-in-flow system for automated analysis of marine microplankton. Mar. Ecol. Prog. Ser. 168: 285–296. doi:10.3354/meps168285

Sieracki, M. E., M. E. Sieracki, M. E. Sieracki, and others. 2010. Optical Plankton Imaging and Analysis Systems for Ocean Observation. Proceedings of OceanObs'09: Sustained Ocean

Observations and Information for Society. Proceedings of the OceanObs'09: Sustained Ocean Observations and Information for Society. European Space Agency. 878–885.

Sournia, A. 1982. Form and Function in Marine Phytoplankton. Biol. Rev. 57: 347–394. doi:10.1111/j.1469-185X.1982.tb00702.x

Start, D., S. McCauley, and B. Gilbert. 2018. Physiology underlies the assembly of ecological communities. Proc. Natl. Acad. Sci. 115: 6016–6021. doi:10.1073/pnas.1802091115

Statzner, B., and L. A. Bêche. 2010. Can biological invertebrate traits resolve effects of multiple stressors on running water ecosystems? Freshw. Biol. 55: 80–119. doi:10.1111/j.1365-2427.2009.02369.x

Stec, K. F., L. Caputi, P. L. Buttigieg, and others. 2017. Modelling plankton ecosystems in the meta-omics era. Are we ready? Mar. Genomics 32: 1–17. doi:10.1016/j.margen.2017.02.006

Stemmann, L., M. Youngbluth, K. Robert, and others. 2008. Global zoogeography of fragile macrozooplankton in the upper 100–1000 m inferred from the underwater video profiler. ICES J. Mar. Sci. 65: 433–442. doi:10.1093/icesjms/fsn010

St-Gelais, N. F., A. Jokela, and B. E. Beisner. 2017. Limited functional responses of plankton food webs in northern lakes following diamond mining. Can. J. Fish. Aquat. Sci. 75: 26–35. doi:10.1139/cjfas-2016-0418

Stuart-Smith, R. D., A. E. Bates, J. S. Lefcheck, and others. 2013. Integrating abundance and functional traits reveals new global hotspots of fish diversity. Nature 501: 539–542. doi:10.1038/nature12529

Sunagawa, S., L. P. Coelho, S. Chaffron, and others. 2015. Structure and function of the global ocean microbiome. Science 348: 1261359. doi:10.1126/science.1261359

Swaffar, S. M., and W. J. O'Brien. 1996. Spines of Daphnia lumholtzi create feeding difficulties for juvenile bluegill sunfish (Lepomis macrochirus). J. Plankton Res. 18: 1055–1061. doi:10.1093/plankt/18.6.1055

Tapolczai, K., A. Bouchez, C. Stenger-Kovács, J. Padisák, and F. Rimet. 2016. Trait-based ecological classifications for benthic algae: review and perspectives. Hydrobiologia 776: 1–17. doi:10.1007/s10750-016-2736-4

Thomas, M. K., C. T. Kremer, and E. Litchman. 2016. Environment and evolutionary history determine the global biogeography of phytoplankton temperature traits. Glob. Ecol. Biogeogr. 25: 75–86. doi:10.1111/geb.12387

Tillin, H. M., J. G. Hiddink, S. Jennings, and M. J. Kaiser. 2006. Chronic bottom trawling alters the functional composition of benthic invertebrate communities on a sea-basin scale. Mar. Ecol. Prog. Ser. 318: 31–45. doi:10.3354/meps318031

Tilman, D. 1994. Competition and Biodiversity in Spatially Structured Habitats. Ecology 75: 2–16. doi:10.2307/1939377

Trakimas, G., R. J. Whittaker, and M. K. Borregaard. 2016. Do biological traits drive geographical

patterns in European amphibians?: Traits of European amphibians. Glob. Ecol. Biogeogr. 25: 1228–1238. doi:10.1111/geb.12479

Troudet, J., P. Grandcolas, A. Blin, R. Vignes-Lebbe, and F. Legendre. 2017. Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7: 9132. doi:10.1038/s41598-017-09084-6

Tyler, E. H. M., P. J. Somerfield, E. V. Berghe, J. Bremner, E. Jackson, O. Langmead, M. L. D. Palomares, and T. J. Webb. 2012. Extensive gaps and biases in our knowledge of a well-known fauna: implications for integrating biological traits into macroecology: Biological knowledge of UK marine fauna. Glob. Ecol. Biogeogr. 21: 922–934. doi:10.1111/j.1466-8238.2011.00726.x

United Nations. 2015. Transforming our world: the 2030 Agenda for Sustainable Development.

U.S. EPA. 2012. Freshwater Biological Traits Database (Final Report).

Usseglio-Polatera, P., M. Bournaud, P. Richoux, and H. Tachet. 2000. Biological and ecological traits of benthic freshwater macroinvertebrates: relationships and definition of groups with similar traits. Freshw. Biol. 43: 175–205. doi:10.1046/j.1365-2427.2000.00535.x

Usseglio-Polatera, P., M. Bournaud, P. Richoux, and H. Tachet. 2000. Biomonitoring through biological traits of benthic macroinvertebrates: how to use species trait databases? Hydrobiologia 422: 153–162. doi:10.1023/A:1017042921298

Usseglio-Polatera, P., Richoux, P., Bournaud, M., and Tachet, H. 2001. A functional classification of benthic macroinvertebrates based on biological and ecological traits: Application to river condition assessment and stream management. Arch. Für Hydrobiol. Suppl. Monogr. Beitr. 139: 53–83.

Valentini, A., P. Taberlet, C. Miaud, and others. 2016. Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Mol. Ecol. 25: 929–942. doi:10.1111/mec.13428

Van Dam, H., A. Mertens, and J. Sinkeldam. 1994. A coded checklist and ecological indicator values of freshwater diatoms from The Netherlands. Netherland J. Aquat. Ecol. 28: 117–133. doi:10.1007/BF02334251

Vanni, M. J., and P. B. McIntyre. 2016. Predicting nutrient excretion of aquatic animals with metabolic ecology and ecological stoichiometry: a global synthesis. Ecology 97: 3460–3471. doi:10.1002/ecy.1582

de Vargas, C., S. Audic, N. Henry, and others. 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 348: 1261605. doi:10.1126/science.1261605

Vasselon, V., F. Rimet, K. Tapolczai, and A. Bouchez. 2017. Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). Ecol. Indic. 82: 1–12. doi:10.1016/j.ecolind.2017.06.024

Verberk, W. C. E. P., H. Siepel, and H. Esselink. 2008. Applying life-history strategies for freshwater macroinvertebrates to lentic waters. Freshw. Biol. 53: 1739–1753. doi:10.1111/j.1365-2427.2008.02036.x

Vilgrain, L., F. Maps, M. Picheral, M. Babin, J.-O. Irisson, and S.-D. Ayata. under review. Trait-based approach on zooplankton in situ images reveals contrasted ecological patterns along ice melt dynamics. Limnol. Oceanogr. LO-20-0190.

Villar, E., L. Cabrol, and L.-E. Heimbürger-Boavida. 2019. Widespread microbial mercury methylation genes in the global ocean. bioRxiv 648329. doi:10.1101/648329

Villéger, S., N. W. H. Mason, and D. Mouillot. 2008. New Multidimensional Functional Diversity Indices for a Multifaceted Framework in Functional Ecology. Ecology 89: 2290–2301. doi:10.1890/07-1206.1

Villon, S., M. Chaumont, G. Subsol, S. Villéger, T. Claverie, and D. Mouillot. 2016. Coral Reef Fish Detection and Recognition in Underwater Videos by Supervised Machine Learning: Comparison Between Deep Learning and HOG+SVM Methods. Advanced Concepts for Intelligent Vision Systems. Springer International Publishing. 160–171.

Violle, C., B. J. Enquist, B. J. McGill, L. Jiang, C. H. Albert, C. Hulshof, V. Jung, and J. Messier. 2012. The return of the variance: intraspecific variability in community ecology. Trends Ecol. Evol. 27: 244–252. doi:10.1016/j.tree.2011.11.014

Violle, C., M.-L. Navas, D. Vile, E. Kazakou, C. Fortunel, I. Hummel, and E. Garnier. 2007. Let the concept of trait be functional! Oikos 116: 882–892. doi:10.1111/j.0030-1299.2007.15559.x

Violle, C., P. B. Reich, S. W. Pacala, B. J. Enquist, and J. Kattge. 2014. The emergence and promise of functional biogeography. Proc. Natl. Acad. Sci. 111: 13690–13696. doi:10.1073/pnas.1415442111

Weiss, K. C. B., and C. A. Ray. 2019. Unifying functional trait approaches to understand the assemblage of ecological communities: synthesizing taxonomic divides. Ecography 42: 2012–2020. doi:10.1111/ecog.04387

Weithoff, G., and B. E. Beisner. 2019. Measures and Approaches in Trait-Based Phytoplankton Community Ecology – From Freshwater to Marine Ecosystems. Front. Mar. Sci. 6. doi:10.3389/fmars.2019.00040

Weitz, J. S., and S. A. Levin. 2006. Size and scaling of predator–prey dynamics. Ecol. Lett. 9: 548–557. doi:10.1111/j.1461-0248.2006.00900.x

Willby, N. J., V. J. Abernethy, and B. O. L. Demars. 2000. Attribute-based classification of European hydrophytes and its relationship to habitat utilization. Freshw. Biol. 43: 43–74. doi:10.1046/j.1365-2427.2000.00523.x

Winemiller, K. O., D. B. Fitzgerald, L. M. Bower, and E. R. Pianka. 2015. Functional traits, convergent evolution, and periodic tables of niches. Ecol. Lett. 18: 737–751. doi:10.1111/ele.12462

Woodward, G., D. C. Speirs, and A. G. Hildrew. 2005. Quantification and Resolution of a Complex, Size-Structured Food Web, p. 85–135. In Advances in Ecological Research. Elsevier.

WoRMS Editorial Board. 2019. World Register of Marine Species. WoRMS Editor. Board

2019 World Regist. Mar. Species Available Httpwwwmarinespeciesorg VLIZ Accessed 2019-07-12 Doi1014284170.

Wright, I. J., P. B. Reich, M. Westoby, and others. 2004. The worldwide leaf economics spectrum. Nature 428: 821–827. doi:10.1038/nature02403

Xu, Y., T. Stoeck, D. Forster, Z. Ma, L. Zhang, and X. Fan. 2018. Environmental status assessment using biological traits analyses and functional diversity indices of benthic ciliate communities. Mar. Pollut. Bull. 131: 646–654. doi:10.1016/j.marpolbul.2018.04.064

Zhao, T., S. Villéger, S. Lek, and J. Cucherousset. 2014. High intraspecific variability in the functional niche of a predator is associated with ontogenetic shift and individual specialization. Ecol. Evol. 4: 4649–4657. doi:10.1002/ece3.1260

Acknowledgements

This paper emerged from informal discussions about trait-based ecology between aquatic ecologists, especially through visit of SDA at TAKUVIK, ISYEB and LIEC during her sabbatical periods. SM was funded by a postdoctoral grant from Sorbonne Université while SDA sabbaticals were funded by Sorbonne Université and CNRS. The authors declare no conflict of interests.

Authors contribution statement

SM, SDA and ML proposed the initial idea and wrote the original draft. FL, AB and EF contributed to writing some subsections of the original draft. EF performed the mental map. All authors contributed to and reviewed the different portions and versions of the manuscript according to their respective scientific expertise.

Data accessibility

Code and resource for the bibliographic network (Supplementary Figure A.4) is available at: https://github.com/severine13/Biblio-Functional-traits. The mental map presenting the functional trait typology is available online at http://doi.org/10.5281/zenodo.3635898.

Supplementary information

A database of 2,476 publications related to trait-based approaches was extracted from a search on Web of Science (TS=(functional trait OR trait-based) AND TS=(aquatic OR marine OR ocean OR coastal OR deep-sea OR pelagic OR benthic OR fresh water OR lake OR river OR limnology)) AND DOCUMENT TYPES: (Article) – on 11.29.2018). Publications were classified into 8 main scientific fields grouped from their Web of Science research areas. A similar search on Web of Science of all freshwater and marine ecology articles was carried out and the percentage of articles related to trait-based approaches computed over time. These informations have been illustrated in Figure A.5. The network (Figure A.4) was created using VOSviewer version 1.6.8 (van Eck and Waltman 2010). To create the network, data were cleaned using a thesaurus file, merging keywords into more general terms and homogenizing them (removing plurals, merging synonyms). For example “trait plasticity”, “trait evolution”, “trait ecology”, “trait adaptation”, “quantitative trait”, “biological trait approach” are closely related terms and were merged into “trait”. This step is important to increase the overlap between publications. The analysis method relied upon an incidence

matrix of 13,337 keywords, appearing in at least 26 publications. Based on this threshold, 100 top-keywords of the highest co-occurrence were identified. A network was built based on co-occurrence calculations (i.e. nodes correspond to keywords, edges correspond to the strength of the links between keywords) and for its representation the diameter of the corresponding circle was plotted as proportional to their total occurrence in the 2,476 publications. The association strength method (also called proximity index, for computation see van Eck and Waltman 2009) was used to normalize the strength of the links between keywords. There were 3 clusters (blue, orange and purple), identified as minimizing the distance between nodes (for the weighted and parameterized variant of modularity-based clustering method used for bibliographic networks, see Waltman et al. 2010), with a minimum size defined as 20 keywords for easier interpretation.

Trait-based modeling

Trait-based ecological models can simulate functional traits as a continuum of trait values according to particular trade-offs (Follows and Dutkiewicz 2011). In such models, the numerical analog of an "individual" is the numerical entity defined by a unique vector of parameters that constitute its numerical traits. These models then simulate the response of individual phenotypic plasticity to environmental forcing and frequently rely on optimization for maximizing individual fitness (Smith et al. 2011). When parameters can vary to represent genotypic variability, these models can also be used to model trait adaptation and evolution (such as temperature optimum trait of phytoplankton growth, e.g. Grimaud et al. 2015). Most of these trait-based models focused on protists, as it may be easier to describe in a mechanistic way the functional traits and associated trade-offs of unicellular organisms (Merico et al. 2009; Follows and Dutkiewicz 2011). However, trait-based models were also developed for multicellular organisms such as copepods (Prowe et al. 2019) or even fish (e.g. Andersen et al. 2016). In these models, the traits that are the most commonly simulated are morphological and/or physiological traits, in particular traits that relate to size and allometric relationships (e.g. Hartvig et al. 2011; Acevedo-Trejos et al. 2015; Andersen et al. 2016; Blanchard et al. 2017) and/or to resource acquisition, including assimilation rates (e.g. Fiksen et al. 2013) or light-harvesting vs nutrient-harvesting investment for protists (with usually a focus on pigment-related traits, e.g. (Litchman et al. 2007; Hickman et al. 2010) or more recently on mixotrophy (e.g. Chakraborty et al. 2017; Leles et al. 2018). Some models have also focused on modeling life history traits, such as copepod dormancy (Maps et al. 2014) or bivalve fecundity and age at maturity (Sarà et al. 2013). Another type of trait-based model relies on traits to define functional groups whose dynamics are explicitly simulated, such as the Plankton Ecology Group (PEG) models (Sommer et al. 1986, 2012) or the Plankton Functional Type (PFT) models (Le Quere et al. 2005; Hood et al. 2006). In these models, traits are usually fixed parameters (e.g. size, nutrient assimilation rates, trophic regime) that are used in a mechanistic approach of ecological and physiological processes relying on trade-offs between the fundamental functions of the organisms (e.g. Smith et al. 2014). Functional types based on size, feeding modes and ecosystem engineering (bioturbation) have also been used for modeling benthic fauna (e.g. Chardy and Dauvin 1992; Rosenberg 2001; Alexandridis et al. 2017). Dynamic Energy Budget (DEB) models (Nisbet et al. 2000) could also be seen as trait-based models that rely on the DEB theory (Kooijman and Kooijman 2000) and its trade-offs among energy allocation.

Several models actually borrow from both approaches by defining, for example, a few groups of organisms for which life-cycle strategies or peculiar trade-offs are hardwired, while allometric relationships to the size master trait allow for inter-individual or inter-specific variability within the functional groups (e.g. Ward et al. 2012). Finally, statistical trait-based models (involving statistical relationships rather than differential equations based on mechanistic assumptions) have also been developed, for instance for relating functional traits to body trace metal concentrations in freshwater invertebrates (Peter et al. 2018) . Trait-based models were used for simulating the distribution of traits at the community scale (Andersen and Beyer 2006), emergent trait biogeography (Follows et al. 2007; Record et al. 2013), size spectrum (e.g. Andersen et al. 2016), competition among species and/or seasonal dynamics (Merico et al. 2009; Terseleer et al. 2014; Leblanc et al. 2018), impact of traits on biogeochemical cycles (Stamieszkin et al. 2015; Coles et al. 2017), or impact of human pressures such as fishing on trait and trophic structure (Andersen and Pedersen, 2010). Traitbased models can also be used for estimating unknown traits, for instance from phylogeny (Bruggeman 2011) or for estimating the inter- and intra-specific variability of traits (Maps et al. 2014). A description of some examples of trait-based models developed in marine ecology can be found in Kiørboe et al. (2018), with a special focus on optimality-based resource acquisition in unicellular plankton and size-based trophic dynamics of fish communities. To our knowledge, mechanistic trait-based models developed for benthic organisms remain relatively scarce (e.g. Sarà et al. 2013) and have considered size, ecosystem engineering, adult motility, fecundity and dispersal. A few statistical trait-based models have also been developed for these organisms (e.g. Peter et al. 2018). With the exception of a few studies (e.g. Maps et al. 2014; Banas and Campbell 2016), the intra-specific variability of the traits is poorly or not taken into account, mainly because of the lack of empirical information on this variability. A first perspective for trait-based modeling could then be to tackle this question of inter-individual variation of traits and its impact on ecosystem structure and functioning. Trait-based models can for instance be used to quantify the impact of environmental changes on the intra- and inter-specific variability of functional traits (e.g. inter-individual variability of lipid content and inter-specific variability of size of Arctic copepods in the Barents Sea, Renaud et al. 2018). A second perspective would be the development of a new generation of trait-based models using -omics data (Mock et al. 2016; Stec et al. 2017; Coles et al. 2017). Metatranscriptomic data could for instance be used to identify physiological traits of phytoplankton, combined to a mechanistic model of the phytoplankton cell, and used to construct a trait-based global marine ecosystem model (Mock et al. 2016). Emergent communities of marine microbes (from bacteria to phytoplankton) have also been predicted by directly simulating their metagenomes and metatranscriptomes (Coles et al. 2017). Such models could then be used for estimating functional trait plasticity across ecosystems. The idea of improving ecosystem models using - omics is not new (Hood et al. 2006), but functional trait-based approaches could constitute the common framework needed for modellers, observers, molecular biologists, and ecologists working in limnology and oceanography.

Supplementary references (not cited in main text)

Acevedo-Trejos, E., G. Brandt, J. Bruggeman, and A. Merico. 2015. Mechanisms shaping size structure and functional diversity of phytoplankton communities in the ocean. Sci. Rep. 5: 8918.

doi:10.1038/srep08918

Alexandridis, N., J. M. Dambacher, F. Jean, N. Desroy, and C. Bacher. 2017. Qualitative modelling of functional relationships in marine benthic communities. Ecol. Model. 360: 300–312. doi:10.1016/j.ecolmodel.2017.07.021

Andersen, K. H., and J. E. Beyer. 2006. Asymptotic Size Determines Species Abundance in the Marine Size Spectrum. Am. Nat. 168: 54–61. doi:10.1086/504849

Andersen, K. H., and M. Pedersen,. 2010. Damped trophic cascades driven by fishing in model marine ecosystems. Proc. R. Soc. B Biol. Sci. 277: 795–802. doi:10.1098/rspb.2009.1512

Blanchard, J. L., R. F. Heneghan, J. D. Everett, R. Trebilco, and A. J. Richardson. 2017. From Bacteria to Whales: Using Functional Size Spectra to Model Marine Ecosystems. Trends Ecol. Evol. 32: 174–186. doi:10.1016/j.tree.2016.12.003

Bruggeman, J. 2011. A Phylogenetic Approach to the Estimation of Phytoplankton Traits1. J. Phycol. 47: 52–65. doi:10.1111/j.1529-8817.2010.00946.x

Chakraborty, S., L. T. Nielsen, and K. H. Andersen. 2017. Trophic Strategies of Unicellular Plankton. Am. Nat. 189: E77–E90. doi:10.1086/690764

Chardy, P., and J.-C. Dauvin. 1992. Carbon flows in a subtidal fine sand community from the western English Channel: a simulation analysis. Mar. Ecol. Prog. Ser. 81: 147–161.

Díaz, S., A. Purvis, J. H. C. Cornelissen, G. M. Mace, M. J. Donoghue, R. M. Ewers, P. Jordano, and W. D. Pearse. 2013. Functional traits, the phylogeny of function, and ecosystem service vulnerability. Ecol. Evol. 3: 2958–2975. doi:10.1002/ece3.601

van Eck, N. J., and L. Waltman. 2009. How to normalize cooccurrence data? An analysis of some well-known similarity measures. J. Am. Soc. Inf. Sci. Technol. doi:10.1002/asi.21075

van Eck, N. J., and L. Waltman. 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84: 523–538. doi:10.1007/s11192-009-0146-3

Fiksen, Ø., M. J. Follows, and D. L. Aksnes. 2013. Trait-based models of nutrient uptake in microbes extend the Michaelis-Menten framework. Limnol. Oceanogr. 58: 193–202. doi:10.4319/lo.2013.58.1.0193

Follows, M. J., and S. Dutkiewicz. 2011. Modeling Diverse Communities of Marine Microbes. Annu. Rev. Mar. Sci. 3: 427–451. doi:10.1146/annurev-marine-120709-142848

Follows, M. J., S. Dutkiewicz, S. Grant, and S. W. Chisholm. 2007. Emergent biogeography of microbial communities in a model ocean. Science 315: 1843–1846. doi:10.1126/science.1138544

Grimaud, G. M., V. Le guennec, S.-D. Ayata, F. Mairet, A. Sciandra, and O. Bernard. 2015. Modelling the effect of temperature on phytoplankton growth across the global ocean. IFAC-Pap. 48: 228–233. doi:10.1016/j.ifacol.2015.05.059

Hartvig, M., K. H. Andersen, and J. E. Beyer. 2011. Food web framework for size-structured populations. J. Theor. Biol. 272: 113–122. doi:10.1016/j.jtbi.2010.12.006

Hickman, A. E., S. Dutkiewicz, R. G. Williams, and M. J. Follows. 2010. Modelling the effects of chromatic adaptation on phytoplankton community structure in the oligotrophic ocean. Mar. Ecol. Prog. Ser. 406: 1–17. doi:10.3354/meps08588

Kooijman, S. A. L. M., and S. A. L. M. Kooijman. 2000. Dynamic Energy and Mass Budgets in Biological Systems, Cambridge University Press.

Le Quere, C., S. P. Harrison, I. Colin Prentice, and others. 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Glob. Change Biol. doi:10.1111/j.1365-2486.2005.1004.x

Leblanc, K., B. Quéguiner, F. Diaz, and others. 2018. Nanoplanktonic diatoms are globally overlooked but play a role in spring blooms and carbon export. Nat. Commun. 9: 953. doi:10.1038/s41467-018-03376-9

Leles, S. G., L. Polimene, J. Bruggeman, J. Blackford, S. Ciavatta, A. Mitra, and K. J. Flynn. 2018. Modelling mixotrophic functional diversity and implications for ecosystem function. J. Plankton Res. 40: 627–642. doi:10.1093/plankt/fby044

Merico, A., J. Bruggeman, and K. Wirtz. 2009. A trait-based approach for downscaling complexity in plankton ecosystem models. Ecol. Model. 220: 3001–3010. doi:10.1016/j.ecolmodel.2009.05.005

Mondy, C. P., and P. Usseglio-Polatera. 2014. Using fuzzy-coded traits to elucidate the nonrandom role of anthropogenic stress in the functional homogenisation of invertebrate assemblages. Freshw. Biol. 59: 584–600. doi:10.1111/fwb.12289

Nisbet, R. M., E. B. Muller, K. Lika, and S. a. L. M. Kooijman. 2000. From molecules to ecosystems through dynamic energy budget models. J. Anim. Ecol. 69: 913–926. doi:10.1111/j.1365-2656.2000.00448.x

Record, N. R., A. J. Pershing, and F. Maps. 2013. Emergent copepod communities in an adaptive trait-structured model. Ecol. Model. 260: 11–24. doi:10.1016/j.ecolmodel.2013.03.018

Rosenberg, R. 2001. Marine benthic faunal successional stages and related sedimentary activity. Sci. Mar. 65: 107–119. doi:10.3989/scimar.2001.65s2107

Sara, G., V. Palmeri, V. Montalto, A. Rinaldi, and J. Widdows. 2013. Parameterisation of bivalve functional traits for mechanistic eco-physiological dynamic energy budget (DEB) models. Mar. Ecol. Prog. Ser. 480: 99–117. doi:10.3354/meps10195

Smith, S. L., A. Merico, K. W. Wirtz, and M. Pahlow. 2014. Leaving misleading legacies behind in plankton ecosystem modelling. J. Plankton Res. 36: 613–620. doi:10.1093/plankt/fbu011

Smith, S. L., M. Pahlow, A. Merico, and K. W. Wirtz. 2011. Optimality-based modeling of planktonic organisms. Limnol. Oceanogr. 56: 2080–2094. doi:10.4319/lo.2011.56.6.2080

Sommer, U., R. Adrian, L. De Senerpont Domis, and others. 2012. Beyond the Plankton Ecology Group (PEG) Model: Mechanisms Driving Plankton Succession. Annu. Rev. Ecol. Evol. Syst. 43: 429–448. doi:10.1146/annurev-ecolsys-110411-160251

Sommer, U., Z. M. Gliwicz, W. Lampert, and A. Duncan. 1986. The PEG-model of seasonal succession of planktonic events in fresh waters. Stamieszkin, K., A. J. Pershing, N. R. Record, C. H. Pilskaln, H. G. Dam, and L. R. Feinberg. 2015. Size as the master trait in modeled copepod fecal pellet carbon flux. Limnol. Oceanogr. 60: 2090–2107. doi:10.1002/lno.10156

Terseleer, N., J. Bruggeman, C. Lancelot, and N. Gypens. 2014. Trait-based representation of diatom functional diversity in a plankton functional type model of the eutrophied southern North Sea. Limnol. Oceanogr. 59: 1958–1972. doi:10.4319/lo.2014.59.6.1958

Waltman, L., N. J. van Eck, and E. C. M. Noyons. 2010. A unified approach to mapping and clustering of bibliometric networks. J. Informetr. 4: 629–635. doi:10.1016/j.joi.2010.07.002

Ward, B. A., S. Dutkiewicz, O. Jahn, and M. J. Follows. 2012. A size-structured food-web model for the global ocean. Limnol. Oceanogr. 57: 1877–1891. doi:10.4319/lo.2012.57.6.1877

Supplementary Figures

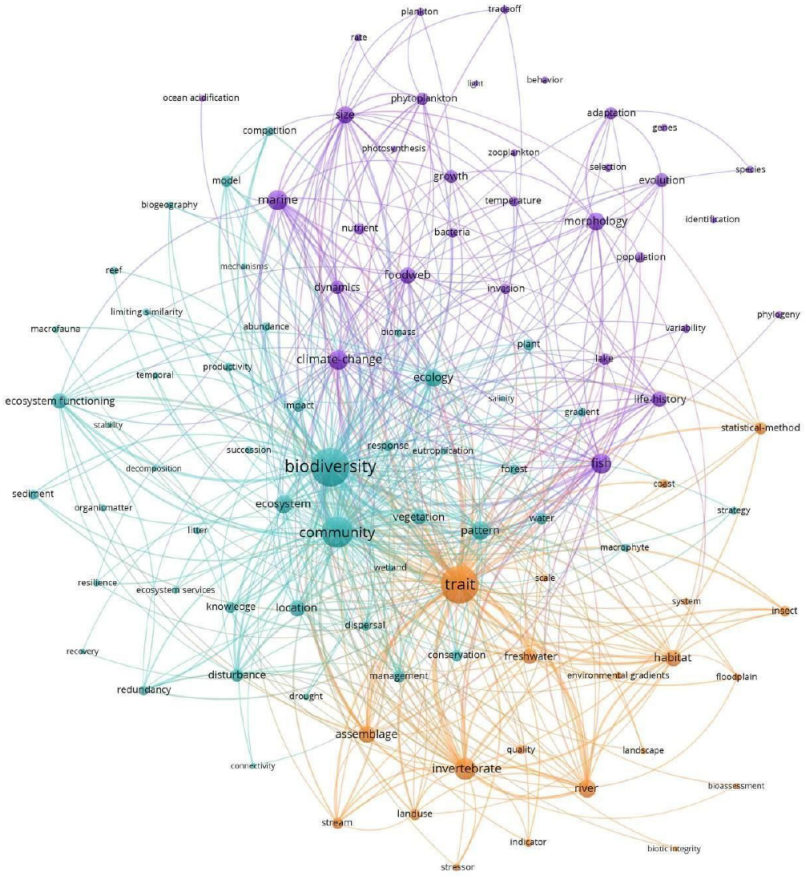

Figure A.4 - Bibliographic network representing the database of publications on traitbased approaches. Three clusters of co-occurring keywords were identified within this network. The first (in orange) and the second (in purple) clusters mainly refer to two different ecosystems. The "freshwater", "river" and "stream" ecosystems drive the first cluster as well as "invertebrate" organisms, while the second cluster refers to "marine" ecosystems with "fish", "zooplankton" and "phytoplankton" organisms. These clusters highlight different ecological questions either linked to "indicators" and "land-use" (cluster 1 in orange) or to "food webs" and "climate-change" (cluster 2 in purple). The third cluster (in blue) is mainly associated with theoretical ecology, grouping "biodiversity", "community", "ecology", "ecological functioning" and "pattern" showing numerous links rather evenly distributed between freshwater and marine ecosystem studies. This third cluster also refers to "sediment" or "organic matter", two components transcending both limnology and oceanography.

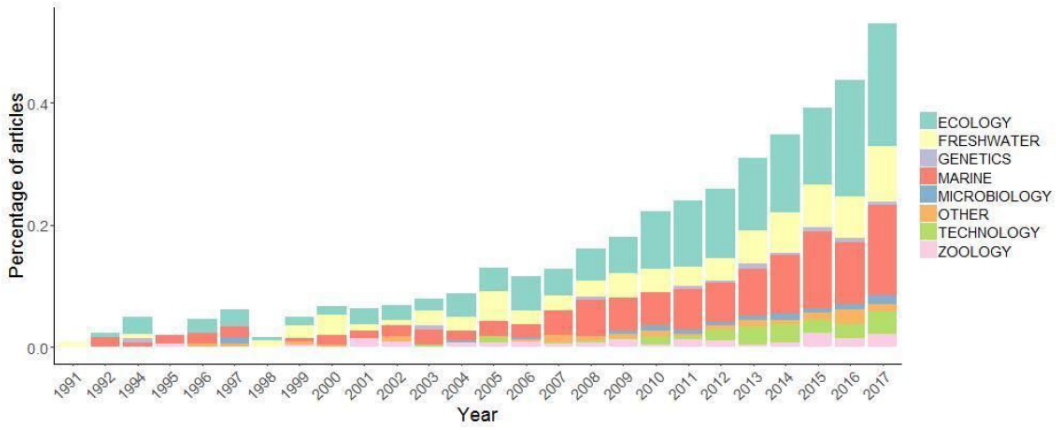

Figure A.5 - The percentage of publications relative to all the ones published in freshwater and marine ecology showing an increase over time, based on a literature survey (described in Supplementary Information). This percentage has increased over time, reaching about 0.5% in 2017. More recently, these two terms began to appear in the Web of Science research fields identified as: "genetics", "microbiology" and "technology" (Figure A.4).

Appendix B

113 transcriptomes of mixotrophic protists

Taxonomy and access numbers for each of the 113 transcriptomes of mixotrophic species identified among the 798 used in Chapter 3.

Supergroup	Phylum	Class	Genus	Species	NCBI ID	ENA ID	Mixotrophy	Mixotype	Reference
Alveolata	Ciliophora	Spirotrichea	Strombidium	rassoulzadegani	1082188	SRR1296873	Yes	GNCM	Haraguchi, L., Jakobsen, H. H., Lundholm, N., & Carstensen, J. (2018). Phytoplankton Community Dynamic: A Driver for Ciliate Trophic Strategies. Front. Mar. Sci, 5, 272.
Alveolata	Ciliophora	Heterotrichea	Climacostomum	virens	49980	SRR1300461	Yes	eSNCM	Reisser, W., Fischer-Defoy, D., Staudinger, J., Schilling, N., & Hausmann, K. (1984). The endosymbiotic unit ofClimacostomum virens andChlorella sp. I. Morphological and physiological studies on the algal partner and its localization in the host cell. Protoplasma, 119(1-2), 93-99.
Alveolata	Ciliophora	Prostomatea	Tiarina	fusa	693140	SRR1296771	Yes	eSNCM	Mordret, S., Romac, S., Henry, N., Colin, S., Carmichael, M., Berney, C., ... & Decelle, J. (2016). The symbiotic life of Symbiodinium in the open ocean within a new species of calcifying ciliate (Tiarina sp.). The ISME journal, 10(6), 1424.
Rhizaria	Foraminifera	Miliolida	Sorites	sp	126664	SRR1296734	Yes	eSNCM	Wecker, P., Fournier, A., Bosserelle, P., Debitus, C., Lecellier, G., & Berteaux-Lecellier, V. (2015). Dinoflagellate diversity among nudibranchs and sponges from French Polynesia: insights into associations and transfer. Comptes Rendus Biologies, 338(4), 278-283.
Alveolata	Ciliophora	Litostomatea	Mesodinium	pulex	283647	SRR1296764	Yes	pSNCM	Tarangkoon, W., & Hansen, P. J. (2011). Prey selection, ingestion and growth responses of the common marine ciliate Mesodinium pulex in the light and in the dark. Aquatic Microbial Ecology, 62(1), 25-38.
Alveolata	Ciliophora	Litostomatea	Myrionecta	rubra	283649	SRR1296700	Yes	pSNCM	Jones, R. I. (2000). Mixotrophy in planktonic protists: an overview. Freshwater Biology, 45(2), 219-226.
Alveolata	Dinoflagellata	Dinophyceae	Dinophysis	acuminata	47934	SRR1296701	Yes	pSNCM	Stoecker, D. K., Hansen, P. J., Caron, D. A., & Mitra, A. (2017). Mixotrophy in the marine plankton. Annual Review of Marine Science, 9, 311-335.
Rhizaria	Foraminifera	Rotaliida	Elphidium	margaritaceum	933848	SRR1300475	Yes	pSNCM	Pillet, L., & Pawlowski, J. (2012). Transcriptome analysis of foraminiferan Elphidium margaritaceum questions the role of gene transfer in kleptoplastidy. Molecular Biology and Evolution, 30(1), 66-69.
Alveolata	Dinoflagellata	Dinophyceae	Akashiwo	sanguinea	143672	SRR1294464	Yes	CM	Park, J., Jeong, H. J., Du Yoo, Y., & Yoon, E. Y. (2013). Mixotrophic dinoflagellate red tides in Korean waters: distribution and ecophysiology. Harmful Algae, 30, S28-S40.
Alveolata	Dinoflagellata	Dinophyceae	Akashiwo	sanguinea	143672	SRR1294462	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Akashiwo	sanguinea	143672	SRR1294463	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Akashiwo	sanguinea	143672	SRR1294461	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Alexandrium	andersonii	327968	SRR1300512	Yes	CM	Lee, K. H., Jeong, H. J., Kwon, J. E., Kang, H. C., Kim, J. H., Jang, S. H., ... & Kim, J. S. (2016). Mixotrophic ability of the phototrophic dinoflagellates Alexandrium andersonii, A. affine, and A. fraterculus. Harmful algae, 59, 67-81.
Alveolata	Dinoflagellata	Dinophyceae	Alexandrium	catenella	2925	SRR1296704	Yes	CM	Legrand, C., & Carlsson, P. (1998). Uptake of high molecular weight dextran by the dinoflagellate Alexandrium catenella. Aquatic microbial ecology, 16(1), 81-86.
Alveolata	Dinoflagellata	Dinophyceae	Alexandrium	monilatum	311494	SRR1296897	Yes	CM	Lee, M. J., Jeong, H. J., Lee, K. H., Jang, S. H., Kim, J. H., & Kim, K. Y. (2015). Mixotrophy in the nematocyst–taeniocyst complex-bearing phototrophic dinoflagellate Polykrikos hartmannii.
Alveolata	Dinoflagellata	Dinophyceae	Alexandrium	monilatum	311494	SRR1296895	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Alexandrium	monilatum	311494	SRR1296896	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Alexandrium	monilatum	311494	SRR1296898	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Alexandrium	temarense	2926	SRR1300222	Yes	CM	Jeong, H. J., Du Yoo, Y., Park, J. Y., Song, J. Y., Kim, S. T., Lee, S. H., ... & Yih, W. H. (2005). Feeding by phototrophic red-tide dinoflagellates: five species newly revealed and six species
Alveolata	Dinoflagellata	Dinophyceae	Alexandrium	temarense	2926	SRR1300221	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Alexandrium	temarense	2926	SRR1296766	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Alexandrium	temarense	2926	SRR1296765	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Amphidinium	carterae	2961	SRR1294392	Yes	CM	Jeong, H. J., Lee, K. H., Du Yoo, Y., Kang, N. S., Song, J. Y., Kim, T. H., ... & Potvin, E. (2018). Effects of light intensity, temperature, and salinity on the growth and ingestion rates of the red-tide mixotrophic dinoflagellate Paragymnodinium shiwhaense. Harmful Algae, 80, 46-54.
Alveolata	Dinoflagellata	Dinophyceae	Amphidinium	carterae	2961	SRR1294391	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Amphidinium	carterae	2961	SRR1296757	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Amphidinium	carterae	2961	SRR1296758	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Amphidinium	carterae	2961	SRR1294393	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Amphidinium	carterae	2961	SRR1294394	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Amphidinium	massartii	160604	SRR1296892	Yes	CM	Meng, A., Corre, E., Probert, I., Gutierrez-Rodriguez, A., Siano, R., Annamale, A., ... & Not, F. (2018). Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network. Molecular ecology, 27(10), 2365-2380.
Alveolata	Dinoflagellata	Dinophyceae	Ceratium	fusus	2916	SRR1300300	Yes	CM	Baek, S. H., Shimode, S., & Kikuchi, T. (2007). Reproductive ecology of the dominant dinoflagellate, Ceratium fusus, in coastal area of Sagami Bay, Japan. Journal of Oceanography, 63(1), 35-45.
Alveolata	Dinoflagellata	Dinophyceae	Ceratium	fusus	2916	SRR1300301	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Gonyaulax	spinifera	66791	SRR1300518	Yes	CM	Stoecker, D. K. (1999). Mixotrophy among Dinoflagellates 1. Journal of Eukaryotic Microbiology, 46(4), 397-401.
Alveolata	Dinoflagellata	Dinophyceae	Gymnodinium	catenatum	39447	SRR1296705	Yes	CM	Jeong, H. J., Du Yoo, Y., Park, J. Y., Song, J. Y., Kim, S. T., Lee, S. H., ... & Yih, W. H. (2005). Feeding by phototrophic red-tide dinoflagellates: five species newly revealed and six species previously known to be mixotrophic. Aquatic microbial ecology, 40(2), 133-150.
Alveolata	Dinoflagellata	Dinophyceae	Heterocapsa	rotundata	89963	SRR1296810	Yes	CM	Jeong, H. J. (2011). Mixotrophy in Red Tide Algae Raphidophytes 1. Journal of Eukaryotic Microbiology, 58(3), 215-222.
Alveolata	Dinoflagellata	Dinophyceae	Heterocapsa	triquetra	66468	SRR1296978	Yes	CM	Legrand, C., Graneli, E., & Carlsson, P. (1998). Induced phagotrophy in the photosynthetic dinoflagellate Heterocapsa triquetra. Aquatic Microbial Ecology, 15(1), 65-75.
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296744	Yes	CM	Glibert, P. M., Burkholder, J. M., Kana, T. M., Alexander, J., Skelton, H., & Shilling, C. (2009). Grazing by Karenia brevis on Synechococcus enhances its growth rate and may help to sustain blooms. Aquatic Microbial Ecology, 55(1), 17-30.
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296952	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296851	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296748	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296749	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296750	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296852	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296743	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296853	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296854	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296714	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karenia	brevis	156230	SRR1296712	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karlodinium	micrum	407301	SRR1300327	Yes	CM	Calbet, A., Bertos, M., Fuentes-Grünewald, C., Alacid, E., Figueroa, R., Renom, B., & Garcés, E. (2011). Intraspecific variability in Karlodinium
Alveolata	Dinoflagellata	Dinophyceae	Karlodinium	micrum	407301	SRR1300325	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Karlodinium	micrum	407301	SRR1300326	Yes	CM	

Supergroup	Phylum	Class	Genus	Species	NCBI ID	ENA ID	Mixotrophy	Mixotype	Reference
Alveolata	Dinoflagellata	Dinophyceae	Lingulodinium	polyedra	160621	SRR1300255	Yes	CM	Jeong, H. J., Du Yoo, Y., Park, J. Y., Song, J. Y., Kim, S. T., Lee, S. H., ... & Yih, W. H. (2005). Feeding by phototrophic red-tide dinoflagellates: five species newly revealed and six species
Alveolata	Dinoflagellata	Dinophyceae	Lingulodinium	polyedra	160621	SRR1300257	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Lingulodinium	polyedra	160621	SRR1300256	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Lingulodinium	polyedra	160621	SRR1300258	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Prorocentrum	lima	39448	SRR1300465	Yes	CM	Meng, A., Corre, E., Probert, I., Gutierrez-Rodriguez, A., Siano, R., Annamale, A., ... & Not, F. (2018). Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network. Molecular ecology, 27(10), 2365-2380.
Alveolata	Dinoflagellata	Dinophyceae	Prorocentrum	micans	2945	SRR1300466	Yes	CM	Jeong, H. J., Du Yoo, Y., Park, J. Y., Song, J. Y., Kim, S. T., Lee, S. H., ... & Yih, W. H. (2005). Feeding by phototrophic red-tide dinoflagellates: five species newly revealed and six species previously known to be mixotrophic. Aquatic microbial ecology, 40(2), 133-150.
Alveolata	Dinoflagellata	Dinophyceae	Prorocentrum	minimum	39449	SRR1296785	Yes	CM	Johnson, M. D. (2015). Inducible mixotrophy in the dinoflagellate Prorocentrum minimum. Journal of Eukaryotic Microbiology, 62(4), 431-443.
Alveolata	Dinoflagellata	Dinophyceae	Prorocentrum	minimum	39449	SRR1296784	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Prorocentrum	minimum	39449	SRR1296788	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Prorocentrum	minimum	39449	SRR1296752	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Prorocentrum	minimum	39449	SRR1296754	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Prorocentrum	minimum	39449	SRR1296787	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Prorocentrum	minimum	39449	SRR1296753	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Pyrocystis	lunula	2972	SRR1300467	Yes	CM	Meng, A., Corre, E., Probert, I., Gutierrez-Rodriguez, A., Siano, R., Annamale, A., ... & Not, F. (2018). Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network. Molecular ecology, 27(10), 2365-2380.
Alveolata	Dinoflagellata	Dinophyceae	Scrippsiella	trochoidea	71861	SRR1296761	Yes	CM	Jeong, H. J., Du Yoo, Y., Park, J. Y., Song, J. Y., Kim, S. T., Lee, S. H., ... & Yih, W. H. (2005). Feeding by phototrophic red-tide dinoflagellates: five species newly revealed and six species
Alveolata	Dinoflagellata	Dinophyceae	Scrippsiella	trochoidea	71861	SRR1296760	Yes	CM	
Alveolata	Dinoflagellata	Dinophyceae	Scrippsiella	trochoidea	71861	SRR1296759	Yes	CM	
Archaeplastida	Chlorophyta	Mamiellophyceae	Mantoniella	antarctica	81844	SRR1300318	Yes	CM	McKie-Krisberg, Z. M., Gast, R. J., & Sanders, R. W. (2015). Physiological responses of three species of Antarctic mixotrophic phytoflagellates to changes in light and dissolved nutrients. Microbial ecology, 70(1), 21-29.
Archaeplastida	Chlorophyta	Mamiellophyceae	Micromonas	commoda	38833	SRR1300457	Yes	CM	Unrein, F., Gasol, J. M., Not, F., Forn, I., & Massana, R. (2014). Mixotrophic haptophytes are key bacterial grazers in oligotrophic coastal waters. The ISME journal, 8(1), 164.
Archaeplastida	Chlorophyta	Mamiellophyceae	Micromonas	polaris	38833	SRR1300367	Yes	CM	
Archaeplastida	Chlorophyta	Mamiellophyceae	Micromonas	pusilla	38833	SRR1300456	Yes	CM	
Archaeplastida	Chlorophyta	Mamiellophyceae	Micromonas	sp	38833	SRR1300458	Yes	CM	
Archaeplastida	Chlorophyta	Mamiellophyceae	Micromonas	sp	38833	SRR1300455	Yes	CM	
Hacrobia	Haptophyta	Prymnesiophyceae	Chrysochromulina	brevifilum	156173	SRR1300415	Yes	CM	McKie-Krisberg, Z. M. (2014). Phagotrophy in photosynthetic eukaryotic microbes from polar environments. Temple University.
Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesium	parvum	97485	SRR1296973	Yes	CM	Liu, Z., Jones, A. C., Campbell, V., Hambright, K. D., Heidelberg, K. B., & Caron, D. A. (2015). Gene expression in the mixotrophic prymnesiophyte, Prymnesium parvum, responds to prey availability. Frontiers in microbiology, 6, 319.
Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesium	parvum	97485	SRR1296769	Yes	CM	
Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesium	parvum	97485	SRR1296917	Yes	CM	
Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesium	parvum	97485	SRR1294411	Yes	CM	
Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesium	parvum	97485	SRR1300223	Yes	CM	
Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesium	parvum	97485	SRR1294412	Yes	CM	
Hacrobia	Haptophyta	Prymnesiophyceae	Prymnesium	parvum	97485	SRR1296710	Yes	CM	
Rhizaria	Cercozoa	Filosa-Chlorarachnea	Bigelowiella	natans	227086	SRR1300405	Yes	CM	Archibald, J. M., Rogers, M. B., Toop, M., Ishida, K. I., & Keeling, P. J. (2003). Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga Bigelowiella natans. Proceedings of the National Academy of Sciences, 100(13), 7678-7683.
Rhizaria	Cercozoa	Filosa-Chlorarachnea	Bigelowiella	natans	227086	SRR1300359	Yes	CM	
Rhizaria	Cercozoa	Filosa-Chlorarachnea	Bigelowiella	natans	227086	SRR1300357	Yes	CM	
Rhizaria	Cercozoa	Filosa-Chlorarachnea	Bigelowiella	natans	227086	SRR1296871	Yes	CM	
Rhizaria	Cercozoa	Filosa-Chlorarachnea	Bigelowiella	natans	227086	SRR1300358	Yes	CM	
Rhizaria	Cercozoa	Filosa-Chlorarachnea	Bigelowiella	natans	227086		Yes	CM	
Stramenopiles	Ochrophyta	Chrysophyceae	Dinobryon	sp	98059	SRR1296865	Yes	CM	Kamjunke, N., Henrichs, T., & Gaedke, U. (2006). Phosphorus gain by bacterivory promotes the mixotrophic flagellate Dinobryon spp. during re-oligotrophication. Journal of plankton research, 29(1), 39-46.
Stramenopiles	Ochrophyta	Chrysophyceae	Dinobryon	sp	98059	SRR1294384	Yes	CM	
Stramenopiles	Ochrophyta	Chrysophyceae	Dinobryon	sp	98059	SRR1296864	Yes	CM	
Stramenopiles	Ochrophyta	Chrysophyceae	Ochromonas	sp	2985	SRR1300383	Yes	CM	Maranger, R., Bird, D. F., & Price, N. M. (1998). Iron acquisition by photosynthetic marine phytoplankton from ingested bacteria. Nature, 396(6708), 248.
Stramenopiles	Ochrophyta	Chrysophyceae	Ochromonas	sp	2985	SRR1300317	Yes	CM	
Stramenopiles	Ochrophyta	Chrysophyceae	Ochromonas	sp	420556	SRR1296863	Yes	CM	
Stramenopiles	Ochrophyta	Chrysophyceae	Ochromonas	sp	420556	SRR1296767	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Chattonella	subsalsa	44440	SRR1300238	Yes	CM	Jeong, H. J. (2011). Mixotrophy in Red Tide Algae Raphidophytes 1. Journal of Eukaryotic Microbiology, 58(3), 215-222.
Stramenopiles	Ochrophyta	Raphidophyceae	Chattonella	subsalsa	44440	SRR1300239	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Chattonella	subsalsa	44440	SRR1300240	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Chattonella	subsalsa	44440	SRR1300237	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Fibrocapsa	japonica	94617	SRR1300377	Yes	CM	Jeong, H. J. (2011). Mixotrophy in Red Tide Algae Raphidophytes 1. Journal of Eukaryotic Microbiology, 58(3), 215-222.
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	2829	SRR1296775	Yes	CM	Jeong, H. J. (2011). Mixotrophy in Red Tide Algae Raphidophytes 1. Journal of Eukaryotic Microbiology, 58(3), 215-222.
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	2829	SRR1296777	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	2829	SRR1296797	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	2829	SRR1296798	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	2829	SRR1296776	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	2829	SRR1296915	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	536046	SRR1296856	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	2829	SRR1296916	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	536046	SRR1296857	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	536046	SRR1296858	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	2829	SRR1296914	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	2829	SRR1296913	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	2829	SRR1296799	Yes	CM	
Stramenopiles	Ochrophyta	Raphidophyceae	Heterosigma	akashiwo	536046	SRR1296859	Yes	CM	

Appendix C

Supplementary tables Faure et al. 2020

Table S1: List of Kegg metabolic pathways detected in protein functional clusters (PFCs). The number of occurrences in the 233,756 PFCs and in the 2,444 PFCs highly linked to environmental gradients (hlePFCs) are given for each pathway. The percentage of occurrences among hlePFCs is also given for each pathway, reflecting how the pathway was selected by our random forest approach.

Table S1

Pathway description	Total occurrences	Occurrences among hlePFCs	Percentage of occurrences among hlePFCs
Biosynthesis of vancomycin group antibiotics	43	2	4.65116279069767
Antigen processing and presentation	66	3	4.54545454545455
Estrogen signaling pathway	66	3	4.54545454545455
IL-17 signaling pathway	66	3	4.54545454545455
Progesterone-mediated oocyte maturation	66	3	4.54545454545455
Prostate cancer	66	3	4.54545454545455
Th17 cell differentiation	66	3	4.54545454545455
Cardiac muscle contraction	182	8	4.3956043956044
Phenylpropanoid biosynthesis	167	7	4.19161676646707
Linoleic acid metabolism	24	1	4.16666666666667
Parkinson disease	215	8	3.72093023255814
Glycosphingolipid biosynthesis - ganglio series	27	1	3.7037037037037
Non-alcoholic fatty liver disease (NAFLD)	225	8	3.55555555555556
Apoptosis - fly	118	4	3.38983050847458
Other types of O-glycan biosynthesis	30	1	3.33333333333333
Huntington disease	276	9	3.26086956521739
Alzheimer disease	371	12	3.23450134770889
Salmonella infection	124	4	3.2258064516129
PI3K-Akt signaling pathway	94	3	3.19148936170213
Phenazine biosynthesis	190	6	3.15789473684211
Various types of N-glycan biosynthesis	32	1	3.125
RNA polymerase	329	10	3.03951367781155
Legionellosis	270	8	2.96296296296296
Shigellosis	141	4	2.83687943262411
Carotenoid biosynthesis	145	4	2.75862068965517
Thermogenesis	591	16	2.7072758037225
NOD-like receptor signaling pathway	338	9	2.66272189349112
Plant-pathogen interaction	304	8	2.63157894736842
Methane metabolism	1563	41	2.62316058861164
Acarbose and validamycin biosynthesis	82	2	2.4390243902439
Type II diabetes mellitus	123	3	2.4390243902439
Ribosome	4603	112	2.43319574190745
Human papillomavirus infection	124	3	2.41935483870968
Protein processing in endoplasmic reticulum	126	3	2.38095238095238
Chloroalkane and chloroalkene degradation	129	3	2.32558139534884
C5-Branched dibasic acid metabolism	653	15	2.29709035222052
Carbon fixation in photosynthetic organisms	1140	26	2.28070175438596
Epithelial cell signaling in Helicobacter pylori infection	44	1	2.27272727272727
beta-Lactam resistance	640	14	2.1875
Photosynthesis	641	14	2.18408736349454
Pathogenic Escherichia coli infection	187	4	2.13903743315508
Chagas disease (American trypanosomiasis)	48	1	2.08333333333333
alpha-Linolenic acid metabolism	49	1	2.04081632653061

Histidine metabolism	936	19	2.02991452991453
Isoquinoline alkaloid biosynthesis	99	2	2.02020202020202
Carbon fixation pathways in prokaryotes	2092	42	2.00764818355641
Pentose phosphate pathway	1358	27	1.98821796759941
Geraniol degradation	458	9	1.96506550218341
African trypanosomiasis	51	1	1.96078431372549
Citrate cycle (TCA cycle)	1499	29	1.9346230820547
Biosynthesis of various secondary metabolites - part 2	104	2	1.92307692307692
MAPK signaling pathway - fly	52	1	1.92307692307692
Prodigiosin biosynthesis	574	11	1.91637630662021
Pertussis	53	1	1.88679245283019
Viral carcinogenesis	159	3	1.88679245283019
Cationic antimicrobial peptide (CAMP) resistance	533	10	1.87617260787993
D-Arginine and D-ornithine metabolism	54	1	1.85185185185185
Fructose and mannose metabolism	925	17	1.83783783783784
Biofilm formation - Pseudomonas aeruginosa	765	14	1.83006535947712
Propanoate metabolism	1921	35	1.82196772514315
Carbon metabolism	6538	119	1.82012847965739
Ethylbenzene degradation	55	1	1.81818181818182
Oxidative phosphorylation	2874	52	1.80932498260264
Biofilm formation - Escherichia coli	613	11	1.79445350734095
Vancomycin resistance	448	8	1.78571428571429
Amino sugar and nucleotide sugar metabolism	1738	31	1.78365937859609
Porphyrin and chlorophyll metabolism	1907	34	1.78290508652333
Glycosphingolipid biosynthesis - globo and isoglobo se	57	1	1.75438596491228
RNA degradation	912	16	1.75438596491228
Biosynthesis of amino acids	8454	148	1.75065057960729
Central carbon metabolism in cancer	345	6	1.73913043478261
2-Oxocarboxylic acid metabolism	1788	31	1.7337807606264
AMPK signaling pathway	176	3	1.70454545454545
beta-Alanine metabolism	647	11	1.70015455950541
Cell cycle - Caulobacter	1194	20	1.6750418760469
Biotin metabolism	1016	17	1.67322834645669
Tuberculosis	180	3	1.66666666666667
Two-component system	3963	66	1.66540499621499
Peptidoglycan biosynthesis	1350	22	1.62962962962963
Cushing syndrome	62	1	1.61290322580645
Polyketide sugar unit biosynthesis	186	3	1.61290322580645
Renal cell carcinoma	62	1	1.61290322580645
Biosynthesis of secondary metabolites	18860	303	1.60657476139979
Tryptophan metabolism	1623	26	1.60197165742452
Biosynthesis of ansamycins	126	2	1.58730158730159
Glyoxylate and dicarboxylate metabolism	2525	40	1.58415841584158
Valine, leucine and isoleucine biosynthesis	1074	17	1.5828677839851
ABC transporters	4189	66	1.57555502506565
Glucagon signaling pathway	381	6	1.5748031496063

Inositol phosphate metabolism	318	5	1.57232704402516
Synthesis and degradation of ketone bodies	510	8	1.56862745098039
HIF-1 signaling pathway	511	8	1.56555772994129
Quorum sensing	3451	54	1.56476383656911
Phenylalanine, tyrosine and tryptophan biosynthesis	1866	29	1.55412647374062
Biosynthesis of antibiotics	13874	215	1.5496612368459
Aminoacyl-tRNA biosynthesis	2650	41	1.54716981132075
Alanine, aspartate and glutamate metabolism	1817	28	1.5410016510732
Longevity regulating pathway	65	1	1.53846153846154
Microbial metabolism in diverse environments	12774	194	1.51870987944262
Butanoate metabolism	1913	29	1.51594354417146
Glycolysis / Gluconeogenesis	1848	28	1.51515151515152
Peroxisome	530	8	1.50943396226415
Fatty acid metabolism	2455	37	1.5071283095723
Fatty acid degradation	1067	16	1.49953139643861
Valine, leucine and isoleucine degradation	1738	26	1.49597238204833
Chlorocyclohexane and chlorobenzene degradation	268	4	1.49253731343284
Folate biosynthesis	1883	28	1.48698884758364
Sulfur relay system	876	13	1.48401826484018
Fatty acid biosynthesis	1550	23	1.48387096774194
Lipopolysaccharide biosynthesis	1281	19	1.4832162373146
Phenylalanine metabolism	1417	21	1.48200423429781
Terpenoid backbone biosynthesis	1284	19	1.4797507788162
Type I diabetes mellitus	68	1	1.47058823529412
Monobactam biosynthesis	479	7	1.46137787056367
Antifolate resistance	411	6	1.45985401459854
Nitrogen metabolism	685	10	1.45985401459854
Pyruvate metabolism	2551	37	1.45041160329283
Flagellar assembly	1534	22	1.43415906127771
Metabolic pathways	45830	652	1.42264891992145
Cysteine and methionine metabolism	2115	30	1.41843971631206
Glycerophospholipid metabolism	1135	16	1.40969162995595
Limonene and pinene degradation	357	5	1.40056022408964
Arginine biosynthesis	1179	16	1.35708227311281
Bacterial secretion system	1916	26	1.35699373695198
Glycine, serine and threonine metabolism	2656	36	1.35542168674699
One carbon pool by folate	1041	14	1.34486071085495
Drug metabolism - other enzymes	967	13	1.34436401240951
Dioxin degradation	75	1	1.33333333333333
Lysine degradation	1126	15	1.33214920071048
Protein export	1660	22	1.32530120481928
Carbapenem biosynthesis	151	2	1.32450331125828
Nucleotide excision repair	908	12	1.3215859030837
Base excision repair	1081	14	1.29509713228492
Styrene degradation	542	7	1.29151291512915
Thiamine metabolism	700	9	1.28571428571429

Novobiocin biosynthesis	312	4	1.28205128205128
Pathways in cancer	555	7	1.26126126126126
Tropane, piperidine and pyridine alkaloid biosynthesis	238	3	1.26050420168067
Insulin resistance	159	2	1.25786163522013
Fluorobenzoate degradation	160	2	1.25
Purine metabolism	3695	46	1.24492557510149
FoxO signaling pathway	81	1	1.23456790123457
Nicotinate and nicotinamide metabolism	1215	15	1.23456790123457
Longevity regulating pathway - worm	742	9	1.21293800539084
Vitamin B6 metabolism	497	6	1.20724346076459
Glycosaminoglycan degradation	250	3	1.2
Ubiquinone and other terpenoid-quinone biosynthesis	669	8	1.19581464872945
Homologous recombination	1840	22	1.19565217391304
Fluid shear stress and atherosclerosis	671	8	1.19225037257824
Primary immunodeficiency	85	1	1.17647058823529
Aminobenzoate degradation	851	10	1.17508813160987
Necroptosis	341	4	1.17302052785924
Pantothenate and CoA biosynthesis	1450	17	1.17241379310345
Lysine biosynthesis	1112	13	1.16906474820144
Riboflavin metabolism	601	7	1.16472545757072
Pyrimidine metabolism	2181	25	1.14626318202659
Benzoate degradation	1225	14	1.14285714285714
Caprolactam degradation	526	6	1.14068441064639
Longevity regulating pathway - multiple species	263	3	1.14068441064639
Cyanoamino acid metabolism	357	4	1.12044817927171
Atrazine degradation	90	1	1.11111111111111
Glucosinolate biosynthesis	182	2	1.0989010989011
Glycerolipid metabolism	643	7	1.08864696734059
Pentose and glucuronate interconversions	559	6	1.07334525939177
Lipoic acid metabolism	190	2	1.05263157894737
Streptomycin biosynthesis	382	4	1.04712041884817
PPAR signaling pathway	383	4	1.0443864229765
DNA replication	1437	15	1.04384133611691
Starch and sucrose metabolism	576	6	1.04166666666667
Sulfur metabolism	1065	11	1.03286384976526
Steroid hormone biosynthesis	291	3	1.03092783505155
Arginine and proline metabolism	1659	17	1.02471368294153
Galactose metabolism	491	5	1.0183299389002
Phosphatidylinositol signaling system	198	2	1.01010101010101
Xylene degradation	100	1	1.00
Selenocompound metabolism	724	7	0.966850828729282
D-Alanine metabolism	211	2	0.947867298578199
Tyrosine metabolism	535	5	0.934579439252336
Mismatch repair	1519	14	0.921658986175115
Adipocytokine signaling pathway	112	1	0.892857142857143
D-Glutamine and D-glutamate metabolism	336	3	0.892857142857143

Insulin signaling pathway	112	1	0.892857142857143
Sphingolipid metabolism	345	3	0.869565217391304
Glutathione metabolism	1160	10	0.862068965517241
Ferroptosis	119	1	0.840336134453782
Meiosis - yeast	126	1	0.793650793650794
Other glycan degradation	126	1	0.793650793650794
Phosphotransferase system (PTS)	127	1	0.78740157480315
Chemical carcinogenesis	387	3	0.775193798449612
Hepatocellular carcinoma	389	3	0.77120822622108
Thyroid hormone synthesis	130	1	0.769230769230769
Metabolism of xenobiotics by cytochrome P450	396	3	0.757575757575758
Biofilm formation - Vibrio cholerae	926	7	0.755939524838013
Ascorbate and aldarate metabolism	269	2	0.743494423791822
Drug metabolism - cytochrome P450	409	3	0.733496332518337
Bacterial chemotaxis	846	6	0.709219858156028
Lysosome	293	2	0.68259385665529
Platinum drug resistance	454	3	0.66079295154185
Phosphonate and phosphinate metabolism	154	1	0.649350649350649
Toluene degradation	171	1	0.584795321637427
Glutamatergic synapse	172	1	0.581395348837209
GABAergic synapse	184	1	0.543478260869565
Biosynthesis of unsaturated fatty acids	195	1	0.512820512820513
RNA transport	206	1	0.485436893203883
Taurine and hypotaurine metabolism	332	1	0.301204819277108
Degradation of aromatic compounds	701	2	0.285306704707561
Alcoholism	33	0	Not detected among hlePFCs
Amoebiasis	8	0	Not detected among hlePFCs
Amphetamine addiction	7	0	Not detected among hlePFCs
Amyotrophic lateral sclerosis (ALS)	50	0	Not detected among hlePFCs
Apelin signaling pathway	10	0	Not detected among hlePFCs
Apoptosis	51	0	Not detected among hlePFCs
Apoptosis - multiple species	31	0	Not detected among hlePFCs
Arabinogalactan biosynthesis - Mycobacterium	46	0	Not detected among hlePFCs
Arachidonic acid metabolism	107	0	Not detected among hlePFCs
Autophagy - animal	11	0	Not detected among hlePFCs
Autophagy - yeast	75	0	Not detected among hlePFCs
Bacterial invasion of epithelial cells	6	0	Not detected among hlePFCs
Basal transcription factors	16	0	Not detected among hlePFCs
Betalain biosynthesis	12	0	Not detected among hlePFCs
Bile secretion	9	0	Not detected among hlePFCs
Biosynthesis of enediyne antibiotics	15	0	Not detected among hlePFCs
Biosynthesis of siderophore group nonribosomal peptic	25	0	Not detected among hlePFCs
Biosynthesis of type II polyketide products	17	0	Not detected among hlePFCs
Biosynthesis of various secondary metabolites - part 3	2	0	Not detected among hlePFCs
Bisphenol degradation	9	0	Not detected among hlePFCs
Bladder cancer	6	0	Not detected among hlePFCs

Calcium signaling pathway	3	0	Not detected among hlePFCs
cAMP signaling pathway	9	0	Not detected among hlePFCs
Carbohydrate digestion and absorption	5	0	Not detected among hlePFCs
Cell cycle	12	0	Not detected among hlePFCs
Cholesterol metabolism	18	0	Not detected among hlePFCs
Choline metabolism in cancer	8	0	Not detected among hlePFCs
Cholinergic synapse	3	0	Not detected among hlePFCs
Chronic myeloid leukemia	2	0	Not detected among hlePFCs
Circadian rhythm	4	0	Not detected among hlePFCs
Cocaine addiction	7	0	Not detected among hlePFCs
Colorectal cancer	31	0	Not detected among hlePFCs
Cutin, suberine and wax biosynthesis	10	0	Not detected among hlePFCs
Cytosolic DNA-sensing pathway	1	0	Not detected among hlePFCs
Dopaminergic synapse	8	0	Not detected among hlePFCs
ECM-receptor interaction	59	0	Not detected among hlePFCs
EGFR tyrosine kinase inhibitor resistance	10	0	Not detected among hlePFCs
Endocytosis	5	0	Not detected among hlePFCs
Epstein-Barr virus infection	31	0	Not detected among hlePFCs
Ether lipid metabolism	29	0	Not detected among hlePFCs
Fanconi anemia pathway	21	0	Not detected among hlePFCs
Fatty acid elongation	1	0	Not detected among hlePFCs
Fc gamma R-mediated phagocytosis	5	0	Not detected among hlePFCs
Flavone and flavonol biosynthesis	3	0	Not detected among hlePFCs
Flavonoid biosynthesis	34	0	Not detected among hlePFCs
Furfural degradation	25	0	Not detected among hlePFCs
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	1	0	Not detected among hlePFCs
GnRH signaling pathway	5	0	Not detected among hlePFCs
Hepatitis B	43	0	Not detected among hlePFCs
Hepatitis C	43	0	Not detected among hlePFCs
Herpes simplex virus 1 infection	42	0	Not detected among hlePFCs
Human cytomegalovirus infection	31	0	Not detected among hlePFCs
Human immunodeficiency virus 1 infection	31	0	Not detected among hlePFCs
Human T-cell leukemia virus 1 infection	20	0	Not detected among hlePFCs
Hypertrophic cardiomyopathy (HCM)	7	0	Not detected among hlePFCs
Influenza A	43	0	Not detected among hlePFCs
Insect hormone biosynthesis	13	0	Not detected among hlePFCs
Isoflavonoid biosynthesis	2	0	Not detected among hlePFCs
Kaposi sarcoma-associated herpesvirus infection	31	0	Not detected among hlePFCs
Leishmaniasis	3	0	Not detected among hlePFCs
Lipoarabinomannan (LAM) biosynthesis	32	0	Not detected among hlePFCs
Mannose type O-glycan biosynthesis	1	0	Not detected among hlePFCs
MAPK signaling pathway - plant	128	0	Not detected among hlePFCs
MAPK signaling pathway - yeast	13	0	Not detected among hlePFCs
Measles	42	0	Not detected among hlePFCs
MicroRNAs in cancer	28	0	Not detected among hlePFCs
Mineral absorption	13	0	Not detected among hlePFCs

mRNA surveillance pathway	32	0	Not detected among hlePFCs
mTOR signaling pathway	10	0	Not detected among hlePFCs
Naphthalene degradation	25	0	Not detected among hlePFCs
Neomycin, kanamycin and gentamicin biosynthesis	20	0	Not detected among hlePFCs
Neuroactive ligand-receptor interaction	18	0	Not detected among hlePFCs
N-Glycan biosynthesis	42	0	Not detected among hlePFCs
Nitrotoluene degradation	48	0	Not detected among hlePFCs
Non-homologous end-joining	20	0	Not detected among hlePFCs
Nonribosomal peptide structures	17	0	Not detected among hlePFCs
Notch signaling pathway	2	0	Not detected among hlePFCs
Osteoclast differentiation	9	0	Not detected among hlePFCs
p53 signaling pathway	31	0	Not detected among hlePFCs
Pancreatic cancer	5	0	Not detected among hlePFCs
Pancreatic secretion	4	0	Not detected among hlePFCs
Parathyroid hormone synthesis, secretion and action	16	0	Not detected among hlePFCs
Penicillin and cephalosporin biosynthesis	81	0	Not detected among hlePFCs
Phagosome	6	0	Not detected among hlePFCs
Phospholipase D signaling pathway	5	0	Not detected among hlePFCs
Photosynthesis - antenna proteins	17	0	Not detected among hlePFCs
Polycyclic aromatic hydrocarbon degradation	106	0	Not detected among hlePFCs
Primary bile acid biosynthesis	48	0	Not detected among hlePFCs
Prion diseases	6	0	Not detected among hlePFCs
Prolactin signaling pathway	21	0	Not detected among hlePFCs
Proteasome	41	0	Not detected among hlePFCs
Protein digestion and absorption	19	0	Not detected among hlePFCs
Proteoglycans in cancer	95	0	Not detected among hlePFCs
Proximal tubule bicarbonate reclamation	41	0	Not detected among hlePFCs
Ras signaling pathway	5	0	Not detected among hlePFCs
Renin-angiotensin system	37	0	Not detected among hlePFCs
Renin secretion	7	0	Not detected among hlePFCs
Retinol metabolism	13	0	Not detected among hlePFCs
Retrograde endocannabinoid signaling	32	0	Not detected among hlePFCs
Rheumatoid arthritis	9	0	Not detected among hlePFCs
Ribosome biogenesis in eukaryotes	221	0	Not detected among hlePFCs
RIG-I-like receptor signaling pathway	4	0	Not detected among hlePFCs
Secondary bile acid biosynthesis	22	0	Not detected among hlePFCs
Serotonergic synapse	7	0	Not detected among hlePFCs
Sesquiterpenoid and triterpenoid biosynthesis	33	0	Not detected among hlePFCs
Small cell lung cancer	31	0	Not detected among hlePFCs
Sphingolipid signaling pathway	17	0	Not detected among hlePFCs
Staphylococcus aureus infection	3	0	Not detected among hlePFCs
Staurosporine biosynthesis	24	0	Not detected among hlePFCs
Steroid biosynthesis	29	0	Not detected among hlePFCs
Steroid degradation	61	0	Not detected among hlePFCs
Stilbenoid, diarylheptanoid and gingerol biosynthesis	34	0	Not detected among hlePFCs
Tetracycline biosynthesis	3	0	Not detected among hlePFCs

Thyroid hormone signaling pathway	1	0	Not detected among hlePFCs
Tight junction	21	0	Not detected among hlePFCs
Toxoplasmosis	31	0	Not detected among hlePFCs
Vibrio cholerae infection	8	0	Not detected among hlePFCs
Viral myocarditis	31	0	Not detected among hlePFCs
Wnt signaling pathway	6	0	Not detected among hlePFCs
Yersinia infection	2	0	Not detected among hlePFCs
Zeatin biosynthesis	83	0	Not detected among hlePFCs

Table S2: Description of the 51 environmental used in the study.

Table S2

VARIABLE ID	TYPE	LEVELS	DESCRIPTION
Season	Qualitative	Winter	Season of sampling
		Spring	
		Summer	
		Autumn	
Season moment	Qualitative	Winter_early	Season moment of sampling
		Winter_middle	
		Winter_late	
		Spring_early	
		Spring_middle	
		Spring_late	
		Summer_early	
		Summer_middle	
		Summer_late	
		Autumn_early	
		Autumn_middle	
		Autumn_late	
Depth	Qualitative	SRF	Qualitative depth, either surface or deep chlorophyll maximum
		DCM	
Marine biome	Qualitative	Coastal Biome	Biome of the sampling station
		Polar Biome	
		Trades Biome	
		Westerlies Biome	
Ocean region	Qualitative	[IO] Indian Ocean (MRGID: 1904)	Ocean region of the sampling station
		[MS] Mediterranean Sea (MRGID:1905)	
		[NAO] North Atlantic Ocean (MRGID:1912)	
		[NPO] North Pacific Ocean (MRGID:1908)	
		[RS] Red Sea (MRGID:4264)	
		[SAO] South Atlantic Ocean (MRGID:1914)	
		[SO] Southern Ocean (MRGID: 1907)	
		[SPO] South Pacific Ocean (MRGID:1910)	
		[ANTA] Antarctic Province (MRGID:21502)	

VARIABLE ID	TYPE	LEVELS	DESCRIPTION
Biogeographical province	Qualitative	[ARAB] Northwest Arabian Sea Upwelling Province (MRGID: 21475)	Biogeographical province of the sampling station, *sensu* Longhurst. The 4 letter code between brackets were used on Figure 3.
		[BENG] Benguela Current Coastal Province (MRGID: 21470)	
		[CAMR] Central American Coastal Province (MRGID: 21494)	
		[CARB] Caribbean Province (MRGID:21466)	
		[CHIL] Chile-Peru Current Coastal Province (MRGID: 21495)	
		[EAFR] Eastern Africa Coastal Province (MRGID:21473)	
		[FKLD] Southwest Atlantic Shelves Province (MRGID: 21469)	
		[GFST] Gulf Stream Province (MRGID:21454)	
		[GUIA] Guianas Coastal Province (MRGID:21463)	
		[ISSG] Indian South Subtropical Gyre Province (MRGID:21472)	
		[MEDI] Mediterranean Sea, Black Sea Province (MRGID: 21465)	
		[MONS] Indian Monsoon Gyres Province (MRGID:21471)	
		[NAST-E] North Atlantic Subtropical Gyral Province (MRGID:21467)	
		[NAST-W] North Atlantic Subtropical Gyral Province (MRGID:21455)	
		[NPST] North Pacific Subtropical and Polar Front Provinces (MRGID:21484)	
		[PEOD] Pacific Equatorial Divergence Province (MRGID: 21489)	
		[PNEC] North Pacific Equatorial Countercurrent Province (MRGID:21488)	
		[REDS] Red Sea, Persian Gulf Province (MRGID:21474)	
		[SATL] South Atlantic Gyral Province (MRGID:21459)	

VARIABLE ID	TYPE	LEVELS	DESCRIPTION
		[SPSG] South Pacific Subtropical Gyre Province, North and South (MRGID: 21486)	
Latitude	Quantitative		Latitude of sampling station
Longitude	Quantitative		Longitude of sampling station
Depth bottom max	Quantitative		Maximum depth of sampling
CO3	Quantitative		CO3 concentration at sampling station
Alkalinity total	Quantitative		Alkalinity at sampling station
Calcite saturation state	Quantitative		Calcite saturation state at sampling station
NO2	Quantitative		NO2 concentration at sampling station
Si	Quantitative		Silicium concentration at sampling station
NO3	Quantitative		NO3 concentration at sampling station
Temperature	Quantitative		Temperature in celsius degrees at sampling station
Salinity	Quantitative		Salinity at sampling station
Oxygen	Quantitative		Oxygen concentration at sampling station
ChlorophyllA	Quantitative		Chlorophyll A concentration at sampling station
Opt backscat coef 470nm	Quantitative		Optical backscattering coefficient, 470nm, includes backscattering by particulate and dissolve matter and water molecules.
Fluorescence	Quantitative		Fluorescence of colored dissolved organic matter
Moon phase prop	Quantitative		Moon phase proportion during sampling
Sunshine duration	Quantitative		Sunshine duration per day during sampling period
Iron 5m	Quantitative		Iron concentration at 5m depth at sampling station
Ammonium 5m	Quantitative		Ammonium concentration at 5m depth at sampling station
NO2 5m	Quantitative		NO2 concentration at 5m depth at sampling station
NO3 5m	Quantitative		NO3 concentration at 5m depth at sampling station

VARIABLE ID	TYPE	LEVELS	DESCRIPTION
Gradient surface temp SST	Quantitative		Horizontal gradient of sea surface temperature at sampling station, during an 8 days period around sampling time
Okubo weiss	Quantitative		Okubo-Weiss parameter at the sampling station
Lyapunov	Quantitative		Maximum Lyapunov exponent at sampling date and station
Residence time	Quantitative		Residence time of the water mass at sampling date and station
Depth euphotic zone	Quantitative		Depth of the euphoric zone at sampling station
Depth mixed layer	Quantitative		Depth of the mixed layer at sampling station
Depth chloro max	Quantitative		Depth of the chlorophyll maximum at sampling station
Depth max Brunt Väisälä	Quantitative		Depth of maximum Brunt Väisälä frequency at sampling station
Depth Max O2	Quantitative		Depth of the oxygen maximum at sampling station
Depth Min O2	Quantitative		Depth of the oxygen minimum at sampling station
Depth nitracline	Quantitative		Depth of the nitracline at sampling station
Carbon flux	Quantitative		Carbon flux at the sampling station
DepthBathy	Quantitative		Bathymetric depth at the sampling station
Coast_Distance	Quantitative		Shortest distance to the coast from sampling station
t_se	Quantitative		Mean temperature seasonal anomaly over the 2005-2012 period at sampling station (1° resolution). File ID : woa13_A5B2_t00_01v2.nc
s_se	Quantitative		Mean salinity seasonal anomaly over the 2005-2012 period at sampling station (1° resolution). File ID : woa13_A5B2_s00_01v2.nc
I_an	Quantitative		Mean density over the 2005-2012 period at sampling station (1° resolution). File ID : woa13_A5B2_I00_01.nc

VARIABLE ID	TYPE	LEVELS	DESCRIPTION
C_se	Quantitative		Mean conductivity seasonal anomaly over the 2005-2012 period at sampling station (1° resolution). File ID : woa13_A5B2_C00_01.nc
o_se	Quantitative		Mean dissolved oxygen seasonal anomaly over all available years at sampling station (1° resolution). File ID : woa13_all_o00_01.nc
O_an	Quantitative		Mean % of oxygen saturation over all available years at sampling station (1° resolution). File ID : woa13_all_O00_01.nc
O_se	Quantitative		Mean % of oxygen saturation seasonal anomaly over all available years at sampling station (1° resolution). File ID : woa13_all_O00_01.nc
i_se	Quantitative		Mean silicate concentration seasonal anomaly over all available years at sampling station (1° resolution). File ID : woa13_all_i00_01.nc
n_se	Quantitative		Mean nitrate concentration seasonal anomaly over all available years at sampling station (1° resolution). File ID : woa13_all_n00_01.nc
p_se	Quantitative		Mean phosphate concentration seasonal anomaly over all available years at sampling station (1° resolution). File ID : woa13_all_p00_01.nc

Appendix D

Curriculum Vitae

Emile Faure

Curriculum Vitae

44 rue de Tolbiac
75013 Paris

+33 6 42 40 21 13

emile.faure@mnhn.fr

Plankton ecology - Bioinformatics - Quantitative analysis - Omics
Numerical ecology - Functional traits - Biogeochemical cycles

Profile

PhD student in marine ecology and bioinformatics at the Laboratoire d'océanographie de Villefranche-sur-Mer (LOV), and the ISYEB, Sorbonne university, Paris. My PhD project is entitled « **Contributions of meta-omics data to the detection and quantification of functional traits in marine planktonic ecosystems** ».

Curriculum

PhD Student, Sorbonne Université – October 2017—Present
Interfaces pour le vivant (IPV) doctoral school.
Co-supervisors: **Sakina-Dorothée Ayata** & **Lucie Bittner**.

Interdisciplinary Masters in Life Sciences, Ecole Normale Supérieure – 2015—2017
Major in **theoretical ecology and modeling**. Ranked **1st over 68** students.
Main classes followed: computational biology, ecological systems modeling, ecology of microbial populations, oceanography, statistics, mathematics for biologists, multivariate statistics (summer school at the oceanographic laboratory of Villefranche sur mer).

Bi-disciplinary licence of biology and mathematics, Station biologique de Roscoff, Université Pierre et Marie Curie – 2012-2015
Selective cursus, diploma obtained with highest distinction (2nd over 15 students).
Third year abroad as an exchange student at California State University Monterey Bay.

Research experiences

M2 internship, Laboratoire d'analyse de données de séquençage haut-débit, Sorbonne Université - January 2017 - June 2017
« From omics to biogeochemical processes modeling in the Ocean »
Co-supervisors: **Sakina-Dorothée Ayata** & **Lucie Bittner**.

M1 internship, Treml Lab, The University of Melbourne - February 2016 - June 2016
« Modeling climate change impacts on multi-species marine populations connectivity across the Indo-Pacific Ocean. »
Supervisor: **Eric Treml**.

Undergraduate voluntary internships
« Modeling the fitness of cry-wolf plants », 6 weeks, ENS, supervisor: **Minus Van Baalen**
« Study of the impact of the hivernal storms of 2013-14 on the biodiversity of macrobenthic populations of Brittany », 2 months, Station biologique de Roscoff, co-supervisors: **Caroline Broudin** & **Eric Thiébaut**

Skills

Language
French: Native **English**: Fluent **German**: Intermediate

Computer sciences
High proficiency: **R**, **UNIX**, **bash**, **awk**, **bioinformatics tools** (Diamond, Prodigal, Salmon, Anvi'o, eggNOG mapper,…).

Intermediate proficiency: **Matlab**, **Python**, **Mathematica**.

Beginner: ArcGIS, C++.

Publications

Faure E, Not F, Benoiston A-S, Labadie K, Bittner L, Ayata S-D (2019).
Mixotrophic protists display contrasted biogeographies in the global ocean.
The ISME Journal.
doi: 10.1038/s41396-018-0340-5.

Faure E, Ayata S-D, Bittner L (*submitted on 04/24/20, in revision following 1st round of reviews*).
Towards omics-based predictions of planktonic functional composition from environmental data.
Nature communications.

Martini S, **Faure E**, Larras F, Boyé A, Aberle N, Archambault P, Bacouillard L, Beisner BE, Bittner L, Castella E, Danger M, Gauthier O, Karp-Boss L, Lombard F, Maps F, Stemann L, Thiébaut E, Usseglio-Polatera P, Vogt M, Laviale M, Ayata SD (*submitted on 09/17/2020 after modifications following the 2nd round of revision*).
Functional trait-based approaches as a common framework for freshwater and marine ecologists.
Limnology and Oceanography.

Conferences & workshops

2020

Faure E, Ayata SD, Bittner L. From genes to functional traits in the global ocean: building de novo plankton functional types from environmental metagenomics data. **AMEMR conference** in Plymouth (**Oral presentation**). Postponed due to corona virus outbreak.
Faure E, Ayata SD, Bittner L. From genes to functional traits in the global ocean: building de novo plankton functional types from environmental metagenomics data. **Gordon Research Conference on marine microbes** in Les Diablerets (Switzerland) (**Poster presentation**). Postponed due to corona virus outbreak.

2019

Faure E, Ayata SD, Bittner L. From genes to functional traits in the global ocean: building de novo plankton functional types from environmental metagenomics data. **IMBER Future Oceans 2** in Brest (**Oral presentation**).
Faure E, Ayata SD, Bittner L. From genes to functional traits in the global ocean: building de novo plankton functional types from environmental metagenomics data. **Fourth workshop on trait-based approaches to ocean life** in Chicheley Hall, Buckinghamshire (**Oral presentation**).
Faure E & Martiny A. How to link genomics with trait-based approaches ? **Fourth workshop on trait-based approaches to ocean life** in Chicheley Hall, Buckinghamshire (**Discussion co-leader**).
Martini S, Faure E, Larras F, Boyé A, Aberle N, Archambault P, Bacouillard L, Beisner BE, Bittner L, Castella E, Danger M, Gauthier O, Karp-Boss L, Lombard F, Maps F, Stemann L, Thiébaut E, Usseglio-Polatera P, Vogt M, Laviale M, Ayata SD. Functional trait-based approaches as a common framework for freshwater and marine ecologists. **Fourth workshop on trait-based approaches to ocean life** in Chicheley Hall, Buckinghamshire (**Poster presentation**).
International summer school on networks and evolution, in Roscoff.
Société française de statistiques workshop on Network approaches, inference and modeling, in Fréjus.

2018

Faure E, Aumont O, Bittner L, Ayata S-D. From omics to biogeochemical modeling in the global ocean. AGU-ASLO **Ocean Sciences Meeting** in Portland, Oregon (**Oral presentation**).

Faure E, Not F, Benoiston A-S, Labadie K, Bittner L, Ayata S-D. Ubiquitous yet contrasted biogeography of mixotrophic protists in the global ocean. **ISME international symposium** in Leipzig (**Oral presentation**).
Faure E, Not F, Benoiston A-S, Labadie K, Bittner L, Ayata S-D. Ubiquitous yet contrasted biogeography of mixotrophic protists in the global ocean. **SFécologie biennial meeting** in Rennes (**Oral presentation**).
Faure E, Not F, Benoiston A-S, Bittner L, Ayata S-D. Linking metabarcoding with environmental context to investigate protists biogeography. MixITiN workshop in Roscoff (**Oral presentation**)

2017
Faure E, Bittner L, Benoiston A-S, Aumont O., Ayata S-D. From omics to biogeochemical processes modeling in the Ocean. **AMEMR conference** in Plymouth (Poster presentation).
EBAME **microbial ecogenomics workshop** in Brest.

Student supervision

Aurélie Pham, **M2 internship**, Master Sciences de la Mer, Sorbonne Universités - January - June 2019
Co-supervision with Lucie Bitter. Evaluation of mixotrophic capacity in dinoflagellates through the use of omics data. Internship validated with the mark of 14.85/20.
Nina Guérin, **M1 internship**, Master Mécanismes du Vivant et Environnement, MNHN - April - June 2019
Co-supervision with Lucie Bittner. Evaluation of dimethylsulfur production by prokaryotic planktonic organisms through the use of omics data. Internship validated (14.97/20).

Teaching

February - March 2020
Modeling workshop for 1st year Masters students (**24 hours**). Focus on planktonic ecology modeling. Included **4 teaching hours**, **20 tutoring hours**, as well as collecting and grading reports for 4 groups of students.

October 2019
Introduction to environmental metagenomics for 2nd year master students (**12 hours**). Presentation of the Tara datasets, practical on sequence homolog detection in the OM-RGC catalog, and quick multivariate analysis of homologs distribution.

January - February 2019
Modeling workshop for 1st year Masters students (**20 hours**). Focus on planktonic ecology modeling. Included **4 teaching hours**, **16 tutoring hours**, as well as collecting and grading reports for 4 groups of students.

December 2018
Part of the jury for **professional insertion orals** of 1st year undergraduate students (**4 hours**).

July 2018
Teaching assistant during the multivariate statistics summer school held in the Villefranche-sur-Mer Oceanographic laboratory (**18 hours**).

Grants and prizes

2019
Korean institute for ocean science and technology (KIOST) grant for attending the IMBER Future Oceans 2 conference (300 €).
Société française de statistiques (SFdS) grant for attending the workshop on network approaches, inference and modeling (695 €).

2018
GDR Génomique Environnementale travel grant to make an oral presentation at the ISME international symposium in Leipzig (500 €).

2017
Interfaces pour le vivant (IPV) doctoral school grant for a three year PhD (>50,000 €).

Public outreach

Press communications presenting/citing my research
https://www.the-scientist.com/news-opinion/mixing-it-up-in-the-web-of-life-65431
http://www.cnrs.fr/fr/le-plancton-mixotrophe-acteur-meconnu-de-la-photosynthese
https://oceans.taraexpeditions.org/m/science/les-actualites/le-plancton-mixotrophe-acteur-meconnu-de-la-photosynthese/

Actions
Part of the *Planktomania* project animators, presenting the planktonic world to schoolchildren through high technology vulgarization materials (3D movies, interactive card games, VR videos, 3D printed models,…).

Oceanographic Cruises

May 2018
MOOSE campaign, 16 days in the mediterranean Sea on IFREMER's *Atalante*, with transects between Toulon, Minorque, Sardinia, and Perpignan. Biological sampling through plankton nets.

Last update : September 30th 2020