



SORBONNE UNIVERSITÉ
ÉCOLE DOCTORALE 129 SCIENCES DE L'ENVIRONNEMENT
LABORATOIRE D'Océanographie de Villefranche - UMR 7093

**Structure de l'écosystème planctonique :
apport des données à haut débit de séquençage et
d'imagerie**

PRÉSENTÉE POUR OBTENIR LE GRADE DE
DOCTEUR DE SORBONNE UNIVERSITÉ PAR

OPHÉLIE DA SILVA

Directeur : Lionel Guidi
Co-directeur : Fabien Lombard

Soutenance prévue à Villefranche-sur-Mer le 30 Septembre 2021

Devant un jury composé de :

URANIA CHRISTAKI	Professeure, Université du Littoral Côte d'Opale	Rapportrice
CORINNE LE QUÉRÉ	Professeure, University of East Anglia	Rapportrice
ÉRIC GOBERVILLE	MCU, Sorbonne Université	Examinateur
KARINE LEBLANC	CR CNRS, Institut Méditerranéen d'Océanologie	Examinatrice
LOUIS LEGENDRE	Professeur Émérite, Sorbonne Université	Examinateur
LIONEL GUIDI	CR CNRS, Sorbonne Université	Directeur
FABIEN LOMBARD	MCU, Sorbonne Université	Co-directeur

"Et toute science, quand nous l'entendons non comme un instrument de pouvoir et de domination, mais comme aventure de connaissance de notre espèce à travers les âges, n'est autre chose que cette harmonie, plus ou moins vaste et plus ou plus riche d'une époque à l'autre, qui se déploie au cours des générations et des siècles, par le délicat contrepoint de tous les thèmes apparus tour à tour, comme appelés du néant, pour se joindre en elle et s'y entrelacer"

Alexandre Grothendieck

Remerciements

Cette thèse est le résultat d'une aventure de trois ans, qui a été enrichissante aussi bien du point de vue professionnel que personnel. Et cela n'aurait pas été possible sans les nombreuses personnes qui ont contribué, de près ou de loin, à sa réalisation.

Je souhaite commencer par remercier Lionel et Fabien qui m'ont fait confiance pour ce projet. Merci beaucoup pour votre accompagnement et la liberté que vous m'avez laissée pour le mener. Lionel, merci pour tes conseils et pour avoir réussi à me guider vers mes aspirations scientifiques. Fabien, merci pour ta disponibilité et ces connaissances sur le plancton que tu m'as partagées. Grâce à toi, j'ai également eu la chance de participer à l'expédition Tara Pacific. Cette transat a été une incroyable expérience. Vous avez été très complémentaires dans mon encadrement et je vous dis donc un grand merci à tous les deux !

Je remercie également Lucie Bittner et Sakina-Dorothee Ayata avec qui j'ai fait mon stage de Master 2, ce qui m'a conduit à cette thèse. J'ai beaucoup aimé travailler avec vous. Merci pour vos encouragements et votre bienveillance. Lucie, merci d'avoir continué à t'impliquer dans le déroulement de ma thèse.

Un grand merci aux membres de mon comité de thèse : Fabrizio d'Ortenzio, Cécile Fauvelot et Amin Madoui. Votre regard externe et vos conseils ont été cruciaux dans le bon déroulement de ma thèse.

Je voudrais aussi remercier les membres de mon jury d'avoir accepté d'évaluer mon travail. Merci à Louis Legendre de présider ma soutenance. Merci Urania Christaki et Corinne Le Quéré de rapporter ce travail, et merci à Éric Goberville et Karine Leblanc d'en être examinateurs. C'est un véritable honneur de vous présenter mon travail et je suis particulièrement enthousiaste à l'idée de recevoir vos commentaires et d'échanger avec vous.

Merci beaucoup aux membres du Laboratoire d'Océanographie de Villefranche et, en particulier, à mon équipe pour leur accueil. J'ai été heureuse de vous rencontrer. C'était un plaisir de pouvoir travailler dans une si bonne ambiance et sur un si beau site. Un grand merci également à tous les membres de l'ABI de m'avoir régulièrement accueillie et pour votre gentillesse. C'était très agréable de venir travailler dans cette bonne humeur générale. Vous avez largement contribué au

bon déroulement de ma thèse et je vous en suis très reconnaissante.

Ce travail n'aurait pas été le même sans les personnes avec lesquelles j'ai directement travaillé. En particulier, merci à Enrico Ser-Giacomi et Alberto Baudena qui ont été au cœur de l'océanographie physique de mon travail. Merci beaucoup pour votre disponibilité. J'ai beaucoup apprécié échanger avec vous. Merci également à Séverine Martini avec qui j'ai pris plaisir à discuter sur les images et les traits morphologiques, et sans qui le travail sur la transparence n'aurait sûrement pas émergé. Je remercie également Damien Éveillard, Samuel Chaffron et Éric Pelletier pour leurs contributions dans les différentes parties de ce travail. Vos conseils ont été très importants.

Merci beaucoup à Laure Mousseau et Rodolphe Lemée, grâce à vous j'ai pu découvrir le monde de l'enseignement. Ça a été une très bonne expérience et j'ai maintenant la chance de pouvoir dire que les faciès de crevette n'ont plus de secret pour moi (ou presque)! Aussi, merci à Didier avec qui j'ai pris beaucoup de plaisir à encadrer les sorties en mer.

Également merci Véronique, pour ta gentillesse et toute l'organisation logistique. Merci également à Anne pour l'ensemble des missions organisées.

Merci à toutes les personnes qui ont rendu ces trois années si agréables. Mes collègues de bureau, bien sûr : Florian, je garde en tête le Paris-Bordeaux. Laure, t'aurais pensé ça possible d'avoir tant de bons souvenirs à JB ? (Et pas que là). Thelma, merci de t'être si bien occupé de la déco de la zone dauphin. Et merci aussi à tous les autres! Flavien, vivement les prochaines sorties. Aurélie, merci pour ces pauses tisanes si sympathiques (et indispensables pour être efficaces). Kévin, merci pour tes conseils avisés. Laura, merci de m'avoir fait découvrir les falaises niçoises! Camille, merci pour les bons moments.

Une pensée particulière pour Laurène, Claire, Pierre et Jonathan.

Et bien sûr, je finirai par remercier ma famille. Merci à ma mère pour son soutien depuis toujours. La base de tout cela, c'est toi. Merci à Jean-Marc, je sais que tu y es pour beaucoup dans ce lien avec la nature. Merci à mes papis et mamies, qui ont su éveiller ma curiosité depuis petite. Merci à ma grand-mère dont les encouragements comptent beaucoup. Merci à Denise et Alain pour votre intérêt et mes si agréables aller-retours du week-end. Sans oublier, un grand merci à ma belle-famille, que j'ai la chance d'avoir si chouette. Et enfin, Jérémie, merci d'avoir été là pendant ces trois années. Et avant ça. Et après ça.

Résumé

Les organismes planctoniques, acteurs clés des écosystèmes, soutiennent les réseaux trophiques et ont un rôle majeur dans les cycles biogéochimiques et la régulation du climat. Tandis que la répartition spatio-temporelle de la diversité planctonique peut être étudiée à plusieurs niveaux, du gène jusqu'à l'écosystème, comprendre les mécanismes qui sous-tendent cette organisation est un défi. En effet, la structure de la diversité résulte de différents processus évolutifs et écologiques qui peuvent agir simultanément sur le vivant. Depuis le début du XXI^{ème} siècle, le milieu océanique fait l'objet d'une surveillance croissante. De nombreuses plateformes d'observation ont été déployées permettant l'acquisition de très nombreuses données couvrant de multiples caractéristiques environnementales. En parallèle, les technologies d'étude du vivant se sont développées, conduisant à un échantillonnage sans précédent des organismes planctoniques. En particulier, les données à haut débit de séquençage et d'imagerie permettent de fournir des informations moléculaires, taxonomiques et fonctionnelles à l'échelle des communautés. L'objectif de cette thèse était d'explorer la structure des écosystèmes planctoniques à l'aide des données à haut débit de séquençage et d'imagerie. Le couplage avec les données environnementales pourrait contribuer à une meilleure compréhension de la répartition spatiale de la diversité planctonique, des espèces jusqu'aux communautés. Dans une première partie, la diversité génétique de protistes a été étudiée à l'échelle de l'espèce. L'hypothèse était que les données métagénomiques pourraient permettre d'accéder à l'organisation de cette diversité mal caractérisée pour les protistes, ainsi qu'aux mécanismes qui la sous-tendent. Dans une deuxième partie, le lien entre diversité génétique et diversité fonctionnelle a été exploré. La transparence a été ciblée. Ce trait fonctionnel est peu exploré à l'échelle des communautés et les bases moléculaires sont mal identifiées. Une approche permettant de faire émerger ce trait des données d'imagerie a été utilisée, ayant conduit à l'exploration de sa biogéographie et ses bases moléculaires. Dans la dernière partie, le haut potentiel de complémentarité entre jeux de données de séquençage, d'imagerie et environnementaux a été exploré, afin de mettre en lumière la structure multi-échelle de l'écosystème planctonique et d'identifier sa structure globale. Enfin, l'ensemble des résultats a été discuté pour mettre en évidence les apports que peuvent fournir ces données à la compréhension des écosystèmes planctoniques, ainsi que les limites auxquelles elles peuvent faire face.

Mots-clés. Plancton, Biogéographie, Meta-omique, Imagerie, Environnement

Abstract

Planktonic organisms are key actors in oceanic ecosystems, which support trophic networks and play a major role in biogeochemical cycles and climate regulation. While the spatio-temporal distribution of planktonic diversity can be investigated at several levels, from the gene to the ecosystem, identifying the underlying mechanisms is challenging. Indeed, the structure of diversity results from different evolutionary and ecological processes that can act simultaneously. Since the beginning of the 21st century, the oceanic environment has been increasingly monitored. Numerous observation platforms have been deployed, leading to the acquisition of a large amount of data for multiple environmental characteristics. At the same time, technologies for studying living organisms have been developed. Thus, an unprecedented sampling of planktonic organisms has taken place. In particular, high-throughput sequencing and imaging data provide molecular, taxonomic and functional information at several biological levels. The objective of this thesis was to explore the structure of planktonic ecosystems using high-throughput sequencing and imaging data. Coupling with environmental data could contribute to a better understanding of the spatial distribution of planktonic diversity, from species to communities. In the first part, the genetic diversity of protists was studied at the species level. The hypothesis was that metagenomics could provide access to the poorly characterized spatial organization of the intraspecific protist genetic diversity, as well as to the mechanisms underlying it. In a second part, the link between genetic diversity and functional diversity was explored. Transparency was targeted. This functional trait is little explored at the community level and its molecular basis is poorly identified. A data-driven approach allowed this trait to emerge from imaging data, leading to the exploration of its biogeography and molecular basis. In the last part, the high potential of complementarity between sequencing, imaging and environmental datasets was explored, in order to highlight the multi-scale structure of the planktonic ecosystem and to identify its global structure. Finally, all the results were discussed to highlight the contributions that these data can provide to the understanding of planktonic ecosystems, as well as the limitations they can face.

Keywords. Plankton, Biogeography, Meta-omics, Imaging, Environment

Table des matières

1	Introduction	1
1.1	L’océan, un abîme multifacette	1
1.1.1	La Planète bleue	1
1.1.2	Un milieu à priori stable et homogène	2
1.1.3	Plancton, ordre et démesure	5
1.2	Diversité et assemblages du vivant dans les écosystèmes planctoniques	8
1.2.1	Organisation et forçages des communautés	8
1.2.2	Estimation de la diversité du vivant	15
1.2.3	Du gène à l’écosystème	18
1.3	Étude des écosystèmes planctoniques à l’ère du <i>Big Data</i>	20
1.3.1	Déluge de données abiotiques...	21
1.3.2	... auquel s’additionne l’avalanche de données biologiques	22
1.3.3	Vers une compréhension holistique	29
1.4	Problématique et objectifs	30
2	Matériel et Méthodes général	35
2.1	Expéditions <i>Tara Oceans Polar Circle</i> et <i>Tara Pacific</i>	35
2.2	Les données biologiques	36
2.2.1	Données issues de séquençage à haut débit	36
2.2.2	Données issues des technologies d’imagerie à haut débit	39
2.3	Les données environnementales	40
2.3.1	Les données biogéochimiques	40
2.3.2	Les données de transport et mélange des masses d’eau	41
2.4	Analyse de données	43
2.4.1	Analyse en composantes principales	43

2.4.2	Méthodes de regroupement (ou classification)	43
2.4.3	Analyse factorielle multiple	44
2.4.4	Régression linéaire	44
2.4.5	Régression par les moindres carrés partiels parcimonieuse	44
3	(Méta)Génomique des populations de protistes	47
3.1	Préambule	47
3.2	Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations . .	49
3.2.1	Introduction	50
3.2.2	Results	53
3.2.3	Discussion	59
3.2.4	Conclusion. Are we ready for population metagenomics?	64
3.2.5	Material and Methods	65
3.3	Conclusion	68
4	Phénotype transparent des communautés planctoniques	71
4.1	Préambule	71
4.2	From morphological traits to molecular basis of transparency in the open ocean	73
4.2.1	Introduction	74
4.2.2	Material and Methods	75
4.2.3	Results	79
4.2.4	Discussion	84
4.2.5	Perspectives	88
4.3	Conclusion	92
5	Structure globale de l'écosystème planctonique	93
5.1	Préambule	93
5.2	Multi-scale exploration of the planktonic ecosystem in the North Atlantic Ocean	95
5.2.1	Introduction	96
5.2.2	Material and Methods	98
5.2.3	Results	102
5.2.4	Community composition along the Atlantic transect	105
5.2.5	Community morphology along the Atlantic transect	105
5.2.6	Global structure of the ecosystem	106

5.2.7	Discussion	112
5.3	Conclusion	116
6	Discussion générale	117
6.1	Décrypter la complexité des écosystèmes planctoniques à partir de données à haut-débit	117
6.1.1	Récapitulatif des principaux résultats	117
6.1.2	Méta-omique et imagerie, des données de diversité complémentaires	119
6.1.3	Les multiples structures de l'écosystème planctonique	122
6.2	Analyse de grands jeux de données	126
6.3	Écologie, une discipline en mutation constante	129
6.3.1	Pluridisciplinarité, spécialisation et collaboration en écologie . . .	129
6.3.2	Le besoin de travailler ensemble, oui mais comment ?	130
	Bibliographie	133
A	Contributions comme co-auteur	171
A.1	Godet et al., Deep Sea Research II, 2020	171
A.2	Richter et al., BioRxiv, 2020	174
A.3	Danko et al., Cell, 2020	178
B	Supplementary Information - Are we ready for population metagenomics?	183
B.1	Methods for preliminary investigations	184
B.1.1	Step 1. Bibliographic study of ecologically relevant protist species in the Mediterranean Sea	184
B.1.2	Step 2. Abundance analysis of the V9 eukaryotic metabarcodes in the TO Mediterranean samples	185
B.1.3	Step 3. Availability of protist reference genome(s) or transcriptome(s)	186
B.2	Metagenomics analysis	188
B.3	Geographic distances	190
B.4	Oceanographic distances	192
B.5	Overview of the study	195
C	Supplementary Information - From trait to molecular basis of the transparency	197
C.1	Morphological descriptors	198
C.2	Environmental organization	200

C.3 Morphological organization	201
D Supplementary Information - Multi-scale exploration of the planktonic ecosystem in the North Atlantic Ocean	203
D.1 Extended Material and Methods - Mathematical formalism	204
D.1.1 Datasets	204
D.1.2 Normalization and standardization	208
D.1.3 Community weighted mean (CWM) of morphological traits	209
D.1.4 Principal component analysis (PCA)	209
D.1.5 Multiple Factorial Analysis (MFA)	209
D.2 Environmental structuring	211
D.3 Multiple factor analysis	212
E Curriculum Vitae	215

Table des figures

1.1	Structure verticale en trois couches de l'océan mondial	3
1.2	Structure verticale des tourbillons	4
1.3	Structure tridimensionnelle des processus de sous-mésoséchelle au niveau de zones frontales	6
1.4	Diversité morphologique du plancton illustrée par Haeckel	6
1.5	Représentation schématique d'une niche	9
1.6	Schéma de la vision déterministe de la diversité	12
1.7	Diversité fonctionnelle	17
1.8	Cascades d'effets directs et indirects des gènes jusqu'aux écosystèmes . . .	20
1.9	Développement chronologique des plateformes d'observation océanogra- phiques	21
1.10	Carte d'altimétrie globale	22
1.11	Observation de l'océan	23
1.12	Image satellite de la couleur de l'océan	24
1.13	Évolution du coût de séquençage	25
1.14	Échantillonnage du plancton par les méthodes d'optique et d'imagerie . .	28
1.15	Biologie des écosystèmes et analyses intégratives des océans mondiaux . .	29
2.1	Expéditions <i>Tara Oceans</i> et <i>Tara Pacific</i>	35
2.2	Distances entre stations successives <i>Tara Oceans</i> et <i>Tara Pacific</i>	37
2.3	Images issues du Flowcam et Zooscan	40
3.1	Diversité eucaryote	48
3.2	Distributions of the horizontal and vertical coverages	54
3.3	Distribution the genetic distances for each species between the 13 Medi- terranean Sea stations	55

Table des figures

3.4	Geographic and environmental characterisation of the Mediterranean Sea stations	56
3.5	Genetic differentiation among the Mediterranean Sea populations	57
4.1	Stratégie de camouflage dans l’océan	72
4.2	Environmental organization of the stations	79
4.3	Principal component analysis performed on the morphological descriptors	81
4.4	Transparency proxies in the global ocean	82
4.5	Latitudinal distribution of functional diversity indices	83
4.6	Taxonomic organization of transparency	84
4.7	Transparency traits according to gene expression of biological processes .	85
4.8	Community composition based on metabarcoding data	89
4.9	Distribution of the node metrics in the co-occurring network according to their transparency	91
5.1	Sampling strategy along the North Atlantic transect of the <i>Tara</i> Pacific expedition	103
5.2	Contrasted biogeochemical and hydrodynamical environmental structuring along the Atlantic transect	107
5.3	Taxonomic composition of plankton communities	108
5.4	Morphological structuring of the communities	109
5.5	Integration of all datasets through a multiple factorial analysis	110
6.1	Importance de la définition taxonomique sur la détection de la diversité .	125
6.2	Utilisation des méthodes d’intégration de données en écologie	128
6.3	Collaboration computationnelle dans les projets de recherche	131
B.1	Distribution of the metabarcode abundances in the Mediterranean Sea . .	185
B.2	Representation of the geographic distances between station 12 and all others	190
B.3	Geographic distances between pairwise stations	191
B.4	Analysis strategy for population metagenomics of marine protists	195
C.1	Distribution of the morphological descriptors	199
C.2	Environmental organization of the stations	200
C.3	Morphological space	201
D.1	Environment structuring along the North Atlantic transect	211
D.2	Graph of variables obtained after proceeding MFA	212

D.3 Projection of the groups in the MFA space 213

Liste des tableaux

1.1	Vue d'ensemble des chapitres	33
2.1	Récapitulatif des diagnostics lagrangiens et eulériens utilisés.	42
3.1	Outputs of the statistical models	58
5.1	Imaging features used to compute CWM traits for plankton communities.	101
B.2	Metagenomic samples used for the analyses	188
B.3	SNPs pipeline	189
B.4	SNPs distribution	189
B.5	Oceanographic distances computed as minimum connection time	193
B.6	Oceanographic distances computed as probabilities of connection	194
C.1	Imaging descriptors	198
D.1	Environmental variables used for the analysis	205
D.2	Metrics characterizing mixing and transport properties of the water masses	207

Introduction

1.1 L'océan, un abîme multifacette

1.1.1 La Planète bleue

La Terre est une des huit planètes qui composent notre système solaire. Tout comme Mercure, Vénus ou Mars, c'est une planète tellurique, formée principalement de roches et à la surface solide. À l'instar de Mars, Jupiter, Saturne et Uranus qui possèdent un ou plusieurs satellites naturels, la Lune tourne autour de la Terre ce qui influence son fonctionnement (*i.e.*, influence gravitationnelle). La Terre présente cependant une atmosphère atypique et elle est caractérisée par la présence d'eau à sa surface dans ses trois états (gazeux, liquide, solide). Les conditions particulières en font une zone habitable compatible avec la vie (Caron, Gauthier, 2007).

L'origine de l'eau sur Terre est encore débattue. Historiquement, il est supposé que des apports externes en eau par des météorites ou des comètes aient permis l'apparition de l'eau sur Terre. Plus récemment, une étude remet en question le postulat initial suivant lequel à sa formation la Terre était constituée de matériaux secs. Les éléments nécessaires à l'apparition de l'eau auraient ainsi pu être présents sur Terre depuis son origine (Piani et al., 2020). L'eau est majoritairement présente sous forme liquide sur Terre. Les océans, les mers et autres étendues aquatiques représentent 75% de la surface terrestre. Associée à des phénomènes optiques (*i.e.*, diffraction, réfraction), cette présence massive d'eau liquide est à l'origine de sa couleur bleue lorsqu'elle est observée depuis l'espace. Elle a ainsi été qualifiée de Planète bleue.

À eux seuls, les océans représentent 97,1% du total de l'hydrosphère, soit 1,34 milliard de km³ (0.1% du volume total de la Terre). La géographie terrestre a largement évolué au cours du temps, sous l'influence des mouvements tectoniques des plaques (Burke, 2011). Ces modifications dans la distribution spatiale des continents et des océans ont modulé les connexions entre les différentes régions au cours du temps (Dietz, Holden, 1970). Actuellement, l'océan global est composé des océans Atlantique, Pacifique, Indien, Arctique et Austral, ainsi que de nombreuses mers telles que la mer Méditerranée ou la mer Baltique. Dès 1872, cet océan a fait l'objet d'une campagne océanographique mondiale, à bord du *Challenger*. Cependant, il reste actuellement inégalement étudié.

L'océan est un environnement difficile d'accès. Il peut être soumis à des vents violents ou de hautes vagues, comme dans les régions des 40^{ème} hurlants et 50^{ème} rugissants dans l'océan Austral. Il peut être couvert de glace de mer (ou banquise), le rendant peu accessible comme au niveau de l'océan Arctique. Les distances spatiales sont également très importantes. Par exemple, l'océan Pacifique représente à lui seul plus de 43% de la surface des océans (165 millions de km²). Le point maritime le plus éloigné de toute terre est d'ailleurs situé dans l'océan Pacifique sud, à 2670 km des côtes. Mais les difficultés d'exploration ne résident pas qu'en surface. Les océans et mers présentent une profondeur moyenne de 3682 m avec un maximum de près de 11 km au niveau de la fosse des Mariannes. Cette échelle verticale ajoute une difficulté à leur étude. Les grandes profondeurs ne sont explorées que depuis le XX^{ème} siècle grâce aux développements technologiques d'engins résistants aux conditions extrêmes de pression (*e.g.*, *Remotely Operated Vehicle*, Humphris, 2009). Ainsi, par opposition à l'environnement de surface, les profondeurs supérieures à quelques centaines de mètres restent moins étudiées (Costello et al., 2018; Martin et al., 2020).

1.1.2 Un milieu à priori stable et homogène

À première vue, l'océan paraît homogène, stable et sans barrière. Pourtant, c'est un environnement dynamique et contrasté. À large échelle, deux grands types de circulation sont mis en place par des mécanismes différents. Les variations de densité, médiées par des variations de température et de salinité, vont générer des phénomènes de convection entraînant des mouvements d'eau verticaux qui sont le moteur de la circulation thermohaline. D'autre part, les vents à la surface de l'océan combinés à la force de Coriolis sont à l'origine d'une circulation de surface caractérisée par la présence de larges gyres. La circulation mondiale, portée par ces circulations à large échelle forcées par la densité et les vents, forme le tapis roulant de l'océan (*i.e.*, *the Global Conveyor Belt*).

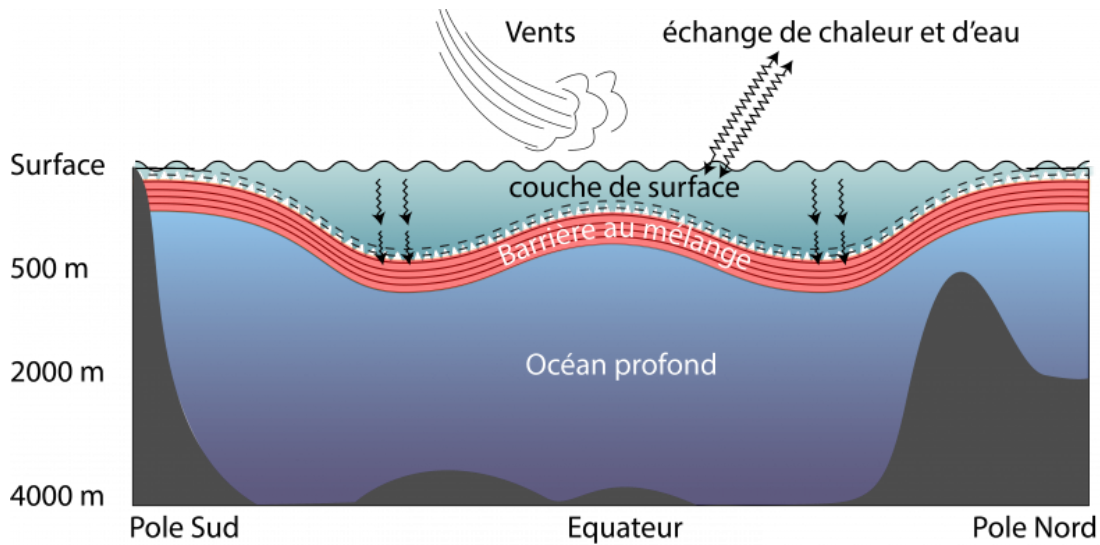


FIGURE 1.1 – Structure verticale en trois couches de l'océan mondial (© Jean-Baptiste Sallée, LOCEAN CNRS/MNHN/IRD/Sorbonne Université).

Elle est à l'origine de la distribution de la chaleur et des nutriments à l'échelle globale (Ganachaud, Wunsch, 2000). Des masses d'eau aux propriétés physico-chimiques différentes sont mises en contact ce qui génère une stratification de l'océan (Figure 1.1, Sallée et al., 2021). Ainsi, l'océan est décomposable en couches de densité différentes. La couche de surface, moins dense (*i.e.*, plus chaude et/ou moins salée), est bien mélangée sous l'action des vents. Cette couche homogène est aussi appelée couche de mélange. Sa profondeur varie suivant la latitude et les saisons (de quelques mètres à quelques centaines de mètres, Boyer Montégut de et al., 2004). La couche profonde, formée d'eau plus dense, est stable et peu dynamique. Entre les deux couches se situe la pynocline qui est une zone d'augmentation rapide de la densité. La pynocline peut coïncider avec la thermocline (*i.e.*, variation brutale de température) ou la halocline (*i.e.*, variation brutale de salinité) et peut-être définie de plusieurs façons (Boyer Montégut de et al., 2004). La pynocline forme une barrière entre la couche de mélange pauvre en nutriments et l'océan profond riche en nutriments. En hiver, l'approfondissement de la couche de mélange (*i.e.*, stratification moins intense) permet son enrichissement en nutriments. Il existe d'autres paramètres présentant des variations suivant la profondeur. Une diminution exponentielle de la quantité de la lumière et une altération de sa qualité (due à une absorption différentielle suivant les longueurs d'onde) avec la profondeur sont observées (Kirk, 1977; Johnsen, Sosik, 2004). Ainsi, le long de la colonne d'eau, des conditions physico-chimiques et de lumière très hétérogènes sont trouvées. L'océan mondial n'est

donc pas uniforme à large échelle. Il est dynamique et compartimenté. Cette description représente cependant une vision moyenne de l’océan et il existe de la variabilité à plus fine échelle.

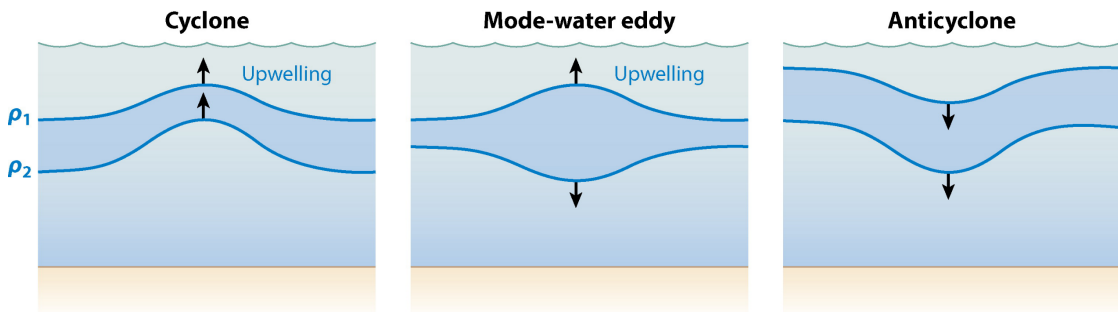


FIGURE 1.2 – Structure verticale des tourbillons. Les isopycnes sont déplacées suivant le type de tourbillons (depuis McGillicuddy, 2016).

L’écoulement de l’océan est turbulent. Cette turbulence apporte de l’hétérogénéité dans le paysage océanique à large échelle (Lozier, 2010). Elle est caractérisée par des processus dynamiques de plus petites échelles spatio-temporelles telles que la mésoéchelle (10-100 km, semaine-mois) et la sous-mésoéchelle (100 m-10 km, heures-semaines). L’activité de mésoéchelle se traduit par la présence de nombreux tourbillons qui résultent d’instabilités de courant large échelle (McGillicuddy, 2016). Leur dynamique est bien établie et leur écoulement est caractérisé par les équations de l’équilibre quasi-géostrophique. Ces tourbillons perturbent localement la couche de mélange soit en modifiant ses caractéristiques de transport, soit en modulant la disponibilité en nutriments et en lumière. L’agitation tourbillonnaire brasse les eaux et modifie la dispersion des particules. La plupart des tourbillons, du fait de leur dynamique non linéaire (*i.e.*, vitesse géostrophique du tourbillon supérieure à la vitesse de propagation), piègent le fluide en leur centre (Chelton et al., 2011). Ainsi, le cœur des tourbillons, qui correspond au milieu de leur formation, est isolé peut être transporté sur de longues distances (Zhang et al., 2014). Les fluides piégés peuvent donc présenter des caractéristiques différentes de celles du milieu environnant (*e.g.*, tourbillons formés à partir des méandres du Gulf Stream). Sur le plan vertical, la présence de tourbillons affecte la répartition des isopycnes et engendre des phénomènes de pompage vertical (Klein, Lapeyre, 2009; McGillicuddy, 2016). Les tourbillons cycloniques (*i.e.*, rotation antihoraire, dans l’hémisphère nord) sont associés à une remontée des isopycnes qui se traduit par un bombement de la surface, alors que les tourbillons anticycloniques (*i.e.*, rotation horaire, dans l’hémisphère nord) présentent une descente des isopycnes et une dépression de la surface (Figure 1.2). Un cas particulier

est celui des tourbillons d'eaux modales caractérisés par le soulèvement de la pycnocline saisonnière et l'abaissement de la pycnocline principale (*i.e.*, perturbations en forme de lentilles). Les tourbillons cycloniques et d'eaux modales provoquent une remontée des nutriments vers la surface, alors que les anticycloniques emportent vers le fond les eaux appauvries de surface. D'autres mécanismes impliquant les interactions avec les vents ou l'atmosphère modulent les impacts de ces tourbillons sur la structure des masses d'eau (*e.g.*, review de McGillicuddy, 2016).

Le cisaillement de l'écoulement par les tourbillons de mésoéchelles génère à sous-mésoéchelle la présence de filaments et de fronts de densité. Il existe cependant d'autres processus de formation de structures de sous-mésoéchelle (McWilliams, 2016). La turbulence de sous-mésoéchelle conduit généralement à une augmentation de la stratification, une couche de mélange plus fine et une diminution du mélange vertical. Les filaments sont associés à des phénomènes de brassage qui vont structurer la dispersion (McWilliams et al., 2009; Mahadevan, 2016). Les fronts de densité correspondent à des zones de contact entre masses d'eau aux propriétés différentes. Ils sont associés à un courant de jet horizontal et une circulation secondaire verticale (Figure 1.3). Cette circulation verticale est préférentiellement localisée dans la couche de mélange. Elle peut cependant atteindre la pycnocline et engendrer un courant local de nutriments (Lévy et al., 2018). Ces processus de mésoéchelle et sous-mésoéchelle, en modulant le transport des masses d'eau et les conditions locales en nutriments, influencent la distribution des organismes dans l'océan et en particulier celle du plancton.

1.1.3 Plancton, ordre et démesure

Le terme plancton (Hensen, 1887) regroupe tous les organismes qui vivent dans la colonne d'eau et qui ne peuvent se mouvoir activement contre les courants. Leur déplacement est donc dominé par le transport passif via les courants qui les dispersent. Ces organismes sont toutefois capables de déplacement actif (*e.g.*, migrations verticales) et peuvent présenter des structures morphologiques liées au déplacement (*e.g.*, flagelle, cils). La définition du plancton n'ayant pas de base phylogénétique (*i.e.*, pas d'ancêtre commun), le plancton est composé d'organismes très diversifiés des points de vue morphologique, taxonomique, physiologique ou trophique.

Le plancton, qui constitue la majeure partie de la biomasse des océans, regroupe des éléments génétiques mobiles (*e.g.*, virus), des procaryotes (*i.e.*, bactéries, archées), des eucaryotes unicellulaires (*i.e.*, protistes) et des eucaryotes multicellulaires (*i.e.*, métazoaires). Il présente une magnitude 6 de taille (de quelques nanomètres jusqu'à plusieurs mètres),

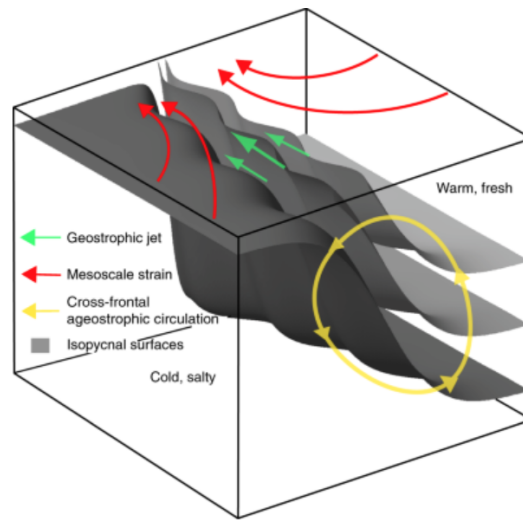


FIGURE 1.3 – Structure tridimensionnelle des processus de sous-mésoscale au niveau de zones frontales qui résultent de gradients de densité latéraux. Flèches rouges : tourbillon de mésoscale, flèches vertes : front, flèches jaunes : courant agéostrophique généré (depuis Lévy et al., 2018)

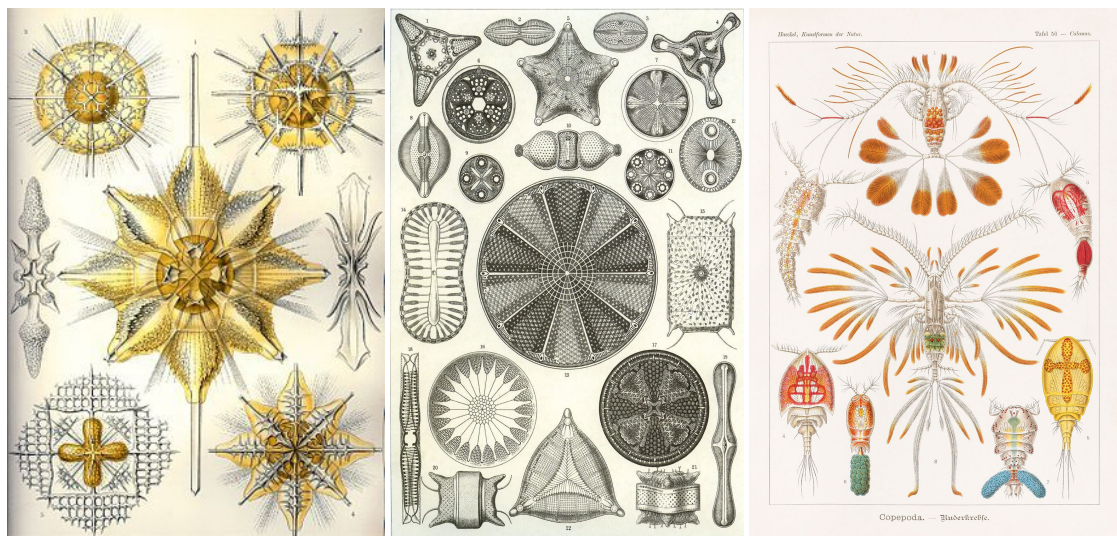


FIGURE 1.4 – Illustration schématique de la diversité morphologique des organismes planctoniques selon Haeckel dans 'Les formes de la nature' (1904). Adaptations à la vie pélagique.

des morphologies variées (Figure 1.4) et de nombreuses interactions inter-organismes (*e.g.*, prédation, symbiose). Les organismes planctoniques peuvent être autotrophes (*i.e.*, production de matière organique à partir de matière minérale), hétérotrophes (*i.e.*, pro-

duction de matière organique à partir de matière organique, zooplancton) ou mixotrophes (*i.e.*, capacité d'autotrophie et d'hétérotrophie, majorité des protistes; Stoecker et al., 2017). Le phytoplancton, regroupe les organismes planctoniques réalisant la photosynthèse, représente 1% de la biomasse photosynthétique terrestre, mais est responsable de 40% de la production d'oxygène et 50% de la production primaire nette (Field et al., 1998).

De par leurs caractéristiques physiologiques et interactions écologiques, les organismes planctoniques sont profondément impliqués dans les grands cycles biogéochimiques qui régissent le climat. Par exemple, ils interviennent dans trois des quatre mécanismes qui modulent les flux de carbone dans l'océan : (1) la pompe à carbone biologique (qui inclut la fixation du CO₂ dissous par la photosynthèse jusqu'à séquestration en profondeur du carbone faisant intervenir différents mécanismes; Boyd et al., 2019), (2) la pompe à carbonate (*i.e.*, export de carbone via la sédimentation de coquilles carbonatées d'organismes planctoniques calcifiants; Legendre et al., 2015) et (3) la pompe à carbone microbienne (*i.e.*, décomposition du carbone organique dissous labile en formes réfractaires via des processus de reminéralisation; Legendre et al., 2015). Le fonctionnement de la pompe à carbone est influencé par la structure des communautés planctoniques (Beaugrand et al., 2010; Guidi et al., 2016). Le plancton a également un rôle clé dans les réseaux trophiques qu'il soutient. Des changements de communautés de producteurs primaires modifient en cascade le réseau trophique jusqu'aux niveaux supérieurs (*i.e.*, prédateurs, poissons; Richardson et al., 2009). Des réorganisations spatio-temporelles de la distribution des communautés planctoniques pourraient donc affecter l'ensemble des réseaux trophiques (Edwards, Richardson, 2004).

Tout comme les conditions environnementales, la distribution du plancton n'est pas homogène dans l'océan et présente des variations spatio-temporelles (*e.g.*, Continuous Plankton Recorder Survey Team, 2004; Godet et al., 2020; Annexe A.1). Les communautés de plancton ont changé au cours des temps géologiques, au gré des variations climatiques (Lowery et al., 2020) ou de l'évolution de nouveaux acteurs de l'écosystème planctonique (Falkowski et al., 2004). Les changements climatiques actuels induisent aussi de larges modifications de communautés (*e.g.*, Beaugrand, 2004; Beaugrand et al., 2002; Jonkers et al., 2019), notamment marqués de migrations vers les hautes latitudes (Barton et al., 2016; Brun et al., 2019). Mais ces changements peuvent aussi être saisonniers ou inter-annuels (d'Alcalà et al., 2004; Beaugrand, 2005; Benedetti et al., 2019). Ainsi la composition des communautés planctoniques dans les océans n'est pas uniforme dans l'espace et dans le temps. Leur distribution et les processus à l'origine des variations

de la diversité peuvent alors être explorés.

1.2 Diversité et assemblages du vivant dans les écosystèmes planctoniques

La biogéographie est la discipline qui s'intéresse à la répartition spatio-temporelle de l'abondance et de diversité du vivant. Il est maintenant bien établi que les organismes planctoniques présentent des structures spatio-temporelles et ce, dans les multiples lignées phylogénétiques allant des organismes unicellulaires (*e.g.*, diatomés, Casteleyn et al., 2010; Godhe et al., 2016; Malviya et al., 2016; dinoflagellés, Okolodkov, Dodge, 1996; Casabianca et al., 2012; Lowe et al., 2012; haptophytes, Bittner et al., 2013b; Read et al., 2013; mamiellales, Monier et al., 2016; Vannier et al., 2016; Leconte et al., 2020; rhizaria, Meilland et al., 2016; Richirt et al., 2021) aux organismes pluricellulaires (*e.g.*, copépodes, Bonnet et al., 2005; Beaugrand et al., 2002; Madoui et al., 2017; cnidaires, Lucas et al., 2014; Leoni et al., 2021; Gueroun et al., 2021; euphausiacées, Letessier et al., 2011; Godet et al., 2020). L'identification des processus à l'origine de la distribution de l'abondance et de la diversité des organismes sont eux moins bien caractérisés (Logares, 2011; Hanson et al., 2012; Kraft et al., 2015). Différents cadres conceptuels permettent d'étudier ces processus et, dans cette partie, les processus écologiques façonnant la diversité et la composition des communautés seront ciblés (*i.e.*, échelle interspécifique). Ainsi, plusieurs théories de la biodiversité ont été formulées pour expliquer les assemblages des communautés et leur répartition spatio-temporelle. Ces théories, qui reposent sur des principes différents, ne sont pas exclusives les unes des autres. Elles font intervenir des processus aléatoires et d'autres basés sur trois types de forçages : les contraintes abiotiques, les contraintes biotiques et les contraintes à la dispersion.

1.2.1 Organisation et forçages des communautés

Environnement

Deux visions concernant l'influence de l'environnement sur les assemblages de communautés s'opposent mais sont non contradictoires. La première considère que tous les individus sont équivalents, comparables du point de vue de leur survie et mortalité. C'est la théorie neutre qui fournit une vision stochastique de la répartition de la biodiversité (Hubbell, 2001; Chave, 2004). Ainsi, les assemblages de communautés résulteraient de l'équilibre entre processus démographiques (*i.e.*, extinction, spéciation) aléatoires. La théorie neutre de la biodiversité a été vivement controversée, notamment pour

son hypothèse d'individus fonctionnellement identiques. Des études ont toutefois montré qu'elle pouvait expliquer la composition et l'abondance d'espèces au sein de communautés (Chust et al., 2013). Elle serait particulièrement pertinente dans le cadre de communautés riches en espèces avec de nombreuses espèces rares pour lesquelles les processus neutres deviennent inévitables.

Par opposition, la deuxième vision est déterministe. Elle repose sur le concept de niche écologique qui a été initialement défini par Roswell Johnson en 1910, puis reformulé en 1957 par Hutchinson dont la définition est aujourd'hui largement utilisée (Chase, Leibold, 2009). La niche multidimensionnelle de Hutchinson est un hypervolume à n-dimensions dans lequel chaque dimension correspond à une variable de l'environnement. Ce volume représente les gammes de valeurs des caractéristiques abiotiques dans lesquelles un organisme peut survivre, se développer et se reproduire (Figure 1.5).

Pour chaque variable environnementale, la gamme de valeurs dans laquelle peut survivre un organisme peut être plus ou moins grande. Elle définit une tolérance autour d'un optimum. Par exemple, il a été mesuré en laboratoire que *Skeletonema costatum* peut survivre et se développer dans une gamme de température allant de 5 à 25°C, alors que *Nitzschia frigida* préfère des températures comprises entre -1.8 à 2°C (Suzuki, Takahashi, 1995). Dans cet exemple, la température conditionne donc la diversité des milieux dans lesquels ces diatomés peuvent être présentes.

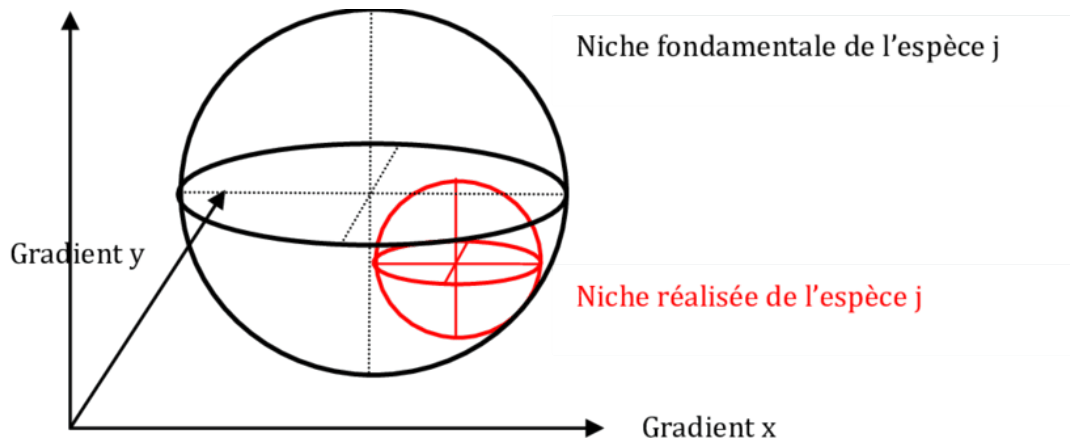


FIGURE 1.5 – Représentation schématique d'une niche à trois dimensions (depuis Boissesson, 2014). Chaque dimension représente un paramètre environnemental. L'ensemble des gammes de valeurs pour tous les paramètres compatibles avec la survie et le développement de l'espèce est représenté par la sphère noire. C'est la niche fondamentale. La niche réalisée en rouge tient compte des interactions avec les autres espèces qui coexistent et qui réduit l'espace de niche dans lequel l'espèce peut réellement persister dans l'habitat.

Interactions biotiques

Les organismes vivent en communauté. Ils ne sont pas isolés les uns des autres mais, au contraire, interagissent entre eux au sein de leur milieu. Ces interactions, directes ou indirectes, qui influencent la survie des organismes, sont diverses. Elles comprennent par exemple, la symbiose qui inclut le mutualisme (*i.e.*, bénéfice réciproque), le commensalisme (*i.e.*, bénéfice/effet neutre) ou le parasitisme (*i.e.*, bénéfice/préjudice), mais aussi les relations de prédateur/proies (*i.e.*, bénéfice/préjudice). Les conséquences des relations entre organismes sur les assemblages de communautés sont formulées dans la théorie de la coexistence (Chesson, 2000).

La théorie de la coexistence met l'accent sur les phénomènes de compétition qui peuvent exister entre les organismes (Chesson, 2000; HilleRisLambers et al., 2012; Barabás et al., 2018; Chesson, 2018). L'impact de la compétition pour la ressource sur les assemblages de communautés dépend à la fois de la niche fondamentale et de la valeur sélective (*i.e.*, *fitness*) des organismes considérés (HilleRisLambers et al., 2012; Kraft et al., 2015; Barabás et al., 2018). Ainsi, la coexistence des organismes est d'autant favorisée que leurs niches fondamentales sont différentes. La différence de niche est donc un effet stabilisateur de la coexistence. Au contraire, si les organismes présentent des niches fondamentales identiques, un phénomène de compétition exclusive conduisant à la survie d'un organisme vis-à-vis d'un autre sera observé (Hardin, 1960; Armstrong, McGehee, 1980). De plus, des valeurs sélectives très différentes (*i.e.* différence dans l'efficacité d'exploitation de ressources communes ou dans la tolérance d'un ennemi commun) conduisent à une compétition par dominance (HilleRisLambers et al., 2012; Kraft et al., 2015; Barabás et al., 2018).

En raison de ces interactions, les organismes n'occupent pas la totalité de leur niche fondamentale. Ils utilisent un sous-ensemble, appelé niche réalisée (Figure 1.5). Ainsi, les conditions environnementales ou biotiques peuvent conditionner la survie et le développement des organismes et donc les assemblages de communautés.

Dans le milieu aquatique, la diversité apparente du plancton est longtemps restée une énigme. En effet, suivant le principe d'exclusion compétitive, la coexistence d'espèces qui partagent la même niche fondamentale devrait conduire à des phénomènes d'exclusion. En particulier, le nombre d'espèces est supposé ne pas pouvoir dépasser le nombre de ressources limitantes. Or, une grande diversité est observée pour le plancton marin, en particulier en conditions limitantes. Des organismes partageant une même niche, peuvent coexister. Ainsi, ce phénomène de compétition exclusive a été remis en question. C'est

le ‘Paradoxe du plancton’ (Hutchinson, 1961). Il est supposé que la dynamique complexe de compétition génère un régime oscillatoire et chaotique propice à la coexistence de plusieurs espèces même en conditions de ressources limitantes (Huisman, Weissing, 1999; Scheffer et al., 2003). La dynamique de fine échelle qui restructure constamment l’environnement contribue également au maintien de cette diversité (Bracco et al., 2000; Scheffer et al., 2003).

Dispersion

La dispersion est un facteur contraignant la répartition spatiale des organismes en conditionnant leur capacité à atteindre une localisation et à s’y établir (*i.e.*, survie, développement et reproduction possibles), à partir de leur emplacement initial (Pineda et al., 2007; McManus, Woodson, 2012). Une métacommunauté est alors définie comme un ensemble de communautés spatialement séparées en interaction via les processus de dispersion (Leibold et al., 2004). Les relations entre groupes d’individus géographiquement séparés, ou connectivité, influencent donc la structure spatiale de la diversité.

La dispersion des organismes planctoniques permet de connecter des environnements locaux et donc le passage à une échelle spatiale plus grande. Elle repose en partie sur les processus de transport passif par les masses d’eau (*i.e.*, phénomènes d’advection et de diffusion), qui peuvent être modulés par les comportements de nage verticaux des organismes (McManus, Woodson, 2012; Chaput et al., 2019). La capacité d’un organisme à s’établir dans un habitat atteint par transport peut dépendre à la fois de processus déterministes que des processus neutres.

D’une part, les caractéristiques environnementales locales peuvent être favorables ou incompatibles avec sa persistance de l’organisme dans le milieu (théorie des niches). C’est la filtration environnementale qui, au sens strict, correspond au processus par lequel les conditions environnementales empêchent un organisme de s’établir et se développer dans le lieu atteint (Kraft et al., 2015). La persistance de l’organisme dans le milieu atteint dépend également des organismes déjà présents (théorie de la coexistence). L’ensemble du processus est appelé filtration par habitat (basée sur la niche réalisée, Figure 1.5; Kraft et al., 2015). La distinction entre la filtration environnementale et la filtration par habitat est difficile à établir en conditions naturelles, et de nombreuses études ont conclu à tort identifier des processus de filtration environnementale (Kraft et al., 2015; Cadotte, Tucker, 2017). Finalement, dans la vision déterministe de la biodiversité, l’organisation des communautés résulte de l’application successive de filtre à la dispersion, filtre par l’environnement et filtre biotique (Figure 1.6.).

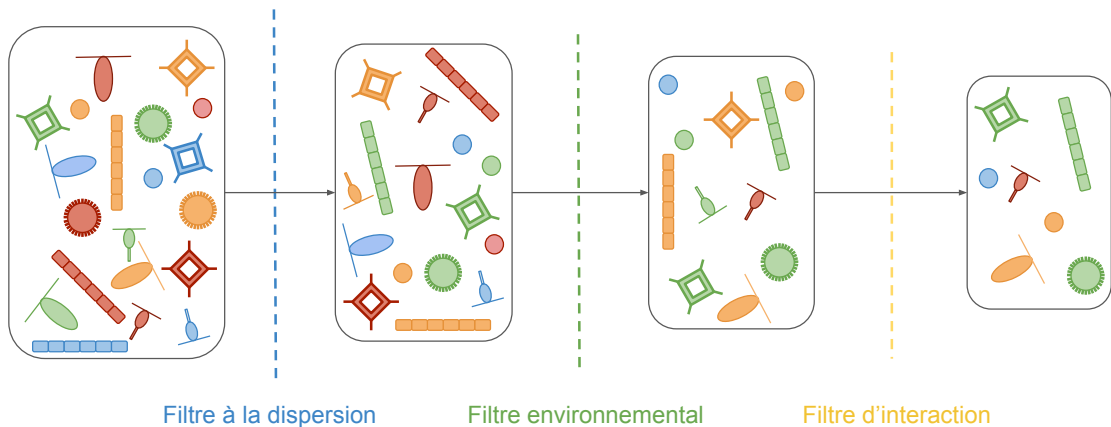


FIGURE 1.6 – Schéma représentant la vision déterministe des filtres en série hiérarchique qui, depuis un ensemble de résidents locaux potentiels (pool régional), va sélectionner ceux qui se trouvent réellement dans le milieu (adapté depuis Cadotte, Tucker, 2017). Ici le filtre environnemental correspond à la première étape = le transport n’est pas possible. Le filtre à la dispersion correspond au fait que dans la nouvelle localité les conditions abiotiques vont ou non être favorables (basé sur la niche des organismes). Enfin le filtre d’interaction correspond au fait que les espèces présentes peuvent être ou non plus compétitives et donc permettre l’implantation.

La dispersion peut aussi résulter de processus neutres. Bien qu’une filtration environnementale importante basée sur la sélection de niche tendra à diminuer l’influence des processus de dérive stochastique (Chase, Leibold, 2009), la part entre les processus neutres et les processus de niche dans les patterns d’abondance des communautés n’est pas claire (*e.g.*, Cottenie, 2005).

Historiquement, l’influence des conditions environnementales sur la biogéographie des organismes planctoniques unicellulaires a été très étudiée (De Wit, Bouvier, 2006). Il a longtemps été considéré que ces organismes n’étaient pas limités par la dispersion (Finlay, 2002). Ainsi, l’organisation spatiale des communautés résulterait de la sélection par l’environnement comme formulé par le paradigme basé sur la théorie des niches ‘Everything is everywhere, but, the environment selects’ (Baas-Becking hypothesis; De Wit, Bouvier, 2006; O’Malley, 2008). D’après ce paradigme, tous les organismes sont cosmopolites et l’environnement sélectionne les dominants. La majorité de la diversité serait donc latente et difficilement observable en raison des capacités de détection des méthodes d’étude. Cependant, cette diversité latente pourrait croître dès que les conditions deviennent favorables. Toutefois, les processus sélectifs ne sont pas les seuls à façonner

la répartition de la diversité microbienne qui résulteraient de plusieurs processus (Chust et al., 2013; Hellweger et al., 2014; Fondi et al., 2016; Logares et al., 2020; Ward et al., 2021).

Effets confondants

L'étude des caractéristiques biologiques permet de fournir des indices sur les différents processus qui structurent la diversité. En particulier, il est supposé qu'à partir d'un pool d'espèces donné, les mêmes espèces sont retrouvées dans un même environnement. C'est le principe à la base des modèles de niches. L'environnement sélectionne les espèces qui peuvent tolérer les conditions abiotiques locales. Ainsi, si un trait, qui est une caractéristique mesurable à l'échelle des individus et qui influence leur valeur sélective (ou *fitness* ; Violle et al., 2007), est lié à la tolérance abiotique, alors il est supposé y avoir une convergence phénotypique dans l'environnement (Kraft et al., 2015). Les individus d'un environnement partageront un même phénotype. De même, des individus retrouvés dans des environnements distants aux caractéristiques abiotiques similaires auront le même phénotype. Si ces traits sont liés à un signal phylogénétique, il y aura alors convergence phylogénétique en plus de la convergence phénotypique. Ainsi, l'étude de l'assemblage des communautés est souvent basée sur l'identification de convergence de traits. Cependant, l'étude des processus de filtration environnementale basée sur la convergence a été critiquée (Kraft et al., 2015; Cadotte, Tucker, 2017). En effet, la convergence peut résulter d'autres processus tels que les mécanismes de coexistence d'après laquelle la coexistence est favorisée lorsque les niches des espèces ne se recoupent pas. Ainsi, différencier les effets de la filtration environnementale de ceux de la coexistence est compliqué et des cadres d'étude ont été proposés (*e.g.*, Letten et al., 2017).

Autres modèles de distribution de la diversité des métacommunautés

Dans un contexte de métacommunautés, l'effet de masse peut expliquer des patterns d'abondance et de diversité (*i.e.*, *source-sink dynamics*, Waller et al., 2018). Dans ce cadre, une dispersion intense d'une parcelle favorable au développement des organismes vers une autre dont les conditions sont incompatibles avec la croissance et la survie, peut aboutir à la persistance des organismes dans le milieu défavorable. Ce type de relation est pertinent dans le cadre des organismes benthiques au stade larvaire planctonique pour lesquels il est important d'identifier les parcelles source qui permettent le maintien des communautés à l'échelle régionale (*e.g.*, White, Samhuri, 2011). Cependant, elle peut aussi gouverner l'abondance locale d'organismes au cycle de vie entièrement planctonique

(*i.e.*, holoplancton). Par exemple, l'étude de la dynamique d'invasion du cténoptère *Mnemiopsis leidyi* dans la mer du nord, a permis d'identifier les communautés sources des communautés puits (Bolte et al., 2013). De manière générale, cette théorie a reçu une attention croissante dans le cadre des mesures de gestion et de définition de zones de conservation prioritaires (Hansen, 2011; Furrer, Pasinelli, 2016).

Afin d'expliquer la composition et l'abondance des communautés, la théorie de la dynamique des îlots (ou *patch dynamics*, Luisa, 2012) a également été formulée. Elle considère de multiples îlots identiques connectés entre eux. Les organismes les moins compétitifs s'y dispersent plus rapidement qu'ils ne s'éteignent dans d'autres. Les patterns d'abondance et de diversité sont donc modélisés au sein des îlots par la balance entre compétition et colonisation. Cette théorie permet notamment d'expliquer la dynamique verticale du zooplancton (Genin et al., 1994). En effet, les organismes zooplanctoniques effectuent des migrations nyctémérales (*i.e.*, *diel vertical migration*) en réponse à la pression de prédation. Ils s'alimentent en surface de nuit et trouvent refuge en profondeur de jour (Hays, 2003). Ces deux théories s'ajoutent à celles précédemment formulées impliquant des processus neutres et des processus déterministes.

Démêler les mécanismes qui façonnent les communautés est donc un défi. En particulier, il existe de très nombreuses théories écologiques pour expliquer la composition et l'abondance des espèces au sein des communautés. En 1994, 120 hypothèses ont été décrites pour expliquer la diversité des espèces (Palmer, 1994). Toutefois, réconcilier et unifier les différentes théories écologiques est compliqué des points de vue mathématique et conceptuel (Chase, 2005). Vellend (2010) a proposé de regrouper les différentes théories sous quatre grands types de processus : la spéciation qui génère de nouvelles espèces, la sélection qui témoignent des changements déterministes basés sur la valeur sélective des espèces, la dérive qui regroupe les modifications stochastiques et la dispersion qui représente les mouvements des espèces. Toutes les théories écologiques précédemment formulées peuvent être conceptualisées d'après l'importance qu'elle donne à chacun de ces processus (Vellend, 2010). Cette classification en quatre processus écologiques (*i.e.*, spéciation, sélection, dérive et dispersion) peut permettre de réconcilier les différentes approches écologiques et l'émergence d'un cadre général.

Afin d'étudier les schémas de distribution du vivant et de comprendre les mécanismes qui en sont à l'origine, il est nécessaire de quantifier cette diversité et d'identifier sa variation, afin de faire le lien avec les forçages potentiels. À cette fin, différentes approches se sont

développées.

1.2.2 Estimation de la diversité du vivant

Pour étudier la distribution de la diversité et déchiffrer les mécanismes qui peuvent la façonner, il est nécessaire d'utiliser des indices permettant de l'évaluer. La biodiversité englobe tous les niveaux de variation du vivant. Les indices peuvent donc être estimés à différents niveaux (*e.g.*, moléculaire, taxonomique, fonctionnel) et à différentes échelles, des individus jusqu'aux communautés.

Indices de diversité

La structure d'un système et donc sa variabilité peuvent être évalués à partir de la distribution de classes définies (*e.g.*, espèces, familles) et de leur importance. Cette importance peut être quantifiée via l'abondance ou, lorsque cela est pertinent, par la biomasse qui est un meilleur indicateur de fonctionnalité d'une classe car dépendant de l'activité métabolique (Brown et al., 2004). L'estimation de la diversité peut se faire à trois échelles (Whittaker, 1960). La diversité α correspond à la diversité dans un milieu échantillonné. C'est une mesure de diversité locale, à l'intérieur d'un système délimité. La diversité β compare la diversité entre systèmes locaux et permet d'estimer les phénomènes de remplacement et d'imbrication (Baselga, 2010). Enfin, la diversité γ correspond au taux d'addition de nouvelles classes lors d'échantillonnages multiples. Elle rend compte de la variabilité locale de la diversité au sein d'un même système à l'échelle régionale.

Il existe une multitude d'indices permettant d'estimer la diversité. Ces estimateurs sont basés sur une ou plusieurs composantes de la diversité. En effet, la diversité est classiquement décomposée en trois composantes indépendantes entre elles : la richesse, l'équitabilité et la divergence (*e.g.*, Mason et al., 2005). Plus la richesse, l'équitabilité et la divergence augmentent, plus le système est diversifié. Les définitions de ces composantes dépendent de l'approche utilisée. Ici, deux approches sont considérées : l'approche taxonomique et l'approche fonctionnelle.

Approche taxonomique

L'approche taxonomique se concentre sur l'identité des classes du système. Dans cette approche, la richesse correspond au nombre de classes qui coexistent. L'équitabilité quantifie la régularité de la distribution de ces classes. Ainsi, à nombre de classes constant, un système dont les classes sont présentes en même proportion sera plus diversifié qu'un

système ayant une classe dominante. Enfin, la divergence mesure la différence entre classes (distance ou dissimilarité). Cette mesure de divergence peut être pondérée ou non par la fréquence des classes. Dans une approche taxonomique, toutes les classes sont équidistantes. Ainsi, la divergence par paire de classes est constante. Pourtant, plus deux classes partagent un ancêtre commun lointain, moins elles sont similaires du point de vue phylogénétique. Il est possible de prendre en compte cette distance entre classe pour estimer la diversité phylogénétique (Graham, Fine, 2008; Cavender-Bares et al., 2009). Ainsi, la diversité taxonomique ou phylogénétique permet de caractériser les assemblages biologiques. Pour explorer le fonctionnement associé à ces assemblages, il est possible d'étudier leur diversité fonctionnelle.

Approche fonctionnelle

La diversité fonctionnelle est la composante de la diversité qui influe sur le fonctionnement du système (Díaz, Cabido, 2001; Petchey, Gaston, 2006). Elle repose sur le concept de traits fonctionnels qui sont le support d'un espace fonctionnel dans lequel la diversité fonctionnelle est estimée. En effet, l'espace fonctionnel est un espace euclidien multidimensionnel dont chacune des n dimensions correspond à un trait fonctionnel permettant de positionner les classes à partir de leurs valeurs de traits. Dans l'approche fonctionnelle, la richesse correspond au volume de l'espace occupé par toutes les classes dans l'espace fonctionnel. L'équitabilité est la régularité des distributions d'abondance relative des classes dans l'espace fonctionnel. Et la divergence quantifie la proportion de l'abondance totale due aux classes ayant les valeurs de traits les plus extrêmes (Figure 1.7, Mason et al., 2005; Mouchet et al., 2010; Mouillot et al., 2013). D'autres composantes de la diversité fonctionnelle ont été définies mais ne sont pas détaillées ici (Mouillot et al., 2013).

Les traits fonctionnels se répartissent dans plusieurs catégories : les traits d'histoire de vie, les traits morphologiques, les traits physiologiques et les traits comportementaux (Litchman, Klausmeier, 2008; Martini et al., 2020). Dans le cadre des organismes planctoniques, plusieurs traits fonctionnels ont été étudiés. Par exemple, la stratégie de nutrition est un trait physiologique et comportemental qui détermine la survie et le développement des organismes dans leur environnement en modulant l'acquisition des ressources (Martini et al., 2020). Chez les copépodes, la nutrition passive (basée sur la stratégie embusquée, ou *ambush feeding*) présente une organisation latitudinale (Prowse et al., 2019, non observée par Brun et al., 2016). Cette stratégie est dominante aux hautes latitudes et peut avoir des répercussions sur l'ensemble du réseau trophique (Prowse et al., 2018).

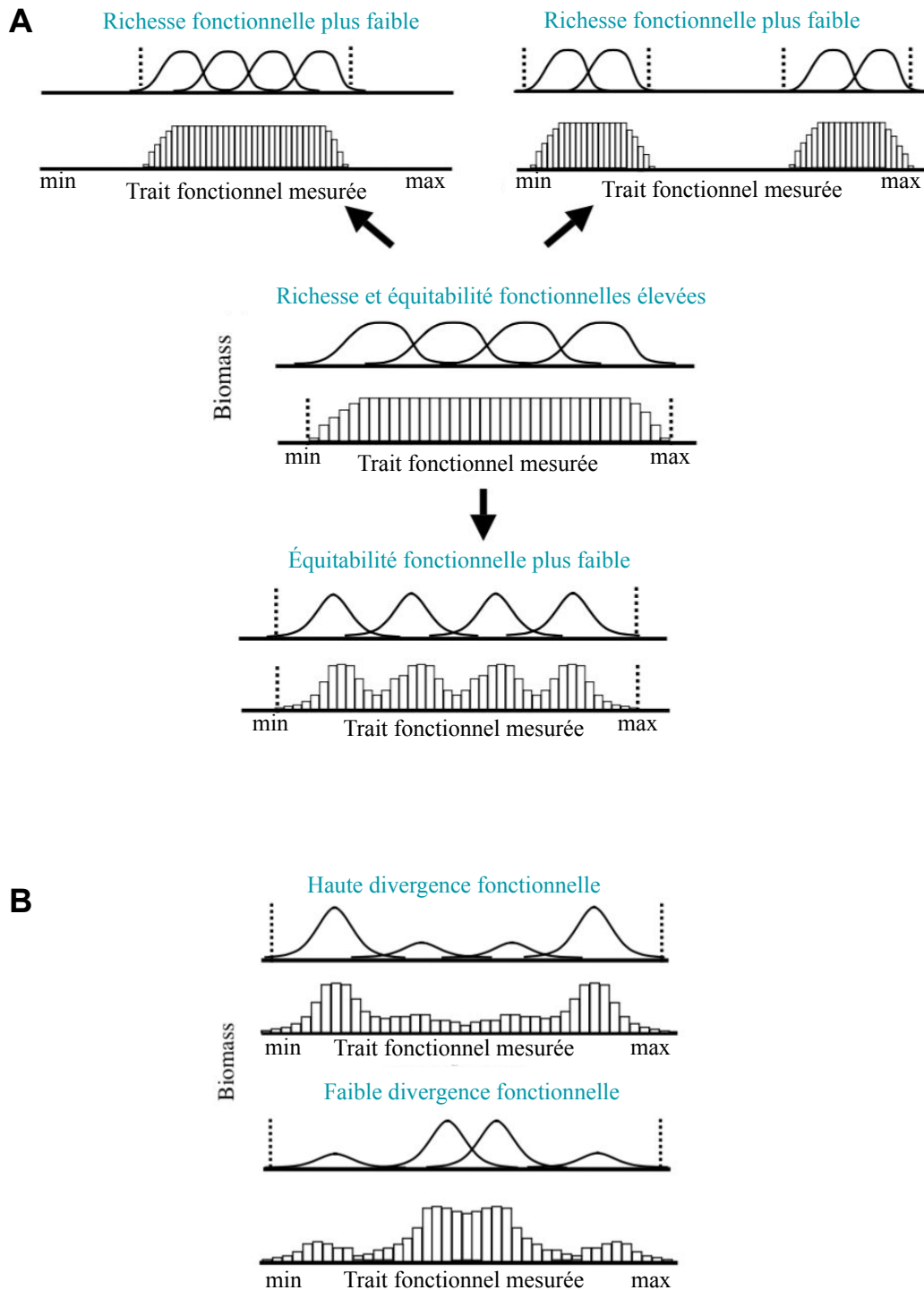


FIGURE 1.7 – (A) Variation de la richesse et de l'équitabilité fonctionnelle. (B) Variation de la divergence fonctionnelle. Les courbes de densité indique l'abondance des espèces suivant les valeurs de traits. Les histogrammes indiquent l'abondance des organismes suivant les valeurs de traits. Adapté depuis Mason et al. (2005). 17

La capacité à produire des formes dormantes est un trait d’histoire de vie qui influence la survie des organismes (Martini et al., 2020). Ce trait est courant chez les organismes phytoplanctoniques et leur permet de persister lors de conditions défavorables (Litchman, Klausmeier, 2008). Ce trait conditionne, par exemple, la biogéographie des communautés de diatomés arctiques (Fragoso et al., 2018). Parmi les traits morphologiques, la taille est largement étudiée. Elle intervient dans l’acquisition des ressources, la croissance, la reproduction et la survie (Litchman, Klausmeier, 2008; Litchman et al., 2013; Martini et al., 2020). La taille présente une structure latitudinale inversement liée à la température (Brun et al., 2016). D’autres traits morphologiques restent beaucoup moins étudiés. Identifier les répartitions spatiales des traits fonctionnels pourrait permettre de mieux comprendre les mécanismes sous-jacents à l’assemblage des communautés (Barton et al., 2013).

Ainsi, les études de diversité se focalisent traditionnellement sur des approches taxonomiques, phylogénétiques ou fonctionnelles. Plus récemment, la diversité génétique fait l’objet d’une attention croissante. Bien qu’elle ait été étudiée à l’échelle des communautés (Richter et al., 2020, Annexe A.2), la diversité génétique est généralement caractérisée au niveau intraspécifique. Les processus qui contribuent à cette diversité, appelés processus évolutifs (ou microévolutifs), sont : la mutation, la sélection, le flux de gènes et la dérive génétique. Il existe donc un parallèle étroit entre les processus évolutifs et les processus écologiques dont la distinction peut être difficile (Vellend, 2010; Hanson et al., 2012). La diversité génétique intraspécifique repose sur la présence de variants (*i.e.*, modification de bases dans le matériel génétique) et conditionne la capacité des organismes à s’adapter à leur environnement, ce qui est crucial dans un contexte de changement climatique. La diversité peut donc être étudiée à de multiples niveaux et tous ces niveaux sont importants pour comprendre la structure et le fonctionnement des écosystèmes.

1.2.3 Du gène à l’écosystème

La période actuelle, nommée Anthropocène, est caractérisée par l’impact direct ou indirect de l’activité humaine sur les écosystèmes. La biodiversité subit de nombreuses pressions dont les effets varient suivant les organismes (Chivers et al., 2017). Toutes les facettes de la biodiversité sont touchées, que ce soit taxonomique, phylogénétique, fonctionnel ou génétique. En effet, des réorganisations spatiales des communautés ont été observées (Hays et al., 2005; Beaugrand, 2005; Beaugrand et al., 2009; Brun et al., 2019) et des d’autres sont prédites en lien avec le changement climatique (Benedetti et al.,

2018). Il a également été établi que la diversité subit une érosion massive. Par exemple, une perte de 6% de la diversité intraspécifique a été mesurée (Leigh et al., 2019). Les modifications de la diversité peuvent altérer le fonctionnement des écosystèmes et par extension des services rendus par la nature aux sociétés humaines (Worm et al., 2006; Tilman et al., 2014; Gattuso et al., 2015; Pecl et al., 2017).

Un lien étroit existe en effet entre biodiversité et fonctionnement des écosystèmes. La biodiversité conditionne la productivité, la stabilité et la capacité à être envahie des écosystèmes (Tilman et al., 2014). Les modifications de la biodiversité sont un des facteurs majeurs du changement de fonctionnement des écosystèmes (Tilman et al., 2014). Traditionnellement, les changements de biodiversité sur le fonctionnement des écosystèmes sont étudiés au niveau interspécifique. Pourtant, il a récemment été montré que les changements de diversité intraspécifiques pouvaient avoir des conséquences de même ampleur, voire plus importantes (Des Roches et al., 2018). Or, par leur fonctionnement, les écosystèmes soutiennent de nombreux services dont bénéficient les sociétés humaines. Ils sont par exemple une source alimentaire, un réservoir de molécules d'intérêt médical et lieu de nombreuses activités de loisirs. L'altération de la diversité à ses multiples niveaux peut conduire à une modification du fonctionnement des écosystèmes et par extension des services rendus par la nature aux sociétés humaines (Worm et al., 2006; Tilman et al., 2014; Gattuso et al., 2015; Pecl et al., 2017). Pour comprendre le fonctionnement des écosystèmes et les menaces auxquelles ils font face, il est donc nécessaire de s'intéresser à la biodiversité dans son ensemble, du gène jusqu'à l'écosystème.

Comprendre le lien entre la diversité et le fonctionnement des écosystèmes est un réel défi. En effet, les écosystèmes sont des systèmes complexes hiérarchiques définis à différentes échelles spatiales et temporelles et entretenant des interactions étroites entre et au sein des différents niveaux biologiques et avec l'environnement (Figure 1.8, Stange et al., 2021). Au niveau génétique, les effets d'un variant au niveau de la communauté ou de l'écosystème dépendent de sa fréquence dans la population, de la fréquence des autres variants et des interactions avec les variants des espèces co-occurrentes (Stange et al., 2021). Cette diversité génétique, qui constitue la capacité biologique potentielle, conditionne le phénotype et donc la survie et le développement des organismes (Booy et al., 2000). À l'inverse, l'influence de l'environnement sur la diversité génétique est médiée par le phénotype (Stange et al., 2021). Les effets en cascade auxquels s'ajoutent les interactions (entre et au sein des niveaux biologiques) déterminent la structure et le fonctionnement de la communauté. C'est la capacité biologique réalisée. Comprendre comment

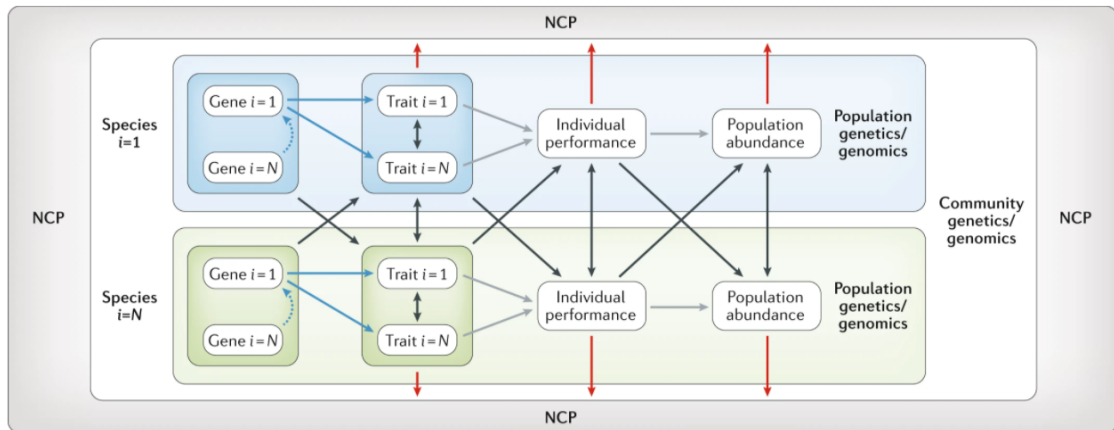


FIGURE 1.8 – Cascades d’effets directs et indirects, des gènes jusqu’aux écosystèmes, qui façonnent les services rendus par la Nature à l’Homme (Nature’s Contributions to People, NCP ; d’après Stange et al., 2021). Les flèches rouges représentent les influences directes sur les services rendus par la Nature à l’Homme. Les flèches grises représentent les influences indirectes. Chaque gène et ses variants (i.e. allèles) peut influencer un ou plusieurs traits (flèches bleues). Les gènes peuvent interagir entre eux pour influencer un trait (flèches bleues pointillées). Les flèches noires représentent les interactions qui ont lieu au sein de la communauté. Les gènes d’espèces différentes peuvent interagir par le biais des phénotypes. Les effets peuvent ensuite se répercuter au sein et entre les espèces.

le contenu génétique des communautés correspond à la composition taxonomique, au répertoire métabolique et à l’expression phénotypique est un défi. Cela permettrait de prédire les phénotypes des systèmes qui soutiennent le fonctionnement des écosystèmes à partir de l’information moléculaire. À cette fin, un échantillonnage systématique est nécessaire (Doney et al., 2004; Whitham et al., 2006; DeLong, 2009). Il a été rendu possible par le développement de nouvelles techniques de suivi et d’échantillonnage ces dernières décennies.

1.3 Étude des écosystèmes planctoniques à l’ère du *Big Data*

Actuellement, l’Homme exerce de nombreuses pressions directes et indirectes sur les écosystèmes océaniques. Des mutations de la structure, du fonctionnement et de la dynamique de ces écosystèmes ont déjà été observées et d’autres sont prédites. Cela pourra s’accompagner de modifications des services rendus par la Nature à l’Homme (e.g., Pecl et al., 2017; Stange et al., 2021). Afin de mieux comprendre les pressions subies, prédire les états futurs et mettre en place des solutions de conservation efficaces, une surveillance accrue des océans a été mise en place. Elle se traduit par une acquisition croissante de

données.

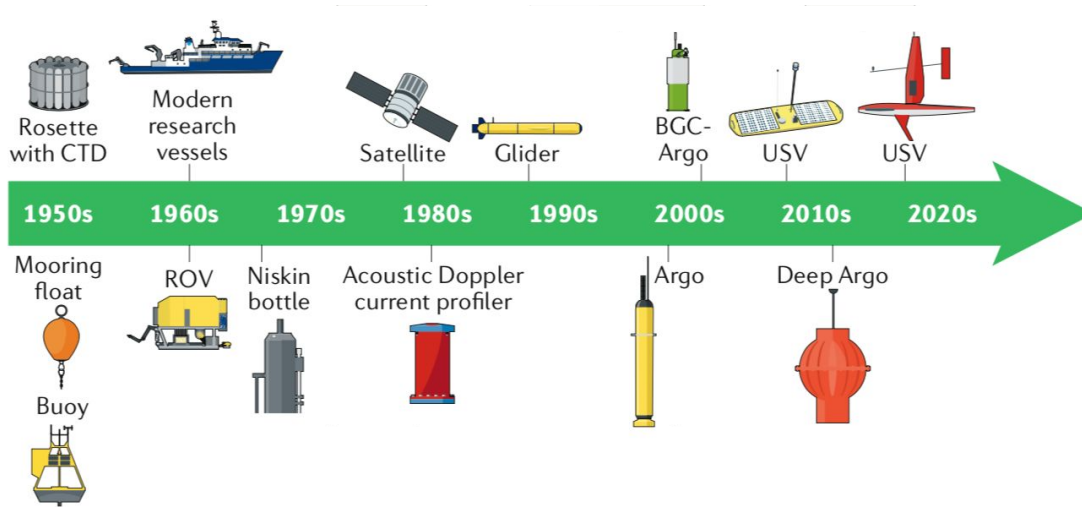


FIGURE 1.9 – Développement chronologique des plateformes d’observation océanographiques (modifié d’après Chai et al., 2020). BGC-Argo, Biogeochemical-Argo; CTD, conductivity, temperature depth; ROV, Remotely Operated underwater Vehicle; USV, Unmanned Surface Vehicle.

1.3.1 Déluge de données abiotiques...

La compréhension de l’environnement océanique évolue avec le développement des méthodes d’observation. Les données collectées par les plateformes d’observation historiques, telles que campagnes en mer ou les mouillages, ne conduisent pas à une compréhension complète de l’ensemble des processus complexes et multiéchelles de l’océan. Afin de combler ces lacunes, de nouvelles technologies ont été développées dès la fin du XX^{ème} siècle (Figure 1.9; Chai et al., 2020). En particulier, l’observation de l’océan a été largement stimulée par l’essor des satellites et des plateformes mobiles autonomes telles que les flotteurs et gliders. Ces nouvelles plateformes ont engendré une augmentation considérable de la quantité de données acquises sur le milieu océanique au début des années 2000 (Chai et al., 2020; Bates et al., 2018). Tout en complétant les méthodes d’observation de campagnes océanographiques qui fournissent des données de très bonne qualité, l’utilisation de ces nouvelles technologies a conduit à une augmentation de la couverture spatiale, verticale et temporelle de l’océan et donc d’une caractérisation plus fine de la variabilité environnementale à plusieurs échelles spatio-temporelles.

Le développement, depuis la fin des années 70, de la télédétection supportée par les satellites a largement contribué à la vision turbulente de l’océan. Les données altimé-

triques ont révélé la présence globale de tourbillons (Figure 1.10, Chelton et al., 2011) qui vont piéger et transporter les eaux en leur cœur (caractéristique de non linéarité, Chelton et al., 2011). Le temps de rétention dans les tourbillons varie suivant le type de tourbillon et la profondeur, ce qui pourrait avoir un impact sur les communautés planctoniques piégées (Condie, Condie, 2016). Les satellites permettent une bonne connaissance de la couche de surface mais sont limités pour la connaissance de la colonne d'eau. D'autres instruments ont été développés et permettent d'accéder à ces informations verticales (Figure 1.11, Whitt et al., 2020).

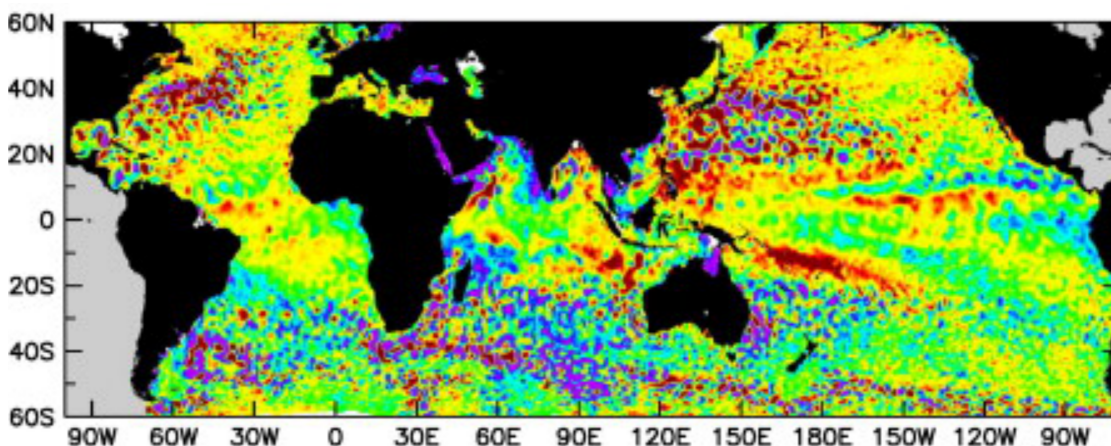


FIGURE 1.10 – Carte d'altimétrie globale (28 août 1996) construite à partir des données TOPEX/Poseidon (T/P) et des données fusionnées T/P et ERS-1 dans la série de référence AVISO. De nombreux tourbillons mésoéchelles peuvent être identifiés et suivis. Extrait de Chelton et al. (2011)

Parmi les plateformes autonomes, plusieurs réalisent des profils verticaux comme les flotteurs Argo, les gliders et les autres plateformes autonomes (*e.g.*, les observatoires câblés). Elles font partie de programmes de larges envergures (Rudnick, 2016; Roemmich et al., 2019; Testor et al., 2019; Claustre et al., 2020). Elles génèrent de nombreuses données mais sont limitées à certaines variables d'intérêt (*e.g.*, température, salinité). La mesure de variables biologiques reste marginale et souvent ne prend pas en compte la diversité (*e.g.*, chlorophylle-a).

1.3.2 ... auquel s'additionne l'avalanche de données biologiques

Les technologies précédentes (*i.e.*, satellites, flotteurs et gliders) permettent, en plus de l'environnement physico-chimique, de suivre les communautés planctoniques. Dans l'océan du large, la couleur de l'eau est principalement dépendante du phytoplancton

1.3. Étude des écosystèmes planctoniques à l'ère du *Big Data*

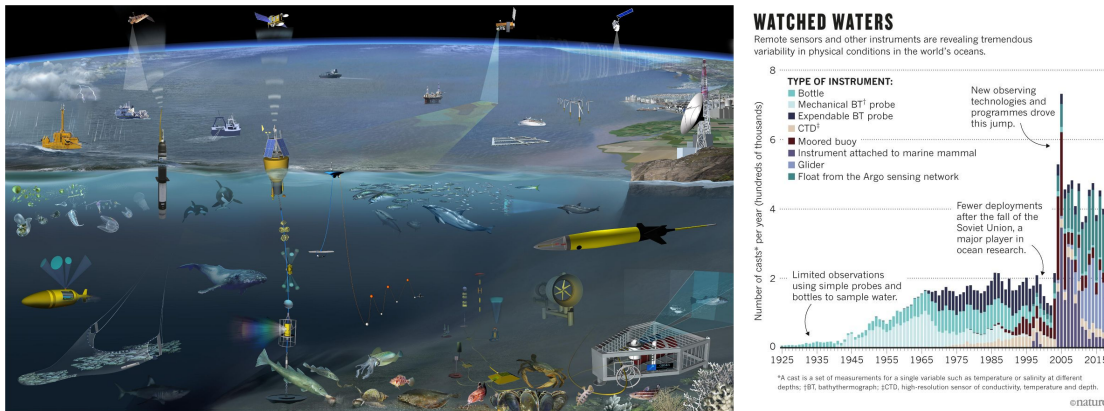


FIGURE 1.11 – Observation de l'océan. Sur la gauche, les nombreuses plateformes autonomes et de télédétection qui composent le système d'observation des océans sont représentées (depuis Whitt et al., 2020). Différentes technologies sont représentées dont, par exemple, les satellites, les flotteurs, les bouées d'amarrage, les flotteurs Argo. Sur la droite, l'histogramme représente l'évolution temporelle de l'instrumentation utilisée pour la surveillance des océans (depuis Bates et al., 2018).

dont le contenu pigmentaire modifie les radiations du spectre lumineux. Les données associées à la couleur de l'eau permettent d'inférer les concentrations en chlorophylle-a, proxy de la biomasse phytoplanctonique (McClain, 2009). En effet, ces données ont également permis de montrer l'importance de la turbulence dans la distribution disparate (i.e. en patch ou en filament, Figure 1.12) du plancton (d'Ovidio et al., 2010; Lévy et al., 2018). Elles ont aussi permis d'identifier les zones frontales comme des régions de fortes productivités en raison des processus de mélange qui s'y produisent (Allen et al., 2005; Lévy et al., 2018). Les distributions horizontale et verticale des communautés phytoplanctoniques peuvent également être estimées (McClain, 2009; Groom et al., 2019). Les Maximum de chlorophylle profonds sont détectables par analyse des profils BCG-Argo (Mignot et al., 2014; Uitz et al., 2010; Ricour et al., 2021). Les méthodes acoustiques permettent d'étudier le zooplancton et peuvent être acquises depuis des gliders (Chai et al., 2020). Cependant, l'objectif de cette partie n'est pas de fournir une vision exhaustive des différentes méthodes d'étude des communautés planctoniques. Deux types de données ayant subi un développement important ces dernières décennies et permettant de fournir des informations complémentaires sur les communautés planctoniques sont ciblées : les données de séquençage et les données d'imagerie.

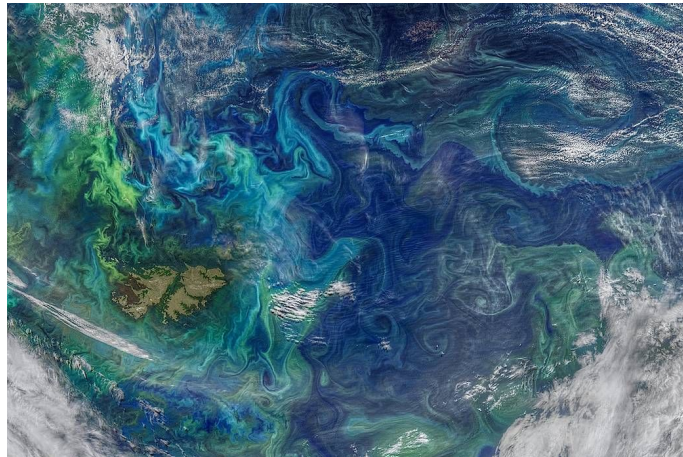


FIGURE 1.12 – Image satellite de la couleur de l’océan. La dynamique de sous-mésoéchelle incluant les tourbillons et les filaments est à l’origine de la distribution hétérogène du phytoplancton (© NASA)

Données de séquençage

La structure en double hélice de l’ADN a été découverte en 1953 par Jim Watson et Francis Crick. Chaque brin est constitué d’un enchaînement de nucléotides (ou bases) dont la séquence code l’information génétique. Il a rapidement été d’intérêt médical de déchiffrer le génome (ensemble des gènes) de l’Homme afin de relier profil génétique et maladie. Le séquençage est la technique permettant de déterminer la succession de nucléotides constituant un fragment d’ADN. Les premières techniques de séquençage ont été décrites en 1977 avec deux techniques développées en parallèle : la méthode chimique de Maxam-Gilbert (Maxam, Gilbert, 1977) et la méthode enzymatique de Sanger (Sanger et al., 1977), bien que les premières descriptions de séquençage aient été faites en 1973. La méthode de Sanger, aujourd’hui encore utilisée, a rapidement été préférée à celle de Maxam-Gilbert. Elle a été le support d’un progrès considérable et a permis de séquencer le premier génome humain en 2003 dans le cadre du Human Genome Project initié 13 ans plus tôt (Lander et al., 2001). Bien que précise, c’est une méthode longue et coûteuse (coût total du projet : \$2.7 milliards). Son application est donc limitée pour les grands génomes. De nouvelles méthodes de séquençage ont été développées ces 20 dernières années, conduisant à un séquençage de plus en plus massif du vivant.

En 2005, le séquençage à haut débit (HTS : *High Throughput Sequencing* ou NGS : *Next Generation Sequencing*) a été mis au point (Goodwin et al., 2016). Cette technologie repose sur un principe de séquençage massif parallélisé permettant d’augmenter le débit

(*i.e.*, la quantité de bases par jour et par machine). L'augmentation du débit est accompagnée d'une diminution du coût de séquençage (Figure 1.13). En 2014, un génome humain a été séquençé en quelques jours pour \$1000. Outre ce séquençage de génome ou transcriptome complet, il est possible de ne séquençer qu'une région spécifique. Ce séquençage ciblé, ou séquençage des amplicons, est à la base de l'approche par barcoding. Ces courtes séquences permettent de déterminer la taxonomie des organismes et d'étudier leurs relations phylogénétiques (Hebert et al., 2003). Quelque soit l'approche (génomique, transcriptomique ou par barcode), le séquençage et en particulier les technologies HTS ont permis un véritable essor de la biologie moléculaire. Ces technologies ont été appliquées à l'étude de l'environnement par le séquençage des communautés qui composent les milieux.

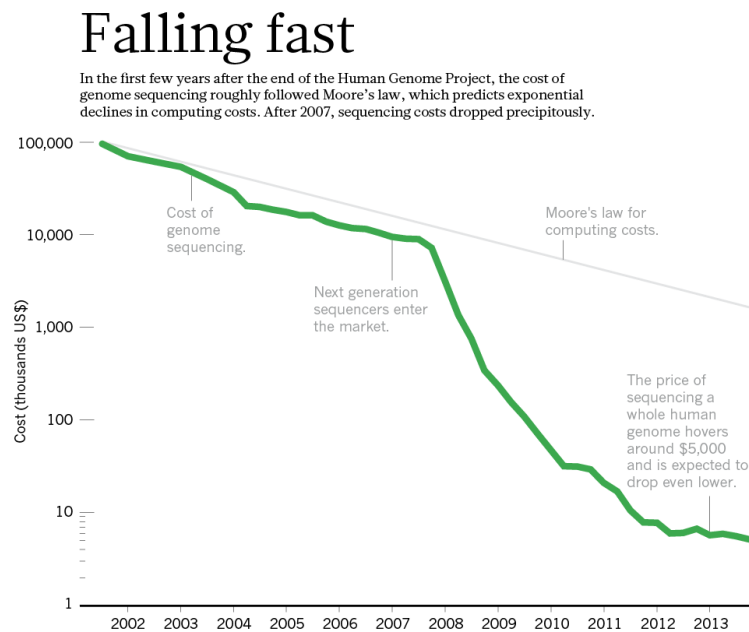


FIGURE 1.13 – Évolution du coût de séquençage. La loi de Moore prédit, de manière empirique, que le rapport en puissance d'un ordinateur et son prix double tous les 18 mois. Ainsi, le coût des machines diminue tandis que leur capacité augmente. Les technologies de séquençage étant liées aux capacités informatiques, la loi de Moore était utilisée pour prédire l'évolution du coût de séquençage. Le coût réel de séquençage s'éloigne, depuis 2008, de cette loi empirique (depuis Check Hayden, 2014).

Le séquençage à haut débit est le support des méthodes dites environnementales (ou méta-omiques). La méta-omique consiste au séquençage de l'ensemble du matériel génétique d'une communauté. Suivant le matériel séquençé, il s'agira de métagénomique

(ADN), métatranscriptomique (ARN) ou métabarcoding (séquences marqueur) qui informent respectivement sur les gènes présents, les gènes exprimés et les taxons présents (à noter qu'il existe d'autres données méta-omiques, *e.g.*, métaprotéomiques, qui ne sont pas abordées ici). Les communautés peuvent être issues de milieux variés. Dans le milieu terrestre, les études ont par exemple portées sur les interactions plantes-pollinisateurs via la collecte des pollens présents sur les insectes (*e.g.*, Pornon et al., 2017), la caractérisation des communautés du sol (*e.g.*, Arthropodes, Oliverio et al., 2018; Champignons, Schmidt et al., 2013) ou celles de l'air (de Groot et al., 2021), ou encore la structure spatiale des microbiomes en zone urbaine (Danko et al., 2021, Annexe A.3). La première campagne de séquençage des communautés planctoniques de milieu océanique a débuté en 2003 (*Global Ocean Sampling*, Venter et al., 2004) et reposait sur la méthode Sanger. Depuis les techniques de séquençage HTS ont été le support d'autres expéditions de grande échelle (*e.g.*, *Tara Oceans*, Karsenti et al., 2011; Malaspina, Duarte, 2015; *Tara Pacific*, Gorsky et al., 2019), ainsi que des suivis temporels (*e.g.*, zooplancton à BATS, Blanco-Bercial, 2020; *Ocean Sampling Day*, Kopf et al., 2015), conduisant à une amélioration des connaissances sur les communautés planctoniques.

Les données de métabarcoding ont révélé qu'une grande part de la diversité eucaryote reste inconnue (De Vargas et al., 2015). Les données métagénomiques et métatranscriptomiques ont permis de reconstruire les gènes eucaryotes présents dans l'environnement et d'estimer leur abondance et expression (Carradec et al., 2018). L'expression des gènes peut permettre d'identifier les fonctions écologiques d'organismes non cultivables (Carradec et al., 2018; Labarre et al., 2020) et d'étudier les processus d'adaptation (Salazar et al., 2019). Des processus généraux ont également été mis en évidence dans la structure des communautés. Les espèces rares présentent un renouvellement spatial suivant une loi puissance générale (*i.e.*, pas de variation de l'exposant), ce qui suggère un mécanisme unique large échelle (Ser-Giacomi et al., 2018). Des facteurs environnementaux ont également été identifiés. La diminution de la diversité vers les hautes latitudes pourrait être liée aux modifications de température et de disponibilités en nutriments (Ibarbalz et al., 2019). Ces facteurs pourraient influencer les capacités adaptatives des communautés (Salazar et al., 2019). Dans le compartiment microbien, il a été mis en évidence que des processus écologiques différents structurent les communautés procaryotes et picoeucaryotes (Logares et al., 2020). Enfin, la dynamique de l'océan a aussi été identifiée comme forçant la structure des communautés que ce soit à large échelle (Richter et al., 2020) ou à plus petite échelle, comme dans les tourbillons rétentifs dont le coeur crée des conditions différentes de celle du milieu dans lequel ils sont inclus et conduisent à

des communautés spécifiques (Villar et al., 2015). Enfin, plusieurs études ont montré une réponse différente des communautés suivant la taille des organismes qui les composent (*e.g.*, Ibarbalz et al., 2019; Richter et al., 2020; Sommeria-Klein et al., 2020). La taille, et par extension la morphologie des communautés et sa relation avec la structure des communautés peuvent être plus largement explorées à partir des données d'imagerie.

Données d'imagerie

Historiquement, les méthodes d'observation des organismes planctoniques reposaient principalement sur de la microscopie (*e.g.*, classique, inversée, à épifluorescence). Elles présentent l'intérêt d'être très précises dans la quantification et l'identification mais demandent beaucoup de temps. De nouvelles technologies automatisées ont progressivement été développées. Elles ont largement diminué le temps d'analyse des organismes et seront par la suite qualifiées de technologies d'imagerie à haut débit. Elles permettent d'acquérir des images de tous les organismes d'un échantillon soit après collecte par filet (*e.g.*, Flowcam, Sieracki et al., 1998; Zooscan, Gorsky et al., 2010), soit directement in situ (*e.g.*, *Underwater Vision Profiler*, Picheral et al., 2010; Zooglider, Ohman et al., 2019). Plusieurs de ces technologies ont été utilisées lors d'expéditions océanographiques (*e.g.*, Pesant et al., 2015; Gorsky et al., 2019). D'autres peuvent être combinées avec des plateformes autonomes (*e.g.*, UVP sur glider, Chai et al., 2020). Ces différents outils permettent une acquisition sur la composition et les phénotypes des communautés sans précédent et à large échelle spatio-temporelle.

L'ensemble des technologies d'imagerie permet d'imager individuellement (possiblement par post-traitement informatique) tous les organismes échantillonnés d'une communauté. Suivant les outils utilisés, les images peuvent présenter des caractéristiques différentes en termes de coloration (niveau de gris ou codage RGB) et de résolution. Bien que chaque technologie d'imagerie à haut débit soit spécifique d'une gamme de taille, l'ensemble des outils permet de couvrir la quasi-totalité du spectre de taille planctonique (Figure 1.14).

L'utilisation de l'imagerie pour caractériser les communautés planctoniques présente plusieurs avantages. À chaque organisme, une assignation taxonomique peut être attribuée (de niveau résolutif limité) permettant l'exploration de la diversité des communautés (Ibarbalz et al., 2019). Contrairement aux données méta-omiques de nature compositionnelle (Gloor et al., 2017), les méthodes d'imagerie permettent d'avoir une vision quantitative des communautés (Lombard et al., 2019). Ainsi, abondance et biomasse peuvent être estimées. Cela a par exemple permis d'identifier que les Rhizaria constituent une part importante de la biomasse protiste, ce qui peut avoir une influence sur

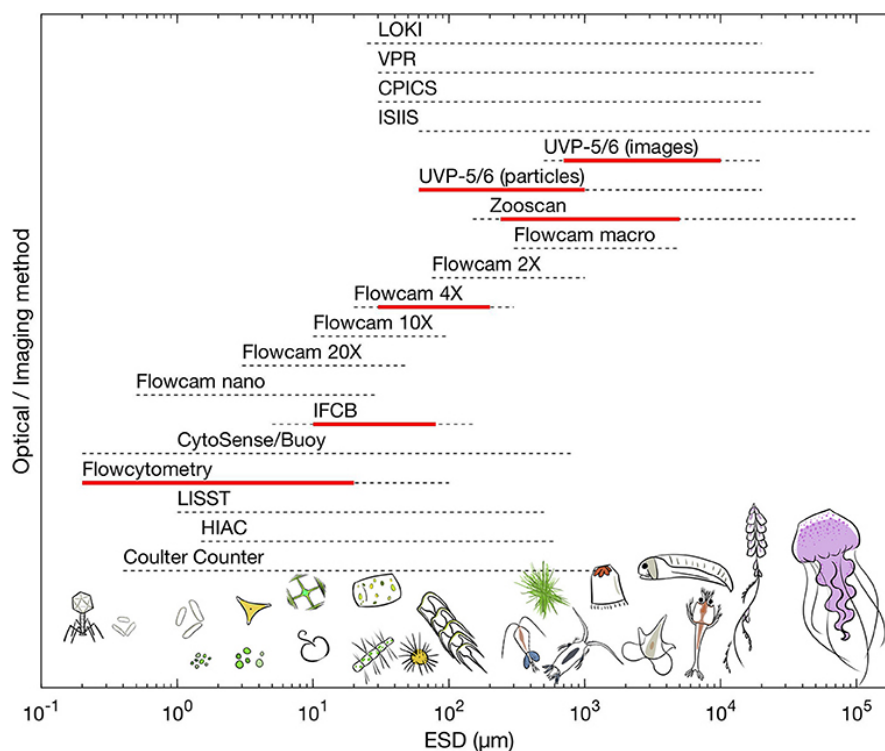


FIGURE 1.14 – Échantillonnage du plancton par les méthodes d'optique et d'imagerie. 19 méthodes sont représentées. Pour chacune, la gamme de taille (en diamètre sphérique équivalent, ESD) pouvant être échantillonnée est indiquée (pointillés : gamme totale, rouge : gamme pour obtenir des informations quantitatives). L'ensemble de ces méthodes permet de couvrir la quasi-totalité des organismes planctoniques (Lombard et al., 2019)

les flux de carbone (Biard et al., 2017). À partir des particules non vivantes imagées, il a été suggéré que l'intensité de l'export de carbone est en relation étroite avec la structure des communautés planctoniques (Guidi et al., 2016). Les images permettent d'extraire des informations sur les traits fonctionnels (*sensus* Violle et al., 2007). La caractérisation morphologique des particules de neige marine, basée sur des traits autres que la taille, pourrait permettre de mieux quantifier l'efficacité de la pompe à carbone biologique (Trudnowska et al., 2021). Par exemple, une autre analyse basée uniquement sur la morphologie des organismes a permis de révéler l'influence de la dynamique de la glace de mer sur la structure spatiale des communautés de Copépodes (Vilgrain et al., 2021). Cette méthodologie s'affranchit du besoin d'assignation taxonomique ou fonctionnelle et donc de la dépendance aux bases de données de référence ou aux connaissances d'expert (nécessaire, par exemple, pour inférer des fonctions aux données issues de séquençage, *e.g.*, Faure et al., 2019).

Les données acquises par prélèvement in situ, télédétection, plateformes autonomes, séquençage ou imagerie, tendent à couvrir de plus en plus l'intégralité du milieu océanique et de manière plus résolutive. Elles pourraient donc permettre de décrypter la complexité des écosystèmes qui impliquent des processus s'opérant à différentes échelles spatio-temporelles. Un des défis consiste donc en l'extraction de connaissance à partir de ces vastes quantités de données hétérogènes.

1.3.3 Vers une compréhension holistique

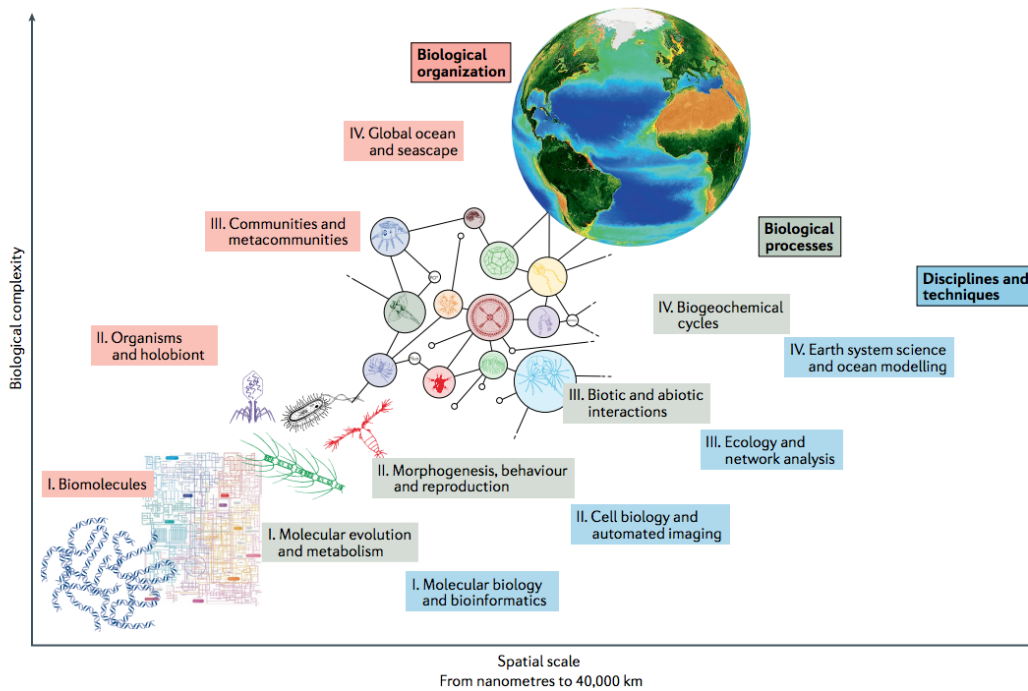


FIGURE 1.15 – Biologie des écosystèmes et analyses intégratives des océans mondiaux. Les données massives collectées permettent d'accéder aux différents niveaux biologiques (depuis Sunagawa et al., 2020)

Les écosystèmes sont des systèmes complexes dont les constituants sont en interaction étroite les uns avec les autres. Les propriétés émergentes de ces écosystèmes ne sont pas le reflet des propriétés individuelles de ses constituants. Pour comprendre leur structure et leur fonctionnement, il est donc nécessaire de les étudier dans leur globalité, en intégrant l'ensemble des informations qui les caractérisent : la diversité génétique, morphologique, fonctionnelle dans son contexte environnemental (Figure 1.15, Karsenti et al., 2011; Schuwirth et al., 2019; Sunagawa et al., 2020). Les données acquises permettent de caractériser ces différentes facettes des écosystèmes mais présentent de nombreux défis

quant à l'extraction de l'information. Aux protocoles d'analyse standards, qui ne sont plus applicables, vont se substituer des approches dirigées par les données où l'analyse et l'interprétation vont permettre d'établir de nouvelles hypothèses. L'extraction de l'information, basée sur les données et leur exploration, mobilise à la fois des approches empiriques, théoriques et informatiques (Hey et al., 2009). C'est la science à forte intensité de données (ou *data-intensive science*), qui génère une réelle mutation dans la recherche afin d'exploiter le potentiel de ces jeux de données volumineux (Karsenti et al., 2011; Hampton et al., 2013). Toutefois les études systémiques multiéchelles qui sont rendues possibles pourraient mener à des avancées considérables sur notre compréhension holistique du vivant (Karsenti et al., 2011; Sunagawa et al., 2020).

1.4 Problématique et objectifs

Dans l'introduction, l'écosystème océanique a été présenté en mettant en évidence son hétérogénéité biogéochimique et hydrodynamique. Parmi les organismes vivants dans ce milieu, le plancton eucaryote, primordial pour le fonctionnement des écosystèmes, a été ciblé. Après avoir introduit brièvement ses caractéristiques, les différents forçages qui s'exercent sur lui et qui conditionnent les assemblages de communautés ont été présentés. Dans une dernière partie, il a été montré que le milieu océanique fait face à une surveillance croissante ayant abouti à une explosion du nombre de données et de type de données collectés. Ces données pourraient permettre de mieux comprendre la structuration des écosystèmes et les mécanismes à son origine, qui sont encore mal caractérisés.

L'objectif de cette thèse de doctorat est donc de réaliser une étude détaillée de la structure de l'écosystème planctonique en s'intéressant à la question suivante :

Comment est structuré l'écosystème planctonique, des populations aux communautés, et quels sont les facteurs qui façonnent son organisation ?

Cette question peut être évaluée pour de multiples échelles biologiques et spatiales grâce à l'utilisation des données à haut débit de séquençage et d'imagerie, ainsi qu'aux données environnementales biogéochimiques et hydrodynamiques. Ainsi, trois questions ont été déclinées de cette problématique générale. Elles ont permis d'étudier la structure de l'écosystème planctonique de l'espèce jusqu'à la communauté, en englobant la diversité génétique, la diversité fonctionnelle et la diversité taxonomique. Les différents axes fournissent des pistes sur les contraintes agissant sur les organismes et donc sur les sources potentielles de contrôle des écosystèmes étudiés (*i.e.*, filtre de dispersion, environnemen-

tal ou d'interaction).

1. *Les données méta-omiques peuvent-elles servir de support à l'étude de la diversité génomique de l'ensemble des espèces de l'écosystème ?*

Dans un premier temps, j'ai voulu voir si les données métagénomiques pouvaient permettre de mieux comprendre la structuration de la diversité à l'échelle intraspécifique.

2. *Les données à haut débit permettent-elles de faire un lien entre génotype et phénotype des communautés ?*

Le phénotype résulte de l'expression de la diversité génétique. Ces deux niveaux d'organisation sont en étroite interaction. Dans une deuxième partie, j'ai voulu caractériser le lien existant entre diversité génétique et diversité fonctionnelle à l'échelle des communautés.

3. *Les données à haut débit peuvent-elles contribuer à une vision intégrée des écosystèmes planctoniques ?*

La structure et le fonctionnement des écosystèmes résultent des multiples niveaux de structuration qui interagissent ensemble. Cependant ces différents niveaux de structuration sont souvent étudiés de manière découplée. Dans une dernière partie, j'ai étudié l'écosystème planctonique dans son ensemble, en analysant conjointement les différents niveaux de structuration de l'écosystème afin de caractériser la structure globale qui en résulte.

Pour répondre à ces questions, le travail a été structuré en trois cas d'étude correspondant aux chapitres décrits ci-dessous (Table 1.1).

Vue d'ensemble des chapitres

- **(Méta)Génomique des populations de protistes**

Les variations génétiques intraspécifiques sont structurées dans le temps et dans l'espace. Les approches à l'échelle du génome permettent de les caractériser finement. Dans le milieu océanique, la répartition des variations génétiques intraspécifiques est biaisée vers les macro-organismes à intérêt commercial ou ciblé par des mesures de protection. Celle des protistes est bien moins connue. Les données métagénomiques peuvent être une opportunité pour étudier la diversité génétique intraspécifique à l'échelle du génome de ces organismes. Cette approche pourrait

permettre de comprendre les processus qui s'appliquent sur toute espèce de l'environnement et donc quantifier l'importance relative des processus évolutifs et/ou écologiques.

- **Phénotype transparent dans les communautés planctoniques**

Des caractéristiques individuelles des organismes émergent des propriétés mesurables à l'échelle des communautés. C'est, par exemple, le cas de la transparence qui forme la deuxième principale structure morphologique retrouvée au sein des communautés planctoniques. Cette transparence planctonique est peu étudiée à l'échelle des communautés. Elle conditionne pourtant la survie des organismes, notamment en influençant les relations de proies-prédateurs. Le but de cette étude est de caractériser la transparence des communautés planctoniques à l'échelle globale, afin de la mettre en lien avec la diversité génétique, l'environnement et les interactions biotiques.

- **Structure des communautés planctoniques dans l'Atlantique Nord**

La structure des communautés présente des variations spatio-temporelles qui sont régies par de multiples processus dont l'influence et l'importance relative est difficile à caractériser. Ces processus sont imbriqués et se produisent de manière simultanée à de multiples échelles. Différentes sources de données peuvent donc permettre de caractériser les structures qui en résultent. Dans cette partie, la structure globale de l'écosystème planctonique a été évaluée le long d'un transect dans l'océan Atlantique Nord à partir de données environnementales, de metabarcoding et d'imagerie. Identifier la redondance ou la complémentarité des informations peut permettre de mieux comprendre les facteurs qui structurent les communautés et les échelles spatio-temporelles auxquelles ils prennent place.

Axe	Diversité génomique	Lien entre phénotype et génomique	Vision intégrée de l'écosystème
Cas d'étude	Métagénomique des populations (Chap.3)	Phénotype et Expression (Chap.4)	Structure des communautés (Chap.5)
Campagne	Tara Oceans	Tara Oceans	Tara Pacific
Région	Méditerranée	Océan global	Atlantique Nord
Organismes	Protistes	Multicellulaires	Protistes jusqu'au multicellulaires
Biologie	Métagénomique	Méatranscriptomique Réseau de co-occurrence Imagerie	Metabarcoding Imagerie
Environnement	MedAtlas Mean connection time Probabilité de connexion	Tara Oceans	Tara Pacific Diagnostics lagrangiens
Analyses	Analyses en composantes principales Modèles linéaires	Analyses en composantes principales Analyse du réseau de corrélations pondérées	Analyses en composantes principales Régression par les moindres carrés partiels Analyse factorielle multiple

TABLE 1.1 – Vue d'ensemble des chapitres. Les différentes sources de données et principales méthodes utilisées sont récapitulées.

Matériel et Méthodes général

2.1 Expéditions *Tara Oceans Polar Circle* et *Tara Pacific*

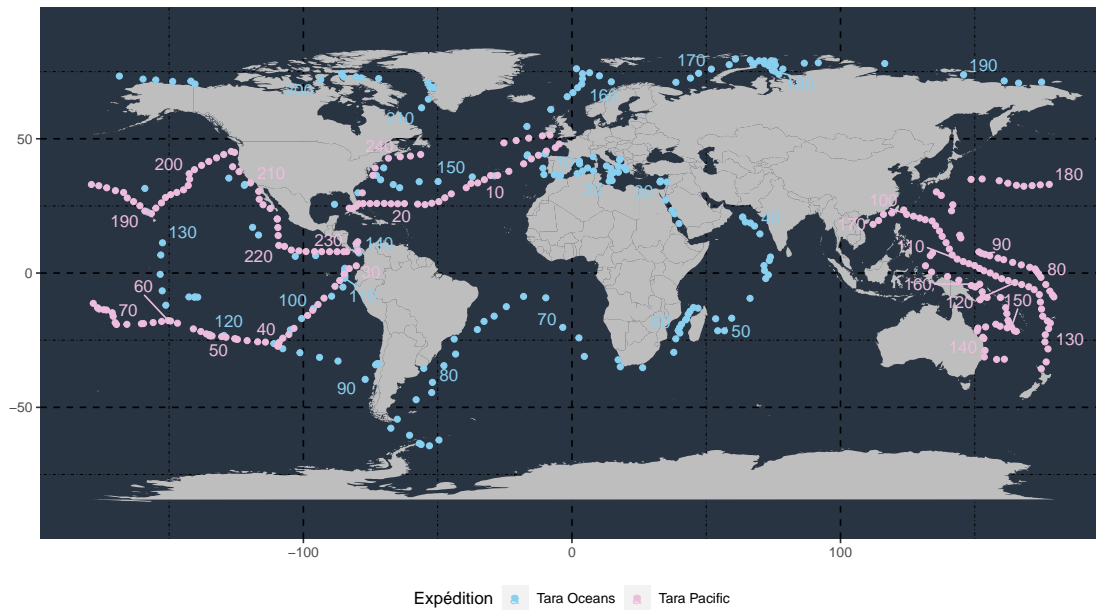


FIGURE 2.1 – Expéditions *Tara Oceans* et *Tara Pacific*. Itinéraires des échantillonnages et distances (en km) entre deux stations successives. La couleur des points indique l'expédition (rose : *Tara Oceans*, bleu : *Tara Pacific*)

Au cours de l'expédition *Tara Oceans* (comprenant *Tara Oceans Polar Circle*), qui s'est

déroulée de septembre 2009 à décembre 2013, 210 stations réparties dans l’océan Atlantique, l’océan Pacifique, l’océan Indien, l’océan Arctique, la mer Méditerranée et la mer Rouge ont été échantillonnées (157 + 53 stations arctiques, Figure 2.1). *Tara Oceans* avait pour but de permettre l’étude du plancton dans le milieu océanique en lien avec l’environnement, ce qui est un enjeu crucial dans le contexte de changement climatique. Plus récemment, l’expédition *Tara Pacific* a eu lieu. Elle avait pour cible principale les récifs coralliens. Les coraux sont des composants clés des écosystèmes de part leur rôle de nurseries et frayères. Mais étant des organismes calcaires fixés à leur substrat, ils sont menacés par l’acidification des océans qui fragilise leur squelette et par le changement climatique puisqu’ils ne peuvent migrer en phase avec les modifications de leur environnement. Bien que ces coraux soient principalement situés dans des zones insulaires, un échantillonnage de pleine mer a également été effectué. Au total, 249 stations de pleine mer ont été échantillonnées d’avril 2016 à novembre 2018 (Figure 2.1). La couverture spatiale associée à l’échantillonnage océanique lors de l’expédition *Tara Pacific* est plus restreinte. Seuls le bassin Atlantique Nord et l’océan Pacifique ont été échantillonnés lors de l’expédition. Cependant, la résolution de l’échantillonnage était plus fine que lors de l’expédition *Tara Oceans*, avec une distance moyenne entre deux stations successives (Figure 2.2) passant de 501 (*Tara Oceans*) à 353 km (*Tara Pacific*). Ainsi, il est supposé que des processus de plus fine échelle pourraient être détectés avec les jeux de données associés à cette deuxième expédition. Tandis que l’échantillonnage était exclusivement de surface lors de *Tara Pacific*, plusieurs profondeurs ont été échantillonnées pendant *Tara Oceans* dont la surface, la profondeur du *Deep Chlorophyll Maximum* et le moins connu mésopélagique. En outre, les protocoles appliqués lors de l’expédition *Tara Oceans* ont été mis en œuvre pour *Tara Pacific* (Gorsky et al., 2019).

2.2 Les données biologiques

Les jeux de données biologiques utilisés étaient exclusivement issus des expéditions *Tara Oceans* et *Tara Pacific* (Alberti et al., 2017; Gorsky et al., 2019). Ce travail s’est concentré sur la partie eucaryote de la diversité biologique.

2.2.1 Données issues de séquençage à haut débit

Échantillonnage par fractions de taille

Au cours de ces expéditions, les communautés planctoniques ont été échantillonnées par fractions de taille. Les données eucaryotes *Tara Oceans* couvrent une gamme de taille

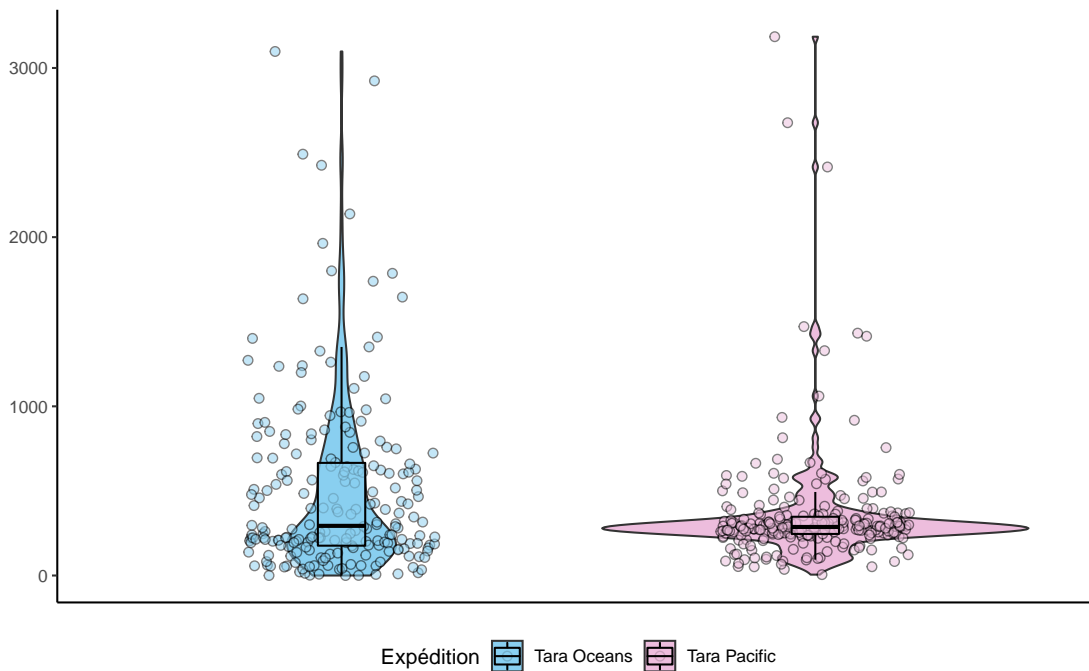


FIGURE 2.2 – Distance entre deux stations successives lors des expéditions *Tara Oceans* (bleu) et *Tara Pacific* (rose).

allant 0,8 à 2000 μm répartie dans quatre fractions de taille (0,8-5 μm ou 0,8-3 μm pour les stations arctiques, 5-20 μm ou 3-20 μm pour les stations arctiques, 20-180 μm et 180-2000 μm) et les échantillons *Tara Pacific* couvrent la gamme de taille 0,2-2000 μm distribuée dans quatre fractions de taille (0,2-3 μm , 3-20 μm , 20-2000 μm et 300-2000 μm).

Jeux de données de Metabarcoding

Des données de metabarcoding ont été générées. Au total, 405002 OTUs eucaryotes constituent le jeu de données de metabarcoding *Tara Oceans* et 147016 OTUs eucaryotes constituent le jeu de données *Tara Pacific*. À noter que tous les échantillons *Tara Pacific* n'ont pas encore été traités et donc que les valeurs indiquées ici ne sont pas définitives. La méthodologie appliquée pour générer les matrices d'abondance des OTUs aux stations *Tara Oceans* et *Tara Pacific* est semblable (De Vargas et al., 2015; Gorsky et al., 2019). L'ADN contenu dans les échantillons prélevés par fractions de taille a été extrait. La boucle hypervariable V9 du gène ARNr 18S a été amplifiée par PCR, puis séquencé par la technologie de séquençage haut débit Illumina. Les séquences d'ADNr

18S V9 obtenues sont appelées métabarcodes. Afin de limiter l'impact des erreurs de séquençage, les métabarcodes ont été filtrés pour ne conserver que ceux de bonne confiance (*i.e.*, ≥ 3 exemplaires au total et retrouvés dans au moins deux échantillons). Les métabarcodes ont été regroupés en OTUs par une approche Swarm (Mahé et al., 2015). Cette méthode présente l'avantage de former des OTUs robustes qui ne dépendent ni des métabarcodes d'entrée, ni d'une paramétrisation de regroupement arbitraire (Mahé et al., 2015). L'assignation taxonomique des OTUs a été attribuée par comparaison des séquences avec des barcodes de référence. La base de données de référence V9_PR2 pour *Tara Oceans* et EukRibo pour *Tara Pacific*. Les bases de références sont formées par les barcodes 18S de la base de données PR2 avec modification de la structure initiale et ajout des données 16S de la base de données Silvia. En effet, la région V9 du gène de l'ARNr 18S est une séquence marqueur des eucaryotes et présente des variabilités permettant de discriminer les taxons entre eux. Cependant elle amplifie également des séquences de certaines bactéries (ex. Cyanobactéries) qui ont donc pu être retirées des jeux de données eucaryotes sur la base des assignations taxonomiques. Dans chaque échantillon, les abondances des OTUs ont été estimées à partir du nombre d'exemplaires de métabarcodes. La méthodologie complète pour générer les jeux de données est détaillée dans De Vargas et al. (2015).

Jeux de données métagénomiques et métatranscriptomiques

Seules les données de métagénomique et de métatranscriptomique associées à l'expédition *Tara Oceans* ont été utilisées. Depuis les échantillons collectés par fraction de taille, ADN et ARN des échantillons ont été extraits. Les bibliothèques associées ont été créées, puis séquencées par la technologie de séquençage haut débit Illumina. Les séquences obtenues à l'issue du séquençage sont appelées reads (ou lectures). Les échantillons métagénomiques contenant les reads bruts ont été utilisés pour l'étude du chapitre 3 (Table 1.1).

À partir des reads, il est possible de reconstruire les gènes dont ils sont issus. Puis, par comparaison avec les bases de données de référence, il est possible d'attribuer à ces gènes une assignation taxonomique ou fonctionnelle. Ainsi, à partir d'une sous partie des échantillons de métatranscriptomique *Tara Oceans*, le catalogue de gènes eucaryotes Marine Atlas of *Tara Oceans* Unigenes (MATOU ; Carradec et al., 2018) a été constitué. Brièvement, les reads métatranscriptomiques ont été assemblés en séquences plus longues, appelées contigs, afin de reconstruire les transcrits. Ces contigs ont été regroupés sur la base de leur ressemblance nucléique. Pour chaque groupe, le plus long contig a été

retenu comme représentant et forme un unigène. Pour chaque échantillon, l'abondance et l'expression de chaque unigène a été estimée par alignement des reads métagénomiques et métatranscriptomiques respectivement. Ce sont ces matrices d'abondance qui ont été utilisées dans le chapitre 4 (Table 1.1).

2.2.2 Données issues des technologies d'imagerie à haut débit

Pendant les expéditions *Tara Oceans* et *Tara Pacific*, des données d'imagerie ont été acquises fournissant des informations quantitatives sur la taxonomie et la morphologie des organismes qui composent les communautés. Les jeux de données d'imagerie qui ont été utilisés au cours de ce travail ont été acquis grâce à deux technologies différentes qui permettent d'imager des échantillons de plancton collectés à l'aide de filets (Figure 2.3).

La technologie ZooScan est un scanner à plancton qui permet un vignettage de l'ensemble des organismes présents dans un échantillon (Gorsky et al., 2019). Le ZooScan a été utilisé au cours des deux expéditions. Parmi les fractions collectées, nous nous sommes intéressés à celle $>200 \mu\text{m}$ (filet WP2) de l'expédition *Tara Oceans* pour le chapitre 4 et à l'unique fraction $>300 \mu\text{m}$ (filet HSN) collectée lors de *Tara Pacific* pour le chapitre 5. Ainsi, le mésoplancton qui est dominé par des organismes multicellulaires est imagé par ZooScan.

La technologie Le FlowCam (Sieracki et al., 1998) a été utilisée uniquement pendant l'expédition *Tara Pacific*. Il a été embarqué à bord, permettant une acquisition des images tout au long de l'expédition. Contrairement au ZooScan, il s'agit d'un système d'imagerie en flux dans lequel chaque particule est prise en photo. La fraction de taille imagée par cette technologie est comprise entre $20\text{-}200 \mu\text{m}$ (filet decknet). Les échantillons associés au Flowcam sont donc associés au microplancton et dominés par les unicellulaires eucaryotes.

Toutes les images acquises et leurs métadonnées sont hébergées sur la plateforme web EcoTaxa (<https://ecotaxa.obs-vlfr.fr>). L'identification taxonomique de chaque image est prédite via un algorithme combinant random forest et réseaux de neurones convolutifs. Toutes les prédictions sont ensuite validées ou corrigées par un opérateur expert en taxonomie.

Les images acquises par les deux technologies sont des images noir et blanc. Ainsi, chaque image peut être considérée comme une matrice dont chaque élément est un nombre compris entre 0 (noir) et 255 (blanc). Deux types de descripteurs peuvent donc être calculés pour chaque objet à partir de ces images : des descripteurs associés aux niveaux

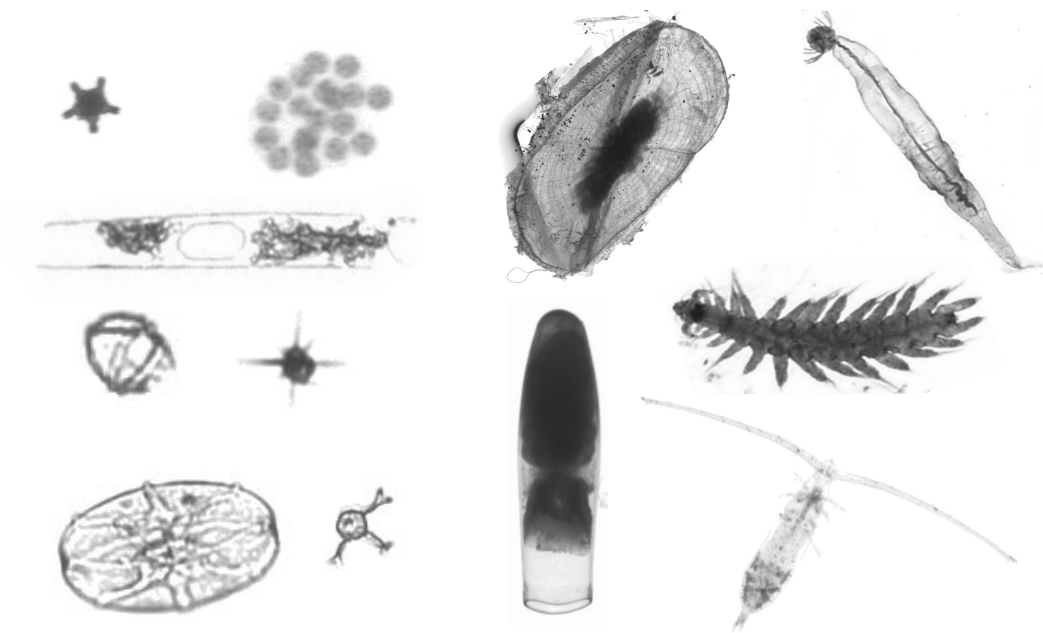


FIGURE 2.3 – Exemple de données acquises par les technologies d’imagerie Flowcam et ZooScan.

de gris de l’objet et des descripteurs associés aux dimensions de l’objet. Au total, 50 descripteurs sont automatiquement calculés dans l’environnement EcoTaxa (Picheral et al., 2017).

2.3 Les données environnementales

Lors des expéditions *Tara*, une grande variété d’environnements physiques et biogéochimiques ont été traversés, allant de régions froides eutrophes à chaudes oligotrophes. Ainsi, les données environnementales permettent de caractériser les conditions physiques et biogéochimiques de l’échantillonnage, et peuvent donc permettre d’établir des liens entre la structuration biologique observée et les contraintes abiotiques qui peuvent la forcer. Les données utilisées sont de deux types : d’une part, des données sur les conditions physiques et biogéochimiques des masses d’eau et d’autre part, des données sur les conditions de transport et de mélange des masses d’eau.

2.3.1 Les données biogéochimiques

Les données biogéochimiques regroupent des données de nature très différentes. Elles comprennent, en autres, (1) des données physiques telles que la température, la salinité,

la densité, (2) des données chimiques telles que la concentration en macro- et micronutriments ou la concentrations en carbonates, ou encore (3) des données associées à la biologie telle la concentration en chlorophylle-a, la concentration en carbone organique particulaire. Pour l'expédition *Tara Oceans*, les données biogéochimiques sont accessibles publiquement sur la plateforme de stockage de données *Pangaea* (Diepenbroek et al., 2002). Elles se décomposent de plusieurs fichiers qui fournissent soit les métadonnées (*i.e.*, provenance des échantillons qui inclut les informations de campagnes, de stations et de collecte, Pesant et al., 2015), soit les variables environnementales réparties dans plusieurs jeux de données. Les données biogéochimiques *Tara Pacific* sont accessibles publiquement sur le répertoire de données Zenodo. Chaque fichier regroupe métadonnées et variables mesurées (Pesant et al., 2015).

2.3.2 Les données de transport et mélange des masses d'eau

Dans l'introduction, l'impact potentiel des processus de fine échelle sur les conditions du milieu et la biologie a été présenté. Les données informant sur l'hydrodynamisme des masses d'eau étaient disponibles pour les échantillons de l'expédition *Tara Oceans* (Ardyna et al., 2017) mais pas pour ceux de l'expédition *Tara Pacific*. Ainsi, les données publiques de *Tara Pacific* ont donc été complétées par des données sur les conditions de transport et de mélange des masses d'eau. Pour cela, différents diagnostics eulériens et lagrangiens ont été calculés et les simulations permettant de les générer ont été réalisées par Alberto Baudena (post-doctorant LOV). Un récapitulatif de ces diagnostics est fourni dans la table 2.1.

De plus, les caractéristiques de mélange et de transport des masses d'eau conditionnent également la dispersion des organismes planctoniques. Dans le chapitre 3, la connectivité entre stations a été estimée à partir de deux métriques. La première correspond au temps moyen de connexion entre paires de stations et indique le nombre de jours que met une particule présente à une station donnée pour arriver dans une seconde station donnée. Les données utilisées étaient publiques (Berline et al., 2014). La seconde métrique estime, pour un couple de station (A,B), la probabilité qu'une particule initialement présente à A arrive dans la station B à un temps défini t . Ces données de probabilité de connexion ont été calculées par Enrico Ser-Giacomi (post-doctorant LOCEAN) Les matrices de distance obtenues par les deux méthodes sont des matrices asymétriques.

Diagnostic	Unité	Type	Interprétation
Vitesse absolue	m.s^{-1}	Eulérien	Il s'agit de la vitesse horizontale en 2 dimensions (i.e. incluant les composantes zonale et méridionale du champ de vitesse)
Énergie cinétique	$\text{m}^2.\text{s}^{-2}$	Eulérien	L'énergie cinétique correspond à l'énergie des particules due à leur mouvement.
Okubo-Weiss	jours^{-2}	Eulérien	Le diagnostic d'Okubo-Weiss est utilisé pour indiquer la position par rapport à un tourbillon. Les valeurs positives indiquent que les particules sont hors d'un tourbillon, alors que les valeurs négatives indiquent qu'elles sont dedans.
Vorticité	jours^{-1}	Eulérien	La vorticité est utilisée pour détecter les sens de rotation des tourbillons. Si les valeurs sont positives, le tourbillon est cyclonique (ou anti-horaire). Si les valeurs sont négatives, il est anticyclonique (ou horaire). Si les valeurs sont nulles, il n'y a pas de tourbillon.
Temps de rétention	jours	Eulérien	Le temps de rétention correspond au temps pendant lequel une particule ou parcelle d'eau reste piégée à l'intérieur d'un tourbillon.
Divergence	jours^{-1}	Lagrangien	Son interprétation est la même que pour la divergence eulérienne mais sur une période de temps temps antérieure.
Finite time Lyapunov exponent	jours^{-1}	Lagrangien	Les exposants de Lyapunov permettent d'identifier des zones turbulentes telles que les zones frontales. Plus leur valeur augmente, plus la probabilité de trouver un front est grande.
Betweenness		Lagrangien	Ce diagnostic permet d'identifier des régions de type goulot d'étranglement de circulation. Plus les valeurs sont élevées, plus les particules d'eau qui traversent la localisation proviennent d'endroits différents et se dirigent vers de multiples destinations.

TABLE 2.1 – Récapitulatif des diagnostics lagrangiens et eulériens utilisés.

2.4 Analyse de données

2.4.1 Analyse en composantes principales

L'Analyse en composantes principales (ou ACP, Legendre, Legendre, 2012) est une méthode d'analyse multivariée exploratoire. Elle s'applique sur une matrice composée de n observations et p variables quantitatives et vise à fournir une description synthétique de ces données. De nouvelles variables, appelées composantes principales, sont construites successivement et sont des combinaisons linéaires des variables initiales. Elles maximisent l'information (inertie) non résumée par les composantes principales précédemment construites. Ainsi, ces variables sont orthogonales et représentent les axes de maximum de variance du jeu de données. Dans le nouvel espace (ou espace réduit), les ressemblances entre individus et les liaisons entre variables peuvent être explorées. Les ACP ont été réalisées sur des données standardisées. Ainsi, les variables ont la même variance (égale à 1) et donc la même importance dans l'analyse (importance est proportionnelle à l'écart-type).

2.4.2 Méthodes de regroupement (ou classification)

Les méthodes de regroupement visent à définir des groupes d'observations homogènes du point de vue de l'ensemble des variables qui les décrivent (*i.e.*, typologie des observations). Parmi les différentes méthodes de regroupement, la classification ascendante hiérarchique produit une structure arborescente décrivant les relations entre individus et groupes d'individus. La construction de l'arbre fait intervenir une distance entre individus (*e.g.*, euclidienne, bray-curtis) et une distance entre groupes d'individus (*e.g.*, distance minimale ou lien simple, distance maximale ou lien complet). Dans les études présentées ici, le lien de Ward a été utilisé. Il agrège successivement les individus ou groupes d'individus en minimisant la diminution de l'inertie interne et présente l'avantage d'éviter les phénomènes de chaîne. Une fois obtenu, il est élagué afin d'obtenir une partition des individus. Cette partition en k groupes tend à maximiser la variabilité (*i.e.*, inertie) intergroupe et minimise celle intragroupe. Pour choisir le nombre de groupes, les pertes d'inertie suivant le nombre de groupes sont étudiées. L'objectif est de définir suffisamment de groupes pour que chaque groupe soit homogène sans en créer trop, ce qui conduirait à définir des groupes qui ne sont pas réellement différenciés. Les méthodes de regroupement ont été utilisées sur les résultats des ACP afin de construire les groupes sur le maximum d'information.

2.4.3 Analyse factorielle multiple

L'analyse factorielle multiple (AFM, Escofier, Pages, 1994) est une méthode d'analyse statistique multivariée qui permet d'analyser des jeux de données organisés en groupes de variables. Chaque groupe de variable doit être de même nature (*i.e.*, quantitatif ou qualitatif). Elle est basée sur deux étapes. Dans un premier temps, une ACP (ou analyse des correspondances multiples si les variables sont qualitatives - non présenté ici) est réalisée sur chaque groupe de variables. Cela permet d'obtenir pour chaque groupe de données la première valeur propre. Puis, une ACP est réalisée sur l'ensemble des variables pondérées par la racine de la première valeur propre du groupe auquel elles appartiennent. Cela permet d'équilibrer l'influence de chaque groupe dans l'analyse, tout en maintenant leur structure. Par analogie avec l'ACP, la première composante principale de l'AFM maximise l'inertie projetée de toutes les variables. L'AFM permet d'estimer les relations existantes entre les observations, les liaisons entre variables, mais aussi l'influence des groupes de variables sur la structure des observations. Un avantage de cette méthode est de permettre de caractériser la similitude ou les différences fournies par chaque groupe de variables.

2.4.4 Régression linéaire

Les modèles de régression linéaire ont pour but d'établir la relation entre une variable unique à expliquer et une ou plusieurs variables explicatives. Ces modèles prédictifs permettent de tester si le modèle explique mieux les données qu'un modèle aléatoire (test global de Fisher), de sélectionner le modèle expliquant le mieux les données (*e.g.*, via des tests de modèles emboîtés) et de quantifier l'influence de chaque variable explicative sur la variable à prédire.

2.4.5 Régression par les moindres carrés partiels parcimonieuse

La régression par les moindres carrés partiels parcimonieuse (ou sPLS pour *sparse Partial Least Squares*) est une méthode de régression linéaire (Chung, 2010), permettant d'estimer la relation entre deux jeux de données multidimensionnels. L'un des jeux de données est à expliquer et l'autre est explicatif (à noter que la sPLS peut également être canonique et dans ce cas, les jeux de données auront un rôle symétrique dans l'analyse). La relation entre les jeux de données est estimée à partir d'une matrice de corrélation. Cette méthode permet de prendre en charge des variables colinéaires, bruitées et ayant des valeurs manquantes. La sPLS réalise une sélection de variables dans les deux ensembles de données, ce qui accroît l'interprétabilité des résultats. La sélection est basée

sur le critère de pénalisation LASSO. Si aucune contrainte n'est fournie sur le nombre de dimensions à conserver, alors la sPLS est équivalente à la PLS (Wold et al., 2001).

(Méta)Génomique des populations de protistes

3.1 Préambule

Dans l'introduction, il a été souligné que la biodiversité englobe la variabilité biologique à tous les niveaux d'organisation du vivant, du gène à l'écosystème. Parmi eux, la diversité génétique intraspécifique est primordiale car elle détermine le potentiel évolutif et adaptatif des espèces. Cette diversité génétique, qui fait référence à variabilité nucléotidique au sein des gènes, résulte de divers types d'événements, comme les mutations (*i.e.*, changement d'un nucléotide pour un autre) ou les insertions-délétions (*i.e.*, ajout ou suppression d'un ou plusieurs nucléotides). Ces événements sont à l'origine de différentes versions des gènes, appelées allèles, dont les fréquences au sein des populations peuvent être estimées. La comparaison des fréquences alléliques entre les sous-populations permet de déterminer si la diversité génétique est partagée ou s'il y a une accumulation de variation génétique au sein d'une sous-population, entraînant une différenciation génétique entre elles (Bohonak, 1999; Nei, 1973; Weir, Cockerham, 1984; Wright, 1969). L'existence d'une distribution hétérogène des fréquences alléliques entre sous-populations indiquent un flux génétique faible entre elles, alors que l'absence de différenciation génétique suggère un flux génétique fort et une panmixie complète (Weir, Cockerham, 1984; Wright, 1969)(Weir et Cockerham, 1983; Wright, 1969). Étudier la distribution des fréquences alléliques au regard des facteurs externes peut fournir des indices sur les processus actuels et historiques qui la façonnent (Sobel et al., 2010).

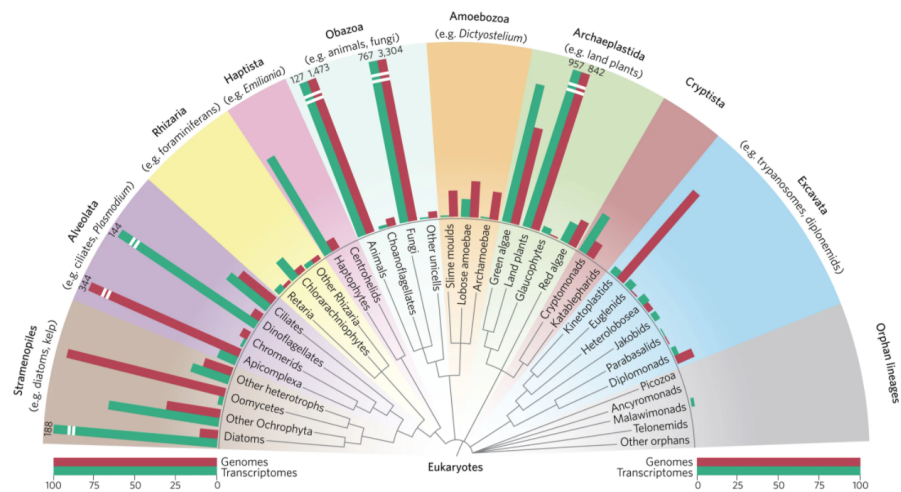


FIGURE 3.1 – Diversité eucaryote suivant les relations phylogénétiques (cladogramme). Le nombre de projets de séquençage de génomes et de transcriptomes répertoriés comme terminés ou en cours dans la base de données Genomes Online (février 2017) sont indiqués par les histogrammes. Depuis Sibbald, Archibald (2017)

Dans le vivant, la vision de la diversité génétique est biaisée suivant les taxons. Contrairement aux organismes bactériens dont de nombreux génomes complets de divers taxons sont annotés (Wu et al., 2009), le nombre de génomes ou transcriptomes eucaryotes disponibles est beaucoup plus faible. De plus, au sein des eucaryotes, les métazoaires ont été beaucoup plus ciblés (Figure 3.1). Ainsi, la distribution spatiale de la diversité génétique intraspécifique des protistes a été beaucoup moins étudiée et les mécanismes qui la sous-tendent sont encore peu connus. Parmi les approches d'étude de la diversité génétique intraspécifique, celles à l'échelle du génome peuvent permettre de détecter de nombreux variants génétiques et donc conduire à une estimation fiable de la variation génétique (Allendorf et al., 2010). Ainsi, les données métagénomiques, en fournissant le contenu génétique de tous les organismes présents dans une communauté pourraient constituer une véritable opportunité pour mieux comprendre comment s'organise la diversité génétique intraspécifique des protistes.

L'objectif de ce chapitre est donc de développer une méthodologie générale pour quantifier la différenciation génétique de différentes espèces de protistes à partir des données métagénomiques, puis de la mettre en lien avec les différents facteurs externes qui pourraient la contraindre. Cette étude a été menée dans la Mer Méditerranée qui présente de nombreux gradients environnementaux et une circulation dynamique (Millot, Taupier-Letage, 2005; Ayata et al., 2018). Le reste de ce chapitre correspond au manuscrit intitulé

3.2. Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations

Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations.

3.2 Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations

Ophélie Da Silva^{1,2}, Enrico Ser-Giacomi^{3,4}, Jade Leconte⁵, Éric Pelletier^{5,6}, Cécile Fauvelot^{1,7}, Amin Madoui⁵, Lionel Guidi^{1,6}, Fabien Lombard^{1,6,8}, Sakina-Dorothee Ayata^{1,2,*}, Lucie Bittner^{2,8,*}

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

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

3 - Sorbonne Université, UMR 7159 CNRS-IRD-MNHN, LOCEAN-IPSL, 75005 Paris, France

4 - Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 54-1514 MIT, Cambridge, MA 02139, USA

5 - Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

6 - Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans-GOSEE, Paris, France

7 - Institut de Recherche pour le Développement (IRD), UMR ENTROPIE, Nouméa, New Caledonia

8 - Institut Universitaire de France (IUF), Paris, France

* - Contributed equally

Abstract

Since a decade, high-throughput sequencing has transformed the study of marine planktonic communities and has highlighted the extent of protist diversity in these ecosystems. Still, little is known about their genetic diversity at the species-scale, and about their major speciation mechanisms. As data from large scale samplings are now available, we postulate that metagenomics could contribute to deciphering the processes shaping protist genetic differentiation in the marine realm. As a proof of concept, we focused on

the Mediterranean Sea and chose three abundant protist species : *Bathycoccus prasinos*, *Pelagomonas calceolata* and *Phaeocystis cordata*. We compared the genetic differentiation of each species in light of geographical, environmental and oceanographic distances. We highlighted that isolation-by-environment may more shape the genetic differentiation of *Bathycoccus prasinos* whereas *Phaeocystis cordata* may be more impacted by isolation-by-distance. At present time, the use of metagenomics to accurately estimate the genetic differentiation of protists remains challenging since resulting coverages are lower compared to traditional population surveys. However we started here to shed light on ecological and evolutionary processes occurring within natural marine populations, and paved the way for future protist population studies exploiting metagenomics.

Keywords : Metagenomics ; Genetic differentiation ; Protists ; Marine plankton ; Mediterranean Sea.

3.2.1 Introduction

Single-celled eukaryotes or protists are major contributors to the planktonic diversity in the oceans (Moon-van der Staay et al., 2001; De Vargas et al., 2015). They encompass a myriad of lifestyles (from free-living to colonies), trophic modes (phototrophy, heterotrophy and mixotrophy), as well as morphological characteristics such as naked cell, calcareous or siliceous skeleton, or presence of spines (Caron et al., 2012, 2017). Despite their key roles in the functioning of marine pelagic ecosystems, both as primary and secondary producers, hence affecting trophic dynamics and global biogeochemical cycles (Biard et al., 2016; Dolan, 2018), they have long been and are still under-explored especially from a genomic point of view (del Campo et al., 2014; Sibbald, Archibald, 2017).

The scarcity of reference genomic data naturally results in a misunderstanding of the processes that underpin the contemporary distribution of genetic diversity in natural populations (Logares, 2011; Lebreton et al., 2012). Protists are supposed to have huge population sizes and the potential ability for long-distance dispersal (Dolan, 2005; Watts et al., 2013). Both properties have been hypothesized to reduce the chances of evolutionary diversification (Finlay, 2002). In comparison with most macroorganisms, protists are thus rather expected to have little opportunity for allopatric divergence and to show low levels of spatial genetic structure. Last decade studies, based on high-throughput sequencing from natural communities via metabarcoding, have revealed a high diversity of protist species, as well as endemic or cosmopolitan species (Bittner et al., 2013a; Logares et al., 2014; De Vargas et al., 2015; Forster et al., 2015; Malviya et al., 2016).

3.2. Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations

Moreover, the very few genomic studies based on protist ‘model micro-organisms’ such as *Emiliania huxleyi* (Iglesias-Rodríguez et al., 2006; Read et al., 2013) or *Ostreococcus tauri* (Blanc-Mathieu et al., 2017) have highlighted a large intraspecific diversity in the ecosystems. In that respect, even if dispersal is high, protist population structure has been frequently described at local (Evans et al., 2005), regional (Casteleyn et al., 2009) and even at global (Casteleyn et al., 2010) scales, and several processes were reported as promoters of their diversification (Sjöqvist et al., 2015).

In the marine realm, gene flow and connectivity among planktonic populations is driven by three major forces : marine currents, abiotic (*i.e.*, physico-chemical) environmental conditions and biotic factors. Physical processes such as seascape topography or oceanographic currents were suggested to contribute to the structuring of populations through the genetic isolation from more global populations (Nagai et al., 2007, 2009; Casabianca et al., 2012). Currents driving the oceanographic connectivity could lead to asymmetric migration patterns and spatial differentiation of populations (Godhe et al., 2013; Riginos et al., 2016), which may promote adaptation of populations to local environmental conditions (Kawecki, Holt, 2002; Sjöqvist et al., 2015). Genetic differentiation may also be driven by natural selection through environmental conditions such as silicate and nitrate/nitrite concentrations (Gao et al., 2019), changes in salinity (Sjöqvist et al., 2015; Godhe et al., 2016), light or temperature (Mena et al., 2019; Latorre et al., 2021).

To date population genetic studies have generally focused only on a restricted number of single protist species or on sparse genomic markers. With the recent advances in high-throughput sequencing, single nucleotide polymorphisms (SNPs, a nucleotide variation at a specific position in the genome) analysis is emerging as a powerful approach to infer population genetic differentiation among natural populations. SNPs show advantages such as a high abundance randomly distributed in genomes, low mutation rate and low false genotyping rate, while presenting a good statistical power and being relatively cost-effective (Morin, McCarthy, 2007; Selkoe et al., 2016). Generally, SNP detection methods use long reference sequences, from a genome or transcriptome, to map short sequences (reads) obtained by high-throughput sequencing. In this way, recent studies started to provide a metagenome-level description of the ecological preferences of a few protists (Vannier et al., 2016; Seeleuthner et al., 2018; Leconte et al., 2020). However, except a recent global analysis using proxy of species obtained from a genomic reference-free method (metavariant species; Laso-Jadart et al., 2021), no study has tried to explore in a systematic way (as soon as reference data are available) genetic differentiation of protists from metagenomes, nor to propose guidelines to implement the study of the

ecological and evolutionary processes shaping their diversity at species-scale.

To address this issue, we chose to focus on the Mediterranean Sea, which is an ideal location to study the population genetics of protists. It is a semi-enclosed marginal sea of the North Atlantic Ocean that is characterized by tortuous coastlines, with several environmental gradients despite a highly dynamic circulation (Millot, Taupier-Letage, 2005; Ayata et al., 2018). A large number of population genetics studies have demonstrated how dispersal and environment impact macroorganisms' population genetic diversity in the Mediterranean Sea (*e.g.*, the striped red mullet in Dalongeville et al., 2018). On the contrary, few studies have focused on protists, highlighting the existence of a genetic structure between the Atlantic Ocean and the Mediterranean Sea (Lowe et al., 2012) or into the Mediterranean Sea due to circulation patterns (Casabianca et al., 2012). However, as in all other ocean areas, protists have been much less studied. During fall 2009, the Tara schooner sailed in the Mediterranean Sea and collected plankton by size fractions from 0.8 μm to 2 mm. Metagenomic datasets were generated through the Tara Oceans project (Alberti et al., 2017; Carradec et al., 2018), which enable to investigate few highly-abundant components of the ecosystems (*e.g.*, the *Oithona nana* copepod, Madoui et al., 2017; SAR11 Bacteria, Delmont et al., 2019). Protists constituted major components too (Carradec et al., 2018), but the corresponding data were not exploited yet. The available Tara Oceans datasets constitute a unique opportunity to test if metagenomics can, at the present time, support population genetics studies for protists.

The objective of our study is to exploit the currently available metagenomic data for characterizing genetic differentiation of protists in the Mediterranean Sea. We focus on three planktonic species with the aim of studying how their genetic differentiation relates with external drivers, such as geography, environmental conditions and oceanographic circulation. We hypothesize that the genetic differentiation is greater for distant populations (Wright, 1943) and/or for populations in different hydrological environments (Wang, Bradburd, 2014). We address the question if current metagenomic data could be used for estimating (1) how the genetic differentiation is organised at species-scale and (2) which are its main drivers. Finally we discuss how current metagenomics could support the mining of genetic diversity and provide new resources for studying population genetics of the, so far, least-studied but abundant and ecologically relevant protist species.

3.2.2 Results

Genetic distances from metagenomic samples based on a selection of three distinct protist species

Preliminary investigations based on (1) a bibliographic study of ecologically relevant protist species in the Mediterranean Sea, (2) an abundance analysis of eukaryotic metabarcodes in the Tara Oceans (TO) Mediterranean samples (De Vargas et al., 2015), and (3) the availability of protist reference genome(s) or transcriptome(s), led us to select three phylogenetically distinct, ubiquitous, planktonic species (Supp. Mat. B.1) : *Bathycoccus prasinus* (Chlorophyta, Mamiellales; Eikrem, Throndsen, 1990), *Pelagomonas calceolata* (Ochrophyta, Pelagomonadales; Andersen et al., 1993) and *Phaeocystis cordata* (Haptophyta, Phaeocystales; Zingone et al., 1999). A reference genome was used for *B. prasinus* (Moreau et al., 2012), whereas, at the time of our analysis, only transcriptomes were publicly available for *P. calceolata* and *P. cordata* (Keeling et al., 2014; Johnson et al., 2019). The cumulative length of these assemblies varied by two fold from *P. cordata* to *P. calceolata* (respectively 9.4 and 21 Mb), while the length of *B. prasinus* assembly was intermediate (15 Mb). Based on the cell size of these organisms (Supp. Mat. B.1), we focused on the 0.8-5 μm size fraction of the TO dataset, consisting of 13 metagenomics samples collected at the surface layer from 13 stations in the Mediterranean Sea (accession number PRJEB4352, Carradec et al., 2018; see Supp. Mat. B.2 for details). We mapped the metagenomic reads (on average 185 million reads/sample) on reference assemblies and obtained horizontal and vertical coverages between species (Figure 3.2) (percentage of the assembly covered by at least one read and mean number of reads aligned at each position of the assembly, respectively). *B. prasinus* and *P. calceolata* displayed higher horizontal coverages than *P. cordata* despite their longer references size (maximal horizontal coverage, *B. prasinus* : 98.1%, *P. calceolata* : 68.2%, *P. cordata* : 19.6%). The mean vertical coverage was also sparser for *B. prasinus* and for *P. calceolata* than for *P. cordata* (min-max mean vertical coverage, *B. prasinus* : 0.085-17.6 X, *P. calceolata* : 0.119-7.2 X, *P. cordata* : 0.51-2.47 X). *B. prasinus* showed its highest coverages at station 6 and station 5.

SNPs were detected from aligned and filtered reads (Methods and Supp. Mat. B.2). *B. prasinus*, *P. calceolata* and *P. cordata* showed contrasted total numbers of SNPs over reference size ratio (respectively 3.4, 51.4 and 0.5 SNPs/Mb). In average, only 9.26%, 9.67% and 18.74% of the total number of SNPs were observed in each station for *B. prasinus*, *P. calceolata* and *P. cordata* respectively. Moreover, a filtration on the vertical

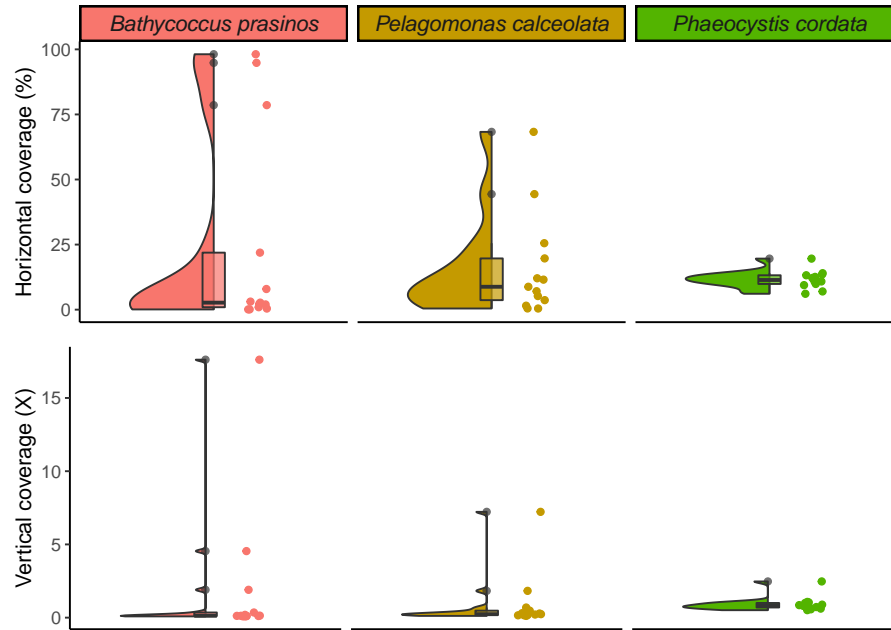


FIGURE 3.2 – Distributions of the horizontal and mean vertical coverages for each species within the 13 Mediterranean Sea stations

coverage of 4X (Methods and Supp. Mat. B.2) led to the removal of a certain number of SNPs and consequently pairwise F_{ST} between all stations pairs were computable for *P. cordata*, whereas for *P. calceolata* the station 30 had to be removed, and for *B. prasinos* 20 pairwise F_{ST} were not computable. We obtained 58 pairwise genetic distances (pairwise F_{ST}) for *B. prasinos*, 66 for *P. calceolata* and 78 for *P. cordata*.

Our results show the Mediterranean Sea populations of *B. prasinos* had a stronger global genetic differentiation (global F_{ST} =0.136) than *P. calceolata* (global F_{ST} =0.066) and *P. cordata* (global F_{ST} =0.045) (Figure 3.3). *B. prasinos* also showed the most contrasted genetic differentiation (pairwise F_{ST}) ranging from little (0.012) to very high genetic differentiation (0.476) (Figure 3.3). *B. prasinos* was also the only species showing very high genetic differentiation values (seven F_{ST} values > 0.25). *P. calceolata* showed little to high genetic differentiation (F_{ST} ranging from 0.019 to 0.181) whereas *P. cordata* had the lowest F_{ST} values (ranging from 0.026 to 0.069). The maximum of F_{ST} for *P. cordata* was in little genetic differentiation class, whereas *P. calceolata* and *B. prasinos* peaks were in the moderate genetic differentiation class (Figure 3.3).

3.2. Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations

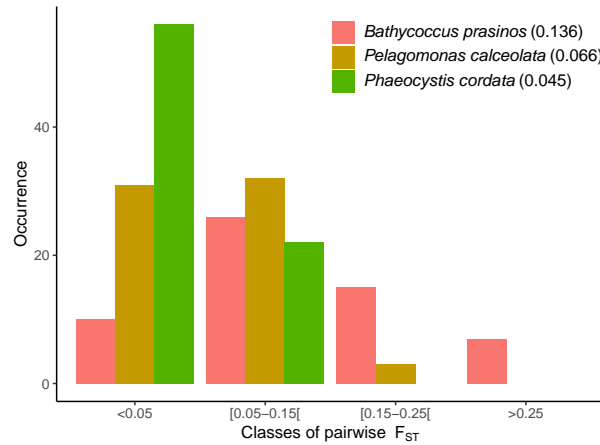


FIGURE 3.3 – Distribution the genetic distances for each species between the 13 Mediterranean Sea stations. The pairwise F_{ST} were grouped into four classes of genetic differentiation (<0.05 : little, [0.05-0.15] : moderate, [0.15-0.25[: high, >0.25 : very high) following Hartl and Clark (1997).

Geographic, environmental and oceanographic distances inferred from the Mediterranean Tara Oceans stations

Geographic distances were computed as the minimal distances between two stations without crossing lands. Estimated geographic distances ranged between 33.55 km and 3624.49 km (mean : 1154.18 km, median : 1037.79 km ; Figure 3.4A, Supp. Mat. B.3).

A Principal Component Analysis (PCA) was performed on the environmental conditions of the Mediterranean TO stations, and then the environmental distances between stations were estimated from the Euclidian distances between stations in the PCA space. In order to strengthen the statistics, we involved in the analysis the 13 stations corresponding to our metagenomic samples and 9 additional TO stations also sampled in the Mediterranean Sea (Figure 3.4B). Three significant axes were identified : the first PCA axis (Dim1, 48.6% of the total variance) distinguished the warmer and more oligotrophic stations (Dim1<0) from the colder and richer ones (Dim1>0) ; the second PCA axis (Dim2, 18.8% of the total variance) separated the saltier and ammonium-rich stations (Dim2>0) from the less salty and ammonium-poor stations (Dim2<0) ; and the third PCA axis (Dim3, 15.2% of the variance) divided the silicate-rich stations (Dim3<0) from the phosphate and nitrite-rich ones (Dim3>0). Environmental distances were calculated as the variance-weighted Euclidean distances on the first three dimensions (*i.e.*, significant according to the Kaiser-Gutmann criterion).

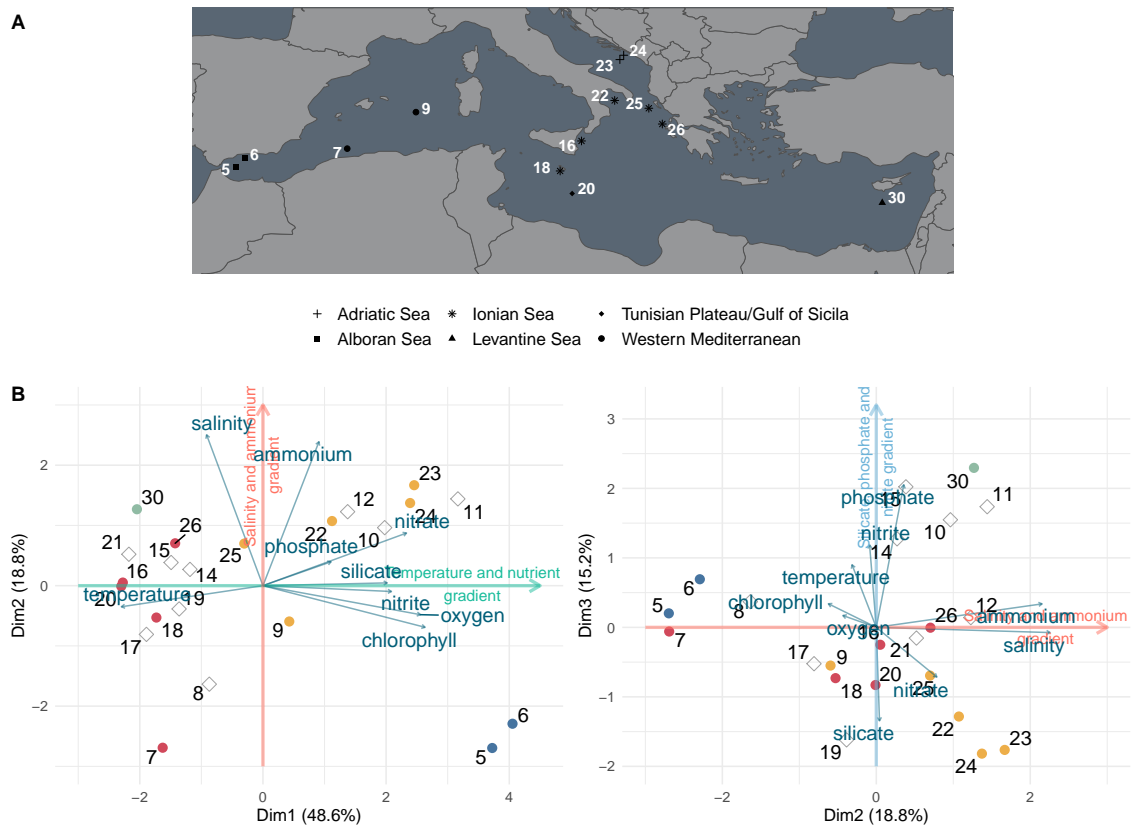


FIGURE 3.4 – Geographic and environmental characterisation of the Mediterranean Sea stations. (A) Geographical location of the 13 TO stations studied with metagenomics. Geographical entities are based on the Marine Ecoregions of the World (Spalding et al., 2007) and are represented by shapes. Cross : Adriatic sea, square : Alboran sea, star : Ionian sea, triangle : Levantine sea, diamond : Tunisian Plateau/Gulf of Sicila, circle : Western Mediterranean. (B) Principal Component Analysis of the environmental variables retrieved at 22 stations (the 13 TO stations studied with metagenomics, reported with colored dots, and nine extra stations, reported with empty diamonds). Stations are indicated by numbers (with increasing numbers from west to east, following the Tara Oceans cruise trajectory), environmental variables are represented with blue arrows. The three first significant axes (Dim1, Dim2, Dim3) are presented with their corresponding percentages of variance.

Oceanographic distances were inferred from Lagrangian modelling both as the mean connection times (data from Berline et al., 2014) and as connection probabilities among stations (Ser-Giacomi et al., 2015). They ranged between 46.15 km and 545.94 km (mean : 219.89 km, median : 239.9 km) and 2.75×10^{-6} and 9.07×10^{-3} (mean : 8.03×10^{-4} , median : 2.85×10^{-4}) respectively (Supp. Mat. B.4). The oceanographic distance matrices were

3.2. Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations

asymmetric and some stations were never connected.

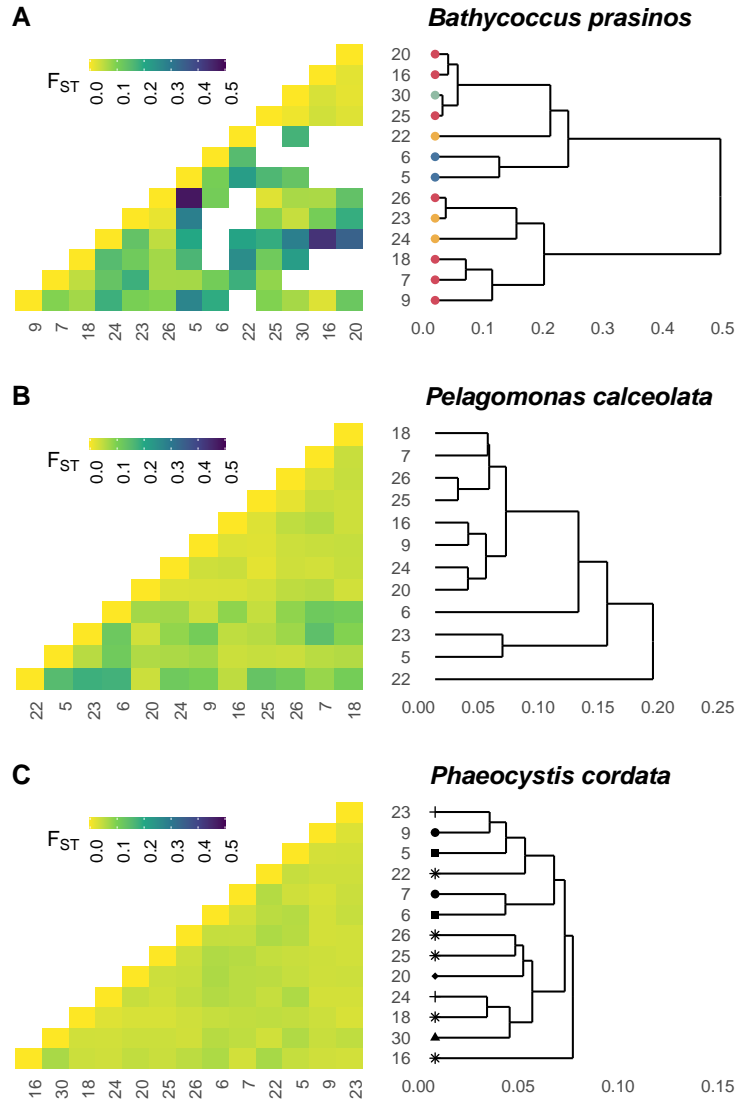


FIGURE 3.5 – Genetic differentiation among the Mediterranean Sea populations. Heatmaps of genetic distances (pairwise F_{ST}) with associated dendrograms obtained by hierarchical clustering for (A) *B. prasinos*, (B) *P. calceolata* and (C) *P. cordata*. For *B. prasinos*, missing values (due to absence of common SNPs) were replaced by the mean value of F_{ST} of the stations. In the dendrograms, stations were highlighted with symbols (as in Fig. 3.4A) when genetic distances were correlated to geographic distances and with colored dots (as in Fig. 3.4B) when the genetic distances were correlated to environmental distances.

Deciphering drivers of genetic differentiation

For each of the three planktonic protist species, the link of genetic differentiation with geography, environment and oceanographic circulation was assessed through linear regression models. We showed with Fisher tests that our linear models better explained the genetic differentiation of *B. prasinus* and *P. cordata* than the null models (Table 3.1). On the contrary, our models did not provide a better fit than the null model for *P. calceolata*. The optimal selection model procedure led to the selection of environmental distances for *B. prasinus* and of geographic distances for *P. cordata* (Table 3.1). Selected models explained 15.24% of the genetic differentiation for *B. prasinus* and 12.04% for *P. cordata*. The genetic differentiation of *B. prasinus* significantly increased with environmental distances, whereas the genetic differentiation of *P. cordata* significantly increased with geographic distances.

Species	Selected vs null model	Selected model	R ²
<i>Bathycoccus prasinus</i>	$pvalue = 0.0024$	$\beta_{env} = 0.0382$ $pvalue_{env} = 0.0024$	15.24%
<i>Pelagomonas calceolata</i>	$pvalue = 0.3436$		
<i>Phaeocystis cordata</i>	$pvalue = 0.0019$	$\beta_{geo} = 4.624 \cdot 10^{-6}$ $pvalue_{geo} = 0.0019$	12.04%

TABLE 3.1 – Outputs of the statistical models testing the impact of geography, environment, and ocean circulation on the genetic distances of the three planktonic species. P-value of the overall significance test is provided (selected vs null model). For each selected model, the value of the regression coefficient β and the associated p-values are indicated for the selected variables (env : environmental distances, geo : geographic distances), as well as the determination coefficient R²

The stronger effect of external drivers in genetic differentiation was observed for *B. prasinus* and was positively correlated to the environment. *B. prasinus*' highest genetic differentiations were observed between station 5 (Alboran Sea) and station 26 (in the Ionian Sea) and between stations 16 and 24 (Figures 3.4B and 3.5A), while the smallest genetic differentiations were observed among stations 16, 20, 25 and 30, all located in the Eastern basin. This group of stations was genetically distant from stations 23 and 24 (Adriatic Sea), and also from stations 5 and 6 (Alboran Sea) and also station 22 (Ionian Sea) (Figure 3.5A). These two groups of stations (16, 20, 25, 30 vs. 5, 6, 22, 23, 24) are environmentally separated by the nutrient-temperature gradient (Dim1, Figure 3.4B), with the former stations characterized by warmer and more oligotrophic conditions (Easter Basin of the Mediterranean Sea).

3.2. Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations

For *P. cordata*, the highest genetic differentiations were observed between stations from the Western Basin (stations 5, 6, 7, 9, Alboran Sea and Western Mediterranean) and from the Eastern Basin (stations 18, 20, 24, 25, 26, 30, Ionian Sea, Tunisian Plateau/Gulf of Sicilia, Adriatic Sea and Levantine Sea) (Figures 3.4A and 3.5B). Exceptions are however observed, with spatially close stations that are genetically distant for this species (*e.g.*, stations 9 and 23, or stations 18 and 24), confirming that the geographical distances are not the only factor driving the genetic differentiation of this species.

3.2.3 Discussion

Our population metagenomics approach applied to marine planktonic protists intended to offer new insights into the genetic differentiation of understudied organisms and to elucidate the main abiotic drivers shaping it. The genetic differentiation of three species of marine protists, which are a priori all ubiquitous at the global scale, was explored through metagenomics in the Mediterranean Sea, a model sea for studying biodiversity and the effect of global changes (Lejeusne et al., 2010; Piroddi et al., 2017).

Main challenges of the *population metagenomics* approach

We here developed a computational method in order to implement population genetic studies from metagenomics data. This pipeline, which was designed according to the FAIR principles (Wilkinson et al., 2016), can be transposed and used with any protist lineage as long as reference sequences are available. The number of species that could be investigated with our pipeline is thereby dependent on the number of reference assemblies that are available. Historical sequencing priorities in favor of species of biomedical and biotechnological interest has led to a bias in the molecular public databases in which relatively few protist references are currently available (del Campo et al., 2014; Keeling, Campo del, 2017; Sibbald, Archibald, 2017). In the last decade, an increasing number of sequencing initiatives have tried to compensate this bias (*e.g.*, MMETSP, Keeling et al., 2014; Johnson et al., 2019), but most protist lineages remain under-represented (del Campo et al., 2014; Sibbald, Archibald, 2017). Thus, even if the resource available is increasing, there are still relatively few assemblies that may support protist population genetics studies based on metagenomic data (*e.g.*, about 200 genomes in the Genome Online Database in February 2017, mostly parasitic; Sibbald, Archibald, 2017). So far, at most a few dozen species are relevant for a focus on the Mediterranean Sea. It is nonetheless very likely that the ongoing increasing availability of references will further pave the way for future similar population metagenomics analyses.

Although our three targeted species are reported as globally ubiquitous, based on morphological observations (GBIF.org, 2021), and abundant in the Mediterranean Sea, based on metabarcoding data (Supp. Mat. B.1), the reference sequences selected here may include biases, as references do not come from the exact same geographic location as the metagenomic data. The *Pelagomonas calceolata* reference strain (RCC969) was sampled in the southern Pacific Ocean (Le Gall et al., 2008), which is, as the Mediterranean Sea, an oligotrophic area. However, variations between *Pelagomonas* strains have already been reported, notably driven by different light-acclimation strategies (Kulk et al., 2012; Worden et al., 2012). The two other references, *Phaeocystis cordata* (strain RCC1383) and *Bathycoccus prasinus* (strain RCC1105), were both isolated in the Mediterranean Sea, respectively from the Tyrrhenian Sea (Zingone et al., 1999) and from Banyuls' bay (north-western Mediterranean Sea, Moreau et al., 2012). Yet our 13 metagenomics samples were not collected into these areas. The relatively low mapping coverages obtained here might thus reflect the fact that wild lineages differ significantly from the current reference lineages, which are the one in culture or the ones that have been punctually isolated from the environment and which only constitute a small fraction, poorly indicative of natural populations (Worden et al., 2012; Bittner et al., 2013a; Laso-Jadart et al., 2021). Using metagenomic-assembled references, also referred to as “metagenomic assembled genomes” (MAGs), could therefore partly circumvent this issue. MAGs have been built for marine viruses, archaea, bacteria (Tully et al., 2017, 2018; Delmont et al., 2018) or more recently for microbial eukaryotes, either for genomes (Delmont et al., 2021) or for transcriptomes (Vorobev et al., 2020). However, MAGs result from “consensus assemblies” and could thus integrate the variability of several sampled organisms or populations. The fact that they represent “molecular units”, present and abundant in the natural environment is certainly the main advantage of these references (*i.e.*, because they are built from the selection of co-occurring, abundant metagenomic reads). However, the correspondence of these “molecular units” with real “biological units” (here species) is not guaranteed (Shaiber, Eren, 2019) and the biological scale at which the study is carried out (species, genus, biological association/holobiont) may remain uncertain. Finally, the representativity of the current reconstructed metagenomic species is high, but likely far from the phylogenetic diversity highlighted with metagenomics studies based on classical *de novo* assemblies (Carradec et al., 2018 versus Delmont et al., 2021). In addition, some highly abundant but genomically complex lineages (*e.g.*, Dinoflagellates) still fail to be reconstructed from the environment (Delmont et al., 2021), mainly due to the globally shallowed sequencing depth generally achieved for environmental samples.

3.2. Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations

The main difference between traditional population genetics studies and our metagenomics-based approach is that metagenomic data contains all the community signal. This hence imposes a first analytical procedure during which the sequences of the targeted species have to be extracted from the bulk data. Even if the TO metagenomic samples were obtained by the filtration of large seawater volumes and were sequenced to a consequent depth (160 million reads per sample; (Pesant et al., 2015; Alberti et al., 2017), it has been previously shown that the sampling effort for the smallest planktonic fraction did not result in a saturation of the eukaryotic gene space (Carradec et al., 2018). Our targeted species, even if theoretically abundant in the Mediterranean Sea, might thus be underrepresented in the metagenomic samples. We however made the assumption that the direct mapping of metagenomic reads on references (instead of their use through *de novo* assembly of long sequences) could inform us on the genetic structure of protists populations (Vannier et al., 2016; Leconte et al., 2020). We therefore tuned the filtering parameters of our population metagenomics approach in order to ensure the detection of good quality variants of our targeted species. Firstly, during the read recruitment step, we removed alignments with less than 95% mean identity in order to obtain a proper genome abundance estimate of the targeted species. This identity threshold is in accordance with previous estimations based on Chlorophyta lineages (Blanc-Mathieu et al., 2017; Leconte et al., 2020) and should ensure that all reads recruited could be assumed to belong to the same species, despite intraspecific variation. Future new genomic references on Haptophyta and Ochrophyta will allow to refine this filtering step, meanwhile we pragmatically studied their genomic signal with the same parameters. Secondly, defining a mapping coverage threshold is not straightforward. In studies based on model organisms, *i.e.*, mainly human, mice and a few others Metazoa, as well as pathogenic Bacteria, at least 30 X vertical coverage thresholds are usually expected (Sims et al., 2014; Davide, Donati, 2017), while in our study vertical coverages ranged between 0.085 X and 17.6 X. However low coverage thresholds (*e.g.*, 4 X) have already been used to analyze metagenomic data (*e.g.*, on copepods, Madoui et al., 2017), allowing the first population genomic inferences based on metagenomic data for non-model planktonic species. Therefore, in spite of relatively low coverages and consequent F_{ST} estimates based on a restricted amount of data, we believe that our results are valuable to describe first trends for genomic differentiation from natural populations of protists.

Protist genetic differentiation in the Mediterranean Sea and its different drivers

Previous studies on marine organisms highlighted that physical distances, estimated from geographic distances (Dalongeville et al., 2018) or oceanographic distances (Galindo et al., 2006; White et al., 2010; Dalongeville et al., 2018), shape the genetic variation of marine species at species-scale. Environmental characteristics, such as temperature gradient, or kelp cover, may also be of major importance in structuring marine population (Kawecki, Holt, 2002; Selkoe et al., 2010). Most seascape genetics studies rely however on the study of pluricellular and macroscopic organisms (Selkoe et al., 2016). Thereby, by analogy with these studies, we conducted a comparative study highlighting that *Bathycoccus prasinus* (Chlorophyta), *Pelagomonas calceolata* (Ochrophyta) and *Phaeocystis cordata* (Haptophyta) exhibit distinct spatial patterns, forced by different external constraints. However, our statistical models did not explain a large part of the genetic differentiation for *B. prasinus* and *P. cordata*, and none for *P. calceolata*; our results must be hence interpreted with caution. For *B. prasinus*, gene flow appeared stronger among populations from similar environments, which suggested an isolation-by-environment pattern (Wang, Bradburd, 2014). Several mechanisms may also explain the genetic differentiation due to environmental forcing, especially local adaptation, non-random mating due to adaptation or phenotypic plasticity (Sexton et al., 2014). Nonetheless, Vannier et al. (2016) already suggested that environmental conditions such as temperature and light may influence the distribution of *Bathycoccus*, which argue in favor of our isolation-by-environment hypothesis and that the Baas Becking hypothesis “*everything is everywhere but the environment selects.*” (Baas-Becking, 1934) might well rule the genetic differentiation of this lineage. For *P. cordata*, our results suggested that gene flow decreased with geographic distances. This supports a hypothesis of isolation-by-distance for this species (Wright, 1943), in which dispersal constraints would limit gene flow. This could be further explained by the motility capacities of this species. There is no evidence of colonial form for *P. cordata* (Zingone et al., 1999; Rousseau et al., 2007), although organisms of the *Phaeocystis* genus are often found in colonial form (Hamm, 2000). While *P. calceolata* and *P. cordata* are flagellate cells, *B. prasinus* is a non-motile organism (Eikrem, Throndsen, 1990). The flagellar apparatus of *P. calceolata* is however reduced (Heimann et al., 1995). The pattern of isolation-by-distance observed for *P. cordata* might then result from the presence of its flagella, as it may impact its vertical position in the water column and consequently its dispersal. However, such a hypothesis should further be tested in situ. Isolation-by-distance (IBD) has frequently been suggested as an important driver of genomic differentiation in many protist

3.2. Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations

species (Nagai et al., 2007; Casteleyn et al., 2010; Demura et al., 2014) and could act at different scales : *e.g.*, IBD has been described as a driver of genomic differentiation for the diatom *Pseudo-Nitzschia pungens* at global (Casteleyn et al., 2010), regional (Casteleyn et al., 2009) and local scales (Evans et al., 2005). Several studies have underlined that the spatial-scale of sampling is important to detect genetic isolation. For instance, Palumbi (2003) highlighted from simulations that identifying the process of isolation-by-distance is easier at large distances for marine species. Dalongeville et al. (2018) assessed the importance of geographic distances at long-distance spatial scale (>1,000km) and of dispersal constraints at shorter spatial scale (<1,000 km) for structuring the genetic diversity of the red mullet fish in the Mediterranean Sea. For the Dinoflagellates *Oxyrrhis marina* and *O. maritima*, Lowe et al. (2010) identified contrasted genetic structures between the Atlantic and the Mediterranean Sea subpopulations.

Protist dispersal and oceanographic circulation

Surprisingly, oceanographic distances did not explain the genetic differentiation of any of the three studied protist species. Since planktonic protists have potential for long distance dispersal and usually large population sizes (Finlay, 2002; Cermeño, Falkowski, 2009; Norris, 2000), at the regional scale of the Mediterranean Sea, the circulation may not impact the genetic differentiation of marine protists (De Wit, Bouvier, 2006; Martiny et al., 2006; Cermeño, Falkowski, 2009; Gibbons et al., 2013; Hellweger et al., 2014). The transport of marine plankton is conditioned by circulation patterns whose features may act as barriers to dispersal (Thornhill et al., 2008). For instance, the Mediterranean Sea presents several frontal structures that correspond to switches in plankton community composition (Ayata et al., 2018). Thus, oceanographic distances were expected to better represent the physical distances traveled by plankton organisms during their transport than geographic distances (Galindo et al., 2006; White et al., 2010). In our study, the geographic distance was both the best and the unique predictor of the genetic differentiation of *P. cordata*. Yet, the computed geographic distances were correlated to both circulation and environmental distances, and could then integrate both information. As a consequence, both isolation-by-distance and isolation-by-environment may indirectly contribute to the genetic patterns observed for *P. cordata*. In addition, for the three species here studied, the statistical models did not explain most of the genetic differentiation, suggesting that other parameters, such as historic (*e.g.*, population size) or biotic factor (*e.g.*, colonial or symbiotic behaviour), should also be tested in order to improve the prediction of genomic differentiation within each of our protist species.

Another explanation for the absence of effect of ocean circulation on the estimated genetic differentiation comes from the fact that the simulated Lagrangian distances (estimated from the transport of inert particles) may not always correlate to the true distances that a planktonic organism would travel from one station to another. Indeed, several biological characteristics can be identified as potentially impacting the Lagrangian simulations. Firstly, size of organisms is one of the main determinants of plankton dispersal (McManus, Woodson, 2012; Richter et al., 2020). For protists, organism behaviour, like aggregation or colony formation, may impact the organism size and, thus, may affect its transport. Moreover, the modification of the buoyancy, through the formation of colonies and aggregates, passive adaptation or physiological regulation (Febvre-Chevalier, Febvre, 1994), may vary the vertical position of the organisms in the water column. Vertical movements, due to directed swimming towards detected resources (Machemer, 2001) or to avoid predation (Bollens et al., 2012), may further impact the vertical distribution of the organisms. This vertical position is an essential criterion for dispersal modeling, as opposite trajectories may result from depth variation when the direction of ocean circulation varies along the vertical (Ayata et al., 2010; Sundelöf, Jonsson, 2012). During our Lagrangian simulations, transport was modelled for passive particles within the surface layer, without the vertical component (2D framework). Yet, vertical movements of protists up to several hundred μm per second have already been observed (Schuech, Menden-Deuer, 2014). This simplification of the vertical behaviour of planktonic protists may have led to biases in our estimate of oceanographic distances. In addition to these biological simplifications, some technical aspects may also impact the validity of the Lagrangian models to properly simulate dispersal of the targeted species. Indeed, several assumptions of the Lagrangian transport model may affect our Lagrangian distances, such as the size of the cell grid in the physical transport model, the density of the transported particles, or the variability of the ocean circulation (*e.g.*, Ayata et al., 2010). However, daily releases of particles were simulated for several seasons and years, and for several transport duration in order to make the simulations more robust against the circulation variability (Berline et al., 2014; Ser-Giacomi et al., 2015). Thus, it is rather the biological variability that may impact the simulation, even if its effect should decrease with the increase of the transport duration (Monroy et al., 2017).

3.2.4 Conclusion. Are we ready for population metagenomics ?

Metagenomics represents an opportunity to improve knowledge about the genetic differentiation of marine plankton, in particular for protists that play a crucial role in ecosys-

3.2. Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations

tem functioning, but remains poorly investigated mainly due to difficulties in isolation and cultivation together with molecular techniques issues. In this study, the genetic differentiation of three protists species with contrasted life history traits was characterized in the Mediterranean Sea based on both metagenomic samples and reference assemblies. Although weak genetic differentiations were observed, we were able to identify different drivers explaining the observed patterns. Our results suggested that *B. prasinos* evolution was constrained by isolation-by-environment process, whereas *P. cordata* one was constrained by isolation-by-distance process. Although several limitations still remain for the use of metagenomics for population genetic studies, such as the need of a reference assembly close to wild population, population metagenomics seems a promising approach for future studies targeting uncultured but abundant and ecologically important species.

3.2.5 Material and Methods

Our global analysis strategy for population metagenomics is summarized in Supp. Mat. B.5. All scripts are available on <https://github.com/opheliedasilva/popmetag>.

Selection of protist species for genomic population studies

In order to study genetic differentiation within marine protist populations, we chose to exploit metagenomic data collected in the Mediterranean Sea during the Tara Oceans (TO) expedition (Alberti et al., 2017). Plankton organisms ranging from viruses to metazoans have been collected through a dozen of size fractions, and protists were predominant from pico- to microplankton size fractions (*i.e.*, from 0.8 up to 180 μm). Preliminary investigations based on (1) a bibliographic study of ecologically relevant protist species in the Mediterranean Sea, (2) the abundance analysis of the V9 eukaryotic metabarcodes in the TO Mediterranean samples (data from De Vargas et al., 2015), and (3) the availability of protist reference genome(s) or transcriptome(s) (Supp. Mat. B.1), led us to select three ubiquitous species : *Bathycoccus prasinos* (Eikrem, Throndsen, 1990), *Pelagomonas calceolata* (Andersen et al., 1993) and *Phaeocystis cordata* (Zingone et al., 1999). Given the size range of these organisms, we focused on the 0.8-5 μm size fraction of TO data, consisting of 13 metagenomics samples available from the surface layer (accession number PRJEB4352 (Carradec et al., 2018) ; see Supp. Mat. B.1 for details), containing on average 185 million reads.

We implemented a bioinformatic pipeline to extract Single Nucleotide Polymorphisms (SNPs) from metagenomic sequences in comparison to reference sequences (here references correspond to genome or transcriptome). It consists of five steps (detailed in Supp.

mat. B.2, and the whole bioinformatic pipeline is available on github) : (1) checking the quality of the metagenomic reads to remove the low-quality ones (Trimmomatic ; Bolger et al., 2014), (2) mapping the metagenomic reads on the reference assembly to pull out reads of the targeted species (bwa mem ; Li, 2013), (3) filtering these aligned reads, first to remove low complexity sequences and avoid spurious alignment (PRINSEQ ; Schmieder, Edwards, 2011) and second to reduce the recruitment of reads from a closely related species (reads aligned with less than 95% identity were removed, Vannier et al., 2016), (4) detecting genomic variants in comparison with the reference assembly, and (5) filtering the variants in order to only keep the SNPs (*e.g.*, indels were removed). Moreover to minimize false positive, SNPs were filtered based on their vertical coverage (*i.e.*, mean number of reads aligned at each position of the assembly) : we only kept SNPs supported by at least 4 reads but less than $\mu + 2\sigma$ of vertical coverage (μ is the mean and σ is the standard deviation of SNP vertical coverage, in order to remove SNPs also abundant in closely related species). The output of our pipeline corresponds to an abundance table of the SNPs in each station. Based on this output, we computed for each SNP the frequency of alleles in each station. F_{ST} were calculated for each SNP at each pair of stations. As the number of allelic frequencies greatly varied from one locus to another, average frequency used in F_{ST} calculation was always computed for the two stations considered and not for all of them. Finally the genomic differentiation between each pair of stations (pairwise F_{ST}) corresponded to the median F_{ST} and was used as the genetic distance between populations. For *B. prasinos*, 20 pairs of stations had no shared SNP and therefore, no genomic differentiation was computed for these couples of stations. For *P. calceolata*, one station had no common SNPs with all the others and was therefore removed from the analysis. For each species, the global F_{ST} was computed as the mean pairwise F_{ST} .

For each species, we created heatmaps to visualize genetic distances between pairs of stations. Dendrograms, built by hierarchical clustering (ward method), were added on the heatmaps to help identify groups of stations genetically close. The missing genetic distances (20 values out of 78 for *B. prasinos*) have been replaced by the mean value to perform the clustering.

Calculations of geographic, environmental and oceanic distances

Firstly, the latitude and longitude of sampling (metadata from PANGAEA, Pesant et al., 2015) were used to compute geographic distances among pairs of stations (*i.e.*, minimal distances between two stations without crossing the lands). Secondly, the hydrological

3.2. Are we ready for protist population metagenomics? Challenges to decipher drivers of genetic differentiation in marine planktonic populations

and biogeochemical environment from each station was characterized with Medatlas-II climatologies (Fichaut et al., 2003), including surface temperature ($^{\circ}\text{C}$), surface salinity (PSU) and surface concentrations of ammonium (mmol.m^{-3}), oxygen (ml.l^{-1}), nitrate (mmol.m^{-3}), nitrite (mmol.m^{-3}), phosphate (mmol.m^{-3}), silicate (mmol.m^{-3}) and chlorophyll a (mg.m^{-3}). A Principal Component Analysis (PCA) was performed on these normalized and standardized variables (Legendre, Legendre, 2012) and from there, the environmental distances between pair of stations were computed as their Euclidean distances in the PCA space (using only significant axes based on Kaiser-Guttman criterion). Thirdly, biophysical transport by ocean circulation was estimated with Lagrangian model simulations to compute oceanographic distances between stations. As no assumptions about mechanisms of dispersal and underlying model settings for each organism could be established, two types of biophysical models were used to simulate Lagrangian transport in the Mediterranean Sea. The first model provides the mean connection time (MCT, in days) for each pair of stations (Berline et al., 2014), while the second model is based on network-theory to relate dispersion with mixing and transport characteristics of the water masses (Ser-Giacomi et al., 2015). The second model calculates connectivity matrices between two stations, estimated as the probability of connection (PC) for a particle leaving a station to end up in another station in a given time period. To estimate PC over time, three dispersal durations were used (3, 6 and 12 months) and averaged to integrate dispersal characteristics at various temporal scales (seasonal, semestrial, annual circulation). As Lagrangian matrices are asymmetric, we chose the maximum PC and the minimum MCT as oceanographic distances for each pair of stations.

Statistical analyses

The links of geographic, environmental and oceanographic circulation constraints was assessed through linear regressions using the following model :

$$y = \beta_0 + \beta_{geo}x_{geo} + \beta_{env}x_{env} + \beta_{pc}x_{pc} + \beta_{mct}x_{mct} + \epsilon$$

The response variable y corresponds to the normalized genetic distances $F_{ST}/(1-F_{ST})$. The intercept coefficient β_0 is the predicted response variable when pairs of stations are not distant in terms of geography, environment and oceanographic circulation. The geographic (geo), environmental (env) and oceanographic (pc and mct) coefficients are the effects of geographic, environmental and oceanographic distances (respectively x_{geo} , x_{env} , x_{pc} and x_{mct}) on the genetic distances. For each of the species, we selected the optimal model by an exhaustive search procedure using the Bayesian Information Criterion. Fisher tests (null versus selected model; $H_0 : \forall i \in \{geo, env, mct, pc\} \beta_i = 0$ have been

carried out to test the overall significance of the linear models. The variables impacting the genetic differentiation were finally identified with Student tests ($H_0 : \beta_i = 0$). The 0.05 p-value threshold was used.

All analyses (data handling, statistical analyses and tests, graphical representations) were conducted in R (v3.5.0; R Core Team, 2017) using the packages : tidyverse (v1.2.1; Wickham et al., 2019) for data handling; ggplot2 (v3.5.2; Wickham, 2016), ggprel (v0.8.1; Slowikowski, 2020), ggpubr (v0.3.0; Kassambara, 2020), maps (v3.3.0; Becker et al., 2018) for graphs and maps; FactoMineR (v1.42; Lê et al., 2008) for PCA; gdistance (v1.3; Etten van, 2017) for geographical distances computation; leaps (v3.0, Lumley, Lumley, 2013) for statistical models selection.

Conflict of interest statement

The authors declare no conflict of interest.

Acknowledgement

We thank Jean-Olivier Irisson (LOV, SU) for help in calculating distances-by-coast. We also thank all the people involved in Tara Oceans for producing the data we used and making them publicly available. This work was funded mainly by our salary as French State agents and therefore by French taxpayers' taxes. Funding for this research was provided by the ModelOmics project of the Emergence program of Sorbonne Université, and partly supported by the project MEGALODOM, part of the MASTODON program from the MITI, CNRS France. O Da Silva was supported by a PhD Grant from the French Ministry of Higher Education, Research and Innovation for three years. SDA acknowledges the CNRS for her two sabbatical years as visiting researcher at ISYEB. L Bittner acknowledges the Institut Universitaire de France for her 5-year nomination as Junior Member (2020-2025).

Contribution

ODS designed the study and write the original manuscript with input from SDA and LB. SDA and LB directed the study. ESG generated probabilities of connection. JL and EP guided for metagenomic analyses. ODS performed analyses. All authors discussed the results and commented the manuscript.

3.3 Conclusion

L'étude présentée dans ce chapitre évoque comment les métagénomomes construits à partir des données métagénomiques pourraient permettre d'augmenter les références dispo-

nibles et potentiellement, étendre l'approche proposée à des espèces abondantes dans l'environnement. Cependant, l'étude de la diversité génétique peut se faire à d'autres échelles que celle du génome ou transcriptome. En particulier, des marqueurs moléculaires peuvent être utilisés. Ces marqueurs sont des séquences présentes chez l'ensemble des organismes ciblés. Elles peuvent être caractéristique d'un taxon (*e.g.*, Foraminifère, Darling et al., 2008; Diatomée, Casteleyn et al., 2010) ou d'un trait fonctionnel (*e.g.*, photosynthèse, Zeidner et al., 2003; diazotrophie, Thompson et al., 2014). Par ailleurs, d'autres traits fonctionnels n'ont aucun analogue génomique actuellement disponible. Il s'agit pour l'essentiel de traits morphologiques (*e.g.*, taille, transparence). Dans la partie suivante, un trait fonctionnel émergeant des données d'imagerie a été ciblé et nous avons étudié dans quelle mesure ce trait fonctionnel pouvait être relié aux conditions environnementales ou aux différents signaux moléculaires.

Phénotype transparent des communautés planctoniques

4.1 Préambule

Dans l'introduction, il a été indiqué que le phénotype des organismes résulte de l'expression génétique. Le phénotype correspond à un ensemble de traits. Il a été suggéré que les approches par traits pourraient conduire à une plus grande généralisation écologique par rapport aux approches taxonomiques (McGill et al., 2006). En effet, les approches taxonomiques supposent que dans des environnements connectés similaires, des communautés similaires seront retrouvées (Chase, Leibold, 2009). De plus, des espèces non apparentées peuvent partager des traits fonctionnels (*e.g.*, Fukami et al., 2005) qui influencent leur survie et leur développement. Ainsi, l'utilisation de traits fonctionnels peut également constituer une information pertinente pour caractériser la structuration de la communauté et conduire à une meilleure compréhension des processus déterministes d'assemblage (McGill et al., 2006; Porcel et al., 2020). L'approche fonctionnelle s'affranchit de la reconnaissance taxonomique, qui est généralement longue, demande des connaissances expertes et peut être impossible dans le cas d'espèces cryptiques. Toutefois, certains traits peuvent également être difficiles à mesurer, tels que les traits physiologiques nécessitant des expériences en laboratoire (Nock et al., 2016). Les technologies d'imagerie peuvent permettre l'accès à de multiples traits morphologiques. Ce chapitre se concentre sur un trait peu, voire pas étudié à l'échelle des communautés : la transparence, qui influence pourtant la survie des organismes (Johnsen, 2001, 2014; Martini

et al., 2020).

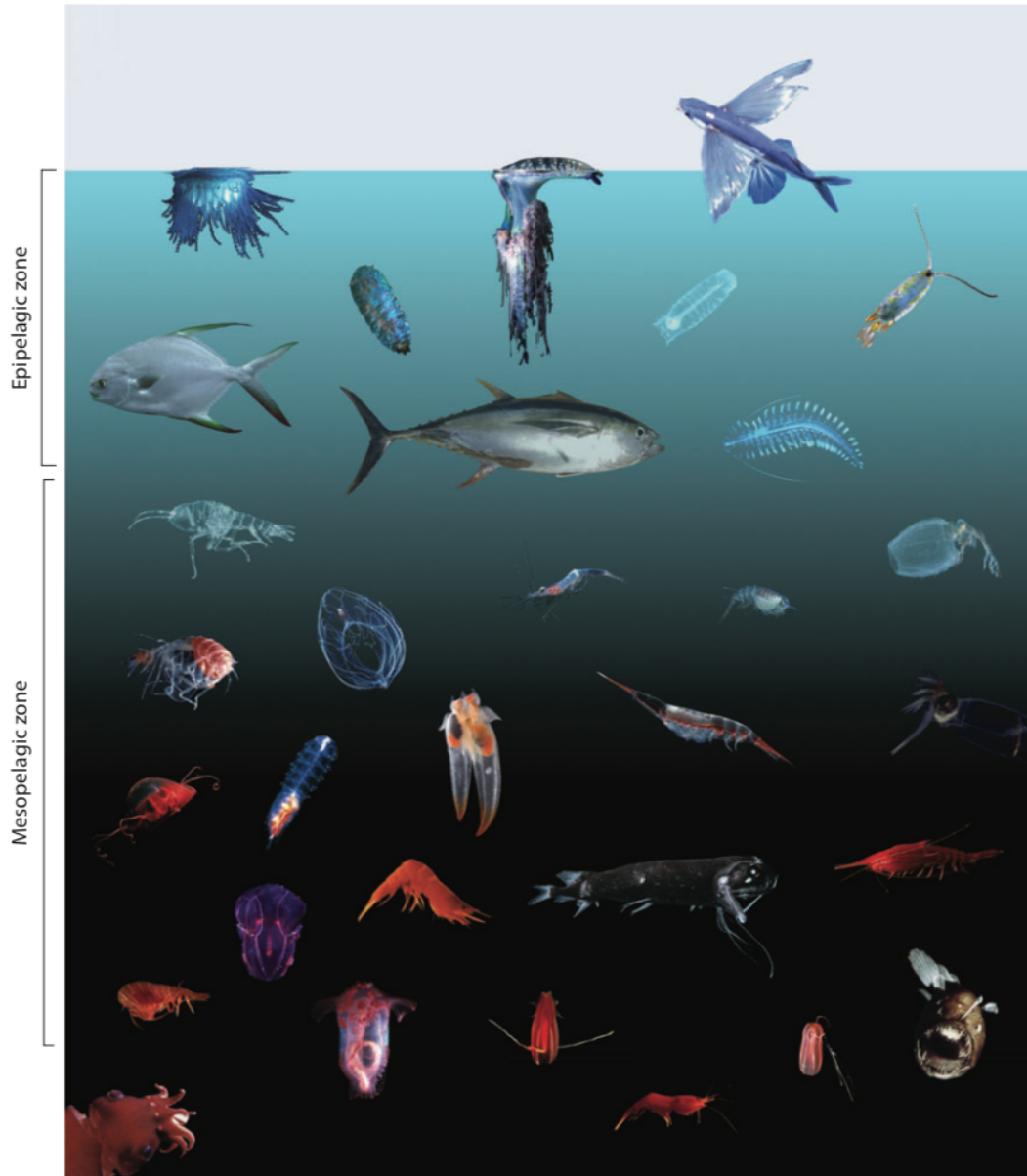


FIGURE 4.1 – Stratégie de camouflage dans l’océan (transparence, miroirs et pigmentation ; contre-illumination non représentée) en fonction de la profondeur. La zone épipelagique de profondeur variable (environ 200 m en eaux claires) est la couche compatible avec la photosynthèse. La zone mésopélagique correspond à la couche où la lumière qui pénètre n’est pas suffisante pour soutenir la photosynthèse (environ 200-1000 m). Extrait de Johnsen (2014)

La transparence est un trait rare en milieu terrestre (*e.g.*, chez certains papillons, Arias et al., 2019), alors qu'elle est répandue dans l'environnement océanique pélagique. En particulier, il a été observé que la transparence se structure verticalement. Les organismes les plus transparents sont dans les eaux les plus superficielles, qui présentent les plus hautes valeurs de radiations lumineuses (Figure 4.1, Johnsen, Widder, 1998; Johnsen, 2014). Ce trait de transparence, en permettant aux organismes de se soustraire à la détection visuelle, est une défense contre la prédation dans le milieu pélagique dépourvu de structures stables permettant de se cacher (*e.g.*, Thetmeyer, Kils, 1995; Johnsen, 2001). La transparence semble donc être un trait capital pour expliquer la structure des communautés. Pourtant, le nombre d'études est limité et aucune à l'échelle des communautés n'a été trouvée. Malgré l'existence de quelques études ayant quantifié en laboratoire la transparence de certaines espèces (basé sur des expériences de transmission lumineuse, *e.g.*, Mackie, Mackie, 1967; Johnsen, Widder, 1998), les mesures disponibles semblent limitées. De plus, aucun marqueur moléculaire de la transparence n'est connu.

Les données d'imagerie pourraient permettre de quantifier la transparence des individus et donc, d'estimer un niveau de transparence à l'échelle des communautés. L'objectif de ce chapitre a été d'identifier un trait de transparence émergeant des données et de le mettre en lien avec les facteurs extérieurs. Les bases moléculaires à l'origine de ce trait ont été explorées. Le reste de ce chapitre correspond à un travail en préparation intitulé *From morphological traits to molecular basis of transparency in the open ocean*.

4.2 From morphological traits to molecular basis of transparency in the open ocean

Authors

Ophélie Da Silva¹, Séverine Martini^{1,2}, Fabien Lombard^{1,3,4}, Lionel Guidi^{1,3}

1 - Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche (LOV), Villefranche-sur-Mer, France

2 - Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM110, 13288, Marseille, France

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

4 - Institut Universitaire de France, Paris, France

4.2.1 Introduction

In the open-ocean, unicellular and multicellular eukaryotic plankton display a myriad of different phenotypes. Its size spans several orders of magnitude, from $<1 \mu\text{m}$ (*Ostreococcus tauri*; Courties et al., 1994; Piganeau, Moreau, 2007) to meters (*e.g.*, *Nemopilema nomurai*, Richardson et al., 2009), which is associated with highly variable metabolic rates (Hirst et al., 2014; Ryabov et al., 2021) and morphological adaptations. Spicules or extensions, as predation defense and fight against sedimentation, are common protists strategies (Vaughn, Allen, 2010; Caron et al., 2012). Plankton can also present calcareous shells or thick cuticles for defense Hamm, Smetacek (2007); Caron et al. (2012). Many planktonic organisms are capable of bioluminescence, involved in the communication signals (*e.g.*, avoid predation, attracting prey, mating; Haddock et al. 2010). External structures such as flagella or cilia allow them to move, even if locomotion is possible without these structures (Chia et al., 1984; Bloodgood, 1995; Heintzelman, 2006). Thus, the morphological diversity of plankton is wide and relates to the constraints of pelagic life. The recent developments of high-throughput imaging technologies provides an unprecedented access to this biological information (Lombard et al., 2019).

The ocean is increasingly imaged, bringing new insights about communities' phenotypes. Previous studies at local or regional scale have shown that communities differ firstly in shape, then in opacity (Vilgrain et al., 2021; Da Silva et al., *in prep.*). Shape is known as a key trait in plankton communities that may induce diversity patterns (Ryabov et al., 2021). Emphasis on size reveals that it is a key factor that both shapes trophic relationships (Woodward et al., 2005; D'Alelio et al., 2016; Ohman, Romagnan, 2016) and impacts ecological processes and biogeographic patterns of planktonic communities (Barton et al., 2013; Richter et al., 2020). Depending on their size, organisms also respond differently to changes in their environment (Barton et al., 2013; Brun et al., 2019). Contrary to the shape, the transparency is more difficult to estimate and has been studied to a lesser extent. However this widely spreaded trait could provide a better understanding of plankton assemblages (Johnsen, 2001).

Transparency is one of the four camouflage strategies found in the featureless pelagic environment (with mirror, pigmentation and counterillumination, Johnsen, 2014). Predominant camouflage strategy varies with depth and transparency is mainly found in surface waters (until 500 m, Johnsen, 2014). It affects interactions within communities by impacting prey-predator relationships (Bagge, 2019; Galloway et al., 2020). In the typically clear open-ocean waters, visual detection depends primarily on contrast (ex-

cept for small organisms) which is modulated by optical phenomena (*i.e.*, absorption and diffusion of light). This whole-body transparency involves every cell and various mechanisms. Cells can have atypical membrane or cytoplasm compositions involving lipid and protein metabolism, respectively (Bagge, 2019). More locally, the body surface can exhibit antireflection coatings and heterogeneities of not well-defined nature that decrease reflectance (Cronin, 2016; Bagge, 2019). Thus, transparency could have specific molecular signatures. However, the morphological, physiological, and metabolic mechanisms of its establishment would benefit to be better understood (Johnsen, 2014; Bagge, 2019).

During the Tara Oceans expedition, eukaryotic communities from picoeukaryotes to metazoans were extensively sampled, creating an unprecedented opportunity for a comprehensive description of the open-ocean ecosystems. Many studies have given meaning to the collected raw biological data. Gene catalogs have been compiled and annotated, providing clues to the genes and functions present and expressed in the environment (Carradec et al., 2018). Interactions between organisms, which modulate the structure of communities and the functioning of ecosystems, have also been inferred (Lima-Mendez et al., 2015; Chaffron et al., 2021). However little attention has been paid on the community phenotype, resulting from the expression of genetic diversity modulated by abiotic and biotic interactions (Stange et al., 2021), and that could be explored from imaging-based morphology. For the first time, transparency will be quantified at the scale of whole communities and replaced in its environmental, biotic and molecular context. The objective of this study is to explore the morphological structure of planktonic communities at global scale to confirm or refute the patterns previously observed at basin-scales. Then, transparency will be targeted to (1) identify potential relationships with environmental factors, (2) detect if it may be related to plankton interaction modifications and (3) establish the molecular bases that could be at its origin.

4.2.2 Material and Methods

Datasets

This study is based on datasets associated with the Tara Oceans circum expedition. More than 200 stations were sampled at different depths and by several size-fractions. Here, we focussed on the metazoan-enriched surface samples since many metazoans exhibit transparency and transparency is the predominant camouflage strategy in the upper ocean.

ENVIRONMENT

Measured and inferred environmental parameters associated with the expedition are available on Pangaea (Pesant et al., 2015). We used 11 parameters informing on the physicochemistry of the water masses (*e.g.*, temperature, salinity), resource availability (*e.g.*, NO₃⁻ concentration, NO₂⁻ concentration, light), biological activity (*e.g.*, chlorophyll-*a* concentration, net primary production) and physical dynamics (mixed-layer depth). Only parameters with less than 10% missing values, which were filled in with the mean parameter value, were selected in order to limit altering the variance of variables (mean was not impacted).

ZOOSCAN IMAGING SYSTEM

Imaging data were used to characterize the morphological diversity of the communities. Samples were collected with a WP2 net (mesh size of 200 µm) and imaged using the Zooscan technology (Gorsky et al., 2010). The morphological descriptors of the objects were computed with the ZooProcess system (Gorsky et al., 2010). The 399,204 objects were automatically annotated at the lowest taxonomic resolution and each annotation was validated or corrected by a taxonomic expert. The objects corresponding to non-living particles were removed from the analysis (155,259 images remaining). The annotated images are hosted in the Ecotaxa infrastructure and available under the accession numbers 377 and 378 (Picheral et al., 2017).

OTU METABARCODING DATASET AND CO-OCCURRING NETWORK

The metabarcoding dataset (V9 18S rDNA) from the Tara Oceans expedition (Ibarbalz et al., 2019) was used to explore the taxonomic composition of communities. This dataset is composed of 1046 samples distributed over 146 stations at several depths and size-fractions (see Alberti et al., 2017 for details). A total of 473303 OTUs were identified. In this study, only surface samples from the 180-2000 µm fraction were analyzed that represented 135 samples and 83288 OTUs.

The co-occurrence network inferred from the Tara Oceans metabarcoding datasets was used to explore potential relationships between organisms (Chaffron et al., 2021). In short, the 115 Tara Oceans stations, having surface and DCM samples for prokaryotic and eukaryotic size-fractions, support the network building. From the normalized OTUs abundance matrices, direct relationships were inferred using FlashWeave method (Tackmann et al., 2019). The network is composed of 20810 nodes and 86026 edges. To be consistent with imaging, only eukaryotes nodes corresponding to the 180-2000 µm size fraction were targeted (2942 nodes) but interactions with all other size classes were analyzed.

EXPRESSION LEVELS OF EUKARYOTIC GENES

From the Tara Oceans metatranscriptomic samples, eukaryotic genes have been reconstructed and annotated. These unigenes are inventoried in the Marine Atlas of Tara Oceans Unigenes (MATOU; Carradec et al., 2018). Metatranscriptomic occurrences, obtained from mapping of metatranscriptomic samples against unigenes, are publicly available (<https://www.genoscope.cns.fr/tara/>). They allow estimation of gene expression and associated functions in the environment. We extracted metatranscriptomic occurrences of samples corresponding to both the 180-2000 μm size-fraction and the surface layer. We used the pFam annotation of each unigene to infer the corresponding GO terms.

Analyses

The analyses were conducted in the R programming environment (v.4.0.3; R Core Team, 2017) and used several packages : tidyverse to manipulate and visualize data (v.1.3.0, Wickham et al., 2019), bestNormalize for yeo-johnson normalization (v1.8.0.9000, <https://github.com/petersonR/bestNormalize>), FactoMineR to perform PCA (v.2.3, Lê et al., 2008), morphr to represent the morphospace (v.0.5.0; <https://github.com/jiho/morphr>), FD for computing functional indices (v.1.0-12, Laliberté et al., 2014).

ENVIRONMENTAL CHARACTERIZATION

A Principal Component Analysis (PCA, Legendre, Legendre, 2012) was carried out to identify in order the main axes (in terms of maximizing variance) that best discriminate the stations based on their environmental characteristics. The PCA was performed on the normalized and standardized dataset composed of 11 environmental variables. The significant principal components, corresponding to linear combinations of the initial variables, were selected based on the Kaiser-Guttman criterion.

MORPHOLOGICAL SPACE

A PCA was also performed to explore the morphological structure of the communities. The dataset is composed of 174,950 images and 38 descriptors (listed in Supplementary Information C.1). To satisfy normality PCA requirement, the descriptors were Yeo-Johnson transformed (Weisberg, 2001; Atkinson et al., 2021). The Kaiser-Guttman criterion was used to select significant principal components mirroring synthetic morphological traits that explain variations in communities' phenotypes. This reduced space is the so-called morphological space hereafter.

FUNCTIONAL DIVERSITY

The morphological space was a functional space for which each dimension was a new functional trait composed by a linear combination of the morphological descriptors. In this functional space, organisms were placed along functional traits according to their functional trait values. For the functional space, functional diversity at each station has here been described based on five functional indices defined as follows. The functional identity is the mean value of functional traits at the community level (Mouillot et al., 2013). The functional richness is the volume of multidimensional functional space occupied by the community (Villéger et al., 2008; Mouillot et al., 2013). The functional divergence is the spreading of the community along the functional space (Mason et al., 2005; Villéger et al., 2008; Mouillot et al., 2013). The functional dispersion is the mean distance to the centroid within the functional space (Laliberté, Legendre, 2010; Mouillot et al., 2013). Finally, the functional evenness is the regularity of the trait value distribution in the community within the functional space (Laliberté, Legendre, 2010; Mouillot et al., 2013).

COMPOSITION AND NETWORK EXPLORATION

V9 18S metabarcoding data from surface samples of the 180-2000 μm size fraction were filtered and only eukaryotes were kept. Then, the relative abundances were calculated to explore the community composition. In addition, nodes of the co-occurrence network corresponding to OTUs found in the 180-2000 μm size-fraction surface samples were characterized. Their degree (*i.e.*, number of adjacent edges) and the transitivity (*i.e.*, probability that the adjacent vertices of a vertex are connected, also called clustering coefficient) were calculated. These coefficients respectively indicate whether the OTU co-occurs to few or many other OTUs and whether it tends to form clusters with the OTUs with which it co-occurs.

MOLECULAR EXPRESSION ANALYSIS

From the metatranscriptomic unigene occurrence datasets, the relative expression of each unigene was computed. The Pfam annotations of unigenes were retrieved and linked with GO terms. To reduce the dimensionality of the dataset, unigenes were grouped according to their GO terms. The Bray-Curtis distances between samples were estimated. Then, a hierarchical clustering (Ward linkage, Murtagh, Legendre, 2014) was performed to group the samples with similar unigenes expression levels. The number of clusters to interpret was chosen based on the Elbow criterion.

4.2.3 Results

Environmental gradients at global-scale

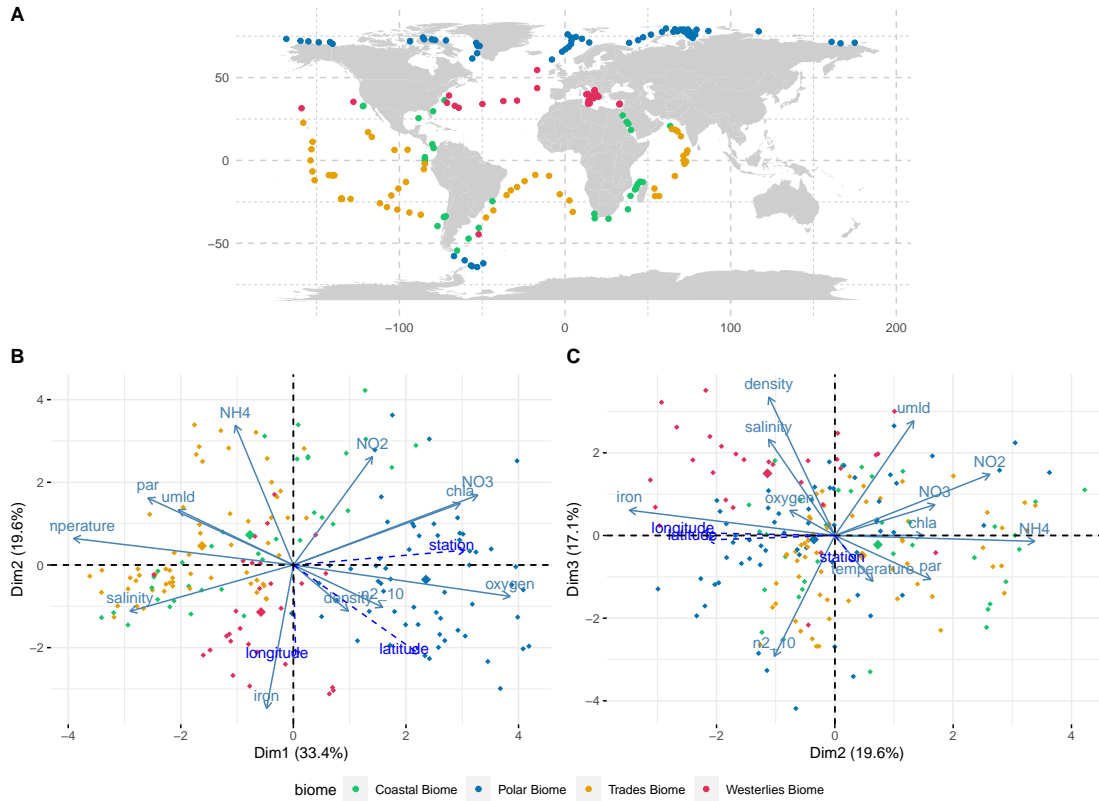


FIGURE 4.2 – Environmental organization of the stations. (A) Geographical positions of the stations and their projections in the plans composed by (B) the two first principal components (PC1 and PC2) and (C) the second and third principal components (PC2 and PC3) of the Principal Component Analysis performed on the environmental variables. Colors of the points represent biomes to which stations belong. Arrows represent projection of environmental variables in the reduced space composed by the three significant principal components (Kaiser-Guttman criterion).

The stations belonged to four biomes (Figure 4.2A). The Polar biome included 53 Arctic stations and 6 Antarctic stations. The Coastal biome included 10 Indian Ocean stations, 4 North Atlantic stations, 4 North Pacific stations, 4 Red Sea stations, 6 South Atlantic stations and 7 South Pacific stations. The Trades biome was composed of 14 Indian Ocean stations, 7 North Pacific stations, 11 South Atlantic stations and 27 South Pacific stations. Finally, the Westerlies biome included 16 Mediterranean stations, 9 North Atlantic stations, 2 North Pacific stations and 1 South Atlantic station. The environmen-

tal conditions at these stations were studied using a PCA based on 12 biogeochemical variables. The environmental PCA revealed three significant axes representing 70.5% of the variance (Figures 4.2B and C). The first principal component PC1 (33.2% of variance) separated richer, more productive and oxygenated waters ($PC1 > 0$) from warmer, saltier, lighter ones ($PC1 < 0$). The PC1 mostly distinguished Polar stations from the Indian Ocean and Red Sea stations (Supplementary Information C.2). The second principal component PC2 (19.6% of variance) segregated nitrogen-enriched waters ($PC2 > 0$) from iron-enriched ones ($PC2 < 0$), separating South Pacific and Southern Oceans stations from Westerlies biome (mostly Mediterranean, Supplementary Information C.2) stations. Finally, the third principal component PC3 (17.7% of variance) separated saltier, denser and with deeper mixed layer depth (MLD) waters ($PC3 > 0$) from fresher and N2-enriched ones ($PC3 < 0$). As expected, the environmental organization was structured at global-scale according to geography. Stations belonging to the same biome or region tended to share more similar biogeochemical characteristics than stations belonging to two different biomes or regions.

Morphological structure of planktonic communities at global-scale

GLOBAL MORPHOLOGICAL STRUCTURE

The imaging-based PCA had six significant principal components. Only the first four principal components, that explained more than 5% of variance, have been described here (more in Supplementary Information C.3). The four first significant principal components represented 83.8% of the morphological variance (Figure 4.3). The first morphological principal component PC1 (38.6% of variance) reflected the organisms' size (*e.g.*, perimeter, area; $PC1 > 0$: large, $PC1 < 0$: small). Mainly descriptors of grey levels contributed to the second principal component PC2 (29% of variance) reflecting the opacity ($PC2 > 0$: less opaque, $PC2 < 0$: more opaque). The sr and kurtosis descriptors mainly contributed to the third principal component PC3 (8.6% of variance). They described standard deviation in comparison to the gray levels ranges (sr descriptor) and the distribution flattening of gray level values (kurtosis descriptor). Thus, PC3 testified to the opacity homogeneity of organisms ($PC3 > 0$: heterogeneous, $PC3 < 0$: homogeneous). Finally, the fourth principal component PC4 (7.6% of variance) discriminated the elongated organisms ($PC4 > 0$) from the circular ones ($PC4 < 0$). Thus, among these four principal components, two corresponded to shape-traits (PC1 and PC4) and the two others were opacity traits (PC2 and PC3). In the following, the analysis will therefore be based on PC2 and PC3, used as proxies for global transparency and transparency

4.2. From morphological traits to molecular basis of transparency in the open ocean

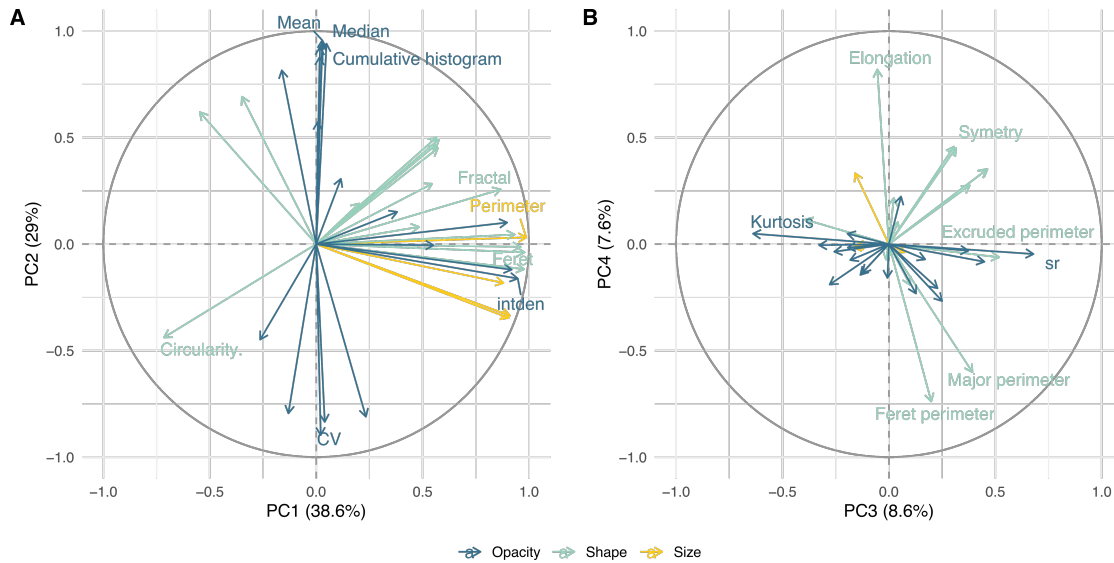


FIGURE 4.3 – Principal component analysis performed on the morphological descriptors. The morphological space is reduced to six significant principal components. Projections on the PCA plans composed by (A) the first and second principal components, (B) third and fourth principal components and (C) fifth and sixth principal components. Morphological descriptors are represented as arrows and the colors indicate type (blue : opacity descriptors, green : shape descriptors, yellow : size descriptors).

heterogeneity of organisms (PC2 and PC3 respectively).

TRANSPARENCY IN THE OPEN-OCEAN

The morphological PC2, proxy of transparency, also captured a size-signal with organisms that tended to be larger for negative PC2 values. PC2 was thus a composite trait between transparency and size. In contrast, PC3 appeared to be more independent of the shape or size of the organisms (Figure 4.4A). The PC2 values are higher in the Westerlies biome (Figure 4.4B). In addition, the Mediterranean Sea, the North Atlantic Ocean and the Red Sea showed higher PC2 values (Figure 4.4C). In these ocean areas, the organisms are more transparent than elsewhere. To the contrary, the North Pacific Ocean presented lower PC2 values. Lower PC3 values were associated with the polar biome, especially to the Arctic Ocean, but also to the Indian Ocean (Figures 4.4B and 4.4C). This suggested that the transparency or opacity of organisms was globally more heterogeneous in these regions. Thus, there appeared to be a geographic structure of transparency.

FUNCTIONAL DIVERSITY OF TRANSPARENCY

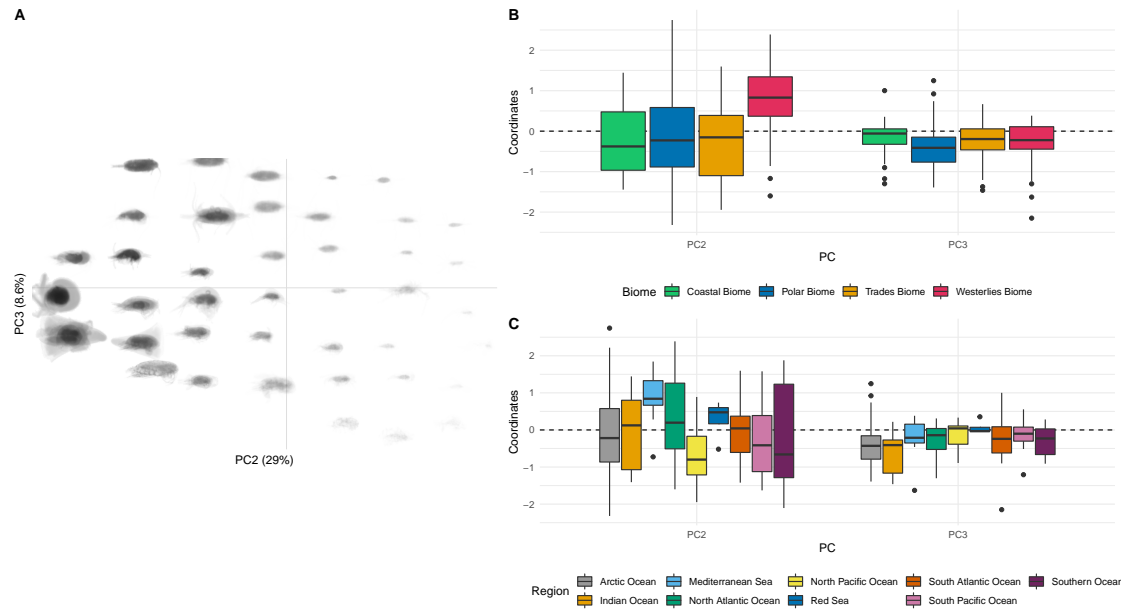


FIGURE 4.4 – Transparency proxies in the global ocean. (A) Morphological space based on the second and third principal components (PC2 and PC3) of the morphological PCA, and coordinates on the PC2 and PC3 for the oceanographic (B) biomes and (C) regions, which are indicated by the colors of the boxplots.

Communities were less transparent in the tropics and subtropics than at high latitudes (Figure 4.5, CWM PC2). The transparency of the communities was more heterogeneous in the subtropics (Figure 4.5, CWM PC3). The communities were therefore rather transparent and uniform at high latitudes, opaque and uniform at low latitudes and transparent and heterogeneous at mid-latitudes. Functional richness was higher in the subtropics meaning that the functional space occupied by organisms was larger. On the other hand, divergence and functional equitability were lower towards the Arctic region. Thus, in this region the proportion of organisms with the most extreme trait values was less important and the distribution of abundances in the functional space was less regular. Finally, the dispersion was lower in the subtropics, which reflected a lower abundance-weighted deviation of trait values from the center of the community’s functional space.

Taxonomic signal of transparency

Depending on the taxonomic groups, the transparency values were different (Figure 4.6). On PC2, Cnidaria (Abylidae, Diphyidae), Tunicata (Fritillariidae, Oikopleuridae) and some protists (Neoceratium, Acantharea, Pyrocystaceae, Collodaria) were rather found

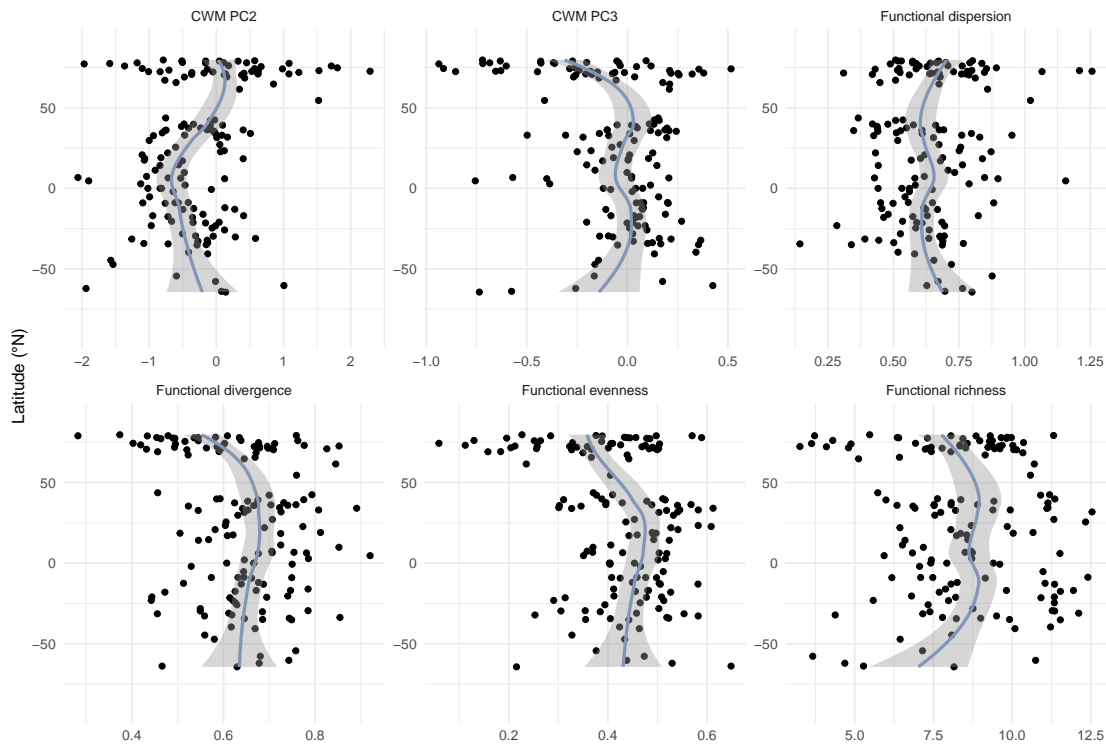


FIGURE 4.5 – Latitudinal distribution of functional diversity indices. Six indices were calculated : Community Weighted Mean based on morphological PC2 (CWM PC2), Community Weighted Mean based on morphological PC3 (CWM PC3), the functional dispersion, functional divergence, the functional evenness and the functional richness.

with positive coordinates whereas mainly Mollusca, Crustacea and Foraminifera were associated with negative coordinates. Moreover, positive coordinates on PC3 mainly corresponded to Copepods (*e.g.*, Centropagiadae, Harpacticoida, Candaciidae, Copepoda, Calanus, Oithonidae). On the contrary, the negative coordinates on PC3 corresponded to gelatinous plankton (*e.g.*, Aglaura, Physonectae, Siphonophorae, Actinaria, Rhopalonema velatum, Dolioda, Cymbulia peroni). Thus, the transparency was structured according to taxonomy.

Molecular exploration of transparency

Each unigene has been assigned a biological process (as well as a molecular function and a biological component) through the Gene Ontology. From the unigenes' relative expressions grouped by biological processes, eight clusters were identified (Figure 4.7A). These clusters grouped similar stations in terms of expression profiles. For each cluster,

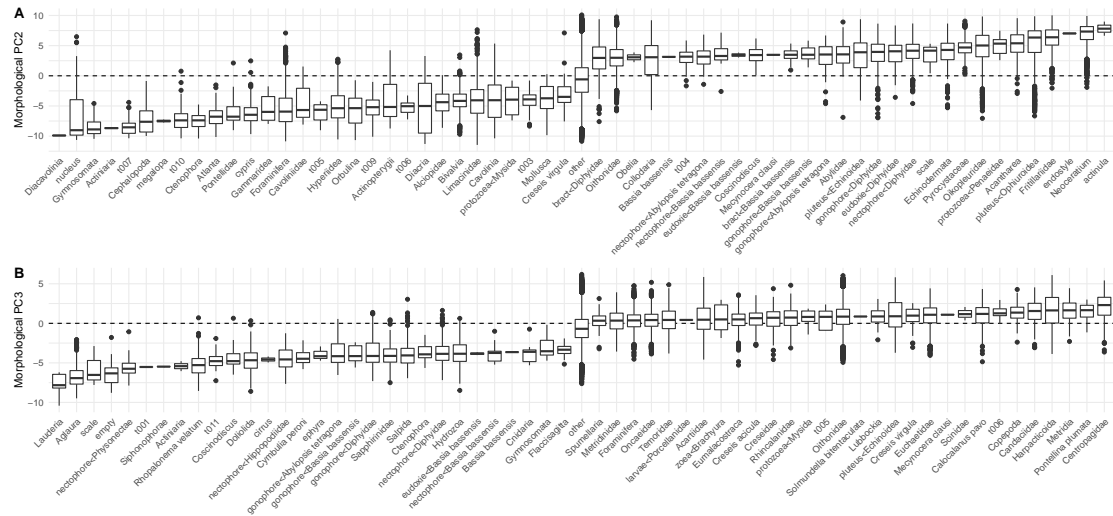


FIGURE 4.6 – Taxonomic organization of transparency. Coordinates of taxonomic groups on the (A) second and (B) third principal components (PC2 and PC3 respectively) of the morphological PCA. Only the 40 taxonomic groups with the most extreme median coordinates (the 20 highest and the 20 lowest) are represented. The other taxonomic groups are merged in the 'other' category.

transparency was investigated (Figure 4.7B for overall transparency, PC2; Figure 4.7C for transparency heterogeneity, PC3). The clusters did not have different community transparency values. Note this analysis was based on the biological processes assigned to unigenes. The same analyses based on both molecular function or biological component annotations led to the same results though.

4.2.4 Discussion

In this study, the morphological structure of the communities was investigated. The communities were differentiated first by their size, then by their opacity, used here as a proxy for transparency which is a largely under-explored trait at the community scale. Here, the spatial organization of community transparency has been described. It showed spatial gradients that could be related to different abiotic and biotic environments. Also, depending on the environment the functional diversity of this trait was variable. Finally, the molecular basis of this trait was explored in order to identify potential molecular markers. But none could be defined at this stage.

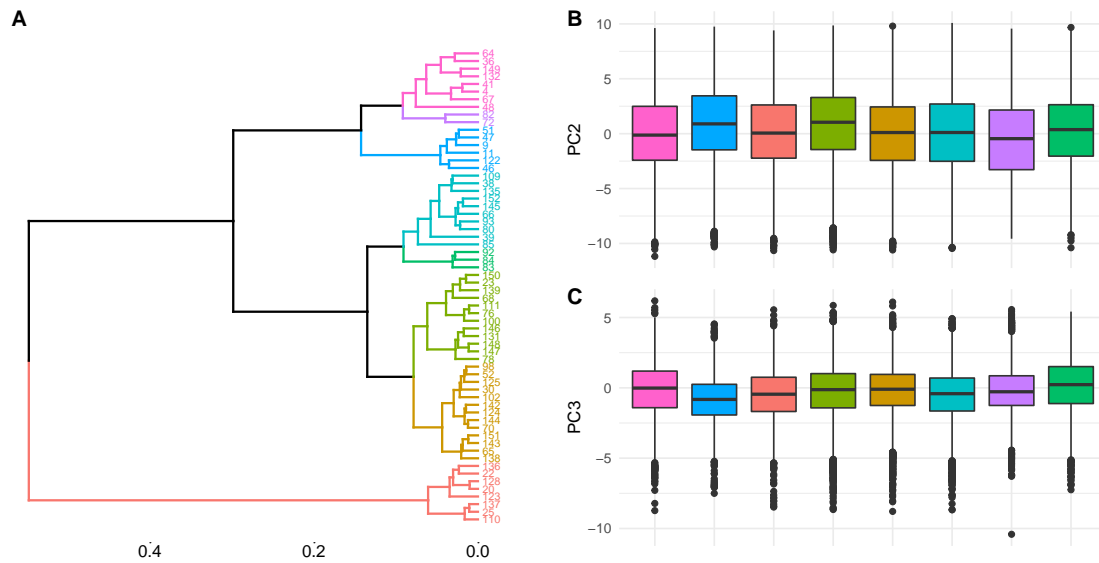


FIGURE 4.7 – Transparency traits according to gene expression of biological processes. (A) Relative unigene expressions grouped by biological processes were used for clustering (Bray-Curtis distances, Ward method). Morphological (B) PC2 and (C) PC3 for each cluster defined in A (Elbow criterion). Colors represent the expression cluster membership.

Main limits

Although transparency is an easily detectable characteristic, it is complex to understand mechanisms generating it. Here we tried to identify the link between transparent phenotype and gene expression. Challenges arose from both genetic data and imaging data. Transparency is widely distributed among planktonic organisms and is found in many lineages (Johnsen, 2001). This morphological feature is assumed to be an evolutionary convergence. Thus, there would have been several independent acquisitions during evolution potentially involving different molecular bases (*e.g.*, transparent collagen for Cnidarians, Cellulose for tunicates; Arenas Gómez et al., 2020). These several transparent mechanisms must coexist within communities and their relative importance will depend on the organisms' abundance that is modulated by the abiotic and biotic constraints of the environment. The coexistence of mechanisms and their variability complexify the signal detection. Moreover, the study between phenotype and gene expression was based on 68 samples that may limit statistical power of the analyses.

The relationship between transparency level and gene expression was studied from the unigenes including non annotated ones. The omics datasets were very large. To reduce

their dimensionality, the genes were clustered according to their functional annotations. But aggregating data based on their annotations rather than their expression profile may have flattened differential expressions. In addition, as many unigenes were not annotated (*i.e.*, unknown genes), this procedure may have led to mask the molecular transparency signals by grouping highly different genes in terms of function, taxonomy and expression levels. Deeper exploration of unknowns is crucial to better understand the molecular basis of transparency (Bernard et al., 2018). There may also be discrepancies between the actual transparency in the water column and the transparency measured by imaging technologies (here Zooscan, Gorsky et al., 2010). Indeed, due to formalin preservation of the samples, quantification of transparency may have been biased (*e.g.*, pigment alteration). Imaging biases to quantify transparency were also detected via taxonomic analysis where some taxa, known to be transparent, were found to be opaque (*e.g.*, the Ctenophora with negative morphological PC2 values). Additionally, some organisms are able to actively adapt their transparency level (*e.g.*, some Cephalopods, Zylinski, Johnsen, 2011). While the associated gene expression could be sampled, the imaging method used for this study would fail. In situ, alive imaging technologies would be more suitable to image this active transparency (*e.g.*, In situ ichthyoplankton imaging system, Cowen, Guigand, 2008; Underwater Vision Profiler, Picheral et al., 2010; Continuous Plankton Imaging and Classification Sensor Grossmann et al., 2015; Zooglider, Ohman et al., 2019; Scripps Plankton Camera system Orenstein et al., 2020).

Transparency in the open-ocean

MORPHOLOGICAL INVESTIGATIONS OF TRANSPARENCY

Here, the spatial organization of transparency and its functional diversity was revealed at global-scale. In particular, we found that the transparency showed a non-linear latitudinal structure with more transparent organisms at subtropics. Transparency often results from a compromise between fight against predation (*i.e.*, camouflage, Johnsen, 2001; Cronin, 2016) and protection against UV radiations (*i.e.*, photoprotection via synthesis of pigments; Hansson, Hylander, 2009; Alvarez et al., 2019; Bashevkin et al., 2019). In the tropics, waters are clear and the light exposure is very important. Polar environments experience dramatic shifts in light exposure with high and long duration sunshine in summer and little to no sunshine in winter. In this environment, the pigment content is likely to be related to acclimation processes rather than community change (Alvarez et al., 2019; Lee et al., 2019). Subtropic areas are characterized by intermediate light characteristics with moderate sunshine. However, in this study, the highest level of

transparency has been estimated in this subtropical environment. As subtropical waters are very clear (*i.e.*, little turbidity), visual detection may be favored. Hence, the transparency related to the absence of pigment may limit predation and could therefore be a common strategy. In polar environments, as waters are exposed to variable and extreme light conditions, the balance between photoprotection and visual detectability could lead to the observed heterogeneous transparency. In tropics, the light constraint may be so intense that organisms should exhibit photoprotective pigments even if it increases the risk of being predated, which may explain communities that appear more opaque.

The spatial organization of transparency found in this study may also be linked to the taxa present in the environment. Indeed, the gelatinous zooplankton, which includes for example most of the Cnidaria, Tunicates or Ctenophores, is characterized by a transparent water body. The gelatinous zooplankton biomass is globally higher in the subtropics and in the North polar region (Lucas et al., 2014), which is consistent with the higher transparency signal detected here. Gelatinous zooplankton are distributed at global scale and face a wide range of environments (Lucas et al., 2014; Lilley et al., 2011). Their distribution depends on physical conditions that may lead to local aggregation phenomena (Graham et al., 2001; Luo et al., 2014). The environmental factors forcing their distribution vary among taxa (Lucas et al., 2014; Luo et al., 2014, 2020). These organisms are long-underestimated but important contributors to the biological carbon pump (Luo et al., 2020). Characterizing community transparency may improve carbon flux estimates.

The polar environment is associated with weak niche differentiation and therefore with strong niche competition. A part of the resources would be unused leading to low productivity, low reliability and high invasion capacity (Mason et al., 2005). Subtropical ecosystems would also present little niche differentiation, but the functional space would be more widely and more equitably used. Communities would be more productive, reliable and less vulnerable to invasion. Finally, tropical ecosystems would be less functionally rich but more divergent and equitable. Functional diversity therefore presents a latitudinal structure. It is however important to note that biases in the functional diversity indices used have been identified and that this functional analysis should be carried out again with indices which do not depend on the structure of the functional space (Fontana et al., 2018).

Here, the transparency was investigated in the surface layer. However, it may be relevant to study the vertical structure of transparency too. Indeed, it has been shown, for

example, that gelatinous communities vary with depth (Raskoff et al., 2005) and that even if the biomass of gelatinous zooplankton is much higher at the surface, its relative abundance in the communities is constant (Lilley et al., 2011). Furthermore, planktonic organisms, which are capable of vertical movement, are not fixed in the layer and can make daily migrations from and to deeper layers (Graham et al., 2001; Hays et al., 2001). These diel vertical migrations vary in amplitude and result from the balance between predation and being predated (Hays, 2003; Pearre, 2003).

MOLECULAR INVESTIGATIONS OF TRANSPARENCY

From the molecular perspective, as no clear correlation was found between imaging-based transparency and gene expression, no molecular marker had been identified for this trait. Thus, while the study of transparency would be limited by molecular approach at the community scale, imaging technologies could be powerful tools to overcome this challenge. The two data sources, imaging and metaomics, are therefore highly complementary and seem to allow the study of different facets of the communities. This argues in favor of the joint study of different datasets to better understand how the communities are structured.

Transparency is difficult to measure at the community level. The measurement of transparency is traditionally done in the laboratory from light refraction and absorption experiments. As transparency is therefore a time-consuming trait to measure, data are available for a subset of the planktonic diversity, leading to a biased view of this trait within the communities. But extending such traditional measurements to a whole community therefore seems very difficult to impossible. Thus, imaging technologies offer a great opportunity to collect a large amount of data about such a hard trait to quantify at the community level and this, for organisms spanning into wide size ranges. In order to provide comparable measurements with previously acquired data, it would be necessary to calibrate and correct the data from quantitative imaging with the data measured in the laboratory. This could make it possible to complete the databases for many taxa.

4.2.5 Perspectives

OTU composition and transparency

Transparency was also explored in regard with community composition based on surface samples of the 180-2000 μm size-fraction from the Tara Oceans V9 18S metabarcoding dataset. Each of the 83288 OTU was assigned a transparent/opaque/other label (3515/13885/65888 OTUs respectively) based on the taxonomic signal identified in Fi-

4.2. From morphological traits to molecular basis of transparency in the open ocean

figure 4.6. Then the relative abundance of each type was calculated (Figure 4.8). High morphological PC2 values were associated with higher transparencies. The stations corresponding to higher mean morphological PC2 values were assumed to be composed of more transparent-type organisms than stations with lower mean morphological PC2 values. This was not observed (Figure 4.8A). For the mean morphological PC3 values at the different stations, which corresponds to the level of gray level heterogeneity, the expected was more complicated to define. However, no relationship seemed to emerge with morphological PC3 either (Figure 4.8B).

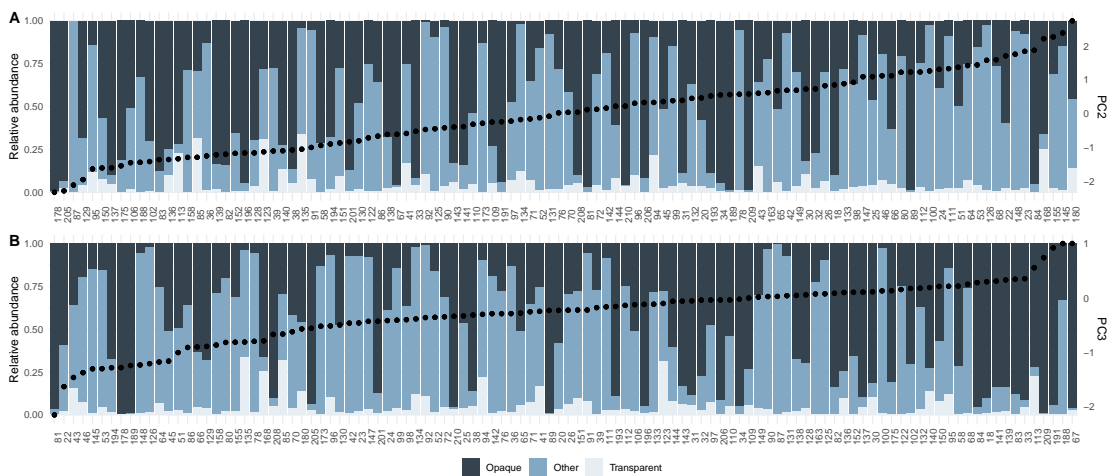


FIGURE 4.8 – Community composition based on metabarcoding data. Colors represent the transparent/opaque/other characteristics of OTUs. Points indicate the mean values of PC2 (top) and PC3 (bottom) at each station. The x-axis is ordered according to the mean PC value (stations by ascending order).

These results are a first step and should be refined. For the moment, only a subpart of the OTUs has been tagged to a transparency assignment. Although some taxa with body composition (especially water content) varying with size (Kiørboe, 2013) may show higher transparency variability, qualitative estimates based on expert knowledge may remain possible. From the assignments of transparency class to all organisms, the best transparency prediction model based on significant morphological principal components could be fitted and selected. Thus, this would (1) quantify the performance of the model and thus validate the use of this imaging technology to measure transparency and (2) estimate a more accurate community-level transparency index. From this more refined estimate, transparency could then be finely related to associated gene expression.

Interactome shaping by the transparency

Since transparency is a defense strategy against predation by visual detection, it is assumed to modulate biotic interactions. Thus, transparency has been explored in the context of co-occurrence networks, proxy of interaction networks. Some nodes of the co-occurring network were targeted based on previous taxonomic results (Figure 4.6). Among the 2942 eukaryote nodes of the 180-2000 μm size-fraction, 597 were assigned as opaque, 125 as transparent and 2223 as other. Two node metrics, the degree (*i.e.*, number of adjacent edges) and the transitivity (*i.e.*, probability that the adjacent vertices of a vertex are connected, also called clustering coefficient), were inspected for characterizing network nodes according to their transparency (Figure 4.9). Data did not meet the assumptions of normality. So Kruskal-Wallis tests were performed to identify whether the groups differed in degree and transitivity. Both tests were significant (degree : p-value of 6.01×10^{-5} , transitivity : p-value of 7.65×10^{-6}). Pairwise comparisons were then performed with Dunn's test to identify the groups that differed. The transparent group differs from the other two in terms of degree and transitivity. The opaque group differs from the other group in terms of transitivity. Thus, transparent organisms seem to live with more organisms (*i.e.*, more co-occurrences) in stable communities (*i.e.*, into clusters of interacting organisms).

Transparency must have a key role in ecosystem structure. For example, gelatinous zooplankton, that exhibit varied trophic modes, maintain complex sharing and competition relationships within niches (Chi et al., 2021). In addition, it has already been established that these taxa links lower (picoplankton) and upper (fishes) trophic levels. In particular, they can exert bottom-up or top-down control on the ecosystems (Pitt et al., 2004; West et al., 2009; Hosia et al., 2015). More generally, as transparency influences visual detection ability, it impacts predation defense. This could explain why transparent taxa must be able to coexist with more other distinct taxa. Thus, transparency seems to be an important factor in the structure of communities (biotic interaction) and to condition the patterns of diversity. It is therefore necessary to study transparency further. A link between transparency and network structure was identified from this first coarse analysis. It would therefore be interesting to further explore this relationship to identify the key network characteristics associated with variations in transparency (Poisot, Gravel, 2014). This could be done by transposing the transparency class assignment as proposed in the previous section. Understanding, on the one hand, the impact of transparency on trophic co-occurrences (proxy of trophic interactions) within communities and, on the other hand, deciphering the environmental forcings on transparency could allow a better

estimate of the vulnerability of community structures to changes in their environment.

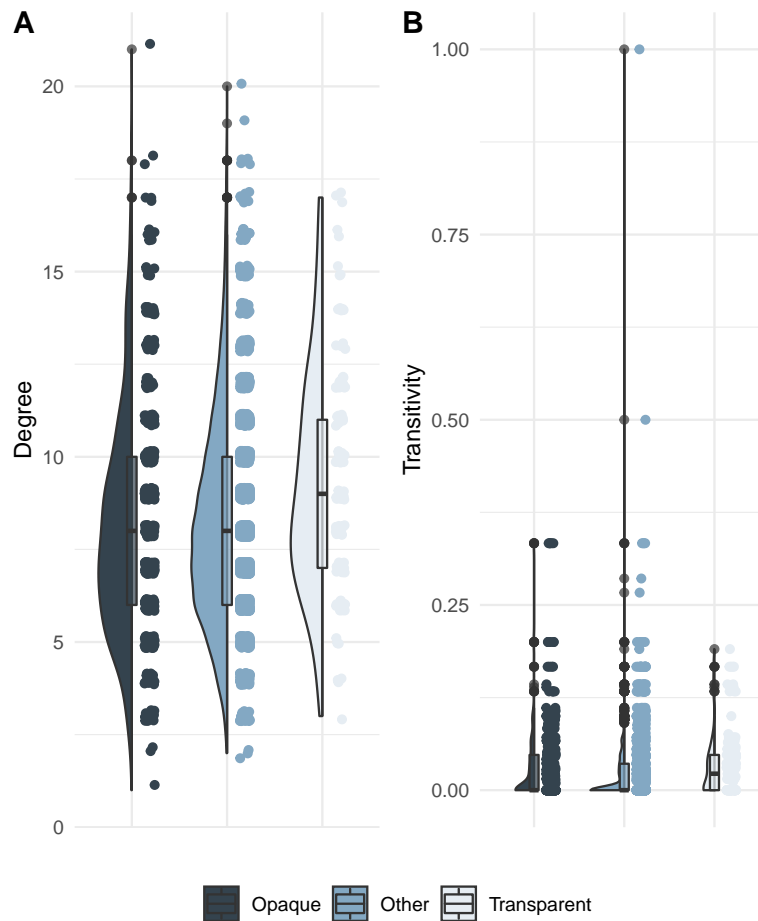


FIGURE 4.9 – Distribution of the node metrics in the co-occurring network according to their transparency. (A) Node degree and (B) transitivity were computed. Colors indicate the Opaque/Transparent/Other label assigned to each node.

To conclude, studying transparency at the community level in the ocean is a real challenge. However, this trait is involved in different ecological processes and, by reflecting the response of organisms to their environment and/or to biotic pressures, this trait could help to explain the structure and therefore the functioning of communities. Imaging technologies appear to be promising to characterize transparency. They are complementary to metaomics data which have here failed in the trait study, but could ultimately allow the identification of new markers. This would therefore make it possible to monitor the genetic diversity of this trait and therefore the phenomena of plasticity or adaptation in

the environment.

Acknowledgements

We thank all the people involved in *Tara Oceans* for producing the data we used and making them publicly available. This work was funded mainly by our salary as French State agents and therefore by French taxpayers' taxes. O Da Silva is supported by a PhD Grant from the French Ministry of Higher Education, Research and Innovation for three years.

Contributions

ODS, SM, LG and FL designed the study. ODS wrote the manuscript. LG and FL directed the study. ODS performed the analysis.

4.3 Conclusion

Dans ce chapitre, un signal de transparence a été caractérisé. Il a été identifié que la transparence semble d'une part, liée à grande échelle à l'environnement et d'autre part, porter un signal taxonomique. Cependant, la relation entre le phénotype transparent et le génotype des communautés n'a pas été établie. Cela peut être dû au caractère polyphylétique de ce trait ou au caractère semi-quantitatif des données méta-omiques. Les données utilisées ici sont donc hétérogènes et complémentaires. En particulier, elles peuvent être structurées différemment à l'échelle des communautés. Dans la dernière partie, les différents types de données, qui fournissent des informations sur des composantes différentes des écosystèmes planctoniques, ont été analysés simultanément, afin d'explorer leur complémentarité.

Structure globale de l'écosystème planctonique

5.1 Préambule

Dans l'introduction, la complexité des écosystèmes a été présentée. Ces écosystèmes sont composés de très nombreuses entités en interaction, caractérisables à de multiples échelles spatio-temporelles et biologiques, et sur lesquelles différents processus évolutifs et écologiques s'exercent. Ainsi, comprendre le fonctionnement et les mécanismes qui structurent les écosystèmes est un réel défi, qui pourrait bénéficier des approches intégratives (Michener and Jones, 2012, Garcia, 2019, Zipkin et al., 2021). Comme souligné dans l'introduction, ces dernières décennies ont été marquées par une augmentation très rapide de la quantité et de la diversité de données acquises. Ces données massives pourraient permettre de mieux comprendre les écosystèmes planctoniques. Dans les deux précédents chapitres, différentes sources de données ont été utilisées. Des structures de l'écosystème différentes ont été identifiées, mettant en évidence que les multiples sources de données contiennent des informations variées mais potentiellement complémentaires. Intégrer l'ensemble de ces données dans une même analyse pourrait fournir un signal plus riche et représentatif de la structure réelle de l'écosystème. Cette étude a été menée dans la région subtropicale de l'Océan Atlantique Nord.

L'océan Atlantique Nord abrite le système *Atlantic Meridional Overturning Circulation* (AMOC), profondément impliqué dans la régulation du climat et dont les modifications

de fonctionnement peuvent avoir un impact sur les écosystèmes océaniques et continentaux à l'échelle globale (Caesar et al., 2018; Liu et al., 2020). Dans sa partie subtropicale, l'océan Atlantique Nord présente un large gyre anticyclonique oligotrophe supporté par plusieurs courants (*i.e.*, Gulf Stream, front subpolaire, courant des Canaries, courant nord-équatorial). Cette région présente des gradients environnementaux et biologiques (Aiken et al., 2000) et des régions stables aux bordures fluctuant spatialement au cours du temps y ont été délimitées (Longhurst, 2010; Reygondeau et al., 2013). De nombreuses études ont été réalisées afin de mieux comprendre cet écosystème. Par exemple, la zone a été échantillonnée pendant plus de 30 ans grâce au *Continuous Plankton Recorder* (CPR), ce qui a permis un suivi temporel de plusieurs groupes planctoniques (Batten et al., 2019). Ces études ont été le socle d'une augmentation du savoir sur la diversité dans la zone et de ses réactions au changement climatique (*e.g.*, Beaugrand et al., 2009; Barton et al., 2016; Vannier et al., 2016; Brun et al., 2019). Cependant, l'échantillonnage CPR présente un biais spécifique aux espèces en raison du comportement d'évitement des organismes, ce qui peut entraîner des problèmes dans la quantification de la composition de la communauté (Batten et al., 2019). De plus, l'échantillonnage CPR peut endommager les organismes (en particulier les gélatineux, Batten et al., 2019), ne fournissant pas d'informations morphologiques précises pour tous les groupes. Les identifications à fine résolution taxonomique peuvent également être difficiles et longues (Batten et al., 2019) et certains organismes écologiquement importants, comme le picoplancton, ne sont pas échantillonnés. Ainsi, le CPR ne semble pas idéal pour étudier les communautés dans leur ensemble. La combinaison des données de séquençage et d'imagerie pourrait permettre de surmonter certaines de ces limites. En les intégrant avec les données environnementales, de multiples niveaux de structuration de l'écosystème planctonique pourraient être explorés simultanément.

L'objectif de ce chapitre est de réaliser l'analyse simultanée des différents types de jeu de données, afin d'identifier la structure globale de l'écosystème. Les sous-parties suivantes correspondent au manuscrit intitulé *Multi-scale exploration of the planktonic ecosystem structuring in the North Atlantic Ocean*.

5.2 Multi-scale exploration of the planktonic ecosystem in the North Atlantic Ocean

Authors

Ophélie Da Silva¹, Alberto Baudena¹, Guillaume Bourdin², Paul Frémont³, Rachel L. Kelly⁴, Louis Caray-Counil¹, Laetitia Jalabert¹, Nicolas Henry^{5,6}, Frédéric Mahé^{7,8}, Julie Poulain⁵, Sarah Romac^{5,6}, Sakina-Dorothee Ayata¹, Chris Bowler^{6,9}, Seth John⁴, Samuel Chaffron^{6,10}, Colomban de Vargas^{5,6}, Damien Éveillard^{6,10}, Olivier Jaillon^{3,6}, Matthew B. Sullivan^{6,11}, Patrick Wincker^{3,6}, Fabien Lombard^{1,6,12}, Lionel Guidi^{1,6}

1 - Sorbonne Universités, CNRS, Laboratoire d’Oceanographie de Villefranche, LOV, F-06230 Villefranche-sur-Mer, France.

2 - School of Marine Sciences, University of Maine, Orono, ME, United States

3 - Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

4 - Department of Earth Sciences, University of Southern California, Los Angeles, CA, USA

5 - Sorbonne Université, CNRS, Station Biologique de Roscoff, UMR 7144, ECOMAP, 29680 Roscoff, France

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

7 - CIRAD, UMR PHIM, F-34398 Montpellier, France

8 - PHIM, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France

9 - Institut de Biologie de l’ENS (IBENS), Département de biologie, Ecole normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

10 - Université de Nantes, Centrale Nantes, CNRS, LS2N, F-44000 Nantes, France

11 - Department of Microbiology, The Ohio State University, Columbus, OH, USA

12 - Institut Universitaire de France (IUF), Paris, France

Abstract

Deciphering the structure of ecosystems is essential for predicting their response to their drivers’ variations. Yet because ecosystem structure results from the interweaving processes co-occurring at several spatio-temporal scales, its study is challenging. Planktonic ecosystems embrace different levels of heterogeneous information. They include biotic components that can be characterized by taxonomic or morphological approaches, and

environmental components including biogeochemical or hydrodynamic properties of water masses. The proper integration of all components is a necessary step to understand planktonic ecosystem better. We propose an integrative methodology to a North Atlantic transect based on the high-resolution *Tara Pacific* imaging and 18S V9 metabarcoding datasets. We have highlighted the existence of two main scales of ecosystem structuring and identified associations between components. Taxonomic composition, morphology of the multicellular-enriched part of communities and biogeochemistry are similarly structured at the basin-scale. However, finer-scale structuring was detected for morphology of the protist-enriched part of the community and for hydrodynamics. Such differences in scales of the components could lead to a highly complex response of the ecosystem and its potential restructuring under climate change. By highlighting the multi-scale structuring of planktonic ecosystems, this study confirms the importance of considering processes occurring over multiple scales. In order to better understand the impacts of these coexisting scales on ecosystem structure, the couplings between the components and the mechanisms by which they interact must be further explored.

5.2.1 Introduction

In natural ecosystems, diversity and abundance patterns result from interactions within and between local communities exchanging individuals (Metacommunity concept, Leibold et al. 2004). For planktonic organisms, dispersal mostly relies on physical transport (advection and diffusion) by ocean currents. In addition, physico-chemical environmental characteristics also shape diversity patterns since they determine survival and fitness of organisms under the environmental filtering process (Kraft et al., 2015). Biotic interactions, on the other hand, are also crucial to understand the structuring of communities because competitive phenomena can lead to the exclusion or dominance of some species in an a priori favorable abiotic environment (*i.e.*, exclusive competition, Kraft et al. 2015). However this process is criticized within the plankton framework since, even in resource limiting conditions, the diversity of plankton tends to remain high in the continuously evolving oceanic environment (Plankton paradox ; Hutchinson 1961; Roy, Chattopadhyay 2007).

Patterns of diversity, abundance and interactions of eukaryotic plankton communities in the ocean have been extensively studied thanks to oceanographic campaigns, and the emergence of new technologies such as environmental sequencing and imaging. Recent global surveys have revealed that a large part of eukaryotic taxa and genes are unknown (De Vargas et al., 2015; Carradec et al., 2018). Interactions between taxa have also

been inferred, leading to the identification of key groups and the importance of biological interactions in the structuring of the community (Lima-Mendez et al., 2015). These different levels of biotic organization have been related to the environment in which organisms evolve. Among others, temperature (Ibarbalz et al., 2019; Salazar et al., 2019) or iron concentrations (Caputi et al., 2019) have been shown to be major drivers of the structuring of plankton communities. At the functional level, diversity declines towards the poles which is critical in a context of climate change (Ibarbalz et al., 2019; Salazar et al., 2019). Finally, the morphology of the organisms seems to condition their dispersal, which impacts their distribution patterns (Villarino et al., 2018; Richter et al., 2020; Sommeria-Klein et al., 2020). All of the above examples illustrate advances of knowledge about planktonic community structuring at global-scale. However, these data have rarely been used for finer-scale community studies, even though the importance of local heterogeneity in their structuring has already been highlighted (d'Ovidio et al., 2010; McGillicuddy, 2016; Lévy et al., 2018). Moreover, most of the studies dealt separately with community abundance, diversity or morphology, and only rarely have attempts been made to link them with the environment or processes of transport and mixing. To our knowledge, to date no study has explored simultaneously all levels of ecosystem structuring. Their interplay thus remains unknown. Yet, ecosystems are complex systems whose emergent properties cannot be studied at the level of the entities composing it (Karsenti, 2008; Ladyman et al., 2013). Thus, such systems could benefit from integrative approaches (Karsenti et al., 2011; Michener, Jones, 2012; Sunagawa et al., 2020; Zipkin et al., 2021), as it is the case in other fields of living sciences (Leonelli, 2013).

The Tara Pacific expedition (2016 to 2018), during which the global ocean was sampled at a near mesoscale spatial resolution for omics and morphological data, was conceived to overcome previous limitations (Gorsky et al., 2019). The different observations collected will allow integrating taxonomy, morphology and environment into a unified framework, leading to a better understanding of the interplay between these different components of the ecosystems. Here we focused on a fine resolution North Atlantic transect realized during this expedition. The Atlantic Ocean is a crucial oceanic basin as it houses the Atlantic Meridional Overturning Circulation system, deeply involved in climate regulation and whose variation can impact surrounding continents as well as global-scale ecosystems (Lenton et al., 2008; Liu et al., 2020). Therefore, the North Atlantic Ocean has been intensively sampled in order to better understand this ecosystem. In particular, numerous Continuous Plankton Recorder (CPR) surveys have been carried out over time, allowing temporal monitoring of several planktonic groups (Reid et al., 2003; Batten et al., 2019).

Among others, important results include evidence that Calanoid copepods are migrating northward in response to temperature variations (Beaugrand et al., 2009). The impacts of copepod community structure on carbon fluxes have also been investigated, and indicate that climate change, by modifying large-bodied copepod distribution, has modified activity of the biological carbon pump in the North Atlantic Ocean (Brun et al., 2019). Only a few studies have targeted protists but some studies suggest the northward migration of diatoms and dinoflagellates (Barton et al., 2016). While CPR data can provide information on fine spatial and temporal variations of some planktonic groups or species, it does not allow the study of the entire plankton community composition because of the sampling method limitations (Batten et al., 2019). In order to fill this gap, here we analysed environmental, metabarcoding and imaging data generated during the Tara Pacific expedition to simultaneously address community composition and morphological structuring of the planktonic ecosystem in the subtropical North Atlantic Ocean.

5.2.2 Material and Methods

Atlantic transect of the Tara Pacific expedition

During the Tara Pacific expedition, which took place from May 2016 to October 2018, biological samples were collected and physicochemical characteristics of the water masses were measured. Data were collected following the Tara Pacific Expedition protocols detailed in Gorsky et al. (2019). In our study, we focused on the outward North Atlantic Ocean transect, comprising 27 stations from [46.667°N ; 6.279°W] sampled on May 29th, 2016, to [24.180°N ; 81.406°W] sampled on July 6th, 2016 (see maps Figure 5.1).

Environmental context : biogeochemistry and hydrodynamics

We used the Tara Pacific biogeochemical context (version 1, DOI :10.5281/zenodo.4068292), including continuous and discrete measurements enriched by satellite and modeling data (Gorsky et al., 2019). Variables with more than 10% of missing values were discarded. For the variables recorded by several instruments, only one representative variable was retained, based on measurement quality (from high to low quality : in situ water sampling, inline system, satellite with a resolution of 1/12° and 1/24°, Copernicus assimilative model). The resulting dataset is composed of 24 environmental variables (D.1). Stations 1, 7 and 8 contained, respectively, 3, 10 and 8 missing values (out of 24) and were discarded from the biogeochemical analysis. Salinity measured with a thermosalinograph was missing at station 9. A linear model was adjusted between this variable and salinity provided by Copernicus, and the missing value was replaced by the predicted one.

Particulate Inorganic Carbon (PIC) variable also showed one missing value which was replaced by the average of PIC values at the previous and at the next station. Data were then normalized with a box-cox transformation and standardized.

In order to characterize the water mixing and transport properties of the area, satellite-derived velocity fields with geostrophic and Ekman components were analysed. From their processing, eight Eulerian and Lagrangian diagnostics were computed, listed as follows (further details are provided in Supp. Info. D.1). Finite Time Lyapunov Exponents (FTLE; Shadden et al. 2005), a Lagrangian quantity associated with mixing and horizontal stirring, were calculated to identify areas of intense frontal or turbulent activity. The Lagrangian betweenness (Ser-Giacomi et al., 2015) was employed to detect the structures which are crucial for the connectivity of the region : high betweenness values identify bottleneck-like regions in which water particles arrive from various origins and are then spread toward several different destinations. The horizontal divergence was used to identify areas of possible upwelling or downwelling. Indeed, this diagnostic detects water convergence and can be used as a proxy of vertical movements in the upper part of the water column. The horizontal divergence was computed both locally (Eulerian divergence) and along water parcel trajectories (Lagrangian divergence; Hernandez-Carrasco et al. 2018). The presence and the direction of rotation of vortexes were evaluated from the Okubo-Weiss metric and from the vorticity, respectively. Both these diagnostics, analogously to the horizontal divergence, were computed in Eulerian and Lagrangian ways. In order to provide an estimation of the time spent by the water parcels inside an eddy, in the months prior to the observations, their retention time was calculated (d'Ovidio et al., 2013). Finally, absolute speed and kinetic energy, providing information about the intensity of the horizontal currents, were estimated. Each diagnostic was computed using different sets of parameters. Every set was tested and, in total, 9 metrics synthesizing the information were kept for the analyses (more details in Supp. Info. D.1). Each of these metrics was log modulus transformed ($L(x) = \text{sign}(x) * \log(|x| + 1)$) and standardized.

Characterization of plankton communities

Biological samples collected during the Tara Pacific expedition span from viruses to metazoans. We investigated the metabarcoding datasets including 97 samples (data available on zenodo, doi :10.5281/zenodo.4073035) from 27 surface stations and four consecutive size fractions, ranging from 0.2 μm to 2000 μm , in which eukaryotes were collected (*i.e.*, 0.2-3 μm , 3-20 μm , 20-2000 μm and 300-3000 μm size-fractions). For each sample, the V9-18S rDNA metabarcodes were sequenced, filtered and clustered into Operatio-

nal Taxonomic Units (OTUs). Taxonomic assignments of OTUs were determined by global alignment against the 18S database EukRibo D.1. Since the primer used for amplification can amplified bacteria, a filtering step based on taxonomic assignments was performed to remove bacteria potentially present in the datasets. The resulting datasets corresponded to raw abundance matrices of OTUs in each sample. As raw abundances were not directly comparable, they underwent a total sum scaling transformation. The resulting relative abundances were represented by histograms to visualize differences in composition between samples based on the most abundant OTUs. In addition, to take into account the compositional nature of the data, a clr-transformation (Aitchison, 1999) was performed on raw abundances, and the dissimilarity between each pair of communities was estimated by the Aitchison distance between samples (*i.e.*, Euclidean distance based on clr-transformed abundances).

Imaging data, generated during the Tara Pacific expedition with Flowcam (20-200 μm ; Gorsky et al. 2019; Sieracki et al. 1998) and Zooscan (300-2000 μm ; Gorsky et al. 2010, 2019) imaging technologies, were also investigated. Due to heterogeneous processing of samples, Flowcam results from stations 1 to 10 were not considered. We analyzed 40 imaging samples distributed into two datasets : one derived from 17 samples analyzed with Flowcam, the other derived from the 23 samples analyzed with Zooscan. As the Flowcam datasets were voluminous (5,000-15,000 objects per sample), the images composing them have been treated differently according to the size of the objects (*i.e.*, organisms). All objects larger than a threshold of 500 pixel area were analyzed, only 30% of the objects smaller than 500 pixel area were taken into account. Taxonomic assignment of each organism was predicted with an automatic recognition algorithm and manually validated or corrected by taxonomic specialists using EcoTaxa (Picheral et al., 2017). The images and their taxonomic assignments are available at EcoTaxa (under project numbers 1700, 1737 and 1744). We discarded every image corresponding to non-living objects (including bad focus, detritus, artifacts, fibers and bubbles) or seaweed. For each object, 12 morphological features automatically estimated in Ecotaxa were used (Table 5.11). Taxonomic assignments from imaging technologies are coarser than metabarcoding ones. However, morphology of individuals are accessible by imaging while it is not by metabarcoding. Thus, these two datatype are complementary and only morphological information were exploited from imaging. The morphological features were normalized by a Yeo Johnson transformation (Yeo, Johnson, 2000) and community weighted means (CWM ; Díaz, Cabido 2001) were computed.

5.2. Multi-scale exploration of the planktonic ecosystem in the North Atlantic Ocean

Name	Type	Description	Ecological meaning
Area	Shape	Surface area of the object	Sedimentation rate, predator/predation, visual predation (Gorsky et al., 2010)
Perimeter	Shape	Length of the outside boundary of the object	Predator/predation (Gorsky et al., 2010)
Circularity	Shape	Circularity (1 : perfect circle)	Nutrient uptake (Gorsky et al., 2010)
Fractal	Shape	Fractal dimension of the object boundary	Sinking rate/turbulences/exchanges rates (Gorsky et al., 2010)
Skeleton area	Shape	Surface of the skeleton area	Sedimentation rate, predator/predation, visual predation (Gorsky et al., 2010)
Horizontal symmetry	Shape	Bilateral horizontal symmetry index	(Romagnan et al., 2016)
Elongation	Shape	Ratio between the primary and the secondary axes of the best fitting ellipse	
Mean	Grey level	Average grey value within the object	Visual predation (Gorsky et al., 2010)
Standard deviation	Grey level	Standard deviation of the grey value within the object	Organelle opaque/stomach detection for visual predators (Gorsky et al., 2010)
Minimum	Grey level	Minimum grey value within the object (0 : black)	Camouflage from predation (Gorsky et al., 2010)
Mean position	Grey level	(mean-max)/range	Camouflage from predation (Gorsky et al., 2010)
Slope	Grey level	Slope of the grey level normalized cumulative histogram of grey values	Camouflage from predation (Gorsky et al., 2010)

TABLE 5.1 – Imaging features used to compute CWM traits for plankton communities.

Data integration

The surface stations are described by numerous datasets that provide information on the biogeochemical and hydrodynamic structure along the transect, as well as the taxonomy and morphology of the communities. A principal component analysis (PCA; Legendre, Legendre 2012) was performed on each dataset. Significant principal components (PCs), according to the Kaiser-Guttman criterion (Guttman, 1954; Jackson, 1993), were conserved and a hierarchical clustering (Ward2 method; Murtagh, Legendre 2014) was carried

out, to provide homogeneous groups of stations (in terms of biophysicochemical, hydrodynamic, taxonomic and morphological characteristics). The final number of groups was chosen based on inertia decomposition (Murtagh, Legendre 2014). After having separately explored the structures of each dataset, we integrated all these data sets in a multiple factor analysis (MFA, Escofier, Pages 1994). The MFA is a multivariate statistical method that allows the simultaneous analysis of heterogeneous groups of variables (here, the 8 different data sources, figure 5.2). By constructing a new consensus space describing the maximum inertia of all the data sets, the MFA provides a synthetic view of the structuring of the individuals (i.e., sampling stations). In addition, relationships between data sources can be characterized. The MFA is composed of two steps. As all the datasets used were quantitative, the step 1 was a PCA on each group of variables. The step 2 consisted in a PCA, performed on the resulting dataset weighted by the first eigenvalues for each group. Two coefficients were used to explore the MFA results. The coefficient Lg was used to quantify the dimensionality of each dataset (i.e. the relationship between a group and itself) and the RV (i.e. rho-vectorial) coefficient was used to quantify the association between groups of data. A mathematical formalization of the study is available in supplementary information (Supp. Info. D.1).

All analyses and visualisations have been performed in the R programming environment (v4.0.3; R Core Team 2017). The tidyverse set of packages was used to manipulate and visualize data (v1.3.0; Wickham et al. 2019; including ggplot2 v3.3.2, and dplyr v1.0.2, packages) and multivariate analyses were performed with FactoMineR (v2.3, Lê et al. 2008). Codes and notebooks associated are available on GitHub (/folder).

5.2.3 Results

In 2016, the Tara boat set sail from Lorient (France) on May 29 to capture variation across the North Atlantic Ocean. It arrived at the entrance of the Gulf of Mexico, off the coast of California and Cuba, on July 6. Biogeochemical, hydrodynamic, metabarcoding and imaging data were collected, measured or modeled at the 27 surface stations constituting the transect. These heterogeneous data were individually analyzed and then simultaneously studied to identify the structuring of the ecosystem from an holistic approach. Geographic distribution and data availability of the different sources are shown in Figure 5.1.

5.2. Multi-scale exploration of the planktonic ecosystem in the North Atlantic Ocean

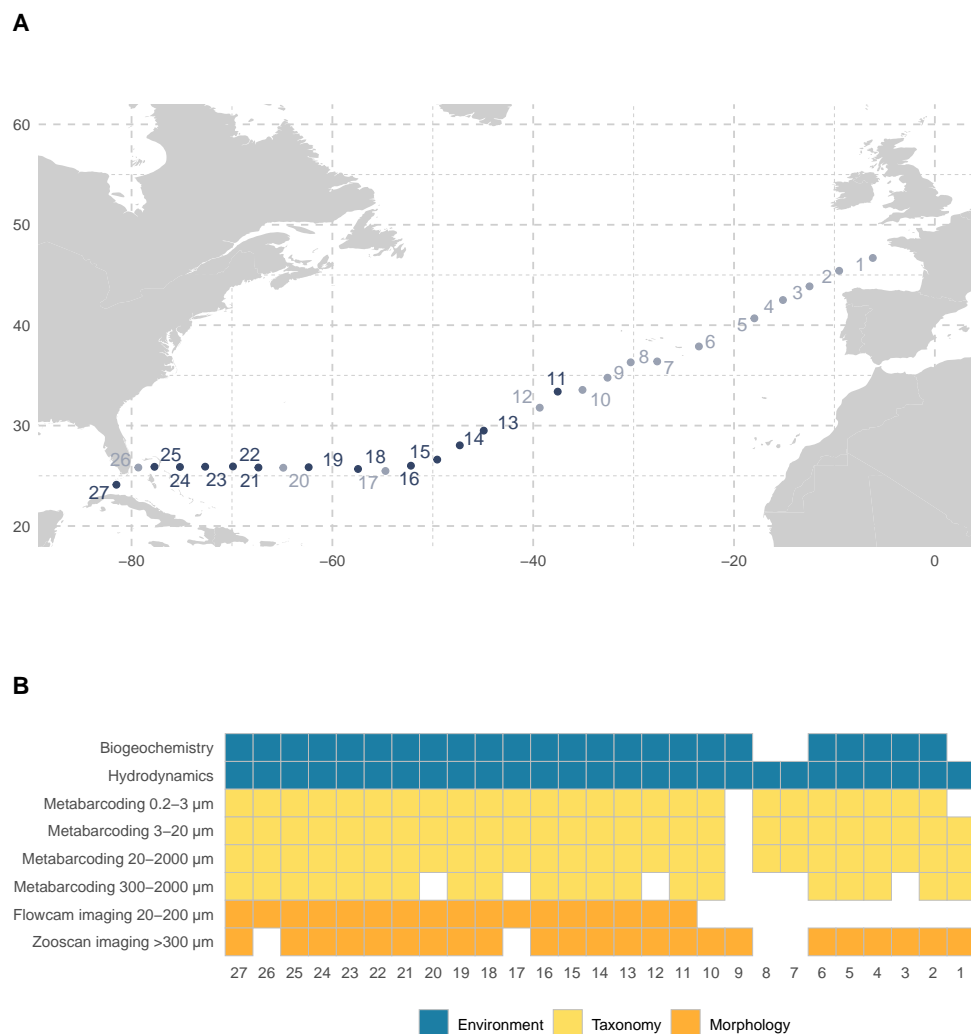


FIGURE 5.1 – Sampling strategy along the North Atlantic transect of the *Tara* Pacific expedition. (A) Map of the station localisation. Colours indicate whether all datasets (corresponding to different data sources) are available (yes : dark blue, no : light blue). (B) Data availability in regard with the data source. Colours represent the type of information for each data source.

Biogeochemical and hydrodynamic environment of the North Atlantic transect

The PCA on the biogeochemical variables revealed 5 significant axes represented 80.5% of the variance. The first principal component PC1 (39.6% of variance) represented a temperate-tropical oligotrophy gradient, separating the warmer, saltier, lighter and more

oligotrophic stations from the colder, fresher and more productive stations (Figure 5.2A). The second principal component PC2 (18.2% of variance) represented an opposition between the center of the gyre and enriched area, segregating windier and more nutrient depleted stations from stations richer in micronutrients and metals (Figure 5.2A). Finally, the four clusters revealed distinct bio-physiochemical zones, named hereafter the temperate productive area, the temperate oligotrophic gyre, the tropical oligotrophic gyre and the rich tropical area. These clusters followed one another spatially along the transect (Figure 5.2B).

The PCA on the hydrodynamical metrics represented 84.9% of variance with 3 significant axes (Figure 5.2C, Supp. Info. D.2) and revealed 6 different clusters. The first principal component PC1 (55.2% of variance) identified dynamic waters, characterized by a strong turbulence (high Lagrangian divergence and betweenness), usually inside eddies (negative Okubo-Weiss, Figure 5.2C). The second principal component PC2 (16.3% of variance) represented steadier and less turbulent waters (Figure 5.2C). Stations 26 and 27, located in the Gulf Stream current neighbourhood, are very different from the others and each constitutes their own cluster. They were characterized by high absolute velocity and kinetic energy and thus, a very high dynamism, but differed according to the divergence and vorticity, both higher at station 26, as well as the presence of a frontal structure (high Finite time Lyapunov exponent values), higher at station 27 (Figure 5.2C). Station 10 shows similar characteristics as station 27 but to a lesser extent. The other stations were located in 3 distinct areas including areas with longer retention times and preferably in vortices (low Okubo-Weiss values), areas out of vortex (high Okubo-Weiss values) and rather turbulent, and intermediate areas which did not have the characteristics listed above. Contrary to the environmental clustering, the Lagrangian clustering depicted a more heterogeneous spatial representation along the Atlantic transect (Figure 5.2D).

The two first PCs of the biogeochemical PCA captured the basin-scale structuring whereas the other PCs mirrored variability along the transect, likely resulting from finer-scale processes (Supp. Info. D.2). The processes involved in the biogeochemical structuring may be inferred through the investigation of the correlations between the biogeochemical PCs and the hydrodynamic diagnostics (Figure 5.2E). While the hydrodynamic processes showed a moderate east-west opposition, some diagnostics are distributed along the coast-large gradient and concern large-scale current characteristics (e.g., absolute velocity). On a finer scale, the environmental structuring may result from three types of processes, namely retentive phenomena (PC3-retention time correlation : 0.27), low turbulence phenomena (PC4-Finite time Lyapunov exponent correlation and between-

ness correlation : -0.21 and -0.34 respectively), and finally bottleneck phenomena (PC5-betweenness correlation : 0.49).

5.2.4 Community composition along the Atlantic transect

The plankton community compositions were investigated based on metabarcoding datasets, by describing the distribution of the main taxonomic groups in the different size-fractions. We observed plankton community changes throughout the transect that mirrored the biogeochemical changes (Figure 5.3). The 0.2-3 μm size-fraction is mainly composed of taxa belonging to five taxonomic groups : Alveolata, Stramenopiles, Chlorophyta, Rhizaria and Copepoda. Four changes in relative abundances along the transect were observed : 1) a westward decrease in Chlorophyta (mainly Mamiellophyceae and Prasinophyceae clade VII) and Dictyochophyceae (Stramenopiles, mainly Pseudochattonella and Pedinellales), 2) a westward increase in Alveolata (Dinophyceae, MALV-I and MALV-II) and various Stramenopiles (MAST and Chrysophyceae), 3) an increase in the central part of the transect in Collodaria (Rhizaria; mostly Sphaerozoum and other Collodaria) and, finally, 4) a decrease in the central part in Copepoda (mainly Cyclopoida at the north-eastern beginning of the transect and Calanidae, Paracalanidae and Oithonidae at the south-western end). The three other size-fractions are marked by different dominant plankton taxonomic groups, mainly due to Alveolata (principally Dinophyceae, in the 3-20 μm size-fraction), Rhizaria and Copepoda (20-2000 μm and 300-2000 μm size-fractions), but also other Metazoa (300-2000 μm size-fraction). Some changes in the dominance within families were also present, such as in Rhizaria from Acrosphaera in the 20-2000 μm size fraction to Sphaerozoum and Acrosphaera in the 300-2000 μm one, or for Copepoda with Calanidae, Cyclopoida and Paracalanidae in the 20-2000 μm size-fraction and Centropagidae in the 300-2000 μm one. Spatial changes in relative abundances were also observed along the transect (but more abrupt than those of 0.2-3 μm size-fraction, especially for 20-2000 and 300-2000 μm size-fractions) : 1) a westward decrease in Haptophyta and Stramenopiles (3-20 μm size-fraction), 2) a westward increase in Dinophyceae (3-20 μm size-fraction), 3) an increase in the central part of the transect in Rhizaria across the different size-fractions, and 4) a decrease in the central part of Copepoda for all size-fractions.

5.2.5 Community morphology along the Atlantic transect

The PCA on morphological datasets revealed three and two significant axes representing 97.3% and 94.3% of variance, respectively for the smallest and the largest size-fractions

(Figures 5.4A and 5.4B). For the smallest size-fraction, the first principal component (62.8% of total variance) distinguished smaller and more circular communities from larger and more elongated ones while the second principal component (23.7% of total variance) separated lighter from darker communities. For the largest size-fraction, the communities were organized along the first axis (55.6% of total variance) from the lighter and smaller communities to the darker and larger one. The second principal component (38.7% of total variance) separated communities between circular and irregular shapes. The morphological structuring of the communities did not match the environmental structuring of the basin for the smallest size-fraction while it did for the largest one.

A sparse Partial Least Squares (sPLS; Chung 2010) was then performed to select biophysicochemical variables that best predict morphological features of the communities. While for the small size-fraction, PC5 of the biogeochemical PCA, capturing high stratification and low net primary production gradient (Supp. Info. D.2), showed the strongest correlations with morphological features (*e.g.*, organism area, as well as mean, minimum and standard deviation of the grey levels), it was PC1 (*i.e.*, temperate-tropical oligotrophy gradient) and PC2 (*i.e.*, opposition between the center of the gyre and enriched area) for the large size-fraction (*e.g.* area, elongation, skeleton area, Figures 5.4C and 5.4D). Thus, the large size-fraction was mostly correlated with basin-scale heterogeneity, whereas the small size-fraction was mostly correlated to finer-scale heterogeneity. For the large size-fraction, the shape of the communities responded to the biogeochemical PC1 (*e.g.*, area, elongation), whereas both shape and transparency were structured by biogeochemical PC2. Thus, there is a decrease in size from the East to the West, as well as smaller and more transparent organisms in the open-ocean. For the small size-fraction, smaller and more transparent organisms were found with increasing biogeochemical PC5 values.

5.2.6 Global structure of the ecosystem

5.2. Multi-scale exploration of the planktonic ecosystem in the North Atlantic Ocean

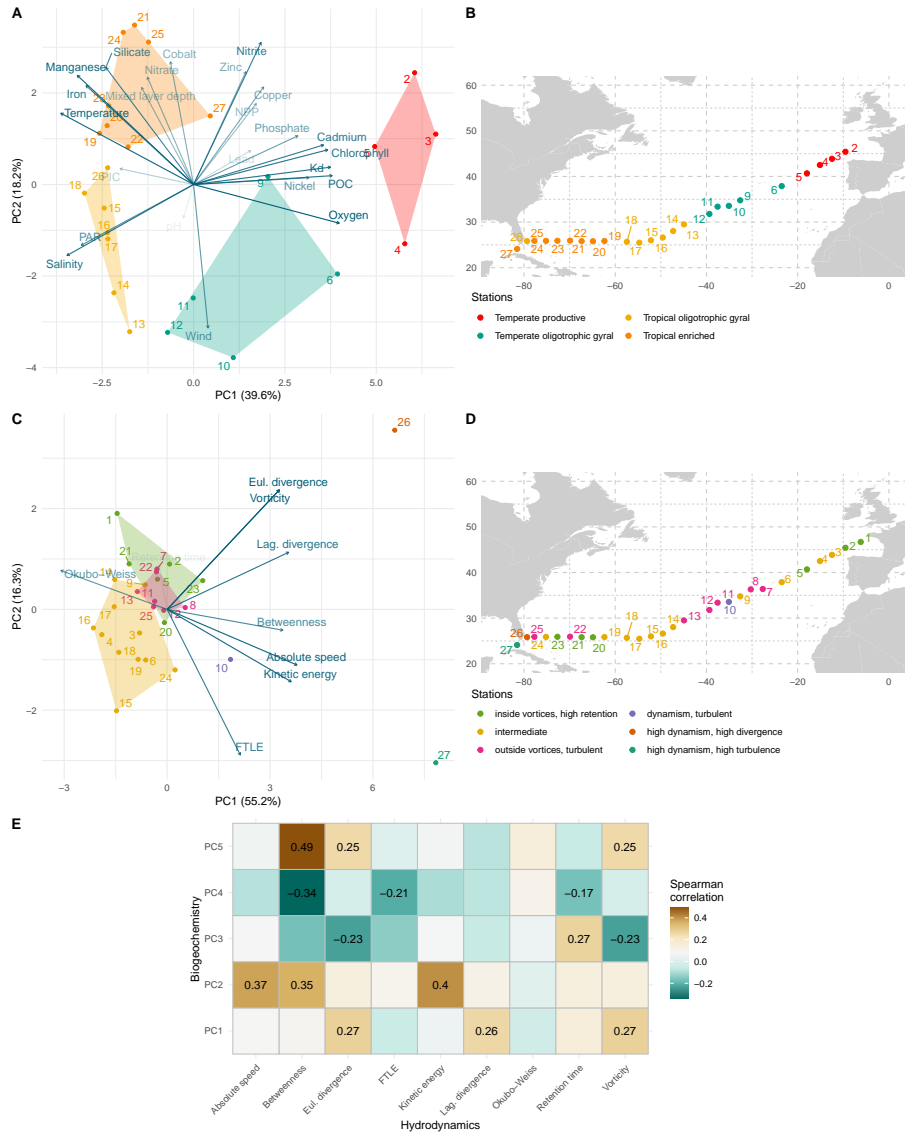


FIGURE 5.2 – Contrasted biogeochemical and hydrodynamical environmental structuring along the Atlantic transect investigated through principal component analyses and ascending hierarchical clusterings (Euclidean distance, Ward2 method). (A) Projection of the stations (dots) and variables (arrows) on the first biogeochemical PCA plan and (B) spatial projection of the stations in the Atlantic basin. (C) Projection of the stations (dots) and variables (arrows) on the first hydrodynamic PCA plan and (D) spatial projection of the stations in the Atlantic basin. Coordinates of the stations on the six first biogeochemical PCs and on three first hydrodynamic PCs were used for clustering stations. Colors indicate to which biogeochemical (A and B) or hydrodynamic (C and D) cluster stations belong in the first PCA plan and along the Atlantic transect. (E) Pearson correlation between biogeochemical PCs and hydrodynamic variables.

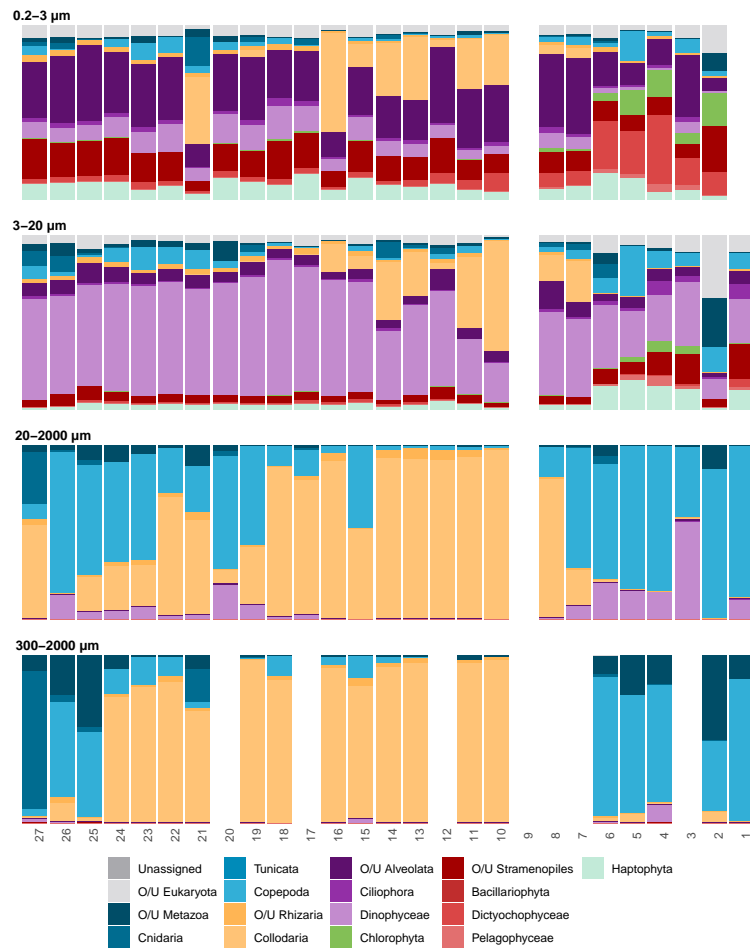


FIGURE 5.3 – Taxonomic composition of plankton communities. Distribution of taxonomic diversity of major lineages at Atlantic stations for size fractions (from top to bottom : 0.2-3 µm, 3-20 µm, 20-2000 µm and 300-2000 µm). Colors indicates lineages. Missing samples are represented by white columns.

5.2. Multi-scale exploration of the planktonic ecosystem in the North Atlantic Ocean

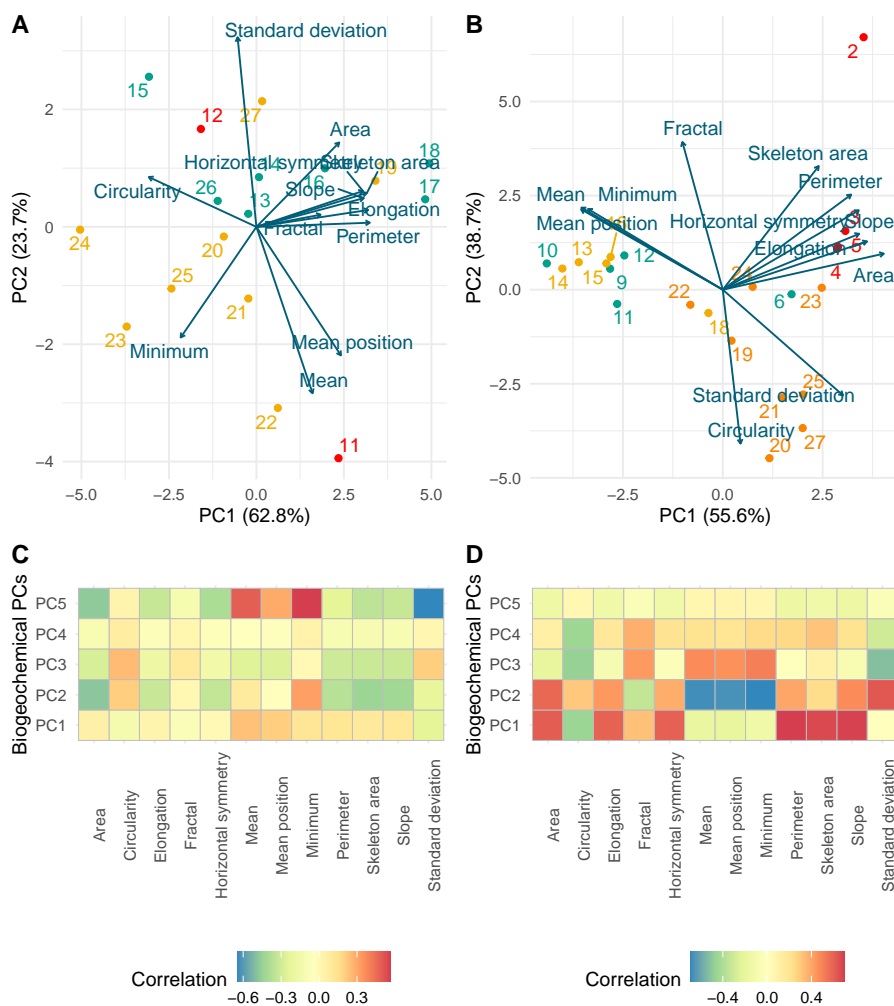


FIGURE 5.4 – Morphological structuring of the communities. Morphological PCA for the two size fractions (A) 20-200 μm and (B) >300 μm . Heatmaps showing pairwise correlations between the environmental PCs and morphological features obtained by a sparse partial least squares (sPLS) model (red : positive, blue : negative, yellow : weak correlation values) for the (A) 20-200 μm and (B) >300 μm size-fractions.

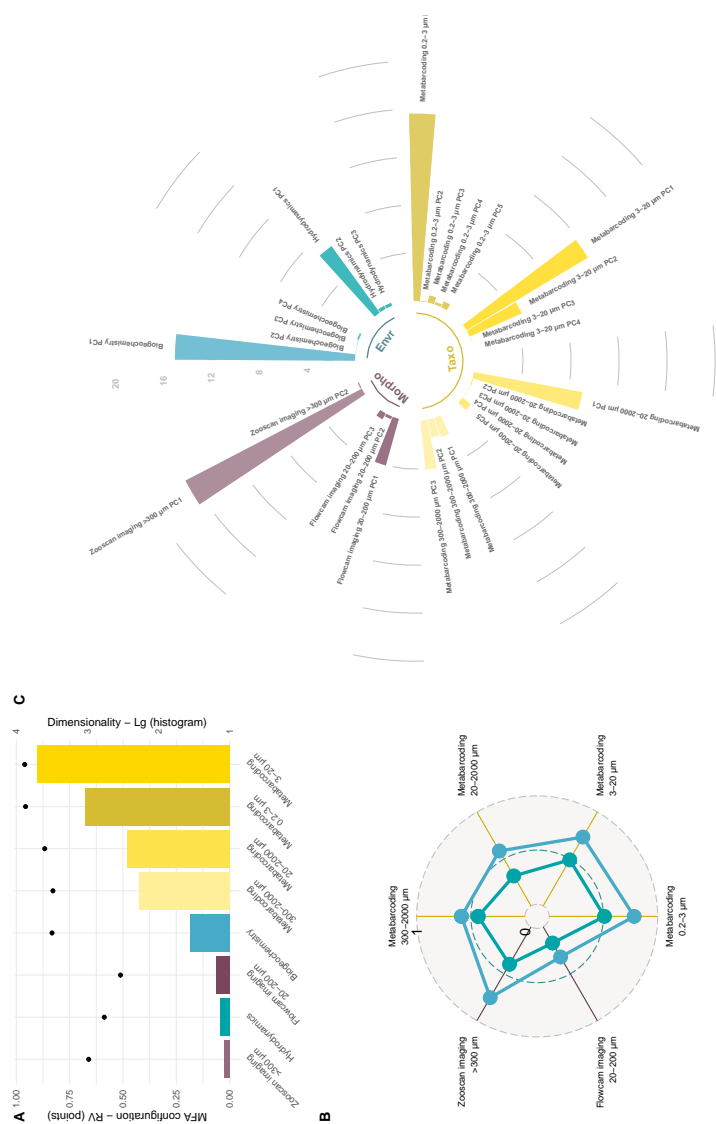


FIGURE 5.5 – Integration of all datasets through a multiple factorial analysis (MFA). (A) Dimensionality of each dataset based on the Lg coefficient (1 : unimodal, higher : increasing dimensionality) is represented with barplot and configuration of each dataset based in the global MFA space on RV coefficient is indicated with the points (the higher a value, the better the projection of the dataset is). (B) Pairwise relationships between biological and environmental datasets examined with RV coefficient (0 : totally different, 1 : identical). (C) Contribution of each principal component to the first dimension of the MFA. For all panels, colors indicate the data source.

A multiple factor analysis (MFA) was used to describe the whole plankton ecosystem and to study the relationships between the information provided by the different datasets (biogeochemistry, hydrodynamics, community composition and morphology). Only 15 of the total 27 stations had an exhaustive dataset, hence were retained for this analysis. The Lg coefficient was used to describe the dimensionality of each dataset (Figure 5.5A). A value of 1 indicates that the inertia of the dataset is carried by the first dimension, while the higher the value, the more dimensions are needed to describe the dataset. The metabarcoding datasets displayed higher Lg values (i.e. 3.0, 3.7, 2.4 and 2.7 from smallest to largest size-fractions) than the others (i.e. 1.6, 1.1, 1.2 and 1.1 for biogeochemistry, hydrodynamics, Flowcam imaging and Zooscan imaging, respectively). Thus, while the information in the metabarcoding datasets was balanced across several of their principal components, most of the environmental and morphological variance was explained by their first principal component. This is consistent with previous results and suggesting that the impact of the reducing number of stations to the analysis is limited.

The RV coefficient was used to explore pairwise relationships between datasets. A RV coefficient equals to 0 indicates a completely different structure of the two datasets, while a RV coefficient equals to 1 indicates that the structure induced on the stations by the two datasets is identical. Firstly, the pairwise relationships between each dataset and the global configuration of the MFA was estimated based on the RV coefficient (Figure 5.5A). The hydrodynamic and imaging datasets exhibited lower RV coefficients (0.51, 0.66 and 0.59 corresponding to hydrodynamics, Flowcam imaging and Zooscan imaging, respectively) than the biogeochemical and metabarcoding datasets (0.83, 0.96, 0.96, 0.87 and 0.83 for biogeochemistry, 0.2-3 μm metabarcoding size-fraction, 3-20 μm metabarcoding size-fraction, 20-2000 μm metabarcoding size-fraction and 300-2000 μm metabarcoding size-fractions, respectively). In particular, the RV coefficient values were very close to 1 for the two smallest metabarcoding size-fractions, indicating that the inertia of these datasets was well represented in the consensus space. Then, pairwise relationships between environmental and biological datasets were assessed based on the RV coefficient (Figure 5.5B). All biological components are more similar to the biogeochemical environment. However, the RV coefficients between the morphology of the small size fraction and the environmental datasets were low (biogeochemistry : 0.32, hydrodynamics : 0.32), indicating that their structuring differed.

Finally, the importance of each dataset to build the consensus space was estimated. For this, the contribution of all principal components of each dataset to the first component of the AFM was estimated (Figure 5.5C ; contribution to the first MFA plan available in

Supp. Info. D.3). High value indicates that the PC to contribute more to the formation of the first dimension of the MFA. The biogeochemical, the two smallest size-fractions metabarcoding and zooscan imaging datasets displayed a first component that strongly contributed to first MFA dimension building (Figure 5.5C). This was less pronounced for the other datasets (e.g. hydrodynamics), or even reversed (e.g. for the 300-2000 μm size-fraction metabarcoding dataset, PC3 contributed the most). Thus, important differences in the structure of the datasets were highlighted, as well as in their contribution to the consensus space.

5.2.7 Discussion

Several data types, reflecting different facets of the ecosystem, have been investigated both individually, to identify separately each structure, and as a whole, to infer their levels of interplay. Briefly, while physical and biogeochemical characteristics of water bodies mainly showed basin-scale structure, transport and mixing characteristics of water bodies varied at finer-scale. Such structuring scales were also found in the biological components, in accordance with both data type and size-fraction. The most noticeable structuring of the ecosystem is at basin-scale and was observed with the biogeochemical dataset, with all size-fractions of community composition and, based on the morphological composition for the $>300 \mu\text{m}$ size-fraction. As such, this structuring reflects the Atlantic basin-scale succession of geographical areas in which biogeochemistry, plankton composition and morphology of the large fraction-size are closely interrelated.

Basin-scale structuring

For the eastern part of the Atlantic transect, waters were rich and productive, related to the deep water mixing events (Longhurst, 2010). In this environment, strict autotrophs associated with nutrient-rich waters, such as Chlorophyta and Haptophyta, were found. Chlorophyta, in particular, were first represented by Mamiellophyceae, then by Prasino-phyceae clade VII, likely mirroring the coast-wide transition (Santos Lopes dos et al., 2017). The toxic phytoflagellate *Pseudochattonella* found in nitrate-rich coastal waters (Chang et al., 2014; Eckford-Soper, Daugbjerg, 2016) was also present. In addition, abundant Copepods were represented by Cyclopoida, Paracalanidae and Centropagidae which are known to be rather distributed in the eastern part of the basin (Continuous Plankton Recorder Survey Team, 2004). Finally, the organisms found in this sector were larger than in the more south-western regions, as it is often observed through temperate to tropical transition.

The central part of the transect was characterized by a poorer environment, related to the anticyclonic gyral circulation generating a depleted nutrient area (Palter et al., 2005). This is an agreement with the abundance of Collodaria, which are symbiotic mixotrophic organisms capable of forming colonies and typical of oligotrophic areas (Biard et al., 2017; Ishitani et al., 2012). In addition, plankton communities were smaller, less circular and more transparent. Indeed, a decrease in body size is linked to higher metabolic rates, because irregular shapes likely increase the surface-to-volume ratio leading to efficient nutrient uptake (Litchman, Klausmeier, 2008; Marañón, 2015) and transparency is an efficient camouflage against visual predation (Johnsen, 2014). These plankton communities thus exhibited advantageous morphology for these clear oligotrophic waters.

The western part of the transect was rather poor but showed higher concentrations of micronutrients and metals, generally linked to the local enrichment by Saharan dust (Colarco et al., 2003). Alveolata were dominating with Dinophyceae known to represent a large part of the phytoplankton (Not et al., 2009; Pierella Karlusich et al., 2020). There were also many MAST, bacterivorous Stramenopiles abundant in oceans (Keeling, Campo del, 2017), that could be linked to presence of diazotroph bacteria due to iron enrichment (Morel, Price, 2003). In parallel, Copepoda became abundant, especially Calanidae, Paracalanidae and Centropagiadae. The availability of micronutrients and trace metals, not limiting photosynthesis (Morel, Price, 2003; Morel et al., 2020), seemed to support a food web based on herbivorous secondary consumers. Finally, in this environment, the organisms tend to be larger, as already observed in high silicate areas (Mousing et al., 2018). This basin-scale pattern is likely related to spatial rather than temporal forcings. Indeed, sampling was performed over one month, during a period without rapid planktonic community turnovers. However, some data, such as the hydrodynamics and the morphology of the small size-fraction, were not structured at the basin-scale level suggesting that processes are at play at other scales.

Structuring at submesoscale

The small size-fraction of the plankton communities (20-200 μm) did not show a basin-scale structured morphology, rather it seemed to be related to the fine-scale heterogeneity of the hydrodynamic environmental context. This heterogeneity could result from mesoscale to sub-mesoscale faster processes (*i.e.*, occurring at scales of the order of 1 to few 100 km and a few days to several weeks). This size-fraction was protist-enriched, while the large size-fraction was enriched in multicellular organisms. These organisms have major differences in their growth and metabolic characteristics. Indeed, protists

are unicellular, showing high metabolic rates and fast life-cycles. Hence, they must react more rapidly to changes in their environment, such as those generated by mesoscale processes (McGillicuddy, 2016). This fine-scale heterogeneity was likely linked to peculiar conditions of betweenness (bottleneck-like area; Ser-Giacomi et al. 2015), vorticity (counter-clockwise eddy) and divergence representing anomalous areas in the basin-wide structure. Plankton communities found in the small size-fraction were smaller and more transparent in these areas corresponding to local conditions with strong connectivity with possible nutrient inputs through local upwelling phenomena (McGillicuddy, 2016). In such areas, a high competitive pressure can be assumed and therefore some organisms, such as smaller and faster-growing ones, might be favored.

Integrated view of the ecosystem

In line with the fourth paradigm of Data Science, which seeks to extract knowledge through massive data mining from different sources (*i.e.*, experiment, simulation, theory; Hey et al. 2009), our exploratory study aimed to detect emerging patterns of planktonic ecosystem structuring from heterogeneous data sources. Therefore, the different patterns observed at the scale of the North Atlantic Ocean were studied together, allowing the identification of a consensus structure, as well as shared or singular characteristics between data sources.

The global analysis provided for each dataset consistent signal with the individual analyses, justifying the relevance of the global analysis on the reduced dataset. In particular, the metabacoding data supported richer information (*i.e.*, the variance of the datasets is organized along several axes of variation) than the environmental and imaging data. The fact that taxonomy did not provide the same information as functional one has already been identified, leading to an increasing use of functional approaches to better understand ecosystem functioning (McGill et al., 2006). Some data sources were therefore structured the same and others differently, which underlined their richness and complementarity for understanding ecosystem structure.

Moreover, it was identified that the main axes of variation for a given datasets may not represent the total variation in the ecosystem. Thus, predicting the axis of variability mostly impacting the ecosystem structure or identifying the most important variables to explain it is challenging. The ecosystem, represented here by the consensus space, did not reflect the sum of individual structures, as predicted by the theory of complex systems (Ladyman et al., 2013). Since all datasets do not contribute equivalently to the ecosystem structure and all variables are not equivalent, the approaches focusing on a

few variables to understand the global structuring of an ecosystem or its functioning can be questioned.

Although these reductionist approaches are essential to give clues about specific patterns, they do not take into account all the inter-scale feedbacks taking place within the systems. Thus, they must provide a restricted understanding of the structure of ecosystems (Zipkin et al., 2021). The linear approach used, here, has therefore allowed to highlight the complex links between data sources and with the emerging structure. It confirms for the oceanic environment, that ecosystem is not equal to the juxtaposition of individual observations from all datasets. It emphasizes that it is crucial to look at the ecosystem as a whole (Karsenti et al., 2011), because by looking at things individually we can focus on elements that ultimately contribute little to the overall structure of the ecosystem.

Persepctives

In this study, multiple heterogeneous data sources informing about processes occurring at different spatio-temporal scales have been used. They have been transformed into a common representation and integrated through a statistical approach into a unique framework. This data integration provided a unified view of the ecosystem (Michener, Jones, 2012). This approach gives a global insight from the different datasets within a single analysis. It allowed analyzing each dataset, their relationships, and their position in the unified view that would not be possible in separate analyses. This emphasized that the oceanographic field is now ready for a better and formal data integration as proposed in Data science. Indeed, marine sciences would benefit from the vast amount of heterogeneous data collected in the past few years for a reliable estimation of ecological processes (Leonelli, 2013; Zipkin et al., 2021). While natural and anthropogenic pressures on ocean life are increasing (Worm et al., 2006; Chivers et al., 2017; Lubchenco et al., 2020), such a holistic approach is more than ever necessary to decipher the structuring and the functioning of these threatened ecosystems (Karsenti et al., 2011; Chust et al., 2017).

Acknowledgement

We acknowledge all participants of the Tara Pacific Expeditions (<https://doi.org/10.5281/zenodo.3777760>). O Da Silva is supported by a PhD Grant from the French Ministry of Higher Education, Research and Innovation for three years.

Contribution

ODS designed the study and wrote the manuscript with several input from LG and

FL, LG and FL directed the study. AB computed Lagrangian diagnostics. GB designed the environmental sampling. JP and SR generated raw metabarcoding data. NH and FM generated the annotated metabarcoding datasets. LCC and LJ annotated imaging data. DE, SC and SDA contributed to conceptualization. ODS performed analyses. AB, GB, RK, SDA, CB, SJ, SC, DE, OJ, MS, PW discussed the results and commented the manuscript.

5.3 Conclusion

Dans ce chapitre, il a été mis en évidence lors de l'analyse conjointe des données, que des informations issues des analyses individuelles de chaque jeu de données étaient retrouvées. Mais une nouvelle vision de l'écosystème a également été fournie dans laquelle les différents jeux de données n'ont pas la même importance. Cela met en avant l'intérêt de tirer profit de sources de données hétérogènes mais complémentaires. Cette analyse constitue une étape vers l'intégration formelle de données en océanographie.

Discussion générale

6.1 Décrypter la complexité des écosystèmes planctoniques à partir de données à haut-débit

6.1.1 Récapitulatif des principaux résultats

Pendant ma thèse, je me suis intéressée aux relations entre structures biologiques et environnementales, afin de mieux comprendre comment s'organise spatialement la diversité planctonique et ce, à différentes échelles biologiques. J'ai exploré les données méta-omiques et d'imagerie à haut débit issues de campagnes d'échantillonnage massif de l'océan pour déchiffrer les relations pouvant exister au sein des communautés. Trois questions principales, identifiées dans l'introduction, ont guidé mon travail.

Question 1. Les données méta-omiques peuvent-elles servir de support à l'étude de la diversité génomique de l'ensemble des espèces de l'écosystème ?

La diversité génétique peut être explorée à l'aide de marqueurs moléculaires ou à l'échelle du génome. Les données métagénomiques peuvent théoriquement permettre l'étude de la diversité génétique de toutes les espèces présentes dans le milieu. Ainsi, elles pourraient constituer une opportunité pour l'étude de la diversité génétique des protistes et des facteurs qui la structurent. Les études précédemment menées se sont principalement basées sur quelques espèces, ne garantissant pas que les mécanismes identifiés soient généralisables à l'intégralité de ce groupe très diversifié. Dans le chapitre 3, j'ai mis en place une analyse générique permettant d'explorer les mécanismes de différenciation

génétique de protistes à partir d'échantillons métagénomiques. J'ai pu montrer que cette démarche est restreinte par le compromis entre abondance de l'espèce dans le milieu, taille du génome et profondeur de séquençage.

Question 2. Les données à haut débit permettent-elles de faire un lien entre génotype et phénotype des communautés ?

Le phénotype influence la structure, le fonctionnement et la dynamique des communautés. Il résulte de l'expression du génotype modulée par les interactions abiotiques et biotiques. L'imagerie permet d'étudier la biogéographie de traits morphologiques spécifiques à l'échelle des communautés. En particulier, j'ai pu identifier que les communautés planctoniques se structurent suivant leur taille et leur transparence. Au cours du chapitre 4, j'ai ciblé la transparence, un trait peu quantifié à l'échelle des communautés et pour lequel aucun marqueur moléculaire n'a été identifié. J'ai supposé que l'analyse conjointe des données de métatranscriptomique et d'imagerie pourrait fournir un support à l'identification de marqueurs moléculaires pouvant conduire à l'analyse de la diversité génétique associée. J'ai mis en évidence que la transparence était spatialement organisée et que l'imagerie était une technologie prometteuse pour l'étude de ce trait, difficilement explorable à l'échelle de la communauté par des approches moléculaires.

Question 3. Les données à haut débit peuvent-elles contribuer à une vision intégrée des écosystèmes planctoniques ?

Dans le chapitre 5, je me suis intéressée à l'écosystème de l'Atlantique Nord qui est un environnement présentant de nombreux gradients. Des explorations individuelles des jeux de données de metabarcoding, d'imagerie et environnementaux ont permis de mettre en évidence plusieurs niveaux de structuration spatiale de la composition taxonomique, de la morphologie et des caractéristiques biogéochimiques et hydrodynamiques, respectivement. Dans un second temps, tous les jeux de données ont été intégrés simultanément afin d'identifier la structure globale de l'écosystème et d'identifier la complémentarité et la redondance des informations fournies par les différentes sources d'information.

Ce travail a été confronté à plusieurs défis. Je vais les présenter et illustrer pourquoi les données de séquençage et d'imagerie à haut débit représentent des opportunités pour mieux comprendre les écosystèmes planctoniques. Je présenterai ensuite quelques perspectives sur leur exploitation et je finirai par mettre en regard les approches basées sur ces données avec les pratiques dans la recherche.

6.1.2 Méta-omique et imagerie, des données de diversité complémentaires

Estimation de la biodiversité taxonomique par metabarcoding

Les études de la biodiversité sont classiquement basées sur des observations directes (*e.g.*, microscopie optique). Ces méthodes permettent l'identification et la mesure précise des individus par un opérateur spécialiste. Cependant, les experts sont de moins en moins nombreux. De plus, ce travail fastidieux est long et difficilement faisable à l'échelle des communautés pour de larges gammes de taille et de nombreux échantillons. Les méthodes à haut débit d'imagerie et de metabarcoding peuvent permettre de surmonter ces limites. Dans ce travail de thèse, la diversité taxonomique des communautés a été explorée à partir des données de metabarcoding. Dans le chapitre 5, ces données ont permis d'estimer la composition des communautés et d'explorer les variations le long d'un transect dans l'océan Nord Atlantique. Dans le chapitre 3, elles ont été utilisées dans une phase de pré-analyse afin d'évaluer les espèces candidates pour l'étude. Toutefois, explorer la diversité en se basant sur le metabarcoding peut être confronté à plusieurs limites.

Le metabarcoding fait intervenir une méthode de séquençage à haut débit basée sur les amplicons. Ainsi, des séquences spécifiques sont ciblées. Elles sont communes à tous les organismes, ont des portions conservées et des portions variables, et présentent l'avantage d'être largement référencées dans les bases de données. Dans notre étude, les eucaryotes étaient ciblés et la séquence marqueur utilisée était le gène de la région hypervariable V9 de ARNr 18S. Ainsi, tous les organismes sont séquencés et peuvent être discriminés suivant l'identité de séquence avec les références pour être taxonomiquement annotés. Cependant, le metabarcoding nécessite une étape d'amplification. À partir de l'échantillon, de nombreux amplicons vont être obtenus par PCR. Or l'amplification PCR peut conduire à des estimations faussées de la diversité via une amplification différentielle de certains taxons (Acinas et al., 2005; Potvin, Lovejoy, 2009). Bien qu'une récente étude a au contraire montré que la diversité estimée par metabarcoding était comparable à celle obtenue à partir de méthodes s'affranchissant de l'étape d'amplification (Obiol et al., 2020), il faut être prudent avec ces données (Loos van der, Nijland, 2021). De plus, la technique de séquençage à haut débit utilisée pour générer les données est le séquençage Illumina. Cette technologie, qui est moins coûteuse et présente moins d'erreurs de séquençage, produit des lectures courtes, d'une centaine de paire de bases. Du fait de leur longueur, seule une partie de l'ARNr 18S peut être ciblée (*e.g.* région hypervariable V4 ou V9). Or, il peut présenter des différences entre les estimations de la diversité suivant

le choix de la région hypervariable (Giner et al., 2016; Grey et al., 2018). En particulier, les estimateurs de la diversité ne sont pas tous impactés de la même manière par le choix du marqueur. Par exemple, la diversité bêta semble plus robuste que la diversité alpha (Grey et al., 2018).

Exploration omique de la diversité

Les techniques d'échantillonnage environnemental méta-omique permettent d'explorer la diversité génétique des communautés en s'affranchissant des limites liées à la culture d'organismes isolés. Cela est d'autant plus important dans un contexte où seule une partie de la diversité est supposée cultivable, notamment pour les unicellulaires. La méta-omique a permis d'explorer plus en profondeur les communautés en permettant l'accès à la biosphère rare. Ces organismes, qui correspondent à une petite portion de la communauté et dont la biogéographie et l'écologie peut différer des organismes dominants, participent au réservoir génétique et fonctionnel des communautés, indispensable à leur résilience face à des changements environnementaux (Lynch, Neufeld, 2015). Or, l'accès à cette biosphère rare dépend des limites de détection des méthodes d'échantillonnage. Ainsi, le développement des méthodes de séquençage permet un accès de plus en plus important à la biosphère rare (Foissner, 2006; Fuhrman, 2009) et modifie la perception de la diversité (qui peut toutefois être biaisée par l'effort d'échantillonnage, Grey et al., 2018). Ces nouvelles méthodologies permettent d'accéder à la structure de cette partie de la diversité (*e.g.*, Ser-Giacomi et al., 2018; Martini et al., 2020; Villa Martín et al., 2020). Ainsi, les études méta-omiques ont commencé à lever le voile sur la grande diversité planctonique dans les océans, que ce soit au niveau taxonomique que génétique (*e.g.* Venter et al., 2004; De Vargas et al., 2015; Carradec et al., 2018).

L'assignation des séquences, qui peut être taxonomique ou fonctionnelle, est basée sur l'homologie avec les séquences connues des bases de données. Toutefois, cette méthode d'assignation est moins robuste que la détermination en laboratoire et il est nécessaire d'être précautionneux pour éviter de biaiser les bases de données. Une grande partie de la diversité révélée reste cependant inconnue que ce soit taxonomiquement et/ou fonctionnellement, notamment pour les unicellulaires. C'est la matière noire microbienne qui pourrait couvrir une part importante de la diversité et des processus des écosystèmes (Rinke et al., 2013; Bernard et al., 2018) et un défi majeur est de réussir à révéler cette diversité (Vanni et al., 2021). Les approches guidées par les données, comme celle menée dans le chapitre 4, peuvent permettre de lever une partie du voile sur les gènes de l'écosystème. La quantité d'inconnues est variable suivant les lignées. Par exemple, la

diversité des bactéries et des macroorganismes a été largement plus étudiée que celle des protistes. Un effort pour mieux explorer et fournir des références pour ces taxons a été entrepris (Keeling et al., 2014; Sibbald, Archibald, 2017). Des approches de reconstruction de génomes à partir des données omiques (*e.g.*, Vorobev et al., 2020; Delmont et al., 2021) permettent d'augmenter les références disponibles et créer des opportunités pour étudier la diversité génétique des organismes présents dans l'environnement, comme réalisé dans le chapitre 3. Il faut cependant veiller à la qualité des références assemblées pour ne pas corrompre les bases de données (Shaiber, Eren, 2019).

Imagerie, d'une quantification de la diversité taxonomique vers une quantification de la diversité fonctionnelle

Les techniques d'imagerie utilisées permettent aussi d'étudier la diversité taxonomique et morphologique d'organismes qui n'auraient pas pu être cultivés. Dans les chapitres 3 et 4, les données d'imagerie utilisées avaient été assignées au niveau taxonomique le plus fin possible. Dans la majorité des cas, cela correspondait au niveau de la famille. Les assignations peuvent toutefois être à un niveau taxonomique plus précis (*e.g.*, genre) pour certains groupes, comme les Copépodes. Considérer des niveaux non uniformes peut conduire à des biais dans l'estimation des indices de diversité taxonomique. Toutefois, ces données présentent l'avantage d'être quantitatives, contrairement aux données métabarcodiques compositionnelles. Ainsi, les abondances brutes entre échantillons peuvent être comparées. Les données d'imagerie et métabarcodiques sont donc complémentaires et elles pourraient tirer des bénéfices mutuels. D'une part, les données métabarcodiques pourraient être calibrées à partir de l'imagerie pour les rendre quantitatives. Et d'autre part, les données de metabarcoding pourraient permettre d'améliorer les prédictions taxonomiques des images par les algorithmes d'apprentissage. Toutefois, les assignations taxonomiques entre les deux sources de données ne sont pas toujours comparables (*i.e.*, résolution taxonomique de l'imagerie globalement moins importante), ce qui constituerait un réel défi. Dans les chapitres 4 et 5, les assignations taxonomiques ont été utilisées pour filtrer les données d'imagerie. En particulier, les particules non vivantes ont été éliminées de l'analyse.

L'autre avantage capital de l'imagerie est l'accès aux caractéristiques morphologiques des individus. Le plus souvent la diversité morphologique est moins importante que celle génétique. L'imagerie a permis d'étudier la diversité fonctionnelle des communautés dans les chapitres 3 et 4. Pour cela, une approche dirigée par les données a été employée, ce qui a permis l'identification de traits spécifiques et l'étude de leur biogéographie.

Toutefois, des traits morphologiques d'intérêt peuvent également être spécifiquement ciblés (*e.g.* la taille). Grâce à l'automatisation des mesures des descripteurs d'image, de nombreux traits sont quantifiés, pour une diversité taxonomique large et potentiellement dans de nombreux environnements. Ainsi, les méthodes d'imagerie à haut débit peuvent également permettre de contribuer à compléter les bases de données fonctionnelles et de mieux comprendre les spécificités écologiques.

Les données méta-omiques et d'imagerie, bien que présentant certaines limites, permettent donc d'explorer la structure des écosystèmes. Dans la partie suivante, je vais mettre en évidence les apports des études qui ont été menées au cours de cette thèse dans le contexte des écosystèmes planctoniques océaniques.

6.1.3 Les multiples structures de l'écosystème planctonique

Imbrication des échelles

Au cours de ce travail, les différentes études ont été menées à plusieurs échelles. Il est bien établi que le système océanique présente une structure multiéchelle qui a été mise en évidence dans de nombreuses études aussi bien du point de vue abiotique que biotique. Les résultats obtenus au cours de ce travail s'inscrivent comme attendu dans ce découpage multiéchelles. Tout d'abord, au niveau global, les données biogéochimiques du chapitre 4 se répartissent suivant les biomes et les régions (Longhurst, 2010). Dans le chapitre 5, la structure spatiale des données biogéochimiques est cohérente suivant les régions précédemment établies dans le bassin atlantique (Longhurst, 2010; Reygondeau et al., 2013). À l'échelle de la mer Méditerranée dans le chapitre 3, les conditions biogéochimiques se regroupent suivant les sous-bassins (Ayata et al., 2018). Ainsi, ces différentes structures sont cohérentes avec la littérature. À plus fine échelle, les résultats sont encore comparables avec la littérature. En effet, l'hydrodynamisme caractérisé dans le chapitre 5 présente une structure fine échelle, ce qui est cohérent avec l'influence bien caractérisée des processus de méso et sous-mésoéchelle sur la circulation océanique (Chelton et al., 2011; McGillicuddy, 2016; McWilliams, 2016). Cependant, l'environnement n'est pas la seule composante des écosystèmes à présenter des structures multiéchelles.

L'imbrication des échelles se retrouve également dans le vivant. Elle a été explorée dans les chapitres 4 et 5. À l'échelle globale, dans le chapitre 4, la transparence est organisée suivant les biomes et les régions. Bien que la biogéographie de ce trait n'ait jamais été caractérisée à l'échelle globale pour les communautés planctoniques, d'autres traits

morphologiques présentent de telles structures. Par exemple, la taille des Copépodes est spatialement organisée suivant la latitude (*e.g.*, Brun et al., 2016). Dans l'étude du chapitre 5, l'échelle de la structure biologique dépend des organismes observés. En effet, les organismes pluricellulaires présentent des répartitions homogènes sur de plus grandes distances que les unicellulaires. Ce résultat paraît a priori contre intuitif car les unicellulaires sont supposés avoir un potentiel de dispersion illimité (Finlay, 2002). Il serait donc attendu qu'ils présentent des patterns stables sur des distances supérieures ou égales à celles d'organismes plus grands. Pourtant, le résultat de ce chapitre est cohérent avec celui de l'étude Richter et al. (2020) menée à l'échelle globale et dans laquelle, à l'intérieur de chaque bassin, une structuration plus fine de la biologie est observée. Notre étude corrèle cette variabilité fine échelle avec des processus hydrodynamiques et suggère qu'elle soit liée à des activités métaboliques plus rapides et cycles de vie plus courts. Ainsi, la taille des organismes apparaît une nouvelle fois comme un facteur déterminant de la structure des communautés.

Les études menées au cours de cette thèse donnent des indices sur les processus pouvant structurer le vivant. Il serait maintenant intéressant d'estimer plus précisément la part de variation biologique découlant de chaque échelle spatiale. Pour cela, des méthodes d'analyses spatiales peuvent être utilisées (Dray et al., 2012). Par exemple, dans une étude semblable à celle menée dans le chapitre 3, les auteurs ont pu montrer que la structure génétique intraspécifique d'espèces de poisson était façonnée par des processus différents suivant l'échelle considérée (*i.e.*, rôle prépondérant de l'isolation géographique et de la dispersion larvaire respectivement aux échelles >1000 km et <1000 km, Dalongeville et al., 2018). De telles analyses pourraient permettre d'identifier les facteurs clés dans la structure des écosystèmes et les échelles associées. Enfin, l'ensemble des études menées au cours de cette thèse s'est intéressé à la structure spatiale de l'écosystème (*i.e.*, sur le plan horizontal). Ainsi, les données utilisées correspondaient systématiquement à des échantillons de surface et de jour. Pourtant, des pré-analyses effectuées dans le chapitre 4 (réalisées par Séverine Martini) ont indiqué que la composition taxonomique des communautés différait de jour et de nuit. Cela peut être mis en lien avec les comportements de migration verticale des organismes en réponse à la pression sélective de prédation (Hays, 2003). La connaissance actuelle s'est largement plus focalisée sur la couche de surface plutôt que celles de profondeur, qui présentent globalement moins de variabilité (Costello et al., 2018). Toutefois, inclure cette dimension verticale pourrait permettre de mieux comprendre les écosystèmes.

Déterminisme de la structure du vivant

L'ensemble de nos études semble montrer une influence de l'environnement sur la structuration planctonique. Ainsi, le facteur environnemental pourrait modeler la diversité génétique (chapitre 3), phénotypique (chapitres 4 et 5) et taxonomique (chapitre 5). Dans les différentes études menées ici, les axes de variance maximale ont été utilisés pour comprendre la structure des composantes abiotiques et biotiques des écosystèmes. Cette approche est pertinente pour se focaliser sur l'information contenue dans les données et en minimisant le bruit (Legendre, Legendre, 2012). Cependant, bien que nous pouvons supposer quelle variable va avoir un rôle important (*i.e.*, celle contribuant le plus à l'axe), le lien entre variables environnementales et variables biologiques n'a pas été établi alors que cette information pourrait donner des pistes intéressantes sur les fonctions écologiques impliquées. Par exemple, une influence importante de la lumière pourrait avoir un lien avec la capacité de photosynthèse ou des phénomènes de photoinhibition (Litchman, Klausmeier, 2008; Alvarez et al., 2019). La concentration en nutriment pourrait avoir un impact sur les niveaux de diazotrophie ou mixotrophie dans les communautés (Litchman, Klausmeier, 2008; Stoecker et al., 2017). Cela permettrait d'une part de comprendre les principaux déterminants des communautés aux différentes échelles (*i.e.*, filtration environnementale), mais aussi d'estimer leur tolérance face aux changements dans leur environnement.

Il est également important de noter que les analyses ne permettent pas d'affirmer que la structure biologique soit uniquement liée aux contraintes environnementales. En effet, les contraintes de dispersion n'ont pas été intégrées aux analyses des chapitres 4 et 5. Ainsi, une partie de la structure biologique observée pourrait être liée à des contraintes de dispersion. De plus, les pressions biotiques peuvent également entrer en jeu (Kraft et al., 2015; Cadotte, Tucker, 2017) et les études réalisées ne permettent pas d'aborder ce point particulier. Il serait néanmoins très intéressant d'explorer ces aspects. De premières analyses basées sur le réseau de co-occurrence dans le chapitre 4 semblent pourtant indiquer que ces interactions biotiques pourraient être une piste intéressante pour mieux comprendre les assemblages biologiques. Construire des indices représentant les structures biotiques (*e.g.*, basés sur des alignements de sous-réseaux, Mandakovic et al., 2018) pourraient peut être permettre de distinguer les signaux biotiques et environnementaux, et ainsi mieux caractériser les processus écologiques qui façonnent la diversité.

La détection des processus écologiques ou évolutifs qui façonnent le vivant reste un défi. D'une part, la résolution taxonomique a un impact sur l'estimation de la diversité

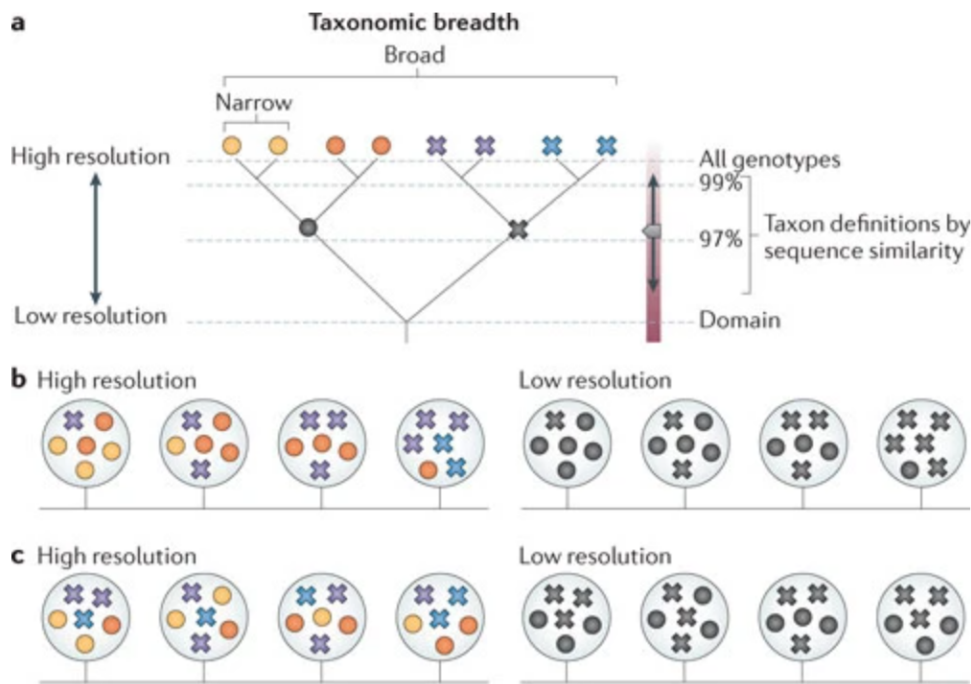


FIGURE 6.1 – Importance de la définition taxonomique sur la détection de la diversité. (a) Les symboles représentent les génotypes. La résolution taxonomique correspond au niveau de variation génétique auquel les taxons sont définis. (b) et (c) sont deux modèles biogéographiques. Le panneau b présente de l'endémisme (croix bleues) qui n'est détectable qu'à haute résolution. Le panneau c présente un cas de diversité qui varie suivant la distance détectable uniquement à haute résolution. Extrait de Hanson et al., 2012

(Figure 6.1). Les approches moléculaires permettent de mieux caractériser la proximité des organismes. Le regroupement en OTU peut aboutir à une séquence unique consensus représentant des milliers d'années d'évolution (Hanson et al., 2012). Cela peut donc complexifier la compréhension des mécanismes évolutifs ou écologiques qui entrent en jeu. D'autre part, le mélange des cadres théoriques pour expliquer les patterns biogéographiques ne facilite pas l'identification des processus (Vellend, 2010). Les terminologies et les approches analytiques différentes se focalisent sur des sous-processus différents. Ainsi, l'utilisation d'un cadre unifié entre échelles biologiques (Vellend, 2010; Hanson et al., 2012) pourrait permettre de faciliter leur compréhension, comme cela a déjà été fait sur des microbiomes (Costello et al., 2012; Logares et al., 2020).

Du gène vers l'écosystème

Dans les différentes études menées, de la structuration a été retrouvée des espèces aux communautés. Au niveau intraspécifique, la diversité a été explorée via les données génétiques. À cette fine échelle biologique, des réponses différentes des organismes ont été mises en évidence. En particulier, les résultats obtenus semblent indiquer que même pour les microorganismes au potentiel de dispersion très important, d'autres processus que ceux basés sur la pression environnementale pourraient entrer en compte, contrairement à l'hypothèse formulée par Baas-Becking (1934). D'autres études sur les microorganismes (procaryotes et eucaryotes) ont montré l'importance des processus autres que la sélection (*i.e.*, limitation de dispersion ou dérive) dans la structure des communautés (Logares et al., 2020).

La diversité intraspécifique a également été prise en compte lors des analyses d'imagerie dans les chapitres 4 et 5. Cependant, l'emphase a été portée sur les structures morphologiques moyennes. Or, dans le chapitre 4, il a été observé que la variabilité des traits pouvait être plus ou moins importante suivant le groupe taxonomique considéré. Cette variabilité peut être un facteur déterminant dans l'assemblage des communautés et la prendre en compte pourrait permettre de mieux comprendre la structuration des communautés (Bolnick et al., 2011; Violle et al., 2012).

Enfin, un aspect capital à étudier sera la mise en lien de la structure et des variations de la diversité, à ses multiples échelles, avec le fonctionnement des écosystèmes. Pour cela, plusieurs stratégies peuvent être envisagées et quelques exemples sont présentés dans la partie suivante.

6.2 Analyse de grands jeux de données

Ces dernières années ont connu une explosion du nombre de données collectées au niveau mondial. Seule une petite portion de ces données vise à valider des hypothèses spécifiques qui sont testées dans le cadre de modèles prédictifs. Ces approches nécessitent des protocoles d'échantillonnage précis permettant d'avoir la puissance de test nécessaire à la validation ou au rejet de l'hypothèse nulle. Les différentes parties de ce travail ont reposé sur des approches corrélatives ou de régression qui, bien que suggérant des liens de causalité entre variables étudiées, ne permettent pas d'établir de lien de causalité. En particulier, la régression décrit la forme d'une relation mais elle n'implique pas nécessairement une relation de causalité. Pour identifier de telles relations, il est nécessaire d'être

en milieu contrôlé et de s'assurer que seul un paramètre explicatif varie. Ainsi, analyse de données et travail expérimental sont très complémentaires. Ici, l'analyse de jeux de données de grande dimension via des méthodes de réduction dimensionnelle comme l'ACP a permis de réduire la complexité des données pour mieux comprendre leur structure et ainsi fournir des pistes sur les liens de causalité pouvant exister.

Les méthodes d'analyse basées sur l'intégration de données, combinant des ensembles de données hétérogènes dans le but d'en extraire des informations significatives, semblent particulièrement prometteuses pour mieux comprendre la structure et le fonctionnement des écosystèmes (Michener, Jones, 2012; Leonelli, 2013; García, 2019; Zipkin et al., 2021). Ces approches connaissent une utilisation croissante (Figure 6.2) et des méthodes sont proposées pour prédire et anticiper les modifications du fonctionnement des écosystèmes (Bohan et al., 2017; Cordier et al., 2017; Reichstein et al., 2019; García, 2019). Certains auteurs s'interrogent d'ailleurs de l'intérêt de comprendre la structure des écosystèmes et mécanismes sous-jacents lorsque leur fonction peut être prédite sans cette connaissance. La compréhension des jeux de données et de leur structure reste cependant essentielle. En effet, les approches intégratives nécessitent en amont de démêler les ambiguïtés sémantiques des données (Dong, Srivastava, 2013). Sans comprendre et s'être intéressé aux données, le risque de laisser passer des résultats aberrants, non qualitatifs, ou redondants est accru, ce qui peut impacter la qualité des modèles. Ainsi, en plus d'une complémentarité dans les données, la complémentarité des approches est essentielle. Dans le futur, les données acquises pourraient être valorisées par des approches d'intégration dynamique dans lesquelles les données sont incorporées en temps réel. Ce type d'approche peut notamment servir de support aux mesures de conservation en fonction de l'évolution des écosystèmes.

Les jeux de données analysés ici sont dans la lignée des données acquises ces dernières décennies. Ces données, dont seule une sous-partie a été exploitée, s'inscrivent dans le mouvement du *Big Data*. En effet, elles présentent les '4V' caractéristiques : volume, variété, véracité et la vélocité. Les données correspondent à un volume important et une multiplicité des sources (*e.g.*, 347 bases de données de biodiversité recensées par Ball-Damerow et al., 2019). Les données sont hétérogènes et portent sur des entités différentes. Les données diffèrent grandement suivant leur source, en termes de résolution et précision. Enfin, les données sont dynamiques, en perpétuelle acquisition et traitement. Le *Big Data*, combiné à des approches intégratives, tend à offrir la possibilité de pouvoir baser sur les données des décisions comme, par exemple, des mesures de préservation de la diversité. Afin d'améliorer la gestion de ces grandes quantités de données, faciliter

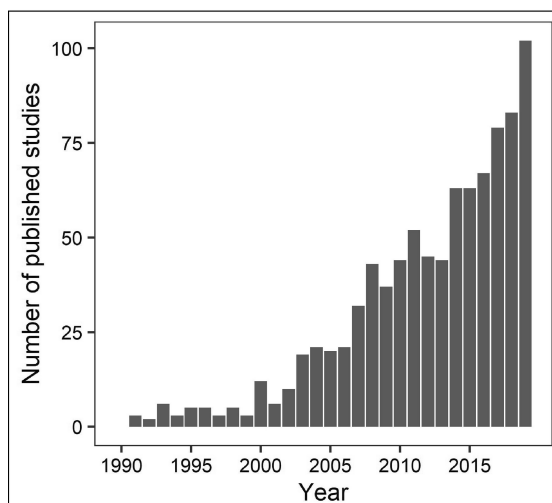


FIGURE 6.2 – Utilisation des méthodes d’intégration de données en écologie. Tandis que les méthodes d’intégration de données étaient très peu utilisées en écologie dans les années 90, elles connaissent un essor considérables depuis le début des années 2000 (extrait de Zipkin et al., 2021).

leur réutilisation et leur intégration dans le but d’exploiter leur potentiel et permettre de nouvelles découvertes, des recommandations sur les données et métadonnées ont été définies. Elles sont décrites par les principes FAIR (*i.e.*, Findable, Accessible, Interoperable et Reusable, Wilkinson et al., 2016). Les données et métadonnées sont identifiées par des identifiants uniques et pérennes. L’accès aux données doit être facilité et possible via des protocoles de communication standards (*e.g.*, FTP pour le transfert de fichier). La description des données est basée sur un vocabulaire standardisé et indique les liens avec d’autres données afin de faciliter leur exploitation. Enfin, afin de favoriser la réutilisation des données, il est préconisé d’utiliser des protocoles standards pour générer les données, informer sur la provenance des données ou indiquer les licences d’utilisation. Cette thèse a été menée en essayant de respecter au mieux ces principes.

Finalement, le déluge de données de ces dernières années a été associé à une profonde modification dans la gestion des données et leur exploitation. La dernière partie présente quelques perspectives sur les pratiques de travail en écologie numérique et computationnelle.

6.3 Écologie, une discipline en mutation constante

6.3.1 Pluridisciplinarité, spécialisation et collaboration en écologie

L'écologie vise à fournir une compréhension du vivant dans son environnement, incluant toutes les relations biotiques et abiotiques qui y prennent place. À cette fin, elle met en relation des disciplines initialement cloisonnées. De par l'objet de son étude, elle est liée à la biologie. Elle est connectée à la géologie car les structures géologiques ou la nature du milieu conditionnent les habitats. Elle est en relation avec la physique, qui régit par exemple l'écoulement des fluides et donc les schémas de circulation. Les réactions moléculaires dépendent de la chimie. Elle implique des composantes climatologique et météorologique. Et l'écologie inclut aussi une dimension politique-sociétale-économique en considérant les services écosystémiques rendus. À cette pluridisciplinarité thématique s'ajoutent la statistique pour l'analyse des données et l'informatique permettant le traitement des données (*e.g.*, bioinformatique, modélisation). Ainsi, mener des études en écologie implique une connaissance hautement pluridisciplinaire, que ce soit du point de vue théorique mais également technique ou méthodologique. De ce fait, de nombreuses collaborations sont mises en place dans le paysage de la recherche. Ma thèse s'est inscrite dans cette dynamique. Elle se situe à l'interface de plusieurs domaines scientifiques et a bénéficié d'interactions entre domaines.

Les collaborations en recherche sont basées sur des initiatives variées. Par exemple, certains groupes de recherche (GDR) visent à comprendre les besoins réciproques entre disciplines. C'est le cas du GDR ecostats qui a organisé un workshop, auquel j'ai assisté, permettant un pont entre intelligence artificielle et écologie. Le but était d'identifier des questions écologiques pertinentes pouvant générer un défi technique, ce qui engendre un bénéfice mutuel pour les deux disciplines (*i.e.* écologie et informatique). Un autre exemple est celui des nombreux Consortium de recherche, comme ceux de *Tara*, qui mettent en relation des chercheurs et des laboratoires aux compétences et expertises très différentes. Ces initiatives créent une véritable dynamique pluridisciplinaire pour aborder de nouveaux défis scientifiques. Toutefois, l'exercice des collaborations reste périlleux. Il nécessite de la confiance et de la communication entre ses différents membres au risque d'échouer. Au contraire, une collaboration fructueuse peut être à l'origine d'une nouvelle façon de concevoir la discipline et potentiellement faire émerger de nouveaux champs de recherche.

6.3.2 Le besoin de travailler ensemble, oui mais comment ?

Les données à haut débit acquises permettent la communication entre disciplines initialement séparées, afin de contribuer à la compréhension holistique des écosystèmes. En raison de leur nature variée, elles font intervenir plusieurs opérateurs dans leur traitement et leur analyse. Par exemple, dans le cadre des expéditions *Tara*, le séquençage des données et l'acquisition des images n'ont pas été réalisés par les mêmes laboratoires. Au niveau du traitement des données, cela peut également être le cas. Certaines tâches sont donc cloisonnées et il est important de maintenir communication et transparence. Cela s'inscrit dans une démarche plus large de reproductibilité et de transparence en science.

Au début des années 2000, il y a eu un séisme dans le monde de la recherche qui a été désigné comme la 'Crise de la reproductibilité'. L'existence de mauvais comportements scientifiques a été mise en lumière. Historiquement, il y a des exemples de malhonnêteté scientifique. Par exemple, en 1974, Dr William Summerlin déclarait faire des transplantations d'organes inter-espèces entre souris blanches et souris à tâches noires. Or, les tâches avaient été artificiellement créées avec un marqueur. Au delà des problèmes de falsification (dont 10% de dissimulation, 20% de conflits d'intérêt et 30% d'entrave à l'intégrité ; Martinson et al., 2005), la crise de la reproductibilité a mis en évidence que de nombreuses études ne pouvaient pas être vérifiées (Baker, 2016). Ce scandale a dépassé la sphère scientifique et atteint le grand public via la presse généraliste. Le *New York Times* a par exemple révélé le nombre croissant d'articles qui étaient rétractés au cours du temps. Parmi ces articles figuraient les études impossibles à reproduire. Contrairement à la réalité, les études de nature computationnelle ont été tout d'abord pensées comme préservées de ce problème de reproductibilité. Or, seulement 26% des analyses ont été estimées reproductibles. Cette reproductibilité est décomposable en plusieurs niveaux : ré-exécution, répétabilité, répliquabilité, reproductibilité et réutilisation (Benureau, Rougier, 2018). En particulier, d'une étude non reproductible à une étude totalement répliquable, il existe un spectre de reproductibilité (Peng, 2011). Depuis cette crise de la reproductibilité, il y a eu un vif intérêt de la communauté scientifique sur les façons de rendre la science plus transparente. Au cours de ce travail, une attention particulière a été portée aux questions de reproductibilité et de transparence.

Pour être reproductible, une étude doit au minimum fournir une description détaillée du travail effectué (*i.e.* pre-reproductible, Stark, 2018). Le partage des études est également encouragé. Comme l'indiquait Philip B. Stark "Science should be 'show me', not 'trust me'; it should be 'help me if you can', not 'catch me if you can'". L'ensemble de la

communauté scientifique doit être impliquée pour permettre une science plus transparente, plus collaborative et à terme plus performante (Stark, 2018; Eren et al., 2021). Le partage doit être encouragé dans le but d'aider la communauté et de permettre de détecter de potentielles erreurs. Dans les deux cas, les retombées sont positives et conduisent à une amélioration des connaissances. Les revues scientifiques peuvent également encourager la transparence. Certaines pourront mettre en place des systèmes de signet pour valoriser les études transparentes. Dans cette optique, différents outils existent pour partager les études scientifiques, faciliter le travail collaboratif et améliorer la reproductibilité (Figure 6.3).

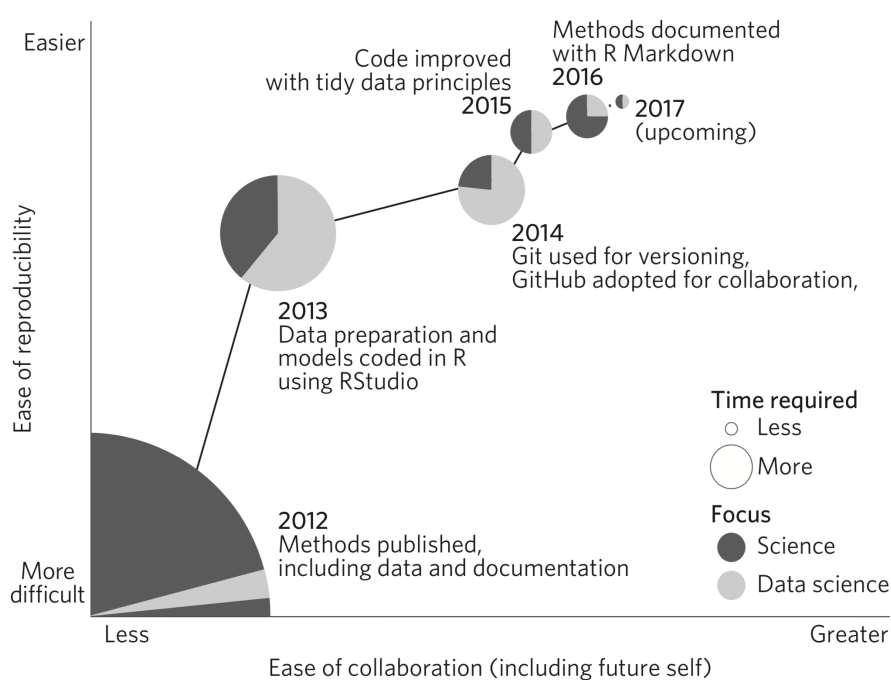


FIGURE 6.3 – Collaboration computationnelle dans les projets de recherche. De nombreux outils ont été développés et peuvent être utilisés dans les projets de recherche. Ainsi, il est de plus en plus facile de collaborer et d'améliorer la reproductibilité lors d'études collaboratives (modifié depuis Lowndes et al., 2017).

Différentes solutions soutiennent le développement d'une recherche plus transparente. Les environnements de développement intégré, tels de RStudio (R) ou PyCharm (python), facilitent la production de code. L'archivage des données et produits d'étude est simplifié via l'émergence de plateforme de dépôts (*e.g.*, Zenodo, FigShare). Les codes à l'origine des résultats peuvent être partagés sur les hébergeurs web (*e.g.*, Git, GitHub). Ils présentent l'avantage de posséder des systèmes de version retraçant les modifications

apportées, favorisent la gestion de projet et les collaborations, et disposent d'espaces d'échange permettant la communication entre développeurs et utilisateurs (*e.g.*, report de bugs, aide). Par analogie avec les laboratoires, l'utilisation de carnets de notes numériques, ou *notebook*, est également encouragée (*e.g.*, R Markdown, Jupiter Notebook ou Emacs). Ils permettent de garder la trace des analyses, des paramètres utilisés et de visualiser les sorties générées. Les workflows, qui informent sur les différents modules d'une analyse et leurs entrées/sorties, sont également très développés (*e.g.*, Galaxy, Goecks et al., 2010; Taverna, Oinn et al., 2004; SnakeMake, Köster, Rahmann, 2012). Ils peuvent être partagés et directement réutilisés. Enfin, avec l'évolution constante des bibliothèques, des logiciels et autres ressources informatiques, le maintien des codes dans le temps est également un défi (Perkel, 2020). Pour cela, des machines virtuelles qui contiennent l'environnement de programmation ou des conteneurs qui contiennent uniquement les dépendances (*e.g.*, conda, docker) permettent une utilisation durable des codes développés. Finalement, de nombreux outils, protocoles ou recommandations sont à disposition pour faciliter et améliorer la recherche en écologie numérique et computationnelle. Ainsi, la recherche en écologie est semblable à son modèle d'étude : en mutation constante dans son environnement.

Bibliographie

- Acinas Silvia G., Sarma-Rupavtarm Ramahi, Klepac-Ceraj Vanja, Polz Martin F.* PCR-Induced Sequence Artifacts and Bias : Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample // *Applied and Environmental Microbiology*. 2005. 71, 12. 8966–8969.
- Aiken J., Rees N., Hooker S., Holligan P., Bale A., Robins D., Moore G., Harris R., Pilgrim D.* The Atlantic Meridional Transect : overview and synthesis of data // *Progress in Oceanography*. 2000. 45, 3. 257–312.
- Aitchison John.* Logratios and natural laws in compositional data analysis // *Mathematical Geology*. 1999. 31, 5. 563–580.
- Alberti Adriana, Poulain Julie, Engelen Stefan, Labadie Karine, Romac Sarah, Ferrera Isabel, Albin Guillaume, Aury Jean-Marc, Belser Caroline, Bertrand Alexis, Cruaud Corinne, Da Silva Corinne, Dossat Carole, Gavory Frédéric, Gas Shahinaz, Guy Julie, Haquelle Maud, Jacoby E'krame, Jaillon Olivier, Lemaître Arnaud, Pelletier Eric, Samson Gaëlle, Wessner Mark, Bazire Pascal, Beluche Odette, Bertrand Laurie, Besnard-Gonnet Marielle, Bordelais Isabelle, Boutard Magali, Dubois Maria, Dumont Corinne, Etedgui Evelyne, Fernandez Patricia, Garcia Espérance, Aiach Nathalie Giordanenco, Guerin Thomas, Hamon Chadia, Brun Elodie, Lebled Sandrine, Lenoble Patricia, Louesse Claudine, Mahieu Eric, Mairey Barbara, Martins Nathalie, Megret Catherine, Milani Claire, Muanga Jacqueline, Orvain Céline, Payen Emilie, Perroud Peggy, Petit Emmanuelle, Robert Dominique, Ronsin Murielle, Vacherie Benoit, Acinas Silvia G., Royo-Llonch Marta, Cornejo-Castillo Francisco M., Logares Ramiro, Fernández-Gómez Beatriz, Bowler Chris, Cochrane Guy, Amid Clara, Hoopen Petra Ten, De Vargas Colomban, Grimsley Nigel, Desgranges Elodie, Kandels-Lewis Stefanie, Ogata Hiroyuki, Poulton Nicole, Sieracki Michael E., Stepanauskas Ramunas, Sullivan Matthew B., Brum Jennifer R., Duhaime Melissa B., Poulos Bonnie T., Hurwitz Bonnie L., Acinas Silvia G., Bork Peer, Boss Emmanuel, Follows Michael, Gorsky Gabriel, Hingamp Pascal, Iudicone Daniele, Karp-Boss Lee, Karsenti Eric, Not Fabrice, Pesant Stéphane, Raes Jeroen, Sardet Christian, Sieracki Michael E., Speich Sabrina, Stemmann Lars, Sullivan Matthew B., Sunagawa Shinichi, Wincker Patrick, Team Genoscope Technical, Coordinators Tara Oceans Consortium.* Viral to meta-zoan marine plankton nucleotide sequences from the Tara Oceans expedition // *Scientific Data*. 2017. 4, 1. 170093.
- Allen John T., Brown Louise, Sanders Richard, Mark Moore C., Mustard Alexander, Fielding Sophie, Lucas Mike, Rixen Michel, Savidge Graham, Henson Stephanie, Mayor Dan.* Diatom

- carbon export enhanced by silicate upwelling in the northeast Atlantic // *Nature*. 2005. 437, 7059. 728–732.
- Allendorf Fred W., Hohenlohe Paul A., Luikart Gordon.* Genomics and the future of conservation genetics // *Nature Reviews Genetics*. 2010. 11, 10. 697–709.
- Alvarez E., Thoms S., Bracher A., Liu Y., Völker C.* Modeling Photoprotection at Global Scale : The Relative Role of Nonphotosynthetic Pigments, Physiological State, and Species Composition // *Global Biogeochemical Cycles*. 2019. 33, 7. 904–926.
- Andersen Robert A., Saunders Gary W., Paskind Michael P., Sexton Julianne P.* Ultrastructure and 18s rRNA gene sequence for *pelagomonas calceolata* gen. Et Sp. nov. and the description of a new algal class, the pelagophyceae classis nov. 1 // *Journal of Phycology*. 1993. 29, 5. 701–715.
- Ardyna Mathieu, d'Ovidio Francesco, Speich Sabrina, Leconte Jade, Chaffron Samuel, Audic Stephane, Garczarek Laurence, Pesant Stephane, Tara Oceans Consortium Coordinators, Tara Oceans Expedition Participants.* Environmental context of all samples from the Tara Oceans Expedition (2009-2013), about mesoscale features at the sampling location. 2017. In : *Tara Oceans Consortium, Coordinators ; Tara Oceans Expedition, Participants (2017) : Registry of all samples from the Tara Oceans Expedition (2009-2013)*. PANGAEA, <https://doi.org/10.1594/PANGAEA.875582>.
- Arenas Gómez Claudia M., Sabin Keith Z., Echeverri Karen.* Wound healing across the animal kingdom : Crosstalk between the immune system and the extracellular matrix // *Developmental Dynamics*. 2020. 249, 7. 834–846.
- Arias Mónica, Mappes Johanna, Desbois Charlotte, Gordon Swanne, McClure Melanie, Elias Marianne, Nokelainen Ossi, Gomez Doris.* Transparency reduces predator detection in mimetic clearwing butterflies // *Functional Ecology*. 2019. 33, 6. 1110–1119.
- Armstrong Robert A., McGehee Richard.* Competitive Exclusion // *The American Naturalist*. 1980. 115, 2. 151–170.
- Atkinson Anthony C., Riani Marco, Corbellini Aldo.* The Box–Cox Transformation : Review and Extensions // *Statistical Science*. 2021. 36, 2. 239 – 255.
- Ayata Sakina-Dorothee, Irisson Jean-Olivier, Aubert Anais, Berline Léo, Dutay Jean-Claude, Mayot Nicolas, Nieblas Anne-Elise, d'Ortenzio Fabrizio, Palmieri Julien, Reygondeau Gabriel, Rossi Vincent, Guieu Cécile.* Regionalisation of the Mediterranean basin, a MERMEX synthesis // *Progress in Oceanography*. 2018. 163. 7–20.
- Ayata Sakina-Dorothee, Lazure Pascal, Thiébaud Éric.* How does the connectivity between populations mediate range limits of marine invertebrates? A case study of larval dispersal between the Bay of Biscay and the English Channel (North-East Atlantic) // *Progress in Oceanography*. 2010. 87, 1-4. 18–36.
- Baas-Becking Lourens Gerhard Marinus.* *Geobiologie ; of inleiding tot de milieukunde*. 1934.
- Bagge Laura E.* Not as clear as it may appear : challenges associated with transparent camouflage in the Ocean // *Integrative and Comparative Biology*. 2019. 6. 1653–1663.
- Baker Monya.* 1,500 scientists lift the lid on reproducibility // *Nature*. 2016. 533, 7604. 452–454.

- Ball-Damerow Joan E., Brenskelle Laura, Barve Narayani, Soltis Pamela S., Sierwald Petra, Bieler Rüdiger, LaFrance Raphael, Ariño Arturo H., Guralnick Robert P.* Research applications of primary biodiversity databases in the digital age // PLOS ONE. 09 2019. 14, 9. 1–26.
- Barabás György, D’Andrea Rafael, Stump Simon Maccracken.* Chesson’s coexistence theory // Ecological Monographs. 2018. 88, 3. 277–303.
- Barton Andrew D, Irwin Andrew J, Finkel Zoe V, Stock Charles A.* Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities // Proceedings of the National Academy of Sciences. 2016. 113, 11. 2964–2969.
- Barton Andrew D., Pershing Andrew J., Litchman Elena, Record Nicholas R., Edwards Kyle F., Finkel Zoe V., Kiørboe Thomas, Ward Ben A.* The biogeography of marine plankton traits // Ecology letters. 2013. 16, 4. 522–534.
- Baselga Andrés.* Partitioning the turnover and nestedness components of beta diversity // Global ecology and biogeography. 2010. 19, 1. 134–143.
- Bashevkin Samuel M., Christy John H., Morgan Steven G.* Photoprotective benefits of pigmentation in the transparent plankton community : a comparative species experimental test // Ecology. 2019. 100, 5. e02680.
- Bates Amanda E., Helmuth Brian, Burrows Michael T., Duncan Murray I., Garrabou Joaquim, Guy-Haim Tamar, Lima Fernando, Queiros Ana M., Seabra Rui, Marsh Robert, Belmaker Jonathan, Bensoussan Nathaniel, Dong Yunwei, Mazaris Antonios D., Smale Dan, Wahl Martin, Rilov Gil.* Biologists ignore ocean weather at their peril. 2018.
- Batten Sonia D., Abu-Alhaija Rana, Chiba Sanae, Edwards Martin, Graham George, Jyothibabu R., Kitchener John A., Koubbi Philippe, McQuatters-Gollop Abigail, Muxagata Erik, Ostle Clare, Richardson Anthony J., Robinson Karen V., Takahashi Kunio T., Verheye Hans M., Wilson Willie.* A global plankton diversity monitoring program // Frontiers in Marine Science. 2019. 6. 321.
- Beaugrand Grégory.* The North Sea regime shift : Evidence, causes, mechanisms and consequences // Progress in Oceanography. 2004. 60, 2. 245–262. Regime shifts in the ocean. Reconciling observations and theory.
- Beaugrand Grégory.* Monitoring pelagic ecosystems using plankton indicators // ICES Journal of Marine Science. 2005. 62, 3. 333–338.
- Beaugrand Grégory, Edwards Martin, Legendre Louis.* Marine biodiversity, ecosystem functioning, and carbon cycles // Proceedings of the National Academy of Sciences. 2010. 107, 22. 10120–10124.
- Beaugrand Gregory, Luczak Christophe, Edwards Martin.* Rapid biogeographical plankton shifts in the North Atlantic Ocean // Global Change Biology. 2009. 15, 7. 1790–1803.
- Beaugrand Grégory, Reid Philip C., Ibañez Frédéric, Lindley J. Alistair, Edwards Martin.* Reorganization of North Atlantic Marine Copepod Biodiversity and Climate // Science. 2002. 296, 5573. 1692–1694.
- Becker Richard A, Wilks Allan R, Brownrigg Ray, Minka Thomas P, Deckmyn Alex.* maps : Draw geographical maps // R package version. 2018. 3, 0.

- Benedetti Fabio, Guilhaumon François, Adloff Fanny, Ayata Sakina-Dorothee.* Investigating uncertainties in zooplankton composition shifts under climate change scenarios in the Mediterranean Sea // *Ecography*. 2018. 41, 2. 345–360.
- Benedetti Fabio, Jalabert Laëtizia, Sourisseau Marc, Becker Beatriz, Cailliau Caroline, Desnos Corinne, Elineau Amanda, Irisson Jean-Olivier, Lombard Fabien, Picheral Marc, Stemmann Lars, Poulaine Patrick.* The Seasonal and Inter-Annual Fluctuations of Plankton Abundance and Community Structure in a North Atlantic Marine Protected Area // *Frontiers in Marine Science*. 2019. 6. 214.
- Benureau Fabien C. Y., Rougier Nicolas P.* Re-run, Repeat, Reproduce, Reuse, Replicate : Transforming Code into Scientific Contributions // *Frontiers in Neuroinformatics*. 2018. 11. 69.
- Berline Léo, Rammou Anna-Maria, Doglioli Andrea, Molcard Anne, Petrenko Anne.* A connectivity-based eco-regionalization method of the Mediterranean Sea // *PloS one*. 2014. 9, 11. e111978.
- Bernard Guillaume, Pathmanathan Jananan S., Lannes Romain, Lopez Philippe, Baptiste Eric.* Microbial Dark Matter Investigations : How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery // *Genome Biology and Evolution*. 2018. 10, 3. 707–715.
- Biard Tristan, Bigeard Estelle, Audic Stéphane, Poulain Julie, Gutierrez-Rodriguez Andres, Pesant Stéphane, Stemmann Lars, Not Fabrice.* Biogeography and diversity of Collodaria (Radiolaria) in the global ocean // *The ISME Journal*. 2017. 11, 6. 1331–1344.
- Biard Tristan, Stemmann Lars, Picheral Marc, Mayot Nicolas, Vandromme Pieter, Hauss Helena, Gorsky Gabriel, Guidi Lionel, Kiko Rainer, Not Fabrice.* In situ imaging reveals the biomass of giant protists in the global ocean // *Nature*. 2016. 532, 7600. 504–507.
- Bittner Lucie, Gobet Angélique, Audic Stéphane, Romac Sarah, Egge Elianne S., Santini Sébastien, Ogata Hiroyuki, Probert Ian, Edvardsen Bente, De Vargas Colomaban.* Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay // *Molecular ecology*. 2013a. 22, 1. 87–101.
- Bittner Lucie, Gobet Angélique, Audic Stéphane, Romac Sarah, Egge Elianne S., Santini Sébastien, Ogata Hiroyuki, Probert Ian, Edvardsen Bente, Vargas Colomaban de.* Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay // *Molecular Ecology*. 2013b. 22, 1. 87–101.
- Blanc-Mathieu Romain, Krasovec Marc, Hebrard Maxime, Yau Sheree, Desgranges Elodie, Martin Joel, Schackwitz Wendy, Kuo Alan, Salin Gerald, Donnadieu Cecile, Desdevises Yves, Sanchez-Ferandin Sophie, Moreau Hervé, Rivals Eric, Grigoriev Igor V., Grimsley Nigel, Eyre-Walker Adam, Piganeau Gwenael.* Population genomics of picophytoplankton unveils novel chromosome hypervariability // *Science advances*. 2017. 3, 7. e1700239.
- Blanco-Bercial Leocadio.* Metabarcoding Analyses and Seasonality of the Zooplankton Community at BATS // *Frontiers in Marine Science*. 2020. 7. 173.
- Bloodgood Robert A.* Chapter 39 Flagellar Surface Motility : Gliding and Microsphere Movements. 47. 1995. 273–279. (Methods in Cell Biology).

- Bohan David A., Vacher Corinne, Tamaddoni-Nezhad Alireza, Raybould Alan, Dumbrell Alex J., Woodward Guy.* Next-Generation Global Biomonitoring : Large-scale, Automated Reconstruction of Ecological Networks // *Trends in Ecology & Evolution.* 2017. 32, 7. 477–487.
- Bohonak Andrew J.* Dispersal, gene flow, and population structure // *The Quarterly review of biology.* 1999. 74, 1. 21–45.
- Boissezon Aurélie.* Distribution et dynamique des communautés de Characées : Impact des facteurs environnementaux régionaux et locaux. 2014.
- Bolger Anthony M., Lohse Marc, Usadel Bjoern.* Trimmomatic : a flexible trimmer for Illumina sequence data // *Bioinformatics.* 2014. 30, 15. 2114–2120.
- Bollens Stephen M., Quenette Joel A., Rollwagen-Bollens Gretchen.* Predator-enhanced diel vertical migration in a planktonic dinoflagellate // *Marine Ecology Progress Series.* 2012. 447. 49–54.
- Bolnick Daniel I., Amarasekare Priyanga, Araújo Márcio S., Bürger Reinhard, Levine Jonathan M., Novak Mark, Rudolf Volker H. W., Schreiber Sebastian J., Urban Mark C., Vasseur David A.* Why intraspecific trait variation matters in community ecology // *Trends in Ecology & Evolution.* 2021/07/27 2011. 26, 4. 183–192.
- Bolte Sören, Fuentes Veronica, Haslob Holger, Huwer Bastian, Thibault-Botha Delphine, Angel Dror, Galil Bella, Javidpour Jamileh, Moss Anthony G, Reusch Thorsten BH.* Population genetics of the invasive ctenophore *Mnemiopsis leidyi* in Europe reveal source–sink dynamics and secondary dispersal to the Mediterranean Sea // *Marine Ecology Progress Series.* 2013. 485. 25–36.
- Bonnet Delphine, Richardson Anthony, Harris Roger, Hirst Andrew, Beaugrand Gregory, Edwards Martin, Ceballos Sara, Diekman Rabea, López-Urrutia Angel, Valdes Luis, Carlotti François, Molinero Juan Carlos, Weikert Horst, Greve Wulf, Lucic Davor, Albaina Aitor, Yahia Nejib Daly, Umani Serena Fonda, Miranda Ana, Santos Antonina dos, Cook Kathryn, Robinson Susan, Fernandez de Puellas Marie Luz.* An overview of *Calanus helgolandicus* ecology in European waters // *Progress in Oceanography.* 2005. 65, 1. 1–53.
- Booy G., Hendriks R. J. J., Smulders M. J. M., Van Groenendael J. M., Vosman B.* Genetic Diversity and the Survival of Populations // *Plant Biol (Stuttg).* 2000. 2, 04. 379–395.
- Boyd Philip W., Claustre Hervé, Levy Marina, Siegel David A., Weber Thomas.* Multi-faceted particle pumps drive carbon sequestration in the ocean // *Nature.* 2019. 568, 7752. 327–335.
- Boyer Montégut Clément de, Madec Gurvan, Fischer Albert S., Lazar Alban, Iudicone Daniele.* Mixed layer depth over the global ocean : An examination of profile data and a profile-based climatology // *Journal of Geophysical Research : Oceans.* 2004. 109, C12.
- Bracco A., Provenzale A., Scheuring I.* Mesoscale vortices and the paradox of the plankton // *Proceedings of the Royal Society of London. Series B : Biological Sciences.* 2000. 267, 1454. 1795–1800.
- Brown James H., Gillooly James F., Allen Andrew P., Savage Van M., West Geoffrey B.* TOWARD A METABOLIC THEORY OF ECOLOGY // *Ecology.* 2004. 85, 7. 1771–1789.

- Brun Philipp, Payne Mark R., Kiørboe Thomas.* Trait biogeography of marine copepods—an analysis across scales // *Ecology Letters*. 2016. 19, 12. 1403–1413.
- Brun Philipp, Stamieszkin Karen, Visser Andre W, Licandro Priscilla, Payne Mark R, Kiørboe Thomas.* Climate change has altered zooplankton-fuelled carbon export in the North Atlantic // *Nature ecology & evolution*. 2019. 3, 3. 416–423.
- Burke Kevin.* Plate tectonics, the Wilson Cycle, and mantle plumes : geodynamics from the top // *Annual Review of Earth and Planetary Sciences*. 2011. 39. 1–29.
- Cadotte Marc W., Tucker Caroline M.* Should environmental filtering be abandoned? // *Trends in ecology & evolution*. 2017. 32, 6. 429–437.
- Caesar L., Rahmstorf S., Robinson A., Feulner G., Saba V.* Observed fingerprint of a weakening Atlantic Ocean overturning circulation // *Nature*. 2018. 556, 7700. 191–196.
- Caputi Luigi, Carradec Quentin, Eveillard Damien, Kirilovsky Amos, Pelletier Eric, Piarella Karlusich Juan J., Rocha Jimenez Vieira Fabio, Villar Emilie, Chaffron Samuel, Malviya Shruti, Scalco Eleonora, Acinas Silvia G., Alberti Adriana, Aury Jean-Marc, Benoiston Anne-Sophie, Bertrand Alexis, Biard Tristan, Bittner Lucie, Boccara Martine, Brum Jennifer R., Brunet Christophe, Busseni Greta, Carratalà Anna, Claustre Hervé, Coelho Luis Pedro, Colin Sébastien, D’Aniello Salvatore, Da Silva Corinne, Del Core Marianna, Doré Hugo, Gasparini Stéphane, Kokoszka Florian, Jamet Jean-Louis, Lejeusne Christophe, Lepoivre Cyrille, Lescot Magali, Lima-Mendez Gipsi, Lombard Fabien, Lukeš Julius, Maillat Nicolas, Madoui Mohammed-Amin, Martinez Elodie, Mazzocchi Maria Grazia, Néou Mario B., Paz-Yepes Javier, Poulain Julie, Ramondenc Simon, Romagnan Jean-Baptiste, Roux Simon, Salvaggio Manta Daniela, Sanges Remo, Speich Sabrina, Sprovieri Mario, Sunagawa Shinichi, Taillandier Vincent, Tanaka Atsuko, Tirichine Leila, Trottier Camille, Uitz Julia, Veluchamy Alaguraj, Veselá Jana, Vincent Flora, Yau Sheree, Kandels-Lewis Stefanie, Searson Sarah, Dimier Céline, Picheral Marc, Coordinators Tara Oceans, Bork Peer, Boss Emmanuel, Vargas Colombar de, Follows Michael J., Grimsley Nigel, Guidi Lionel, Hingamp Pascal, Karsenti Eric, Sordino Paolo, Stemmann Lars, Sullivan Matthew B., Tagliabue Alessandro, Zingone Adriana, Garczarek Laurence, d’Ortenzio Fabrizio, Testor Pierre, Not Fabrice, d’Alcalá Maurizio Ribera, Wincker Patrick, Bowler Chris, Iudicone Daniele.* Community-level responses to iron availability in open ocean plankton ecosystems // *Global Biogeochemical Cycles*. 2019. 33, 3. 391–419.
- Caron David A., Alexander Harriet, Allen Andrew E., Archibald John M., Armbrust E. Virginia, Bachy Charles, Bell Callum J., Bharti Arvind, Dyhrman Sonya T., Guida Stephanie M., Heidelberg Karla B., Kaye Jonathan Z., Metzner Julia, Smith Sarah R., Worden Alexandra Z.* Probing the evolution, ecology and physiology of marine protists using transcriptomics // *Nature Reviews Microbiology*. 2017. 15, 1. 6–20.
- Caron David A., Countway Peter D., Jones Adriane C., Kim Diane Y., Schnetzer Astrid.* Marine protistan diversity // *Annual review of marine science*. 2012. 4. 467–493.
- Caron Jean-Michel, Gauthier Alain.* Planete terre. 2007.
- Carradec Quentin, Pelletier Eric, Da Silva Corinne, Alberti Adriana, Seeleuthner Yoann, Blanc-Mathieu Romain, Lima-Mendez Gipsi, Rocha Fabio, Tirichine Leila, Labadie Karine, Kirilovsky Amos, Bertrand Alexis, Engelen Stefan, Madoui Mohammed-Amin, Méheust Raphaël,*

- Poulain Julie, Romac Sarah, Richter Daniel J., Yoshikawa Genki, Dimier Céline, Kandels-Lewis Stefanie, Picheral Marc, Searson Sarah, Acinas Silvia G., Boss Emmanuel, Follows Michael, Gorsky Gabriel, Grimsley Nigel, Karp-Boss Lee, Krzic Uros, Pesant Stephane, Reynaud Emmanuel G., Sardet Christian, Sieracki Mike, Speich Sabrina, Stemmann Lars, Velayoudon Didier, Weissenbach Jean, Jaillon Olivier, Aury Jean-Marc, Karsenti Eric, Sullivan Matthew B., Sunagawa Shinichi, Bork Peer, Not Fabrice, Hingamp Pascal, Raes Jeroen, Guidi Lionel, Ogata Hiroyuki, Vargas Colomban de, Iudicone Daniele, Bowler Chris, Wincker Patrick, Coordinators Tara Oceans.* A global ocean atlas of eukaryotic genes // *Nature communications*. 2018. 9, 1. 1–13.
- Casabianca Silvia, Penna Antonella, Pecchioli Elena, Jordi Antoni, Basterretxea Gotzon, Vernesi Cristiano.* Population genetic structure and connectivity of the harmful dinoflagellate *Alexandrium minutum* in the Mediterranean Sea // *Proceedings of the Royal Society B : Biological Sciences*. 2012. 279, 1726. 129–138.
- Casteleyn Griet, Adams Nicolaus G., Vanormelingen Pieter, Debeer Ann-Eline, Sabbe Koen, Vyverman Wim.* Natural hybrids in the marine diatom *Pseudo-nitzschia pungens* (Bacillariophyceae) : genetic and morphological evidence // *Protist*. 2009. 160, 2. 343–354.
- Casteleyn Griet, Leliaert Frederik, Backeljau Thierry, Debeer Ann-Eline, Kotaki Yuichi, Rhodes Lesley, Lundholm Nina, Sabbe Koen, Vyverman Wim.* Limits to gene flow in a cosmopolitan marine planktonic diatom // *Proceedings of the National Academy of Sciences*. 2010. 107, 29. 12952–12957.
- Cavender-Bares Jeannine, Kozak Kenneth H., Fine Paul V. A., Kembel Steven W.* The merging of community ecology and phylogenetic biology // *Ecology letters*. 2009. 12, 7. 693–715.
- Cermeño Pedro, Falkowski Paul G.* Controls on diatom biogeography in the ocean // *Science*. 2009. 325, 5947. 1539–1541.
- Chaffron S., Delage E., Budinich M., Vintache D., Henry N., Nef C., Ardyna M., Zayed A.A., Junger P.C., Galand P.E., Lovejoy C., Murray A., Sarmiento H., coordinators Tara Oceans, Acinas S., Babin M., Iudicone D., Jaillon O., Karsenti E., Wincker P., Karp-Boss L., Sullivan M.B., Bowler C., Vargas C. de, Eveillard D.* Environmental vulnerability of the global ocean plankton community interactome // *Science Advances*. 2021. 7, 35. eabg1921.
- Chai Fei, Johnson Kenneth S., Claustre Hervé, Xing Xiaogang, Wang Yuntao, Boss Emmanuel, Riser Stephen, Fennel Katja, Schofield Oscar, Sutton Adrienne.* Monitoring ocean biogeochemistry with autonomous platforms // *Nature Reviews Earth & Environment*. 2020. 1, 6. 315–326.
- Chang F. Hoe, Sutherland Judith E., McVeagh Margaret, Gall Mark.* Molecular phylogeny, pigment composition, toxicology and life history of *Pseudochattonella cf. verruculosa* (Class Dictyochophyceae) from Wellington Harbour, New Zealand // *Harmful Algae*. 2014. 34. 42–55.
- Chaput Romain, Majoris John E., Buston Peter M., Paris Claire B.* Hydrodynamic and biological constraints on group cohesion in plankton // *Journal of Theoretical Biology*. 2019. 482. 109987.
- Chase Jonathan M.* Towards a really unified theory for metacommunities // *Functional Ecology*. 2005. 19, 1. 182–186.

- Chase Jonathan M., Leibold Mathew A.* Ecological niches : Linking classical and contemporary approaches. 2009.
- Chave Jérôme.* Neutral theory and community ecology // Ecology letters. 2004. 7, 3. 241–253.
- Check Hayden Erika.* Technology : The \$1,000 genome // Nature. 2014. 507, 7492. 294–295.
- Chelton Dudley B., Schlax Michael G., Samelson Roger M.* Global observations of nonlinear mesoscale eddies // Progress in Oceanography. 2011. 91, 2. 167–216.
- Chesson Peter.* General Theory of Competitive Coexistence in Spatially-Varying Environments // Theoretical Population Biology. 2000. 58, 3. 211–237.
- Chesson Peter.* Updates on mechanisms of maintenance of species diversity // Journal of Ecology. 2018. 106, 5. 1773–1794.
- Chi Xupeng, Dierking Jan, Hoving Henk-Jan, Lüskow Florian, Denda Anneke, Christiansen Bernd, Sommer Ulrich, Hansen Thomas, Javidpour Jamileh.* Tackling the jelly web : Trophic ecology of gelatinous zooplankton in oceanic food webs of the eastern tropical Atlantic assessed by stable isotope analysis // Limnology and Oceanography. 2021. 66, 2. 289–305.
- Chia Fu-Shiang, Buckland-Nicks John, Young Craig M.* Locomotion of marine invertebrate larvae : a review // Canadian Journal of Zoology. 1984. 62, 7. 1205–1222.
- Chivers William J., Walne Anthony W., Hays Graeme C.* Mismatch between marine plankton range movements and the velocity of climate change // Nature Communications. 2017. 8, 1. 14434.
- Chung Sunduz Dongjunand Keles.* Sparse Partial Least Squares Classification for High Dimensional Data // Statistical Applications in Genetics and Molecular Biology. 2010. 9, 1. 11–75.
- Chust Guillem, Irigoien Xabier, Chave Jerome, Harris Roger P.* Latitudinal phytoplankton distribution and the neutral theory of biodiversity // Global Ecology and Biogeography. 2013. 22, 5. 531–543.
- Chust Guillem, Vogt Meike, Benedetti Fabio, Nakov Teofil, Villéger Sébastien, Aubert Anaïs, Vallina Sergio M., Righetti Damiano, Not Fabrice, Biard Tristan, Bittner Lucie, Benoiston Anne-Sophie, Guidi Lionel, Villarino Ernesto, Gaborit Charlie, Cornils Astrid, Buttay Lucie, Irisson Jean-Olivier, Chiarello Marlène, Vallim Alessandra L., Blanco-Bercial Leocadio, Basconi Laura, Ayata Sakina-Dorothee.* Mare Incognitum : A glimpse into future plankton diversity and ecology research // Frontiers in Marine Science. 2017. 4. 68.
- Claustre Hervé, Johnson Kenneth S., Takeshita Yuichiro.* Observing the Global Ocean with Biogeochemical-Argo // Annual Review of Marine Science. 2020. 12, 1. 23–48. PMID : 31433959.
- Colarco P. R., Toon O. B., Reid J. S., Livingston J. M., Russell P. B., Redemann J., Schmid B., Maring H. B., Savoie D., Welton E. J., Campbell J. R., Holben B. N., Levy R.* Saharan dust transport to the Caribbean during PRIDE : 2. Transport, vertical profiles, and deposition in simulations of in situ and remote sensing observations // Journal of Geophysical Research : Atmospheres. 2003. 108, D19.

- Condie Scott, Condie Ryan.* Retention of plankton within ocean eddies // *Global Ecology and Biogeography*. 2016. 25, 10. 1264–1277.
- Continuous Plankton Recorder Survey Team .* Continuous Plankton Records : Plankton Atlas of the North Atlantic Ocean (1958–1999). II. Biogeographical charts // *Marine Ecology Progress Series*. 2004. 11–75.
- Cordier Tristan, Esling Philippe, Lejzerowicz Franck, Visco Joana, Ouadahi Amine, Martins Catarina, Cedhagen Tomas, Pawlowski Jan.* Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning // *Environmental Science & Technology*. 08 2017. 51, 16. 9118–9126.
- Costello Elizabeth K., Stagaman Keaton, Dethlefsen Les, Bohannan Brendan J. M., Relman David A.* The Application of Ecological Theory Toward an Understanding of the Human Microbiome // *Science*. 2012. 336, 6086. 1255–1262.
- Costello Mark John, Basher Zeenatul, Sayre Roger, Breyer Sean, Wright Dawn J.* Stratifying ocean sampling globally and with depth to account for environmental variability // *Scientific Reports*. 2018. 8, 1. 11259.
- Cottenie Karl.* Integrating environmental and spatial processes in ecological community dynamics // *Ecology Letters*. 2005. 8, 11. 1175–1182.
- Courties Claude, Vaquer André, Troussellier Marc, Lautier Jacques, Chrétiennot-Dinet Marie J., Neveux Jacques, Machado Cordelia, Claustre Hervé.* Smallest eukaryotic organism // *Nature*. 1994. 370, 6487. 255–255.
- Cowen Robert K., Guigand Cedric M.* In situ ichthyoplankton imaging system (ISIIS) : system design and preliminary results // *Limnology and Oceanography : Methods*. 2008. 6, 2. 126–132.
- Cronin Thomas W.* Camouflage : Being Invisible in the Open Ocean // *Current Biology*. 2016. 26, 22. R1179–R1181.
- D’Alelio Domenico, Libralato Simone, Wyatt Timothy, Ribera d’Alcalà Maurizio.* Ecological-network models link diversity, structure and function in the plankton food-web // *Scientific Reports*. 2016. 6, 1. 21806.
- Dalongeville Alicia, Andrello Marco, Mouillot David, Lobreaux Stéphane, Fortin Marie-Josée, Lasram Frida, Belmaker Jonathan, Rocklin Delphine, Manel Stéphanie.* Geographic isolation and larval dispersal shape seascape genetic patterns differently according to spatial scale // *Evolutionary applications*. 2018. 11, 8. 1437–1447.
- Danko David, Bezdán Daniela, Afshin Evan E., Ahsanuddin Sofia, Bhattacharya Chandrima, Butler Daniel J., et al. .* A global metagenomic map of urban microbiomes and antimicrobial resistance // *Cell*. 2021. 184, 13. 3376–3393.e17.
- Davide Albanese, Donati Claudio.* Strain profiling and epidemiology of bacterial species from metagenomic sequencing // *Nature communications*. 2017. 8, 1. 1–14.
- De Vargas Colombari, Audic Stéphane, Henry Nicolas, Decelle Johan, Mahé Frédéric, Logares Ramiro, Lara Enrique, Berney Cédric, Le Bescot Noan, Probert Ian, Carmichael Margaux, Poulain Julie, Romac Sarah, Colin Sébastien, Aury Jean-Marc, Bittner Lucie, Chaffron Samuel, Dunthorn Micah, Engelen Stefan, Flegontova Olga, Guidi Lionel, Horák Aleš, Jaillon*

- Olivier, Lima-Mendez Gipsi, Lukeš Julius, Malviya Shruti, Morard Raphael, Mulot Matthieu, Scalco Eleonora, Siano Raffaele, Vincent Flora, Zingone Adriana, Dimier Céline, Picheral Marc, Searson Sarah, Kandels-Lewis Stefanie, Coordinators Tara Oceans, Acinas Silvia G., Bork Peer, Bowler Chris, Gorsky Gabriel, Grimsley Nigel, Hingamp Pascal, Iudicone Daniele, Not Fabrice, Ogata Hiroyuki, Pesant Stephane, Raes Jeroen, Sieracki Michael E., Speich Sabrina, Stemmann Lars, Sunagawa Shinichi, Weissenbach Jean, Wincker Patrick, Karsenti Eric.* Eukaryotic plankton diversity in the sunlit ocean // *Science*. 2015. 348, 6237.
- De Wit Rutger, Bowler Thierry.* ‘Everything is everywhere, but, the environment selects’; what did Baas Becking and Beijerinck really say? // *Environmental microbiology*. 2006. 8, 4. 755–758.
- DeLong Edward F.* The microbial ocean from genomes to biomes // *Nature*. 2009. 459, 7244. 200–206.
- Delmont Tom O., Gaia Morgan, Hinsinger Damien D., Fremont Paul, Vanni Chiara, Guerra Antonio Fernandez, Eren A. Murat, Kourlaiev Artem, d’Agata Leo, Clayssen Quentin, Villar Emilie, Labadie Karine, Cruaud Corinne, Poulain Julie, Da Silva Corinne, Wessner Marc, Noel Benjamin, Aury Jean-Marc, Coordinators Tara Oceans, Vargas Colomban de, Bowler Chris, Karsenti Eric, Pelletier Eric, Wincker Patrick, Jaillon Olivier.* Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics // *bioRxiv*. 2021.
- Delmont Tom O., Kiefl Evan, Kilinc Ozsel, Esen Ozcan C., Uysal Ismail, Rappe Michael S., Giovannoni Steven, Eren A. Murat.* Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade // *Elife*. 2019. 8. e46497.
- Delmont Tom O., Quince Christopher, Shaiber Alon, Esen Özcan C, Lee Sonny T. M., Rappé Michael S., McLellan Sandra L., Lüscher Sebastian, Eren A. Murat.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes // *Nature microbiology*. 2018. 3, 7. 804–813.
- Demura Mikihide, Nakayama Takuro, Kasai Fumie, Kawachi Masanobu.* Genetic structure of Japanese *Chaetomorpha marina* (Rhodophyta) populations revealed using microsatellite markers // *Phycological research*. 2014. 62, 2. 102–108.
- Des Roches Simone, Post David M., Turley Nash E., Bailey Joseph K., Hendry Andrew P., Kinnison Michael T., Schweitzer Jennifer A., Palkovacs Eric P.* The ecological importance of intraspecific variation // *Nature ecology & evolution*. 2018. 2, 1. 57–64.
- Díaz Sandra, Cabido Marcelo.* Vive la différence : plant functional diversity matters to ecosystem processes // *Trends in ecology & evolution*. 2001. 16, 11. 646–655.
- Diepenbroek Michael, Grobe Hannes, Reinke Manfred, Schindler Uwe, Schlitzer Reiner, Sieger Rainer, Wefer Gerold.* PANGAEA—an information system for environmental sciences // *Computers & Geosciences*. 2002. 28, 10. 1201–1210.
- Dietz Robert S., Holden John C.* Reconstruction of Pangaea : breakup and dispersion of continents, Permian to present // *Journal of Geophysical Research*. 1970. 75, 26. 4939–4956.
- Dolan John R.* An introduction to the biogeography of aquatic microbes // *Aquatic Microbial Ecology*. 2005. 41, 1. 39–48.

- Dolan John R.* Microbial ecology of the oceans. 2018.
- Doney Scott C., Abbott Mark R., Cullen John J., Karl David M., Rothstein Lewis.* From genes to ecosystems : the ocean's new frontier // *Frontiers in Ecology and the Environment*. 2004. 2, 9. 457–468.
- Dong Xin Luna, Srivastava Divesh.* Big data integration // 2013 IEEE 29th International Conference on Data Engineering (ICDE). 2013. 1245–1248.
- Dray S., Péliissier R., Couteron P., Fortin M.-J., Legendre P., Peres-Neto P. R., Bellier E., Bivand R., Blanchet F. G., De Cáceres M., Dufour A.-B., Heegaard E., Jombart T., Munoz F., Oksanen J., Thioulouse J., Wagner H. H.* Community ecology in the age of multivariate multiscale spatial analysis // *Ecological Monographs*. 2012. 82, 3. 257–275.
- Duarte Carlos M.* Seafaring in the 21st Century : The Malaspina 2010 Circumnavigation Expedition // *Limnology and Oceanography Bulletin*. 2015. 24, 1. 11–14.
- Eckford-Soper Lisa, Daugbjerg Niels.* The ichthyotoxic genus *Pseudochattonella* (Dictyochophyceae) : Distribution, toxicity, enumeration, ecological impact, succession and life history – A review // *Harmful Algae*. 2016. 58. 51–58.
- Edwards Martin, Richardson Anthony J.* Impact of climate change on marine pelagic phenology and trophic mismatch // *Nature*. 2004. 430, 7002. 881–884.
- Eikrem W., Throndsen J.* The ultrastructure of *Bathycoccus* gen. nov. and *B. prasinos* sp. nov., a non-motile picoplanktonic alga (Chlorophyta, Prasinophyceae) from the Mediterranean and Atlantic // *Phycologia*. 1990. 29, 3. 344–350.
- Eren A. Murat, Kieft Evan, Shaiber Alon, Veseli Iva, Miller Samuel E., Schechter Matthew S., Fink Isaac, Pan Jessica N., Yousef Mahmoud, Fogarty Emily C., Trigodet Florian, Watson Andrea R., Esen Özcan C., Moore Ryan M., Clayssen Quentin, Lee Michael D., Kivenson Veronika, Graham Elaina D., Merrill Bryan D., Karkman Antti, Blankenberg Daniel, Eppley John M., Sjödin Andreas, Scott Jarrod J., Vázquez-Campos Xabier, McKay Luke J., McDaniel Elizabeth A., Stevens Sarah L. R., Anderson Rika E., Fuessel Jessika, Fernandez-Guerra Antonio, Maignien Lois, Delmont Tom O., Willis Amy D.* Community-led, integrated, reproducible multi-omics with anvi'o // *Nature Microbiology*. 2021. 6, 1. 3–6.
- Escofier Brigitte, Pages Jérôme.* Multiple factor analysis (AFMULT package) // *Computational statistics & data analysis*. 1994. 18, 1. 121–140.
- Etten Jacob van.* R Package *gdistance* : Distances and Routes on Geographical Grids // *Journal of Statistical Software*. feb 2017. 76, 13. 21.
- Evans Katharine M., Kühn Stefanie F., Hayes Paul K.* High levels of genetic diversity and low levels of genetic differentiation in north sea *Pseudo-nitzschia pungens* (Bacillariophyceae) populations // *Journal of Phycology*. 2005. 41, 3. 506–514.
- Falkowski Paul G., Katz Miriam E., Knoll Andrew H., Quigg Antonietta, Raven John A., Schofield Oscar, Taylor F. J. R.* The Evolution of Modern Eukaryotic Phytoplankton // *Science*. 2004. 305, 5682. 354–360.

- Faure Emile, Not Fabrice, Benoiston Anne-Sophie, Labadie Karine, Bittner Lucie, Ayata Sakina-Dorothee.* Mixotrophic protists display contrasted biogeographies in the global ocean // The ISME journal. 2019. 13, 4. 1072–1083.
- Febvre-Chevalier C., Febvre J.* Buoyancy and swimming in marine planktonic protists // Mechanics and Physiology of Animal Swimming. 1994. 13–26.
- Fichaut M., Garcia M. J., Giorgetti A., Iona A., Kuznetsov A., Rixen M., Group Medar.* MEDAR/MEDATLAS 2002 : A Mediterranean and Black Sea database for operational oceanography // Elsevier Oceanography Series. 69. 2003. 645–648.
- Field Christopher B., Behrenfeld Michael J., Randerson James T., Falkowski Paul.* Primary Production of the Biosphere : Integrating Terrestrial and Oceanic Components // Science. 1998. 281, 5374. 237–240.
- Finlay Bland J.* Global dispersal of free-living microbial eukaryote species // science. 2002. 296, 5570. 1061–1063.
- Foissner Wilhelm.* Biogeography and dispersal of micro-organisms : a review emphasizing protists // Acta protozoologica. 2006. 45, 2. 111–136.
- Fondi Marco, Karkman Antti, Tamminen Manu V, Bosi Emanuele, Virta Marko, Fani Renato, Alm Eric, McInerney James O.* “Every gene is everywhere but the environment selects” : global geolocalization of gene sharing in environmental samples through network analysis // Genome biology and evolution. 2016. 8, 5. 1388–1400.
- Fontana Simone, Thomas Mridul Kaniyanthara, Moldoveanu Mirela, Spaak Piet, Pomati Francesco.* Individual-level trait diversity predicts phytoplankton community properties better than species richness or evenness // The ISME Journal. 2018. 12, 2. 356–366.
- Forster Dominik, Bittner Lucie, Karkar Slim, Dunthorn Micah, Romac Sarah, Audic Stéphane, Lopez Philippe, Stoeck Thorsten, Baptiste Eric.* Testing ecological theories with sequence similarity networks : marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms // BMC biology. 2015. 13, 1. 1–16.
- Fragoso Glaucia M., Poulton Alex J., Yashayaev Igor M., Head Erica J. H., Johnsen Geir, Purdie Duncan A.* Diatom biogeography from the Labrador Sea revealed through a trait-based approach // Frontiers in Marine Science. 2018. 5. 297.
- Fuhrman Jed A.* Microbial community structure and its functional implications // Nature. 2009. 459, 7244. 193–199.
- Fukami Tadashi, Martijn Bezemer T., Mortimer Simon R., Putten Wim H. van der.* Species divergence and trait convergence in experimental plant community assembly // Ecology Letters. 2005. 8, 12. 1283–1290.
- Furrer Roman D., Pasinelli Gilberto.* Empirical evidence for source–sink populations : a review on occurrence, assessments and implications // Biological Reviews. 2016. 91, 3. 782–795.
- Galindo Heather M., Olson Donald B., Palumbi Stephen R.* Seascape genetics : a coupled oceanographic-genetic model predicts population structure of Caribbean corals // Current biology. 2006. 16, 16. 1622–1626.

- Galloway James A. M., Green Samuel D., Stevens Martin, Kelley Laura A.* Finding a signal hidden among noise : how can predators overcome camouflage strategies? // *Philosophical Transactions of the Royal Society B : Biological Sciences*. 2020. 375, 1802. 20190478.
- Ganachaud Alexandre, Wunsch Carl.* Improved estimates of global ocean circulation, heat transport and mixing from hydrographic data // *Nature*. 2000. 408, 6811. 453–457.
- Gao Yida, Sassenhagen Ingrid, Richlen Mindy L., Anderson Donald M., Martin Jennifer L., Erdner Deana L.* Spatiotemporal genetic structure of regional-scale *Alexandrium catenella* dinoflagellate blooms explained by extensive dispersal and environmental selection // *Harmful algae*. 2019. 86. 46–54.
- García Cristina.* From ecological indicators to ecological functioning : Integrative approaches to seize on ecological, climatic and socio-economic databases // *Ecological Indicators*. 2019. 107. 105612.
- Gattuso J.-P., Magnan A., Billé R., Cheung W. W. L., Howes E. L., Joos F., Allemand D., Bopp L., Cooley S. R., Eakin C. M., Hoegh-Guldberg O., Kelly R. P., Pörtner H.-O., Rogers A. D., Baxter J. M., Laffoley D., Osborn D., Rankovic A., Rochette J., Sumaila U. R., Treyer S., Turley C.* Contrasting futures for ocean and society from different anthropogenic CO2 emissions scenarios // *Science*. 2015. 349, 6243.
- Genin Amatzia, Greene Charles, Haury Loren, Wiebe Peter, Gal Gideon, Kaartvedt Stein, Meir Eli, Fey Connie, Dawson Jim.* Zooplankton patch dynamics : daily gap formation over abrupt topography // *Deep Sea Research Part I : Oceanographic Research Papers*. 1994. 41, 5. 941–951.
- Gibbons Sean M, Caporaso J Gregory, Pirrung Meg, Field Dawn, Knight Rob, Gilbert Jack A.* Evidence for a persistent microbial seed bank throughout the global ocean // *Proceedings of the National Academy of Sciences*. 2013. 110, 12. 4651–4655.
- Giner Caterina R., Forn Irene, Romac Sarah, Logares Ramiro, de Vargas Colomban, Massana Ramon.* Environmental Sequencing Provides Reasonable Estimates of the Relative Abundance of Specific Picoeukaryotes // *Applied and Environmental Microbiology*. 2016. 82, 15. 4757–4766.
- Gloor Gregory B., Macklaim Jean M., Pawlowsky-Glahn Vera, Egozcue Juan J.* Microbiome Datasets Are Compositional : And This Is Not Optional // *Frontiers in Microbiology*. 2017. 8. 2224.
- Godet Claire, Robuchon Marine, Leroy Boris, Cotté Cédric, Baudena Alberto, Da Silva Ophélie, Fabri-Ruiz Salomé, Lo Monaco Claire, Sergi Sara, Koubbi Philippe.* Matching zooplankton abundance and environment in the South Indian Ocean and Southern Ocean // *Deep Sea Research Part I : Oceanographic Research Papers*. 2020. 163. 103347.
- Godhe Anna, Egardt Jenny, Kleinhans David, Sundqvist Lisa, Hordoir Robinson, Jonsson Per R.* Seascape analysis reveals regional gene flow patterns among populations of a marine planktonic diatom // *Proceedings of the Royal Society B : Biological Sciences*. 2013. 280, 1773. 20131599.
- Godhe Anna, Sjöqvist Conny, Sildever Sirje, Sefbom Josefin, Harðardóttir Sara, Bertos-Fortis Mireia, Bunse Carina, Gross Susanna, Johansson Emma, Jonsson Per R., Khandan Saghar,*

- Legrand Catherine, Lips Inga, Lundholm Nina, Rengefors Karin E., Sassenhagen Ingrid, Suikkanen Sanna, Sundqvist Lisa, Kremp Anke.* Physical barriers and environmental gradients cause spatial and temporal genetic differentiation of an extensive algal bloom // *Journal of Biogeography*. 2016. 43, 6. 1130–1142.
- Goecks Jeremy, Nekrutenko Anton, Taylor James, Team The Galaxy.* Galaxy : a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences // *Genome Biology*. 2010. 11, 8. R86.
- Goodwin Sara, McPherson John D., McCombie W. Richard.* Coming of age : ten years of next-generation sequencing technologies // *Nature Reviews Genetics*. 2016. 17, 6. 333–351.
- Gorsky Gabriel, Bourdin Guillaume, Lombard Fabien, Pedrotti Maria Luiza, Audrain Samuel, Bin Nicolas, Boss Emmanuel, Bowler Chris, Cassar Nicolas, Caudan Loic, Chabot Genevieve, Cohen Natalie R., Cron Daniel, De Vargas Colombar, Dolan John R., Douville Eric, Elineau Amanda, Flores J. Michel, Ghiglione Jean Francois, Haëntjens Nils, Hertau Martin, John Seth G., Kelly Rachel L., Koren Ilan, Lin Yajuan, Marie Dominique, Moulin Clémentine, Moucherie Yohann, Pesant Stéphane, Picheral Marc, Poulain Julie, Pujo-Pay Mireille, Reverdin Gilles, Romac Sarah, Sullivan Mathew B., Trainic Miri, Tressol Marc, Troublé Romain, Vardi Assaf, Voolstra Christian R., Wincker Patrick, Agostini Sylvain, Banaigs Bernard, Boissin Emilie, Forcioli Didier, Furla Paola, Galand Pierre E., Gilson Eric, Reynaud Stéphanie, Sunagawa Shinichi, Thomas Olivier P., Thurber Rebecca Lisette Vega, Zoccola Didier, Planes Serge, Allemand Denis, Karsenti Eric.* Expanding Tara oceans protocols for underway, ecosystemic sampling of the ocean-atmosphere interface during Tara Pacific expedition (2016–2018) // *Frontiers in Marine Science*. 2019. 6. 750.
- Gorsky Gaby, Ohman Mark D, Picheral Marc, Gasparini Stéphane, Stemmann Lars, Romagnan Jean-Baptiste, Cawood Alison, Pesant Stéphane, García-Comas Carmen, Prejger Franck.* Digital zooplankton image analysis using the ZooScan integrated system // *Journal of plankton research*. 2010. 32, 3. 285–303.
- Graham Catherine H., Fine Paul V. A.* Phylogenetic beta diversity : linking ecological and evolutionary processes across space in time // *Ecology letters*. 2008. 11, 12. 1265–1277.
- Graham William M., Pagès Fransesc, Hamner William M.* A physical context for gelatinous zooplankton aggregations : a review // *Jellyfish Blooms : Ecological and Societal Importance*. Dordrecht : Springer Netherlands, 2001. 199–212.
- Grey Erin K., Bernatchez Louis, Cassey Phillip, Deiner Kristy, Deveney Marty, Howland Kimberly L., Lacoursière-Roussel Anaïs, Leong Sandric Chee Yew, Li Yiyuan, Olds Brett, Pfrender Michael E., Prowse Thomas A. A., Renshaw Mark A., Lodge David M.* Effects of sampling effort on biodiversity patterns estimated from environmental DNA metabarcoding surveys // *Scientific Reports*. 2018. 8, 1. 8843.
- Groom Steve, Sathyendranath Shubha, Ban Yai, Bernard Stewart, Brewin Robert, Brotas Vanda, Brockmann Carsten, Chauhan Prakash, Choi Jong-kuk, Chuprin Andrei, Ciavatta Stefano, Cipollini Paolo, Donlon Craig, Franz Bryan, He Xianqiang, Hirata Takafumi, Jackson Tom, Kappel Milton, Krasemann Hajo, Lavender Samantha, Pardo-Martinez Silvia, Mélin Frédéric, Platt Trevor, Santoleri Rosalia, Skakala Jozef, Schaeffer Blake, Smith Marie, Steinmetz Francois, Valente Andre, Wang Menghua.* Satellite Ocean Colour : Current Status and Future Perspective // *Frontiers in Marine Science*. 2019. 6. 485.

- Grossmann Mary M., Gallager Scott M., Mitarai Satoshi.* Continuous monitoring of near-bottom mesoplankton communities in the East China Sea during a series of typhoons // *Journal of Oceanography*. 2015. 71, 1. 115–124.
- Gueroun Sonia K. M., Javidpour Jamileh, Andrade Carlos, Nogueira Natacha, Freitas Mafalda, Canning-Clode João.* Pelagic Cnidaria and Ctenophora diversity patterns and trends in Macaronesia insular systems (NE Atlantic) // *Marine Biodiversity*. 2021. 51, 2. 32.
- Guidi Lionel, Chaffron Samuel, Bittner Lucie, Eveillard Damien, Larhlimi Abdelhalim, Roux Simon, Darzi Youssef, Audic Stephane, Berline Léo, Brum Jennifer R., Coelho Luis Pedro, Espinoza Julio Cesar Ignacio, Malviya Shruti, Sunagawa Shinichi, Dimier Céline, Kandels-Lewis Stefanie, Picheral Marc, Poulain Julie, Searson Sarah, Stemmann Lars, Not Fabrice, Hingamp Pascal, Speich Sabrina, Follows Mick, Karp-Boss Lee, Boss Emmanuel, Ogata Hiroyuki, Pesant Stephane, Weissenbach Jean, Wincker Patrick, Acinas Silvia G., Bork Peer, Vargas Colomban de, Iudicone Daniele, Sullivan Matthew B., Raes Jeroen, Karsenti Eric, Bowler Chris, Gorsky Gabriel, Coordinators Tara Oceans Consortium.* Plankton networks driving carbon export in the oligotrophic ocean // *Nature*. 2016. 532, 7600. 465–470.
- Guttman L.* Some necessary conditions for common-factor analysis // *Psychometrika*. 1954. 19, 2. 149–161.
- Haddock Steven H.D., Moline Mark A., Case James F.* Bioluminescence in the Sea // *Annual Review of Marine Science*. 2010. 2, 1. 443–493. PMID : 21141672.
- Hamm CE.* Architecture, ecology and biogeochemistry of Phaeocystis colonies // *Journal of Sea Research*. 2000. 43, 3-4. 307–315.
- Hamm Christian, Smetacek Victor.* CHAPTER 14 - Armor : Why, When, and How // *Evolution of Primary Producers in the Sea*. Burlington : Academic Press, 2007. 311–332.
- Hampton Stephanie E, Strasser Carly A, Tewksbury Joshua J, Gram Wendy K, Budden Amber E, Batcheller Archer L, Duke Clifford S, Porter John H.* Big data and the future of ecology // *Frontiers in Ecology and the Environment*. 2013. 11, 3. 156–162.
- Hansen A. J.* Contribution of source-sink theory to protected area science // *Sources, sinks and sustainability*. 2011. 339–360.
- Hanson China A., Fuhrman Jed A., Horner-Devine M. Claire, Martiny Jennifer B. H.* Beyond biogeographic patterns : processes shaping the microbial landscape // *Nature Reviews Microbiology*. 2012. 10, 7. 497–506.
- Hansson Lars-Anders, Hylander Samuel.* Effects of ultraviolet radiation on pigmentation, photoenzymatic repair, behavior, and community ecology of zooplankton // *Photochem. Photobiol. Sci.* 2009. 8. 1266–1275.
- Hardin Garrett.* The Competitive Exclusion Principle // *Science*. 1960. 131, 3409. 1292–1297.
- Hays G. C., Kennedy H., Frost B. W.* Individual variability in diel vertical migration of a marine copepod : Why some individuals remain at depth when others migrate // *Limnology and Oceanography*. 2001. 46, 8. 2050–2054.
- Hays Graeme C.* A review of the adaptive significance and ecosystem consequences of zooplankton diel vertical migrations. Dordrecht : Springer Netherlands, 2003. 163–170.

Bibliographie

- Hays Graeme C, Richardson Anthony J, Robinson Carol.* Climate change and marine plankton // Trends in ecology & evolution. 2005. 20, 6. 337–344.
- Hebert Paul D. N., Cywinska Alina, Ball Shelley L., deWaard Jeremy R.* Biological identifications through DNA barcodes // Proceedings of the Royal Society of London. Series B : Biological Sciences. 2003. 270, 1512. 313–321.
- Heimann Kirsten, Andersen Robert A, Wetherbee Richard.* The flagellar development cycle of the uniflagellate *Pelagomonas calceolata* (Pelagophyceae) // Journal of Phycology. 1995. 31, 4. 577–583.
- Heintzelman Matthew B.* Cellular and Molecular Mechanics of Gliding Locomotion in Eukaryotes. 251. 2006. 79–129. (International Review of Cytology).
- Hellweger Ferdi L, Sebille Erik van, Fredrick Neil D.* Biogeographic patterns in ocean microbes emerge in a neutral agent-based model // Science. 2014. 345, 6202. 1346–1349.
- Hensen Victor.* Ueber die Bestimmung des Plankton's oder des im Meere treibenden Materials an Pflanzen und Thieren. 1887.
- Hernandez-Carrasco Ismael, Orfila Alejandro, Rossi Vincent, Garçon Veronique.* Effect of small scale transport processes on phytoplankton distribution in coastal seas // Scientific reports. 2018. 8, 1. 1–13.
- Hey Tony, Tansley Steward, Tolle Kristin.* The fourth paradigm : data-intensive scientific discovery. 2009.
- HilleRisLambers J., Adler P.B., Harpole W.S., Levine J.M., Mayfield M.M.* Rethinking Community Assembly through the Lens of Coexistence Theory // Annual Review of Ecology, Evolution, and Systematics. 2012. 43, 1. 227–248.
- Hirst Andrew G., Glazier Douglas S., Atkinson David.* Body shape shifting during growth permits tests that distinguish between competing geometric theories of metabolic scaling // Ecology Letters. 2014. 17, 10. 1274–1281.
- Hosia Aino, Augustin Christina B., Dinasquet Julie, Granhag Lena, Paulsen Maria L., Riemann Lasse, Rintala Janne-Markus, Setälä Outi, Talvitie Julia, Titelman Josefín.* Autumnal bottom-up and top-down impacts of *Cyanea capillata* : a mesocosm study // Journal of Plankton Research. 2015. 37, 5. 1042–1055.
- Hubbell Stephen P.* The Unified Neutral Theory of Biodiversity and Biogeography. 2001.
- Huisman Jef, Weissing Franz J.* Biodiversity of plankton by species oscillations and chaos // Nature. 1999. 402, 6760. 407–410.
- Humphris Susan E.* Vehicles for deep sea exploration // Elements of Physical Oceanography : A Derivative of the Encyclopedia of Ocean Sciences ; Steele, JH, Thorpe, SA, Thompson, KK, Eds. 2009. 197–210.
- Hutchinson G Evelyn.* The paradox of the plankton // The American Naturalist. 1961. 95, 882. 137–145.

- Ibarbalz Federico M, Henry Nicolas, Brandão Manoela C, Martini Séverine, Busseni Greta, Byrne Hannah, Coelho Luis Pedro, Endo Hisashi, Gasol Josep M, Gregory Ann C., Mahé Frédéric, Rigonato Janaina, Royo-Llonch Marta, Salazar Guillem, Sanz-Sáez Isabel, Scalco Eleonora, Soviadan Dodji, Zayed Ahmed A., Zingone Adriana, Labadie Karine, Ferland Joannie, Marec Claudie, Kandels Stefanie, Picheral Marc, Dimier Céline, Poulain Julie, Pisarev Sergey, Carmichael Margaux, Pesant Stéphane, Acinas Silvia G., Babin Marcel, Bork Peer, Boss Emmanuel, Bowler Chris, Cochrane Guy, de Vargas Colomban, Follows Mick, Gorsky Gabriel, Grimsey Nigel, Guidi Lionel, Hingamp Pascal, Iudicone Daniele, Jaillon Olivier, Kandels Stefanie, Karp-Boss Lee, Karsenti Eric, Not Fabrice, Ogata Hiroyuki, Pesant Stéphane, Poulton Nicole, Raes Jeroen, Sardet Christian, Speich Sabrina, Stemmann Lars, Sullivan Matthew B., Sunagawa Shinichi, Wincker Patrick, Babin Marcel, Boss Emmanuel, Iudicone Daniele, Jaillon Olivier, Acinas Silvia G., Ogata Hiroyuki, Pelletier Eric, Stemmann Lars, Sullivan Matthew B., Sunagawa Shinichi, Bopp Laurent, de Vargas Colomban, Karp-Boss Lee, Wincker Patrick, Lombard Fabien, Bowler Chris, Zinger Lucie.* Global trends in marine plankton diversity across kingdoms of life // *Cell*. 2019. 179, 5. 1084–1097.
- Iglesias-Rodríguez Débora M, Probert Ian, Batley Jacqueline.* Microsatellite cross-amplification in coccolithophores : application in population diversity studies // *Hereditas*. 2006. 143, 2006. 99–102.
- Ishitani Yoshiyuki, Ujié Yurika, Vargas Colomban de, Not Fabrice, Takahashi Kozo.* Phylogenetic Relationships and Evolutionary Patterns of the Order Collodaria (Radiolaria) // *PLOS ONE*. 05 2012. 7, 5. 1–8.
- Jackson Donald A.* Stopping rules in principal components analysis : a comparison of heuristical and statistical approaches // *Ecology*. 1993. 74, 8. 2204–2214.
- Johnsen S., Widder E. A.* Transparency and Visibility of Gelatinous Zooplankton from the Northwestern Atlantic and Gulf of Mexico // *The Biological Bulletin*. 1998. 195, 3. 337–348. PMID : 28297617.
- Johnsen Sönke, Sosik Heidi.* Shedding light on light in the ocean // *Oceanus Magazine*. 2004. 43, 2. 1–5.
- Johnsen Sönke.* Hidden in Plain Sight : The Ecology and Physiology of Organismal Transparency // *The Biological Bulletin*. 2001. 201, 3. 301–318. PMID : 11751243.
- Johnsen Sönke.* Hide and Seek in the Open Sea : Pelagic Camouflage and Visual Countermeasures // *Annual Review of Marine Science*. 2014. 6, 1. 369–392. PMID : 23987915.
- Johnson Lisa K, Alexander Harriet, Brown C Titus.* Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes // *GigaScience*. 2019. 8, 4. g158.
- Jonkers Lukas, Hillebrand Helmut, Kucera Michal.* Global change drives modern plankton communities away from the pre-industrial state // *Nature*. 2019. 570, 7761. 372–375.
- Karsenti Eric.* Self-organization in cell biology : a brief history // *Nature reviews Molecular cell biology*. 2008. 9, 3. 255–262.
- Karsenti Eric, Acinas Silvia G., Bork Peer, Bowler Chris, De Vargas Colomban, Raes Jeroen, Sullivan Matthew, Arendt Detlev, Benzonni Francesca, Claverie Jean-Michel, Follows Mick,*

- Gorsky Gaby, Hingamp Pascal, Iudicone Daniele, Jaillon Olivier, Kandels-Lewis Stefanie, Krzic Uros, Not Fabrice, Ogata Hiroyuki, Pesant Stéphane, Reynaud Emmanuel Georges, Sardet Christian, Sieracki Michael E., Speich Sabrina, Velayoudon Didier, Weissenbach Jean, Winker Patrick, Tara Oceans Consortium the.* A Holistic Approach to Marine Eco-Systems Biology // PLOS Biology. 10 2011. 9, 10. 1–5.
- Kassambara Alboukadel.* ggpubr : 'ggplot2' Based Publication Ready Plots. 2020. R package version 0.3.0.
- Kawecki Tadeusz J, Holt Robert D.* Evolutionary consequences of asymmetric dispersal rates // The American Naturalist. 2002. 160, 3. 333–347.
- Keeling Patrick J., Burki Fabien, Wilcox Heather M., Allam Bassem, Allen Eric E., Amaral-Zettler Linda A., Armbrust E. Virginia, Archibald John M., Bharti Arvind K., Bell Callum J., Beszteri Bank, Bidle Kay D., Cameron Connor T., Campbell Lisa, Caron David A., Cattolico Rose Ann, Collier Jackie L., Coyne Kathryn, Davy Simon K., Deschamps Phillipe, Dyhrman Sonya T., Edvardsen Bente, Gates Ruth D., Gobler Christopher J., Greenwood Spencer J., Guida Stephanie M., Jacobi Jennifer L., Jakobsen Kjetill S., James Erick R., Jenkins Bethany, John Uwe, Johnson Matthew D., Juhl Andrew R., Kamp Anja, Katz Laura A., Kiene Ronald, Kudryavtsev Alexander, Leander Brian S., Lin Senjie, Lovejoy Connie, Lynn Denis, Marchetti Adrian, McManus George, Nedelcu Aurora M., Menden-Deuer Susanne, Miceli Cristina, Mock Thomas, Montresor Marina, Moran Mary Ann, Murray Shauna, Nadathur Govind, Nagai Satoshi, Ngam Peter B., Palenik Brian, Pawlowski Jan, Petroni Giulio, Piganeau Gwenael, Posewitz Matthew C., Rengefors Karin, Romano Giovanna, Rumpho Mary E., Rynearson Tatiana, Schilling Kelly B., Schroeder Declan C., Simpson Alastair G. B., Slamovits Claudio H., Smith David R., Smith G. Jason, Smith Sarah R., Sosik Heidi M., Stief Peter, Theriot Edward, Twary Scott N., Umale Pooja E., Vaultot Daniel, Wawrik Boris, Wheeler Glen L., Wilson William H., Xu Yan, Zingone Adriana, Worden Alexandra Z.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSPP) : illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing // PLoS biology. 2014. 12, 6. e1001889.
- Keeling Patrick J., Campo Javier del.* Marine Protists Are Not Just Big Bacteria // Current Biology. 2017. 27, 11. R541–R549.
- Kjørboe Thomas.* Zooplankton body composition // Limnology and Oceanography. 2013. 58, 5. 1843–1850.
- Kirk JTO.* Attenuation of light in natural waters // Marine and Freshwater Research. 1977. 28, 4. 497–508.
- Klein Patrice, Lapeyre Guillaume.* The Oceanic Vertical Pump Induced by Mesoscale and Submesoscale Turbulence // Annual Review of Marine Science. 2009. 1, 1. 351–375. PMID : 21141041.
- Kopf Anna, Bicak Mesude, Kottmann Renzo, Schnetzer Julia, Kostadinov Ivaylo, Lehmann Katja, Fernandez-Guerra Antonio, Jeanthon Christian, Rahav Eyal, Ulrich Matthias, Wichels Antje, Gerdtz Gunnar, Polymenakou Paraskevi, Kotoulas Giorgos, Siam Rania, Abdallah Rehab Z, Sonnenschein Eva C, Cariou Thierry, O'Gara Fergal, Jackson Stephen, Orlic Sandi, Steinke Michael, Busch Julia, Duarte Bernardo, Caçador Isabel, Canning-Clode João, Bobrova Oleksandra, Marteinsson Viggo, Reynisson Eyjolfur, Loureiro Clara Magalhães,*

- Luna Gian Marco, Quero Grazia Marina, Löscher Carolin R, Kremp Anke, DeLorenzo Marie E, Øvreås Lise, Tolman Jennifer, LaRoche Julie, Penna Antonella, Frischer Marc, Davis Timothy, Katherine Barker, Meyer Christopher P, Ramos Sandra, Magalhães Catarina, Jude-Lemeilleur Florence, Aguirre-Macedo Ma Leopoldina, Wang Shiao, Poulton Nicole, Jones Scott, Collin Rachel, Fuhrman Jed A, Conan Pascal, Alonso Cecilia, Stambler Noga, Goodwin Kelly, Yakimov Michael M, Baltar Federico, Bodrossy Levente, Kamp Jodie Van De, Frampton Dion MF, Ostrowski Martin, Ruth Paul Van, Malthouse Paul, Claus Simon, Deneudt Klaas, Mortelmans Jonas, Pitois Sophie, Wallom David, Salter Ian, Costa Rodrigo, Schroeder Declan C, Kandil Mahrous M, Amaral Valentina, Biancalana Florencia, Santana Rafael, Pedrotti Maria Luiza, Yoshida Takashi, Ogata Hiroyuki, Ingleton Tim, Munnik Kate, Rodriguez-Ezpeleta Naiara, Berteaux-Lecellier Veronique, Wecker Patricia, Cancio Ibon, Vaulot Daniel, Bienhold Christina, Ghazal Hassan, Chaouni Bouchra, Essayeh Soumya, Ettamimi Sara, Zaid El Houcine, Boukhatem Noureddine, Bouali Abderrahim, Chahboune Rajaa, Barrijal Said, Timinouni Mohammed, Otmani Fatima El, Bennani Mohamed, Mea Marianna, Todorova Nadezhda, Karamfilov Ventzislav, Hoopen Petra ten, Cochrane Guy, L'Haridon Stephane, Bizsel Kemal Can, Vezzi Alessandro, Lauro Federico M, Martin Patrick, Jensen Rachelle M, Hinks Jamie, Gebbels Susan, Rosselli Riccardo, Pascale Fabio De, Schiavon Riccardo, Santos Antonina dos, Villar Emilie, Pesant Stéphane, Cataletto Bruno, Malfatti Francesca, Edirisinghe Ranjith, Silveira Jorge A Herrera, Barbier Michele, Turk Valentina, Tinta Tinkara, Fuller Wayne J, Salihoglu Ilkay, Serakinci Nedime, Ergoren Mahmut Cerkez, Bresnan Eileen, Iriberrí Juan, Nyhus Paul Anders Fronth, Bente Edvardsen, Karlsen Hans Erik, Golyshtin Peter N, Gasol Josep M, Moncheva Snezana, Dzhebekova Nina, Johnson Zackary, Sinigalliano Christopher David, Gidley Maribeth Louise, Zingone Adriana, Danovaro Roberto, Tsiamis George, Clark Melody S, Costa Ana Cristina, Bour Monia El, Martins Ana M, Collins R Eric, Ducluzeau Anne-Lise, Martinez Jonathan, Costello Mark J, Amaral-Zettler Linda A, Gilbert Jack A, Davies Neil, Field Dawn, Glöckner Frank Oliver. The ocean sampling day consortium // *Gigascience*. 2015. 4, 1. s13742-015.
- Köster Johannes, Rahmann Sven. Snakemake—a scalable bioinformatics workflow engine // *Bioinformatics*. 2012. 28, 19. 2520–2522.
- Kraft Nathan JB, Godoy Oscar, Levine Jonathan M. Plant functional traits and the multidimensional nature of species coexistence // *Proceedings of the National Academy of Sciences*. 2015. 112, 3. 797–802.
- Kulk Gemma, Vries Pablo de, Poll Willem H van de, Visser Ronald JW, Buma Anita GJ. Temperature-dependent growth and photophysiology of prokaryotic and eukaryotic oceanic picophytoplankton // *Marine Ecology Progress Series*. 2012. 466. 43–55.
- Labarre Aurelie, Obiol Aleix, Wilken Susanne, Forn Irene, Massana Ramon. Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates // *Limnology and Oceanography*. 2020. 65, S1. S149–S160.
- Ladyman James, Lambert James, Wiesner Karoline. What is a complex system? // *European Journal for Philosophy of Science*. 2013. 3, 1. 33–67.
- Laliberté Etienne, Legendre Pierre. A distance-based framework for measuring functional diversity from multiple traits // *Ecology*. 2010. 91, 1. 299–305.
- Laliberté Etienne, Legendre Pierre, Shipley Bill. FD : measuring functional diversity from multiple traits, and other tools for functional ecology. 2014. R package version 1.0-12.

- Lander Eric S., Linton Lauren M., Birren Bruce, Nusbaum Chad, Zody Michael C., Baldwin Jennifer, Devon Keri, Dewar Ken, Doyle Michael, FitzHugh William, Funke Roel, Gage Diane, Harris Katrina, Heaford Andrew, Howland John, Kann Lisa, Lehoczky Jessica, LeVine Rosie, McEwan Paul, McKernan Kevin, Meldrim James, Mesirov Jill P., Miranda Cher, Morris William, Naylor Jerome, Raymond Christina, et al. . Initial sequencing and analysis of the human genome // *Nature*. 2001. 409, 6822. 860–921.
- Laso-Jadart Romuald, O'Malley Michael, Sykulski Adam, Ambroise Christophe, Madoui Mohammed-Amin. How marine currents and environment shape plankton genomic differentiation : a mosaic view from Tara Oceans metagenomic data // *Biorxiv*. 2021.
- Latorre Francisco, Deutschmann Ina M., Labarre Aurélie, Obiol Aleix, Krabberød Anders K., Pelletier Eric, Sieracki Michael E., Cruaud Corinne, Jaillon Olivier, Massana Ramon, Logares Ramiro. Niche adaptation promoted the evolutionary diversification of tiny ocean predators // *Proceedings of the National Academy of Sciences*. 2021. 118, 25.
- Le Gall F, Rigaut-Jalabert F, Marie D, Garczarek L, Viprey M, Gobet Angelique, Vaultot Daniel. Picoplankton diversity in the South-East Pacific Ocean from cultures // *Biogeosciences*. 2008. 5, 1. 203–214.
- Lê Sébastien, Josse Julie, Husson François. FactoMineR : an R package for multivariate analysis // *Journal of statistical software*. 2008. 25, 1. 1–18.
- Le Traon Pierre Yves, Reppucci Antonio, Alvarez Fanjul Enrique, Aouf Lotfi, Behrens Arno, Belmonte Maria, Bentamy Abderrahim, Bertino Laurent, Brando Vittorio Ernesto, Kreiner Matilde Brandt, Benkiran Mounir, Carval Thierry, Ciliberti Stefania A., Claustre Hervé, Clementi Emanuela, Coppini Giovanni, Cossarini Gianpiero, De Alfonso Alonso-Muñoyerro Marta, Delamarche Anne, Dibarboure Gerald, Dinessen Frode, Drevillon Marie, Drillet Yann, Faugere Yannice, Fernández Vicente, Fleming Andrew, Garcia-Hermosa M. Isabel, Sotillo Marcos García, Garric Gilles, Gasparin Florent, Giordan Cedric, Gehlen Marion, Gregoire Marilaure L., Guinehut Stephanie, Hamon Mathieu, Harris Chris, Hernandez Fabrice, Hinkler Jørgen B., Hoyer Jacob, Karvonen Juha, Kay Susan, King Robert, Lavergne Thomas, Lemieux-Dudon Benedicte, Lima Leonardo, Mao Chongyuan, Martin Matthew J., Masina Simona, Melet Angelique, Buongiorno Nardelli Bruno, Nolan Glenn, Pascual Ananda, Pistoia Jenny, Palazov Atanas, Piolle Jean Francois, Pujol Marie Isabelle, Pequignet Anne Christine, Peneva Elisaveta, Pérez Gómez Begoña, Villeon Loic Petit de la, Pinardi Nadia, Pisano Andrea, Pouliquen Sylvie, Reid Rebecca, Remy Elisabeth, Santoleri Rosalia, Siddorn John, She Jun, Staneva Joanna, Stoffelen Ad, Tonani Marina, Vandenbulcke Luc, Schuckmann Karina von, Volpe Gianluca, Wettre Cecilie, Zacharioudaki Anna. From Observation to Information and Users : The Copernicus Marine Service Perspective // *Frontiers in Marine Science*. 2019. 6. 234.
- Lebret Karen, Kritzberg Emma S, Figueroa Rosa, Rengefors Karin. Genetic diversity within and genetic differentiation between blooms of a microalgal species // *Environmental Microbiology*. 2012. 14, 9. 2395–2404.
- Leconte Jade, Benites L Felipe, Vannier Thomas, Wincker Patrick, Piganeau Gwenael, Jaillon Olivier. Genome resolved biogeography of mamiellales // *Genes*. 2020. 11, 1. 66.
- Lee Marcus, Zhang Huan, Sha Yongcui, Hegg Alexander, Ugge Gustaf Ekelund, Vinterstare Jerker, Škerlep Martin, Pärssinen Varpu, Herzog Simon David, Björnerås Caroline, Gollnisch

- Raphael, Johansson Emma, Hu Nan, Nilsson P. Anders, Hulthén Kaj, Rengefors Karin, Langerhans R. Brian, Brönmark Christer, Hansson Lars-Anders.* Low-latitude zooplankton pigmentation plasticity in response to multiple threats // *Royal Society Open Science*. 2019. 6, 7. 190321.
- Legendre Louis, Rivkin Richard B., Weinbauer Markus G., Guidi Lionel, Uitz Julia.* The microbial carbon pump concept : Potential biogeochemical significance in the globally changing ocean // *Progress in Oceanography*. 2015. 134. 432–450.
- Legendre Pierre, Legendre Louis.* Numerical ecology. 2012.
- Leibold M. A., Holyoak M., Mouquet N., Amarasekare P., Chase J. M., Hoopes M. F., Holt R. D., Shurin J. B., Law R., Tilman D., Loreau M., Gonzalez A.* The metacommunity concept : a framework for multi-scale community ecology // *Ecology letters*. 2004. 7, 7. 601–613.
- Leigh Deborah M., Hendry Andrew P., Vázquez-Domínguez Ella, Friesen Vicki L.* Estimated six per cent loss of genetic variation in wild populations since the industrial revolution // *Evolutionary Applications*. 2019. 12, 8. 1505–1512.
- Lejeune Christophe, Chevaldonné Pierre, Pergent-Martini Christine, Boudouresque Charles F, Pérez Thierry.* Climate change effects on a miniature ocean : the highly diverse, highly impacted Mediterranean Sea // *Trends in ecology & evolution*. 2010. 25, 4. 250–260.
- Lenton Timothy M, Held Hermann, Kriegler Elmar, Hall Jim W, Lucht Wolfgang, Rahmstorf Stefan, Schellnhuber Hans Joachim.* Tipping elements in the Earth's climate system // *Proceedings of the national Academy of Sciences*. 2008. 105, 6. 1786–1793.
- Leonelli Sabina.* Integrating data to acquire new knowledge : Three modes of integration in plant science // *Studies in History and Philosophy of Science Part C : Studies in History and Philosophy of Biological and Biomedical Sciences*. 2013. 44, 4. 503–514.
- Leoni Valentina, Bonnet Delphine, Ramírez-Romero Eduardo, Molinero Juan Carlos.* Biogeography and phenology of the jellyfish *Rhizostoma pulmo* (Cnidaria : Scyphozoa) in southern European seas // *Global Ecology and Biogeography*. 2021. 30, 3. 622–639.
- Letessier Tom B., Falkenhaus Tone, Debes Høgni, Bergstad Odd Aksel, Brierley Andrew S.* Abundance patterns and species assemblages of euphausiids associated with the Mid-Atlantic Ridge, North Atlantic // *Journal of Plankton Research*. 2011. 33, 10. 1510–1525.
- Letten Andrew D., Ke Po-Ju, Fukami Tadashi.* Linking modern coexistence theory and contemporary niche theory // *Ecological Monographs*. 2017. 87, 2. 161–177.
- Lévy Marina, Franks Peter JS, Smith K Shafer.* The role of submesoscale currents in structuring marine ecosystems // *Nature communications*. 2018. 9, 1. 1–16.
- Li Heng.* Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM // *arXiv preprint arXiv :1303.3997*. 2013.
- Lilley M. K. S., Beggs S. E., Doyle T. K., Hobson V. J., Stromberg K. H. P., Hays G. C.* Global patterns of epipelagic gelatinous zooplankton biomass // *Marine Biology*. 2011. 158, 11. 2429–2436.

- Lima-Mendez Gipsi, Faust Karoline, Henry Nicolas, Decelle Johan, Colin Sébastien, Carcillo Fabrizio, Chaffron Samuel, Ignacio-Espinosa J. Cesar, Roux Simon, Vincent Flora, Bittner Lucie, Darzi Youssef, Wang Jun, Audic Stéphane, Berline Léo, Bontempi Gianluca, Cabello Ana M., Coppola Laurent, Cornejo-Castillo Francisco M., d'Ovidio Francesco, De Meester Luc, Ferrera Isabel, Garet-Delmas Marie-José, Guidi Lionel, Lara Elena, Pesant Stéphane, Royo-Llonch Marta, Salazar Guillem, Sánchez Pablo, Sebastian Marta, Souffreau Caroline, Dimier Céline, Picheral Marc, Searson Sarah, Kandels-Lewis Stefanie, coordinators Tara Oceans, Gorsky Gabriel, Not Fabrice, Ogata Hiroyuki, Speich Sabrina, Stemmann Lars, Weissenbach Jean, Wincker Patrick, Acinas Silvia G., Sunagawa Shinichi, Bork Peer, Sullivan Matthew B., Karsenti Eric, Bowler Chris, Vargas Colomban de, Raes Jeroen.* Determinants of community structure in the global plankton interactome // *Science*. 2015. 348, 6237.
- Litchman Elena, Klausmeier Christopher A.* Trait-Based Community Ecology of Phytoplankton // *Annual Review of Ecology, Evolution, and Systematics*. 2008. 39, 1. 615–639.
- Litchman Elena, Ohman Mark D., Kjørboe Thomas.* Trait-based approaches to zooplankton communities // *Journal of plankton research*. 2013. 35, 3. 473–484.
- Liu Wei, Fedorov Alexey V, Xie Shang-Ping, Hu Shineng.* Climate impacts of a weakened Atlantic Meridional Overturning Circulation in a warming climate // *Science advances*. 2020. 6, 26. eaaz4876.
- Logares Ramiro.* Population genetics : the next stop for microbial ecologists? // *Open Life Sciences*. 2011. 6, 6. 887–892.
- Logares Ramiro, Audic Stéphane, Bass David, Bittner Lucie, Boutte Christophe, Christen Richard, Claverie Jean-Michel, Decelle Johan, Dolan John R, Dunthorn Micah, Edvardsen Bente, Gobet Angélique, Kooistra Wiebe H.C.F., Mahé Frédéric, Not Fabrice, Ogata Hiroyuki, Pawlowski Jan, Pernice Massimo C., Romac Sarah, Shalchian-Tabrizi Kamran, Simon Nathalie, Stoeck Thorsten, Santini Sébastien, Siano Raffaele, Wincker Patrick, Zingone Adriana, Richards Thomas A., Vargas Colomban de, Massana Ramon.* Patterns of rare and abundant marine microbial eukaryotes // *Current Biology*. 2014. 24, 8. 813–821.
- Logares Ramiro, Deutschmann Ina M., Junger Pedro C., Giner Caterina R., Krabberød Anders K., Schmidt Thomas S. B., Rubinat-Ripoll Laura, Mestre Mireia, Salazar Guillem, Ruiz-González Clara, Sebastián Marta, Vargas Colomban de, Acinas Silvia G., Duarte Carlos M., Gasol Josep M., Massana Ramon.* Disentangling the mechanisms shaping the surface ocean microbiota // *Microbiome*. 2020. 8, 1. 55.
- Lombard Fabien, Boss Emmanuel, Waite Anya M., Vogt Meike, Uitz Julia, Stemmann Lars, Sosik Heidi M., Schulz Jan, Romagnan Jean-Baptiste, Picheral Marc, Pearlman Jay, Ohman Mark D., Niehoff Barbara, Möller Klas O., Miloslavich Patricia, Lara-Lpez Ana, Kudela Raphael, Lopes Rubens M., Kiko Rainer, Karp-Boss Lee, Jaffe Jules S., Iversen Morten H., Irisson Jean-Olivier, Fennel Katja, Hauss Helena, Guidi Lionel, Gorsky Gaby, Giering Sarah L. C., Gaube Peter, Gallagher Scott, Dubelaar George, Cowen Robert K., Carlotti François, Briseño-Avena Christian, Berline Léo, Benoit-Bird Kelly, Bax Nicholas, Batten Sonia, Ayata Sakina Dorothée, Artigas Luis Felipe, Appeltans Ward.* Globally Consistent Quantitative Observations of Planktonic Ecosystems // *Frontiers in Marine Science*. 2019. 6. 196.
- Longhurst Alan R.* Ecological geography of the sea. 2010.

- Loos Luna M. van der, Nijland Reindert.* Biases in bulk : DNA metabarcoding of marine communities and the methodology involved // *Molecular Ecology*. 2021. 30, 13. 3270–3288.
- Lowe Chris D, Martin Laura E, Montagnes David JS, Watts Phillip C.* 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*. 2012. 109, 51. 20998–21003.
- Lowe Chris D, Montagnes David JS, Martin Laura E, Watts Phillip C.* Patterns of genetic diversity in the marine heterotrophic flagellate *Oxyrrhis marina* (Alveolata : Dinophyceae) // *Protist*. 2010. 161, 2. 212–221.
- Lowery Christopher M., Bown Paul R., Fraass Andrew J., Hull Pincelli M.* Ecological Response of Plankton to Environmental Change : Thresholds for Extinction // *Annual Review of Earth and Planetary Sciences*. 2020. 48, 1. 403–429.
- Lozier M. Susan.* Deconstructing the Conveyor Belt // *Science*. 2010. 328, 5985. 1507–1511.
- Lubchenco Jane, Haugan Peter M., Pangestu Mari Elka.* Five priorities for a sustainable ocean economy // *Nature*. 2020. 588. 30–32.
- Lucas Cathy H., Jones Daniel O. B., Hollyhead Catherine J., Condon Robert H., Duarte Carlos M., Graham William M., Robinson Kelly L., Pitt Kylie A., Schildhauer Mark, Regetz Jim.* Gelatinous zooplankton biomass in the global oceans : geographic variation and environmental drivers // *Global Ecology and Biogeography*. 2014. 23, 7. 701–714.
- Luisa Bozzano G.* The ecology of natural disturbance and patch dynamics. 2012.
- Lumley Thomas, Lumley Maintainer Thomas.* Package ‘leaps’ // Regression subset selection. Thomas Lumley Based on Fortran Code by Alan Miller. Available online : <http://CRAN.R-project.org/package=leaps> (Accessed on 18 March 2018). 2013.
- Luo Jessica Y., Condon Robert H., Stock Charles A., Duarte Carlos M., Lucas Cathy H., Pitt Kylie A., Cowen Robert K.* Gelatinous Zooplankton-Mediated Carbon Flows in the Global Oceans : A Data-Driven Modeling Study // *Global Biogeochemical Cycles*. 2020. 34, 9. e2020GB006704.
- Luo Jessica Y., Grassian Benjamin, Tang Dorothy, Irisson Jean-Olivier, Greer Adam T., Guigand Cedric M., McClatchie Sam, Cowen Robert K.* Environmental drivers of the fine-scale distribution of a gelatinous zooplankton community across a mesoscale front // *Marine Ecology Progress Series*. 2014. 510. 129–149.
- Lynch Michael D. J., Neufeld Josh D.* Ecology and exploration of the rare biosphere // *Nature Reviews Microbiology*. 2015. 13, 4. 217–229.
- Machemer Hans.* The swimming cell and its world : Structures and mechanisms of orientation in protists // *European Journal of Protistology*. 2001. 37, 1. 3–14.
- Mackie G. O., Mackie G. V.* Mesogloecal ultrastructure and reversible opacity in a transparent siphonophore // *Vie et milieu*. 1967. 47–72.

- Madoui Mohammed-Amin, Poulain Julie, Sugier Kevin, Wessner Marc, Noel Benjamin, Berline Leo, Labadie Karine, Cornils Astrid, Blanco-Bercial Leocadio, Stemmann Lars, Jamet Jean-Louis, Wincker Patrick.* New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod *Oithona* // *Molecular ecology*. 2017. 26, 17. 4467–4482.
- Mahadevan Amala.* The Impact of Submesoscale Physics on Primary Productivity of Plankton // *Annual Review of Marine Science*. 2016. 8, 1. 161–184. PMID : 26394203.
- Mahé Frédéric, Rognes Torbjørn, Quince Christopher, Vargas Colomban de, Dunthorn Micah.* Swarm v2 : highly-scalable and high-resolution amplicon clustering // *PeerJ*. XII 2015. 3. e1420.
- Malviya Shruti, Scalco Eleonora, Audic Stéphane, Vincent Flora, Veluchamy Alaguraj, Poulain Julie, Wincker Patrick, Iudicone Daniele, Vargas Colomban de, Bittner Lucie, Zingone Adriana, Bowler Chris.* Insights into global diatom distribution and diversity in the world's ocean // *Proceedings of the National Academy of Sciences*. 2016. 113, 11. E1516–E1525.
- Mandakovic Dinka, Rojas Claudia, Maldonado Jonathan, Latorre Mauricio, Travisany Dante, Delage Erwan, Bihouée Audrey, Jean Géraldine, Díaz Francisca P., Fernández-Gómez Beatriz, Cabrera Pablo, Gaete Alexis, Latorre Claudio, Gutiérrez Rodrigo A., Maass Alejandro, Cambiazo Verónica, Navarrete Sergio A., Eveillard Damien, González Mauricio.* Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience // *Scientific Reports*. 2018. 8, 1. 5875.
- Marañón Emilio.* Cell Size as a Key Determinant of Phytoplankton Metabolism and Community Structure // *Annual Review of Marine Science*. 2015. 7, 1. 241–264. PMID : 25062405.
- Martin Adrian, Boyd Philip, Buesseler Ken, Cetinic Ivona, Claustre Hervé, Giering Sari, Henson Stephanie, Irigoien Xavier, Kriest Iris, Memery Laurent, others .* The oceans' twilight zone must be studied now, before it is too late. 2020.
- Martini Séverine, Larras Floriane, Boyé Aurélien, Faure Emile, Aberle Nicole, Archambault Philippe, Bacouillard Lise, Beisner Beatrix E, Bittner Lucie, Castella Emmanuel, Danger Michael, Gauthier Olivier, Karp-Boss Lee, Lombard Fabien, Maps Frédéric, Stemmann Lars, Thiébaud Eric, Usseglio-Polatera Philippe, Vogt Meike, Laviale Martin, Ayata Sakina-Dorothee.* Functional trait-based approaches as a common framework for aquatic ecologists // *Limnology and Oceanography*. 03 2020. 66, 3. 965–994.
- Martinson Brian C., Anderson Melissa S., Vries Raymond de.* Scientists behaving badly // *Nature*. 2005. 435, 7043. 737–738.
- Martiny Jennifer B. Hughes, Bohannan Brendan J. M., Brown James H., Colwell Robert K., Fuhrman Jed A., Green Jessica L., Horner-Devine M. Claire, Kane Matthew, Krumins Jennifer Adams, Kuske Cheryl R., Morin Peter J., Naeem Shahid, Øvreås Lise, Reysenbach Anna-Louise, Smith Val H., Staley James T.* Microbial biogeography : putting microorganisms on the map // *Nature Reviews Microbiology*. 2006. 4, 2. 102–112.
- Mason Norman W. H., Mouillot David, Lee William G., Wilson J. Bastow.* Functional richness, functional evenness and functional divergence : the primary components of functional diversity // *Oikos*. 2005. 111, 1. 112–118.

- Maxam A. M., Gilbert W.* A new method for sequencing DNA // Proceedings of the National Academy of Sciences. 1977. 74, 2. 560–564.
- McClain Charles R.* A Decade of Satellite Ocean Color Observations // Annual Review of Marine Science. 2009. 1, 1. 19–42. PMID : 21141028.
- McGill Brian J, Enquist Brian J, Weiher Evan, Westoby Mark.* Rebuilding community ecology from functional traits // Trends in ecology & evolution. 2006. 21, 4. 178–185.
- McGillicuddy Dennis J.* Mechanisms of Physical-Biological-Biogeochemical Interaction at the Oceanic Mesoscale // Annual Review of Marine Science. 2016. 8, 1. 125–159. PMID : 26359818.
- McManus Margaret Anne, Woodson C Brock.* Plankton distribution and ocean dispersal // Journal of Experimental Biology. 2012. 215, 6. 1008–1016.
- McWilliams J. C., Colas F., Molemaker M. J.* Cold filamentary intensification and oceanic surface convergence lines // Geophysical Research Letters. 2009. 36, 18.
- McWilliams James C.* Submesoscale currents in the ocean // Proceedings of the Royal Society A : Mathematical, Physical and Engineering Sciences. 2016. 472, 2189. 20160117.
- Meilland Julie, Fabri-Ruiz Salomé, Koubbi Philippe, Monaco Claire Lo, Cotte Cédric, Hosie Graham W., Sanchez Sophie, Howa Hélène.* Planktonic foraminiferal biogeography in the Indian sector of the Southern Ocean : Contribution from CPR data // Deep Sea Research Part I : Oceanographic Research Papers. 2016. 110. 75–89.
- Mena Catalina, Reglero Patricia, Hidalgo Manuel, Sintes Eva, Santiago Rocío, Martín Melissa, Moyà Gabriel, Balbín Rosa.* Phytoplankton community structure is driven by stratification in the oligotrophic Mediterranean Sea // Frontiers in microbiology. 2019. 10. 1698.
- Michener William K, Jones Matthew B.* Ecoinformatics : supporting ecology as a data-intensive science // Trends in ecology & evolution. 2012. 27, 2. 85–93.
- Mignot Alexandre, Claustre Hervé, Uitz Julia, Poteau Antoine, D’Ortenzio Fabrizio, Xing Xiaogang.* Understanding the seasonal dynamics of phytoplankton biomass and the deep chlorophyll maximum in oligotrophic environments : A Bio-Argo float investigation // Global Biogeochemical Cycles. 2014. 28, 8. 856–876.
- Millot Claude, Taupier-Letage Isabelle.* Circulation in the Mediterranean sea // The Mediterranean Sea. 2005. 29–66.
- Monier Adam, Worden Alexandra Z., Richards Thomas A.* Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans // Environmental Microbiology Reports. 2016. 8, 4. 461–469.
- Monroy Pedro, Rossi Vincent, Ser-Giacomi Enrico, López Cristóbal, Hernández-García Emilio.* Sensitivity and robustness of larval connectivity diagnostics obtained from Lagrangian Flow Networks // ICES Journal of Marine Science. 2017. 74, 6. 1763–1779.
- Moon-van der Staay Seung Yeo, De Wachter Rupert, Vaultot Daniel.* Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity // Nature. 2001. 409, 6820. 607–610.

- Moreau Hervé, Verhelst Bram, Couloux Arnaud, Derelle Evelyne, Rombauts Stephane, Grimley Nigel, Van Bel Michiel, Poulain Julie, Katinka Michaël, Hohmann-Marriott Martin F., Piganeau Gwenael, Rouzé Pierre, Da Silva Corinne, Wincker Patrick, Peer Yves Van de, Vandepoele Klaas. Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage // *Genome biology*. 2012. 13, 8. 1–16.
- Morel F. M. M., Price N. M. The Biogeochemical Cycles of Trace Metals in the Oceans // *Science*. 2003. 300, 5621. 944–947.
- Morel François M.M., Lam Phoebe J., Saito Mak A. Trace Metal Substitution in Marine Phytoplankton // *Annual Review of Earth and Planetary Sciences*. 2020. 48, 1. 491–517.
- Morin Phillip A, McCarthy Melissa. Highly accurate SNP genotyping from historical and low-quality samples // *Molecular Ecology Notes*. 2007. 7, 6. 937–946.
- Mouchet Maud A., Villéger Sébastien, Mason Norman W. H., Mouillot David. Functional diversity measures : an overview of their redundancy and their ability to discriminate community assembly rules // *Functional Ecology*. 2010. 24, 4. 867–876.
- Mouillot David, Graham Nicholas A. J., Villéger Sébastien, Mason Norman W. H., Bellwood David R. A functional approach reveals community responses to disturbances // *Trends in ecology & evolution*. 2013. 28, 3. 167–177.
- Mousing Erik Askov, Richardson Katherine, Ellegaard Marianne. Global patterns in phytoplankton biomass and community size structure in relation to macronutrients in the open ocean // *Limnology and Oceanography*. 2018. 63, 3. 1298–1312.
- Murtagh Fionn, Legendre Pierre. Ward's hierarchical agglomerative clustering method : which algorithms implement Ward's criterion ? // *Journal of classification*. 2014. 31, 3. 274–295.
- Nagai S, Nishitani G, Sakamoto S, Sugaya T, Lee CK, Kim CH, Itakura S, Yamaguchi M. Genetic structuring and transfer of marine dinoflagellate *Cochlodinium polykrikoides* in Japanese and Korean coastal waters revealed by microsatellites // *Molecular Ecology*. 2009. 18, 11. 2337–2352.
- Nagai Satoshi, Lian Chunlan, Yamaguchi Sanae, Hamaguchi Masami, Matsuyama Yukihiko, Itakura Shigeru, Shimada Hiroshi, Kaga Shinnosuke, Yamauchi Hiroyuki, Sonda Yoshiko, Nishikawa Tetsuya, Kim Chang-Hoon, Hogetsu Taizo. Microsatellite markers reveal population genetic structure of the toxic Dinoflagellate *Alexandrium tamarense* (Dinophyceae) in Japanese coastal waters // *Journal of Phycology*. 2007. 43, 1. 43–54.
- Nei Masatoshi. Analysis of gene diversity in subdivided populations // *Proceedings of the National Academy of Sciences*. 1973. 70, 12. 3321–3323.
- Nock Charles A., Vogt Richard J., Beisner Beatrix E. Functional Traits // *eLS*. 2016. 1–8.
- Norris Richard D. Pelagic species diversity, biogeography, and evolution // *Paleobiology*. 2000. 26, S4. 236–258.
- Not Fabrice, Campo Javier del, Balagué Vanessa, Vargas Colomban de, Massana Ramon. New Insights into the Diversity of Marine Picoeukaryotes // *PLOS ONE*. 09 2009. 4, 9. 1–7.

- Obiol Aleix, Giner Caterina R., Sánchez Pablo, Duarte Carlos M., Acinas Silvia G., Massana Ramon.* A metagenomic assessment of microbial eukaryotic diversity in the global ocean // *Molecular Ecology Resources*. 2020. 20, 3. 718–731.
- Ohman Mark D., Davis Russ E., Sherman Jeffrey T., Grindley Kyle R., Whitmore Benjamin M., Nickels Catherine F., Ellen Jeffrey S.* Zooglider : An autonomous vehicle for optical and acoustic sensing of zooplankton // *Limnology and Oceanography : Methods*. 2019. 17, 1. 69–86.
- Ohman Mark D., Romagnan Jean-Baptiste.* Nonlinear effects of body size and optical attenuation on Diel Vertical Migration by zooplankton // *Limnology and Oceanography*. 2016. 61, 2. 765–770.
- Oinn Tom, Addis Matthew, Ferris Justin, Marvin Darren, Senger Martin, Greenwood Mark, Carver Tim, Glover Kevin, Pocock Matthew R., Wipat Anil, Li Peter.* Taverna : a tool for the composition and enactment of bioinformatics workflows // *Bioinformatics*. 2004. 20, 17. 3045–3054.
- Okolodkov Yuri B., Dodge John D.* Biodiversity and biogeography of planktonic dinoflagellates in the Arctic Ocean // *Journal of Experimental Marine Biology and Ecology*. 1996. 202, 1. 19–27. *Marine Biodiversity*.
- Oliverio Angela M., Gan Huijie, Wickings Kyle, Fierer Noah.* A DNA metabarcoding approach to characterize soil arthropod communities // *Soil Biology and Biochemistry*. 2018. 125. 37–43.
- Orenstein Eric C., Ratelle Devin, Briseño-Avena Christian, Carter Melissa L., Franks Peter J. S., Jaffe Jules S., Roberts Paul L. D.* The Scripps Plankton Camera system : A framework and platform for in situ microscopy // *Limnology and Oceanography : Methods*. 2020. 18, 11. 681–695.
- O'Malley Maureen A.* ‘Everything is everywhere : but the environment selects’ : ubiquitous distribution and ecological determinism in microbial biogeography // *Studies in History and Philosophy of Science Part C : Studies in History and Philosophy of Biological and Biomedical Sciences*. 2008. 39, 3. 314–325.
- Palmer Michael W.* Variation in species richness : Towards a unification of hypotheses // *Folia Geobotanica et Phytotaxonomica*. 1994. 29, 4. 511.
- Palter Jaime B., Lozier M. Susan, Barber Richard T.* The effect of advection on the nutrient reservoir in the North Atlantic subtropical gyre // *Nature*. 2005. 437, 7059. 687–692.
- Palumbi Stephen R.* Population genetics, demographic connectivity, and the design of marine reserves // *Ecological applications*. 2003. 13, sp1. 146–158.
- Pearre Sifford Jr.* Eat and run ? The hunger/satiation hypothesis in vertical migration : history, evidence and consequences // *Biological Reviews*. 2003. 78, 1. 1–79.
- Pecl Gretta T., Araújo Miguel B., Bell Johann D., Blanchard Julia, Bonebrake Timothy C., Chen I-Ching, Clark Timothy D., Colwell Robert K., Danielsen Finn, Evengård Birgitta, Falconi Lorena, Ferrier Simon, Frusher Stewart, Garcia Raquel A., Griffis Roger B., Hobday Alistair J., Janion-Scheepers Charlene, Jarzyna Marta A., Jennings Sarah, Lenoir Jonathan, Linnetved Hlif I., Martin Victoria Y., McCormack Phillipa C., McDonald Jan, Mitchell Nicola J., Mustonen Tero, Pandolfi John M., Pettoelli Nathalie, Popova Ekaterina, Robinson*

- Sharon A., Scheffers Brett R., Shaw Justine D., Sorte Cascade J. B., Strugnell Jan M., Sunday Jennifer M., Tuanmu Mao-Ning, Vergés Adriana, Villanueva Cecilia, Wernberg Thomas, Wapstra Erik, Williams Stephen E.* Biodiversity redistribution under climate change : Impacts on ecosystems and human well-being // *Science*. 2017. 355, 6332.
- Peng Roger D.* Reproducible Research in Computational Science // *Science*. 2011. 334, 6060. 1226–1227.
- Perkel Jeffrey M.* // *Nature*. 2020. 584, 7822. 656–658.
- Pesant Stéphane, Not Fabrice, Picheral Marc, Kandels-Lewis Stefanie, Le Bescot Noan, Gorsky Gabriel, Iudicone Daniele, Karsenti Eric, Speich Sabrina, Troublé Romain, Dimier Céline, Searson Sarah, Acinas Silvia G., Bork Peer, Boss Emmanuel, Bowler Chris, De Vargas Colomban, Follows Michael, Grimsley Nigel, Hingamp Pascal, Jaillon Olivier, Karp-Boss Lee, Krzic Uros, Ogata Hiroyuki, Raes Jeroen, Reynaud Emmanuel G., Sardet Christian, Sieracki Mike, Stemmann Lars, Sullivan Matthew B., Sunagawa Shinichi, Velayoudon Didier, Weissenbach Jean, Wincker Patrick, Coordinators Tara Oceans Consortium.* Open science resources for the discovery and analysis of Tara Oceans data // *Scientific data*. 2015. 2, 1. 1–16.
- Petchey Owen L, Gaston Kevin J.* Functional diversity : back to basics and looking forward // *Ecology letters*. 2006. 9, 6. 741–758.
- Piani Laurette, Marrocchi Yves, Rigaudier Thomas, Vacher Lionel G., Thomassin Dorian, Marty Bernard.* Earth’s water may have been inherited from material similar to enstatite chondrite meteorites // *Science*. 2020. 369, 6507. 1110–1113.
- Picheral M, Colin S, Irisson JO.* EcoTaxa, a tool for the taxonomic classification of images // URL : <http://ecotaxa.obs-vlfr.fr>. 2017.
- Picheral Marc, Guidi Lionel, Stemmann Lars, Karl David M., Iddaoud Ghizlaine, Gorsky Gabriel.* The Underwater Vision Profiler 5 : An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton // *Limnology and Oceanography : Methods*. 2010. 8, 9. 462–473.
- Pierella Karlusich Juan José, Ibarbalz Federico M., Bowler Chris.* Phytoplankton in the Tara Ocean // *Annual Review of Marine Science*. 2020. 12, 1. 233–265. PMID : 31899671.
- Piganeau Gwenaël, Moreau Hervé.* Screening the Sargasso Sea metagenome for data to investigate genome evolution in *Ostreococcus* (Prasinophyceae, Chlorophyta) // *Gene*. 2007. 406, 1. 184–190. *Marine Genomics*.
- Pineda Jesús, Hare Jonathan A, Sponaugle SU.* Larval transport and dispersal in the coastal ocean and consequences for population connectivity // *Oceanography*. 2007. 20, 3. 22–39.
- Piroddi Chiara, Coll Marta, Liqueste Camino, Macias Diego, Greer Krista, Buszowski Joe, Steenbeek Jeroen, Danovaro Roberto, Christensen Villy.* Historical changes of the Mediterranean Sea ecosystem : modelling the role and impact of primary productivity and fisheries changes over time // *Scientific reports*. 2017. 7, 1. 1–18.
- Pitt K. A., Kingsford M. J., Rissik D., Koop K.* Jellyfish modify the response of planktonic assemblages to nutrient pulses // *Marine Ecology Progress Series*. 2004. 351. 1–13.

- Poisot Timothée, Gravel Dominique.* When is an ecological network complex? Connectance drives degree distribution and emerging network properties // *PeerJ*. II 2014. 2. e251.
- Porcel Sol, Chaparro Griselda, Marinone M Cristina, Saad Juan F, Lancelotti Julio, Izaguirre Irina.* The role of environmental, geographical, morphometric and spatial variables on plankton communities in lakes of the arid Patagonian plateaus // *Journal of Plankton Research*. 02 2020. 42, 2. 173–187.
- Pornon André, Andalo Christophe, Burrus Monique, Escaravage Nathalie.* DNA metabarcoding data unveils invisible pollination networks // *Scientific Reports*. 2017. 7, 1. 16828.
- Potvin Marianne, Lovejoy Connie.* PCR-Based Diversity Estimates of Artificial and Environmental 18S rRNA Gene Libraries // *Journal of Eukaryotic Microbiology*. 2009. 56, 2. 174–181.
- Prowe A. E. Friederike, Visser André W., Andersen Ken H., Chiba Sanae, Kjørboe Thomas.* Biogeography of zooplankton feeding strategy // *Limnology and Oceanography*. 2019. 64, 2. 661–678.
- R Core Team . R : A Language and Environment for Statistical Computing.* Vienna, Austria, 2017.
- Raskoff K. A., Purcell J. E., Hopcroft R. R.* Gelatinous zooplankton of the Arctic Ocean : in situ observations under the ice // *Polar Biology*. 2005. 28, 3. 207–217.
- Read Betsy A., Kegel Jessica, Klute Mary J., Kuo Alan, Lefebvre Stephane C., Maumus Florian, Mayer Christoph, Miller John, Monier Adam, Salamov Asaf, Young Jeremy, Aguilar Maria, Claverie Jean-Michel, Frickenhaus Stephan, Gonzalez Karina, Herman Emily K., Lin Yao-Cheng, Napier Johnathan, Ogata Hiroyuki, Sarno Analissa F., Shmutz Jeremy, Schroeder Declan, Vargas Colomban de, Verret Frederic, Dassow Peter von, Valentin Klaus, Peer Yves Van de, Wheeler Glen, Allen Andrew E., Bidle Kay, Borodovsky Mark, Bowler Chris, Brownlee Colin, Mark Cock J., Elias Marek, Gladyshev Vadim N., Groth Marco, Guda Chittibabu, Hadaegh Ahmad, Debora Iglesias-Rodriguez Maria, Jenkins Jerry, Jones Bethan M., Lawson Tracy, Leese Florian, Lindquist Erika, Lobanov Alexei, Lomsadze Alexandre, Malik Shehre-Banoo, Marsh Mary E., Mackinder Luke, Mock Thomas, Mueller-Roeber Bernd, Pagarete António, Parker Micaela, Probert Ian, Quesneville Hadi, Raines Christine, Rensing Stefan A., Riaño-Pachón Diego Mauricio, Richier Sophie, Rokitta Sebastian, Shiraiwa Yoshihiro, Soanes Darren M., Giezen Mark van der, Wahlund Thomas M., Williams Bryony, Wilson Willie, Wolfe Gordon, Wurch Louie L., Dacks Joel B., Delwiche Charles F., Dyhrman Sonya T., Glöckner Gernot, John Uwe, Richards Thomas, Worden Alexandra Z., Zhang Xiaoyu, Grigoriev Igor V., Annotation Consortium *Emiliana huxleyi*.* Pan genome of the phytoplankton *Emiliana* underpins its global distribution // *Nature*. 2013. 499, 7457. 209–213.
- Reichstein Markus, Camps-Valls Gustau, Stevens Bjorn, Jung Martin, Denzler Joachim, Carvalhais Nuno, Prabhat .* Deep learning and process understanding for data-driven Earth system science // *Nature*. 2019. 566, 7743. 195–204.
- Reid PC, Colebrook JM, Matthews JBL, Aiken JCPR, Team Continuous Plankton Recorder.* The Continuous Plankton Recorder : concepts and history, from Plankton Indicator to undulating recorders // *Progress in Oceanography*. 2003. 58, 2-4. 117–173.
- Reygondeau Gabriel, Longhurst Alan, Martinez Elodie, Beaugrand Gregory, Antoine David, Maury Olivier.* Dynamic biogeochemical provinces in the global ocean // *Global Biogeochemical Cycles*. 2013. 27, 4. 1046–1058.

- Richardson Anthony J., Bakun Andrew, Hays Graeme C., Gibbons Mark J.* The jellyfish joyride : causes, consequences and management responses to a more gelatinous future // *Trends in Ecology & Evolution.* 2009. 24, 6. 312–322.
- Richirt J., Schweizer M., Mouret A., Quincharde S., Saad S. A., Bouchet V. M. P., Wade C. M., Jorissen F. J.* Biogeographic distribution of three phylotypes (T1, T2 and T6) of *Ammonia* (foraminifera, Rhizaria) around Great Britain : new insights from combined molecular and morphological recognition // *Journal of Micropalaeontology.* 2021. 40, 1. 61–74.
- Richter Daniel J, Watteaux Romain, Vannier Thomas, Leconte Jade, Frémont Paul, Reygondeau Gabriel, Maillet Nicolas, Henry Nicolas, Benoit Gaëtan, Da Silva Ophélie, Tom O. Delmont, Fernández-Guerra Antonio, Suweis Samir, Narci Romain, Berney Cédric, Eveillard Damien, Gavory Frederick, Guidi Lionel, Labadie Karine, Mahieu Eric, Poulain Julie, Romac Sarah, Roux Simon, Dimier Céline, Kandels Stefanie, Picheral Marc, Searson Sarah, Tara Oceans Coordinators Stéphane Pesant, Aury Jean-Marc, Brum Jennifer R., Lemaitre Claire, Pelletier Eric, Bork Peer, Sunagawa Shinichi, Lombard Fabien, Karp-Boss Lee, Bowler Chris, Sullivan Matthew B., Karsenti Eric, Mariadassou Mahendra, Probert Ian, Peterlongo Pierre, Wincker Patrick, Vargas Colomban de, Ribera d'Alcalà Maurizio, Iudicone Daniele, Jaillon Olivier.* Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems // *BioRxiv.* 2020. 867739.
- Ricour F., Capet A., D'Ortenzio F., Delille B., Grégoire M.* Dynamics of the deep chlorophyll maximum in the Black Sea as depicted by BGC-Argo floats // *Biogeosciences.* 2021. 18, 2. 755–774.
- Riginos Cynthia, Crandall Eric D, Liggins Libby, Bongaerts Pim, Treml Eric A.* Navigating the currents of seascape genomics : how spatial analyses can augment population genomic studies // *Current zoology.* 2016. 62, 6. 581–601.
- Rinke Christian, Schwientek Patrick, Sczyrba Alexander, Ivanova Natalia N., Anderson Iain J., Cheng Jan-Fang, Darling Aaron, Malfatti Stephanie, Swan Brandon K., Gies Esther A., Dodsworth Jeremy A., Hedlund Brian P., Tsiamis George, Sievert Stefan M., Liu Wen-Tso, Eisen Jonathan A., Hallam Steven J., Kyrpides Nikos C., Stepanauskas Ramunas, Rubin Edward M., Hugenholtz Philip, Woyke Tanja.* Insights into the phylogeny and coding potential of microbial dark matter // *Nature.* 2013. 499, 7459. 431–437.
- Roemmich Dean, Alford Matthew H., Claustre Hervé, Johnson Kenneth, King Brian, Moun James, Oke Peter, Owens W. Brechner, Pouliquen Sylvie, Purkey Sarah, Scanderbeg Megan, Suga Toshio, Wijffels Susan, Zilberman Nathalie, Bakker Dorothee, Baringer Molly, Belbeoch Mathieu, Bittig Henry C., Boss Emmanuel, Calil Paulo, Carse Fiona, Carval Thierry, Chai Fei, Conchubhair Diarmuid Ó., d'Ortenzio Fabrizio, Dall'Olmo Giorgio, Desbruyeres Damien, Fennel Katja, Fer Ilker, Ferrari Raffaele, Forget Gael, Freeland Howard, Fujiki Tetsuichi, Gehlen Marion, Greenan Blair, Hallberg Robert, Hibiya Toshiyuki, Hosoda Shigeki, Jayne Steven, Jochum Markus, Johnson Gregory C., Kang KiRyong, Kolodziejczyk Nicolas, Körtzinger Arne, Traon Pierre-Yves Le, Lenn Yueng-Djern, Maze Guillaume, Mork Kjell Arne, Morris Tamaryn, Nagai Takeyoshi, Nash Jonathan, Garabato Alberto Naveira, Olsen Are, Pattabhi Rama Rao, Prakash Satya, Riser Stephen, Schmechtig Catherine, Schmid Claudia, Shroyer Emily, Sterl Andreas, Sutton Philip, Talley Lynne, Tanhua Toste, Thierry Virginie, Thomalla Sandy, Toole John, Troisi Ariel, Trull Thomas W., Turton Jon, Velez-Belchi Pedro Joaquin, Walczowski Waldemar, Wang Haili, Wanninkhof Rik, Waterhouse Amy F., Waterman Stephanie, Watson Andrew, Wilson Cara, Wong Annie P. S., Xu Jianping, Yasuda Ichiro.* On

- the Future of Argo : A Global, Full-Depth, Multi-Disciplinary Array // *Frontiers in Marine Science*. 2019. 6. 439.
- Romagnan Jean Baptiste, Aldamman Lama, Gasparini Stéphane, Nival Paul, Aubert Anaïs, Jamet Jean Louis, Stemann Lars*. High frequency mesozooplankton monitoring : Can imaging systems and automated sample analysis help us describe and interpret changes in zooplankton community composition and size structure—An example from a coastal site // *Journal of Marine Systems*. 2016. 162. 18–28.
- Rousseau Véronique, Chrétiennot-Dinet Marie-Josèphe, Jacobsen Anita, Verity Peter, Whipple Stuart*. The life cycle of *Phaeocystis* : state of knowledge and presumptive role in ecology // *Biogeochemistry*. 2007. 83, 1. 29–47.
- Roy Shovonlal, Chattopadhyay J*. Towards a resolution of ‘the paradox of the plankton’ : A brief overview of the proposed mechanisms // *Ecological complexity*. 2007. 4, 1-2. 26–33.
- Rudnick Daniel L*. Ocean Research Enabled by Underwater Gliders // *Annual Review of Marine Science*. 2016. 8, 1. 519–541. PMID : 26291384.
- Ryabov Alexey, Kerimoglu Onur, Litchman Elena, Olenina Irina, Roselli Leonilde, Basset Alberto, Stanca Elena, Blasius Bernd*. Shape matters : the relationship between cell geometry and diversity in phytoplankton // *Ecology Letters*. 2021. 24, 4. 847–861.
- Salazar Guillem, Paoli Lucas, Alberti Adriana, Huerta-Cepas Jaime, Ruscheweyh Hans-Joachim, Cuenca Miguelangel, Field Christopher M., Coelho Luis Pedro, Cruaud Corinne, Engelen Stefan, Gregory Ann C., Labadie Karine, Marec Claudie, Pelletier Eric, Royo-Llonch Marta, Roux Simon, Sánchez Pablo, Uehara Hideya, Zayed Ahmed A., Zeller Georg, Carmichael Margaux, Dimier Céline, Ferland Joannie, Kandels Stefanie, Picheral Marc, Pisarev Sergey, Poulain Julie, Acinas Silvia G., Babin Marcel, Bork Peer, Boss Emmanuel, Bowler Chris, Cochrane Guy, Vargas Colomban de, Follows Michael, Gorsky Gabriel, Grimsley Nigel, Guidi Lionel, Hingamp Pascal, Iudicone Daniele, Jaillon Olivier, Kandels-Lewis Stefanie, Karp-Boss Lee, Karsenti Eric, Not Fabrice, Ogata Hiroyuki, Pesant Stephane, Poulton Nicole, Raes Jeroen, Sardet Christian, Speich Sabrina, Stemann Lars, Sullivan Matthew B., Sunagawa Shinichi, Wincker Patrick, Acinas Silvia G., Sullivan Matthew B*. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome // *Cell*. 2019. 179, 5. 1068–1083.
- Sallée Jean-Baptiste, Pellichero Violaine, Akhoudas Camille, Pauthenet Etienne, Vignes Lucie, Schmidtke Sunke, Garabato Alberto Naveira, Sutherland Peter, Kuusela Mikael*. Summertime increases in upper-ocean stratification and mixed-layer depth // *Nature*. 2021. 591, 7851. 592–598.
- Sanger F., Nicklen S., Coulson A. R*. DNA sequencing with chain-terminating inhibitors // *Proceedings of the National Academy of Sciences*. 1977. 74, 12. 5463–5467.
- Santos Adriana Lopes dos, Gourvil Priscillia, Tragin Margot, Noël Mary-Hélène, Decelle Johan, Romac Sarah, Vaultot Daniel*. Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters // *The ISME Journal*. 2017. 11, 2. 512–528.
- Scheffer Marten, Rinaldi Sergio, Huisman Jef, Weissing Franz J*. Why plankton communities have no equilibrium : solutions to the paradox // *Hydrobiologia*. 2003. 491, 1. 9–18.

- Schmidt Philipp-André, Bálint Miklós, Greshake Bastian, Bandow Cornelia, Römbke Jörg, Schmitt Imke.* Illumina metabarcoding of a soil fungal community // *Soil Biology and Biochemistry.* 2013. 65. 128–132.
- Schmieder Robert, Edwards Robert.* Quality control and preprocessing of metagenomic datasets // *Bioinformatics.* 2011. 27, 6. 863–864.
- Schuech Rudi, Menden-Deuer Susanne.* Going ballistic in the plankton : anisotropic swimming behavior of marine protists // *Limnology and Oceanography : Fluids and Environments.* 2014. 4, 1. 1–16.
- Schuwirth Nele, Borgwardt Florian, Domisch Sami, Friedrichs Martin, Kattwinkel Mira, Kneis David, Kuemmerlen Mathias, Langhans Simone D., Martínez-López Javier, Vermeiren Peter.* How to make ecological models useful for environmental management // *Ecological Modelling.* 2019. 411. 108784.
- Seeleuthner Yoann, Mondy Samuel, Lombard Vincent, Carradec Quentin, Pelletier Eric, Wessner Marc, Leconte Jade, Mangot Jean-François, Poulain Julie, Labadie Karine, Logares Ramiro, Sunagawa Shinichi, Berardinis Véronique de, Salanoubat Marcel, Dimier Céline, Kandels-Lewis Stefanie, Picheral Marc, Searson Sarah, Acinas Silvia G., Boss Emmanuel, Follows Michael, Gorsky Gabriel, Grimsley Nigel, Karp-Boss Lee, Krzic Uros, Not Fabrice, Ogata Hiroyuki, Raes Jeroen, Reynaud Emmanuel G., Sardet Christian, Speich Sabrina, Stemmann Lars, Velayoudon Didier, Weissenbach Jean, Pesant Stephane, Poulton Nicole, Stepanauskas Ramunas, Bork Peer, Bowler Chris, Hingamp Pascal, Sullivan Matthew B., Iudicone Daniele, Massana Ramon, Aury Jean-Marc, Henrissat Bernard, Karsenti Eric, Jaillon Olivier, Sieracki Mike, Vargas Colomban de, Wincker Patrick, Coordinators Tara Oceans.* Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans // *Nature Communications.* 2018. 9, 1. 310.
- Selkoe Kimberly A, Aloia Cassidy CD, Crandall Eric D, Iacchei Matthew, Liggins Libby, Puritz Jonathan B, Heyden Sophie von der, Toonen Robert J.* A decade of seascape genetics : contributions to basic and applied marine connectivity // *Marine Ecology Progress Series.* 2016. 554. 1–19.
- Selkoe Kimberly A, Watson James R, White Crow, Horin Tal Ben, Iacchei Matthew, Mitarai Satoshi, Siegel David A, Gaines Steven D, Toonen Robert J.* Taking the chaos out of genetic patchiness : seascape genetics reveals ecological and oceanographic drivers of genetic patterns in three temperate reef species // *Molecular ecology.* 2010. 19, 17. 3708–3726.
- Ser-Giacomi Enrico, Rossi Vincent, López Cristóbal, Hernández-García Emilio.* Flow networks : A characterization of geophysical fluid transport // *Chaos : An Interdisciplinary Journal of Nonlinear Science.* 2015. 25, 3. 036404.
- Ser-Giacomi Enrico, Zinger Lucie, Malviya Shruti, De Vargas Colomban, Karsenti Eric, Bowler Chris, De Monte Silvia.* Ubiquitous abundance distribution of non-dominant plankton across the global ocean // *Nature Ecology & Evolution.* 2018. 2, 8. 1243–1249.
- Sexton Jason P, Hangartner Sandra B, Hoffmann Ary A.* Genetic isolation by environment or distance : which pattern of gene flow is most common ? // *Evolution.* 2014. 68, 1. 1–15.
- Shadden Shawn C, Lekien Francois, Marsden Jerrold E.* Definition and properties of Lagrangian coherent structures from finite-time Lyapunov exponents in two-dimensional aperiodic flows // *Physica D : Nonlinear Phenomena.* 2005. 212, 3-4. 271–304.

- Shaiber Alon, Eren A Murat.* Composite metagenome-assembled genomes reduce the quality of public genome repositories // *MBio*. 2019. 10, 3. e00725–19.
- Sibbald Shannon J, Archibald John M.* More protist genomes needed // *Nature ecology & evolution*. 2017. 1, 5. 1–3.
- Sieracki Christian K, Sieracki Michael E, Yentsch Charles S.* An imaging-in-flow system for automated analysis of marine microplankton // *Marine Ecology Progress Series*. 1998. 168. 285–296.
- Sims David, Sudbery Ian, Ilott Nicholas E, Heger Andreas, Ponting Chris P.* Sequencing depth and coverage : key considerations in genomic analyses // *Nature Reviews Genetics*. 2014. 15, 2. 121–132.
- Sjöqvist C, Godhe A, Jonsson PR, Sundqvist L, Krempe A.* Local adaptation and oceanographic connectivity patterns explain genetic differentiation of a marine diatom across the North Sea–Baltic Sea salinity gradient // *Molecular ecology*. 2015. 24, 11. 2871–2885.
- Slowikowski Kamil.* ggrepel : Automatically Position Non-Overlapping Text Labels with 'ggplot2'. 2020. R package version 0.8.2.
- Sobel James M, Chen Grace F, Watt Lorna R, Schemske Douglas W.* The biology of speciation // *Evolution : International Journal of organic evolution*. 2010. 64, 2. 295–315.
- Sommeria-Klein Guilhem, Watteaux Romain, Iudicone Daniele, Bowler Chris, Morlon Hélène.* Global drivers of eukaryotic plankton biogeography in the sunlit ocean // *bioRxiv*. 2020.
- Stange Madlen, Barrett Rowan D. H., Hendry Andrew P.* The importance of genomic variation for biodiversity, ecosystems and people // *Nature Reviews Genetics*. 2021. 22, 2. 89–105.
- Stark Philip B.* Before reproducibility must come preproducibility // *Nature*. 2018. 557, 7706. 613.
- Stoecker Diane K., Hansen Per Juel, Caron David A., Mitra Aditee.* Mixotrophy in the Marine Plankton // *Annual Review of Marine Science*. 2017. 9, 1. 311–335. PMID : 27483121.
- Sunagawa Shinichi, Acinas Silvia G., Bork Peer, Bowler Chris, Acinas Silvia G., Babin Marcel, Boss Emmanuel, Cochrane Guy, Vargas Colomban de, Follows Michael, Gorsky Gabriel, Grimsley Nigel, Guidi Lionel, Hingamp Pascal, Iudicone Daniele, Jaillon Olivier, Kandels Stefanie, Karp-Boss Lee, Karsenti Eric, Lescot Magali, Not Fabrice, Ogata Hiroyuki, Pesant Stéphane, Poulton Nicole, Raes Jeroen, Sardet Christian, Sieracki Mike, Speich Sabrina, Stemmann Lars, Sullivan Matthew B., Wincker Patrick, Eveillard Damien, Lombard Fabien, Pesant Stephane, Sullivan Matthew B., Coordinators Tara Oceans.* Tara Oceans : towards global ocean ecosystems biology // *Nature Reviews Microbiology*. 2020. 18, 8. 428–445.
- Sundelöf Andreas, Jonsson Per R.* Larval dispersal and vertical migration behaviour—a simulation study for short dispersal times // *Marine Ecology*. 2012. 33, 2. 183–193.
- Suzuki Yoshihiro, Takahashi Masayuki.* GROWTH RESPONSES OF SEVERAL DIATOM SPECIES ISOLATED FROM VARIOUS ENVIRONMENTS TO TEMPERATURE // *Journal of Phycology*. 1995. 31, 6. 880–888.

- Tackmann Janko, Matias Rodrigues João Frederico, Mering Christian von.* Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data // *Cell Systems*. 2021/07/26 2019. 9, 3. 286–296.e8.
- Testor Pierre, de Young Brad, Rudnick Daniel L., Glenn Scott, Hayes Daniel, Lee Craig M., Pattiaratchi Charitha, Hill Katherine, Heslop Emma, Turpin Victor, Alenius Pekka, Barrera Carlos, Barth John A., Beaird Nicholas, Bécu Guislain, Bosse Anthony, Bourrin François, Brearley J. Alexander, Chao Yi, Chen Sue, Chiggiato Jacopo, Coppola Laurent, Crout Richard, Cummings James, Curry Beth, Curry Ruth, Davis Richard, Desai Kruti, DiMarco Steve, Edwards Catherine, Fielding Sophie, Fer Ilker, Frajka-Williams Eleanor, Gildor Hezi, Goni Gustavo, Gutierrez Dimitri, Haugan Peter, Hebert David, Heiderich Joleen, Henson Stephanie, Heywood Karen, Hogan Patrick, Houpert Loïc, Huh Sik, E. Inall Mark, Ishii Masso, Ito Shin-ichi, Itoh Sachihiko, Jan Sen, Kaiser Jan, Karstensen Johannes, Kirkpatrick Barbara, Klymak Jody, Kohut Josh, Krahmman Gerd, Krug Marjolaine, McClatchie Sam, Marin Frédéric, Mauri Elena, Mehra Avichal, P. Meredith Michael, Meunier Thomas, Miles Travis, Morell Julio M., Mortier Laurent, Nicholson Sarah, O'Callaghan Joanne, O'Conchubhair Diarmuid, Oke Peter, Pallàs-Sanz Enric, Palmer Matthew, Park JongJin, Perivoliotis Leonidas, Poulain Pierre-Marie, Perry Ruth, Queste Bastien, Rainville Luc, Rehm Eric, Roughan Moninya, Rome Nicholas, Ross Tetjana, Ruiz Simon, Saba Grace, Schaeffer Amandine, Schönau Martha, Schroeder Katrin, Shimizu Yugo, Sloyan Bernadette M., Smeed David, Snowden Derrick, Song Yumi, Swart Sebastian, Tenreiro Miguel, Thompson Andrew, Tintore Joaquin, Todd Robert E., Toro Cesar, Venables Hugh, Wagawa Taku, Waterman Stephanie, Watlington Roy A., Wilson Doug.* Ocean Gliders : A Component of the Integrated GOOS // *Frontiers in Marine Science*. 2019. 6. 422.
- Thetmeyer H., Kils U.* To see and not be seen : the visibility of predator and prey with respect to feeding behaviour // *Marine Ecology Progress Series*. 1995. 126. 1–8.
- Thornhill Daniel J, Mahon Andrew R, Norenburg Jon L, Halanych Kenneth M.* Open-ocean barriers to dispersal : A test case with the Antarctic Polar Front and the ribbon worm *Parborlasia corrugatus* (Nemertea : Lineidae) // *Molecular ecology*. 2008. 17, 23. 5104–5117.
- Tilman David, Isbell Forest, Cowles Jane M.* Biodiversity and ecosystem functioning // *Annual review of ecology, evolution, and systematics*. 2014. 45. 471–493.
- Trudnowska Emilia, Lacour Léo, Ardyna Mathieu, Rogge Andreas, Irisson Jean Olivier, Waite Anya M., Babin Marcel, Stemann Lars.* Marine snow morphology illuminates the evolution of phytoplankton blooms and determines their subsequent vertical export // *Nature Communications*. 2021. 12, 1. 2816.
- Tully Benjamin J, Graham Elaina D, Heidelberg John F.* The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans // *Scientific data*. 2018. 5, 1. 1–8.
- Tully Benjamin J, Sachdeva Rohan, Graham Elaina D, Heidelberg John F.* 290 metagenome-assembled genomes from the Mediterranean Sea : a resource for marine microbiology // *PeerJ*. 2017. 5. e3558.
- Uitz Julia, Claustre Hervé, Gentili Bernard, Stramski Dariusz.* Phytoplankton class-specific primary production in the world's oceans : Seasonal and interannual variability from satellite observations // *Global Biogeochemical Cycles*. 2010. 24, 3.

- Vanni Chiara, Schechter Matthew S., Acinas Silvia G., Barberán Albert, Buttigieg Pier Luigi, Casamayor Emilio O., Delmont Tom O., Duarte Carlos M., Eren A. Murat, Finn Robert D., Kottmann Renzo, Mitchell Alex, Sanchez Pablo, Siren Kimmo, Steinegger Martin, Glöckner Frank Oliver, Fernandez-Guerra Antonio. Unifying the known and unknown microbial coding sequence space // *bioRxiv*. 2021.
- Vannier Thomas, Leconte Jade, Seeleuthner Yoann, Mondy Samuel, Pelletier Eric, Aury Jean-Marc, Vargas Colomban de, Sieracki Michael, Iudicone Daniele, Vaultot Daniel, Wincker Patrick, Jaillon Olivier. Survey of the green picoalga *Bathycoccus* genomes in the global ocean // *Scientific reports*. 2016. 6, 1. 1–11.
- Vaughn Dawn, Allen Jonathan D. The peril of the plankton // *Integrative and Comparative Biology*. 2010. 50, 4. 552–570.
- Vellend Mark. Conceptual Synthesis in Community Ecology // *The Quarterly Review of Biology*. 2010. 85, 2. 183–206. PMID : 20565040.
- Venter J. Craig, Remington Karin, Heidelberg John F., Halpern Aaron L., Rusch Doug, Eisen Jonathan A., Wu Dongying, Paulsen Ian, Nelson Karen E., Nelson William, Fouts Derrick E., Levy Samuel, Knap Anthony H., Lomas Michael W., Nealson Ken, White Owen, Peterson Jeremy, Hoffman Jeff, Parsons Rachel, Baden-Tillson Holly, Pfannkoch Cynthia, Rogers Yu-Hui, Smith Hamilton O. Environmental Genome Shotgun Sequencing of the Sargasso Sea // *Science*. 2004. 304, 5667. 66–74.
- Vilgrain Laure, Maps Frédéric, Picheral Marc, Babin Marcel, Aubry Cyril, Irisson Jean-Olivier, Ayata Sakina-Dorothee. Trait-based approach using in situ copepod images reveals contrasting ecological patterns across an Arctic ice melt zone // *Limnology and Oceanography*. 2021. 66, 4. 1155–1167.
- Villa Martín Paula, Buček Aleš, Bourguignon Thomas, Pigolotti Simone. Ocean currents promote rare species diversity in protists // *Science Advances*. 2020. 6, 29.
- Villar Emilie, Farrant Gregory K., Follows Michael, Garczarek Laurence, Speich Sabrina, Audic Stéphane, Bittner Lucie, Blanke Bruno, Brum Jennifer R., Brunet Christophe, Casotti Raffaella, Chase Alison, Dolan John R., d'Ortenzio Fabrizio, Gattuso Jean-Pierre, Grima Nicolas, Guidi Lionel, Hill Christopher N., Jahn Oliver, Jamet Jean-Louis, Le Goff Hervé, Lepoivre Cyrille, Malviya Shruti, Pelletier Eric, Romagnan Jean-Baptiste, Roux Simon, Santini Sébastien, Scalco Eleonora, Schwenck Sarah M., Tanaka Atsuko, Testor Pierre, Vannier Thomas, Vincent Flora, Zingone Adriana, Dimier Céline, Picheral Marc, Searson Sarah, Kandels-Lewis Stefanie, Coordinators Tara Oceans, Acinas Silvia G., Bork Peer, Boss Emmanuel, Vargas Colomban de, Gorsky Gabriel, Ogata Hiroyuki, Pesant Stéphane, Sullivan Matthew B., Sunagawa Shinichi, Wincker Patrick, Karsenti Eric, Bowler Chris, Not Fabrice, Hingamp Pascal, Iudicone Daniele. Environmental characteristics of Agulhas rings affect interocean plankton transport // *Science*. 2015. 348, 6237.
- Villarino Ernesto, Watson James R., Jönsson Bror, Gasol Josep M., Salazar Guillem, Acinas Silvia G., Estrada Marta, Massana Ramón, Logares Ramiro, Giner Caterina R., Pernice Massimo C., Olivar M. Pilar, Citores Leire, Corell Jon, Rodríguez-Ezpeleta Naiara, Acuña JoséLuis, Molina-Ramírez Axayacatl, González-Gordillo J. Ignacio, Cózar Andrés, Martí Elisa, Cuesta JoséA., Agustí Susana, Fraile-Nuez Eugenio, Duarte Carlos M., Irigoien Xavier, Chust Guillem. Large-scale ocean connectivity and planktonic body size // *Nature communications*. 2018. 9, 1. 1–13.

- Villéger Sébastien, Mason Norman W. H., Mouillot David.* New multidimensional functional diversity indices for a multifaceted framework in functional ecology // *Ecology*. 2008. 89, 8. 2290–2301.
- Violle Cyrille, Enquist Brian J., McGill Brian J., Jiang Lin, Albert Cécile H., Hulshof Catherine, Jung Vincent, Messier Julie.* The return of the variance : intraspecific variability in community ecology // *Trends in Ecology & Evolution*. 2021/07/27 2012. 27, 4. 244–252.
- Violle Cyrille, Navas Marie-Laure, Vile Denis, Kazakou Elena, Fortunel Claire, Hummel Irène, Garnier Eric.* The Biogeochemical Cycles of Trace Metals in the Oceans // *OIKOS*. 2007. 116, 5. 882–892.
- Vorobev Alexey, Dupouy Marion, Carradec Quentin, Delmont Tom O, Annamalé Anita, Wincker Patrick, Pelletier Eric.* Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics // *Genome research*. 2020. 30, 4. 647–659.
- Waller D. M., Mudrak E. L., Rogers D. A.* Do metacommunity mass effects predict changes in species incidence and abundance? // *Ecography*. 2018. 41, 1. 11–23.
- Wang Ian J, Bradburd Gideon S.* Isolation by environment // *Molecular ecology*. 2014. 23, 23. 5649–5662.
- Ward Ben A., Cael B. B., Collins Sinead, Young C. Robert.* Selective constraints on global plankton dispersal // *Proceedings of the National Academy of Sciences*. 2021. 118, 10.
- Watts Phillip C, Lundholm Nina, Ribeiro Sofia, Ellegaard Marianne.* A century-long genetic record reveals that protist effective population sizes are comparable to those of macroscopic species // *Biology Letters*. 2013. 9, 6. 20130849.
- Weir Bruce S, Cockerham C Clark.* Estimating F-statistics for the analysis of population structure // *evolution*. 1984. 1358–1370.
- Weisberg S.* Yeo-Johnson power transformations // *Department of Applied Statistics, University of Minnesota*. 2001. 1. 2003.
- West Elizabeth J., Pitt Kylie A., Welsh David T., Koop Klaus, Rissik David.* Top-down and bottom-up influences of jellyfish on primary productivity and planktonic assemblages // *Limnology and Oceanography*. 2009. 54, 6. 2058–2071.
- White Crow, Selkoe Kimberly A, Watson James, Siegel David A, Zacherl Danielle C, Toonen Robert J.* Ocean currents help explain population genetic structure // *Proceedings of the Royal Society B : Biological Sciences*. 2010. 277, 1688. 1685–1694.
- White J. Wilson, Samhouri Jameal F.* Oceanographic coupling across three trophic levels shapes source–sink dynamics in marine metacommunities // *Oikos*. 2011. 120, 8. 1151–1164.
- Whitham Thomas G., Bailey Joseph K., Schweitzer Jennifer A., Shuster Stephen M., Bangert Randy K., LeRoy Carri J., Lonsdorf Eric V., Allan Gery J., DiFazio Stephen P., Potts Brad M., Fischer Dylan G., Gehring Catherine A., Lindroth Richard L., Marks Jane C., Hart Stephen C., Wimp Gina M., Wooley Stuart C.* A framework for community and ecosystem genetics : from genes to ecosystems // *Nature Reviews Genetics*. 2006. 7, 7. 510–523.

- Whitt Christopher, Pearlman Jay, Polagye Brian, Caimi Frank, Muller-Karger Frank, Copping Andrea, Spence Heather, Madhusudhana Shyam, Kirkwood William, Grosjean Ludovic, Fiaz Bilal Muhammad, Singh Satinder, Singh Sikandra, Manalang Dana, Gupta Ananya Sen, Maguer Alain, Buck Justin J. H., Marouchos Andreas, Atmanand Malayath Aravindakshan, Venkatesan Ramasamy, Narayanaswamy Vedachalam, Testor Pierre, Douglas Elizabeth, Halleux Sebastien de, Khalsa Siri Jodha. Future Vision for Autonomous Ocean Observations // Frontiers in Marine Science. 2020. 7. 697.
- Whittaker Robert Harding. Vegetation of the Siskiyou mountains, Oregon and California // Ecological monographs. 1960. 30, 3. 279–338.
- Wickham Hadley. ggplot2 : Elegant Graphics for Data Analysis. 2016.
- Wickham Hadley, Averick Mara, Bryan Jennifer, Chang Winston, D'Agostino McGowan Lucy, François Romain, Golemund Garrett, Hayes Alex, Henry Lionel, Hester Jim, Kuhn Max, Pedersen Thomas Lin, Miller Evan, Bache Stephan Milton, Müller Kirill, Ooms Jeroen, Robinson David, Seidel Dana Paige, Spinu Vitalie, Takahashi Kohske, Vaughan Davis, Wilke Claus, Woo Kara, Yutani Hiroaki. Welcome to the tidyverse // Journal of Open Source Software. 2019. 4, 43. 1686.
- Wilkinson Mark D., Dumontier Michel, Aalbersberg IJsbrand Jan, Appleton Gabrielle, Arton Myles, Baak Arie, Blomberg Niklas, Boiten Jan-Willem, Silva Santos Luiz Bonino da, Bourne Philip E., Bouwman Jildau, Brookes Anthony J., Clark Tim, Crosas Mercè, Dillo Ingrid, Dumon Olivier, Edmunds Scott, Evelo Chris T., Finkers Richard, Gonzalez-Beltran Alejandra, Gray Alasdair J. G., Groth Paul, Goble Carole, Grethe Jeffrey S., Heringa Jaap, Hoen Peter A. C. 't, Hooft Rob, Kuhn Tobias, Kok Ruben, Kok Joost, Lusher Scott J., Martone Maryann E., Mons Albert, Packer Abel L., Persson Bengt, Rocca-Serra Philippe, Roos Marco, Schaik Rene van, Sansone Susanna-Assunta, Schultes Erik, Sengstag Thierry, Slater Ted, Strawn George, Swertz Morris A., Thompson Mark, Lei Johan van der, Mulligen Erik van, Velterop Jan, Waagmeester Andra, Wittenburg Peter, Wolstencroft Katherine, Zhao Jun, Mons Barend. The FAIR Guiding Principles for scientific data management and stewardship // Scientific Data. 2016. 3, 1. 160018.
- Wold Svante, Trygg Johan, Berglund Anders, Antti Henrik. Some recent developments in PLS modeling // Chemometrics and Intelligent Laboratory Systems. 2001. 58, 2. 131–150. PLS Methods.
- Woodward Guy, Ebenman Bo, Emmerson Mark, Montoya Jose M., Olesen Jens M., Valido Alfredo, Warren Philip H. Body size in ecological networks // Trends in Ecology & Evolution. 2005. 20, 7. 402–409.
- Worden Alexandra Z, Janouskovec Jan, McRose Darcy, Engman Augustin, Welsh Rory M, Malfatti Stephanie, Tringe Susannah G, Keeling Patrick J. Global distribution of a wild alga revealed by targeted metagenomics // Current Biology. 2012. 22, 17. R675–R677.
- Worm Boris, Barbier Edward B., Beaumont Nicola, Duffy J. Emmett, Folke Carl, Halpern Benjamin S., Jackson Jeremy B. C., Lotze Heike K., Micheli Fiorenza, Palumbi Stephen R., Sala Enric, Selkoe Kimberley A., Stachowicz John J., Watson Reg. Impacts of Biodiversity Loss on Ocean Ecosystem Services // Science. 2006. 314, 5800. 787–790.
- Wright Sewall. Isolation by distance // Genetics. 1943. 28, 2. 114.

- Wright Sewall*. Evolution and the genetics of populations : Vol. 2. The theory of gene frequencies. 1969.
- Wu Dongying, Hugenholtz Philip, Mavromatis Konstantinos, Pukall Rüdiger, Dalin Eileen, Ivanova Natalia N., Kunin Victor, Goodwin Lynne, Wu Martin, Tindall Brian J., Hooper Sean D., Pati Amrita, Lykidis Athanasios, Spring Stefan, Anderson Iain J., D'haeseleer Patrick, Zemla Adam, Singer Mitchell, Lapidus Alla, Nolan Matt, Copeland Alex, Han Cliff, Chen Feng, Cheng Jan-Fang, Lucas Susan, Kerfeld Cheryl, Lang Elke, Gronow Sabine, Chain Patrick, Bruce David, Rubin Edward M., Kyrpides Nikos C., Klenk Hans-Peter, Eisen Jonathan A.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea // *Nature*. 2009. 462, 7276. 1056–1060.
- Yeo In-Kwon, Johnson Richard A.* A new family of power transformations to improve normality or symmetry // *Biometrika*. 2000. 87, 4. 954–959.
- Zhang Zhengguang, Wang Wei, Qiu Bo.* Oceanic mass transport by mesoscale eddies // *Science*. 2014. 345, 6194. 322–324.
- Zingone Adriana, Chrétiennot-Dinet Marie-Josèphe, Lange Martin, Medlin Linda.* Morphological and genetic characterization of *Phaeocystis cordata* and *P. jahnii* (Prymnesiophyceae), two new species from the Mediterranean Sea // *Journal of Phycology*. 1999. 35, 6. 1322–1337.
- Zipkin Elise F, Zylstra Erin R, Wright Alexander D, Saunders Sarah P, Finley Andrew O, Dietze Michael C, Itter Malcolm S, Tingley Morgan W.* Addressing data integration challenges to link ecological processes across scales // *Frontiers in Ecology and the Environment*. 2021. 19, 1. 30–38.
- Zylinski Sarah, Johnsen Sönke.* Mesopelagic Cephalopods Switch between Transparency and Pigmentation to Optimize Camouflage in the Deep // *Current Biology*. 2011. 21, 22. 1937–1941.
- d'Alcalà M. Ribera, Conversano F., Corato F., Licandro P., Mangoni O., Marino D., Mazzocchi M. G., Modigh M., Montresor M., Nardella M., others .* Seasonal patterns in plankton communities in a pluriannual time series at a coastal Mediterranean site (Gulf of Naples) : an attempt to discern recurrences and trends // *Scientia marina*. 2004. 68, S1. 65–83.
- d'Ovidio Francesco, De Monte Silvia, Alvain Séverine, Dandonneau Yves, Lévy Marina.* Fluid dynamical niches of phytoplankton types // *Proceedings of the National Academy of Sciences*. 2010. 107, 43. 18366–18370.
- d'Ovidio Francesco, De Monte Silvia, Della Penna Alice, Cotté Cedric, Guinet Christophe.* Ecological implications of eddy retention in the open ocean : a Lagrangian approach // *Journal of Physics A : Mathematical and Theoretical*. 2013. 46, 25. 254023.
- de Groot G. Arjen, Geisen Stefan, Wubs E.R. Jasper, Meulenbroek Liz, Laros Ivo, Snoek L. Basten, Lammertsma Dennis R., Hansen Lars H., Slim Pieter A.* The aerobiome uncovered : Multi-marker metabarcoding reveals potential drivers of turn-over in the full microbial community in the air // *Environment International*. 2021. 154. 106551.
- del Campo Javier, Sieracki Michael E., Molestina Robert, Keeling Patrick, Massana Ramon, Ruiz-Trillo Iñaki.* The others : our biased perspective of eukaryotic genomes // *Trends in ecology & evolution*. 2014. 29, 5. 252–259.

Contributions comme co-auteur

A.1 Godet et al., Deep Sea Reseach II, 2020

Matching zooplankton abundance and environment in the South Indian Ocean and Southern Ocean

Claire Godet^a, Marine Robuchon^{a,b,c}, Boris Leroy^a, Cédric Cotté^d, Alberto Baudena^{d,e}, Ophélie Da Silva^{e,f}, Salomé Fabri-Ruiz^e, Claire Lo Monaco^d, Sara Sergi^d, Philippe Koubbi^{g,h}

a - Laboratoire de Biologie des Organismes et écosystèmes Aquatiques (BOREA), Muséum National d'Histoire Naturelle, CNRS, IRD, Sorbonne Université, Université de Caen Normandie, Université des Antilles, CP 26, 57 Rue Cuvier, 75005, Paris, France

b - Centre d'Ecologie et des Sciences de la Conservation (CESCO), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, CP 135, 57 Rue Cuvier, 75005, Paris, France

c - Joint Research Centre (JRC) of the European Commission, Directorate for Sustainable Resources, Ispra, Italy

d - Sorbonne Université UMR 7159 CNRS-IRD-MNHN, LOCEAN-IPSL, 75005, Paris, France

e - Laboratoire D'Océanographie de Villefranche-sur-Mer, UMR 7093 - CNRS/UPMC, 181 Chemin Du Lazaret, 06230, Villefranche-sur-Mer Cedex, France

f - Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, CP 50, 57

Rue Cuvier, 75005, Paris, France

g - UFR918 Terre, Environnement, Biodiversité. Sorbonne Université, 4 Place Jussieu, 75252, PARIS Cedex 05, France

h - IFREMER Centre Manche Mer Du Nord - 150, Quai Gambetta, 62200, Boulogne-sur-Mer, France

Résumé.

Distinguishing regions based on the geographic distribution of both abiotic factors and living organisms is an old but still actual central issue for biogeographers. In the Southern Ocean, the few existing regionalization studies have been carried out either at very large scales or on the relatively small region around the Sub-Antarctic islands of Kerguelen and the Crozet archipelagos. However, regionalization studies at meso-scales (100–300 km) covering the Indian part of the Southern Ocean and adjacent South Indian Ocean are scarce. These waters, ranging from the Subtropical to the polar region, are home to large populations of well-studied top predators that depend on the biomass of less known mid-trophic level species such as zooplankton. To fill those gaps, our study aims at conducting bioregional analyses of this transition area at the meso-scale based on the distribution of abiotic factors and chlorophyll-a, and to investigate how the abundance of zooplankton varies across the bioregions identified. To that end, we first characterized epipelagic bioregions 30°S in the South Indian Ocean to 65°S in the Southern Ocean and from 40° to 85°E including the islands of Crozet, Kerguelen, Saint-Paul and New Amsterdam. We then determined whether these bioregions correspond to variations in the abundance of zooplankton collected by a Continuous Plankton Recorder. Finally, we analyzed which environmental parameters influence zooplankton abundance. Our analyses evidenced six regions, providing a synthetic overview of a contrasting environment. The spatial variability of zooplankton abundance was explained by most of the environmental variables used in the bioregionalisation and, to a lesser extent, by the bioregions. Copepods are abundant in the colder and physically-energetic regions associated with the Antarctic Circumpolar Current (ACC). Limacina and euphausiids are both abundant in regions characterized by a high concentration of chlorophyll-a, although euphausiids are also abundant in the subtropical region. This work represents a crucial step forward in the integration of living organism distribution in the regionalization of the Indian part of Southern Ocean and adjacent South Indian Ocean. This can, ultimately contribute to the optimization of marine conservation strategies.

Mots clés.

Bioregionalization, Southern Ocean, Indian Ocean, Pelagic ecosystem, Zooplankton, Continuous plankton recorder

Article accessible en ligne.

<https://www.sciencedirect.com/science/article/pii/S0967063720301345>

Contribution.

Dans cette étude, j'ai participé à la mise en place de la méthodologie pour la régionalisation.

A.2 Richter et al., BioRxiv, 2020

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

Daniel J. Richter^{1,2*}, Romain Watteaux^{3,4*}, Thomas Vannier^{5,6,7*}, Jade Leconte^{6,7}, Paul Frémont^{6,7}, Gabriel Reygondeau^{8,9}, Nicolas Maillet¹⁰, Nicolas Henry^{2,7}, Gaëtan Benoit¹¹, Ophélie Da Silva¹², Tom O. Delmont^{6,7}, Antonio Fernández-Guerra^{13,14,15}, Samir Suweis¹⁶, Romain Narci¹⁷, Cédric Berney^{2,7}, Damien Eveillard^{7,18}, Frederick Gavory¹⁹, Lionel Guidi^{20,21}, Karine Labadie¹⁹, Eric Mahieu¹⁹ 7, Julie Poulain⁶, Sarah Romac^{2,7}, Simon Roux²², Céline Dimier^{2,23}, Stefanie Kandels^{24,25}, Marc Picheral^{26,27} 8, Sarah Searson^{26,27}, Tara Oceans Coordinators, Stéphane Pesant^{28,29}, Jean-Marc Aury¹⁹ 9, Jennifer R. Brum^{22,30}, Claire Lemaitre¹¹, Eric Pelletier^{6,7}, Peer Bork^{24,31,32}, Shinichi Sunagawa^{24,33} 10, Fabien Lombard^{12,34}, Lee Karp-Boss³⁵, Chris Bowler^{7,23}, Matthew B. Sullivan^{22,36,37,38,39}, Eric Karsenti^{7,23,25} 11, Mahendra Mariadassou¹⁷, Ian Probert^{2,7}, Pierre Peterlongo¹¹, Patrick Wincker^{6,7} 12, Colomban de Vargas^{2,7**}, Maurizio Ribera d'Alcalà^{3**}, Daniele Iudicone^{3**} \$, Olivier Jaillon^{6,7**} \$

* and \$: equal contributions

** : corresponding authors

Tara Oceans Coordinators :

Silvia G. Acinas⁴⁰, Peer Bork^{24,31,32}, Emmanuel Boss³⁵, Chris Bowler^{7,23}, Guy Cochrane⁴¹ 19, Colomban de Vargas^{2,7}, Gabriel Gorsky¹², Nigel Grimsley^{42,43}, Lionel Guidi^{20,21}, Pascal Hingamp⁴⁴, Daniele Iudicone³ 20, Olivier Jaillon^{6,7}, Stefanie Kandels^{24,25}, Lee Karp-Boss³⁵, Eric Karsenti^{7,23,25}, Fabrice Not^{2,7} 21, Hiroyuki Ogata⁴⁵, Stéphane Pesant^{28,29}, Jeroen Raes^{46,47}, Christian Sardet^{20,48}, Mike Sieracki^{49,50} 22, Sabrina Speich^{51,52}, Lars Stemmann²⁰, Matthew B. Sullivan^{22,36,37,38,39}, Shinichi Sunagawa^{24,33} 23, Patrick Wincker^{6,7}

1 Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain

2 Sorbonne Université, CNRS, Station Biologique de Roscoff, UMR 7144, ECOMAP, 29680 Roscoff, France

3 Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.

4 CEA, DAM, DIF, F-91297 Arpajon Cedex, France

5 Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110, 13288, Mar-

seille, France

6 Génomique Métabolique, Genoscope, Institut de Biologie François Jacob, Commissariat à l'Énergie Atomique, CNRS, Université Evry, Université Paris-Saclay, Evry, France

7 Research Federation for the study of Global Ocean systems ecology and evolution, FR2022/Tara GOsee, Paris, France

8 Changing Ocean Research Unit, Institute for the Oceans and Fisheries, University of British Columbia. Aquatic Ecosystems Research Lab. 2202 Main Mall. Vancouver, BC V6T 1Z4. Canada.

9 Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA.

10 Institut Pasteur - Bioinformatics and Biostatistics Hub - C3BI, USR 3756 IP CNRS - Paris, France.

11 INRIA/IRISA, Genscale team, UMR6074 IRISA CNRS/INRIA/Université de Rennes 1, Campus de Beaulieu, 35042, Rennes, France.

12 Sorbonne Universités, CNRS, Laboratoire d'Océanographie de Villefranche, LOV, F-06230 Villefranche-sur-Mer, France.

13 Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark

14 MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany

15 Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany

16 Dipartimento di Fisica e Astronomia 'G. Galilei' & CNISM, INFN, Università di Padova, Via Marzolo 8, 35131 Padova, Italy.

17 MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France

18 Université de Nantes, Centrale Nantes, CNRS, LS2N, F-44000 Nantes, France

19 Genoscope, Institut de biologie François-Jacob, Commissariat à l'Énergie Atomique (CEA), Université Paris-Saclay, Evry, France

20 Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France.

21 Department of Oceanography, University of Hawaii, Honolulu, Hawaii 96822, USA.

22 Department of Microbiology, The Ohio State University, Columbus, OH 43214, USA

23 Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France.

24 Structural and Computational Biology, European Molecular Biology Laboratory,

Meyerhofstr. 1, 69117 Heidelberg, Germany.

25 Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg Germany.

26 Sorbonne Universités, UPMC Univ Paris 06, UMR 7093 LOV, F-75005, Paris, France.

27 CNRS, UMR 7093 LOV, F-75005, Paris, France.

28 MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany.

29 PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany.

30 Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA, 70808, USA

31 Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany.

32 Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany.

33 Institute of Microbiology, Department of Biology, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

34 Institut Universitaire de France (IUF), Paris, France 35 School of Marine Sciences, University of Maine, Orono, Maine 04469, USA.

36 EMERGE Biology Integration Institute, The Ohio State University, Columbus, Ohio 43210, USA

37 Center of Microbiome Science, The Ohio State University, Columbus, Ohio 43210, USA

38 The Interdisciplinary Biophysics Graduate Program, The Ohio State University, Columbus, Ohio 43210, USA

39 Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus OH 43214 USA

40 Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM), CSIC, Barcelona, Spain.

41 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome 83 Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

42 CNRS, UMR 7232, BIOM, Avenue Pierre Fabre, 66650 Banyuls-sur-Mer, France.

43 Sorbonne Universités Paris 06, OOB UPMC, Avenue Pierre Fabre, 66650 Banyuls-sur-Mer, France.

44 Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UM 110, 13288, Marseille, France

45 Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan.

46 Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.

47 VIB Center for Microbiology, Herestraat 49, 3000 Leuven, Belgium.

48 CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France.

49 National Science Foundation, Arlington, VA 22230, USA. 50 Bigelow Laboratory for Ocean Sciences East Boothbay, ME, USA.

51 Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France.

52 Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue 95 Lhomond, 75231 Paris Cedex 05, France.

Résumé.

Biogeographical studies have traditionally focused on readily visible organisms, but recent technological advances are enabling analyses of the large-scale distribution of microscopic organisms, whose biogeographical patterns have long been debated. Here we assessed the global structure of plankton geography and its relation to the biological, chemical and physical context of the ocean (the ‘seascape’) by analyzing metagenomes of plankton communities sampled across oceans during the Tara Oceans expedition, in light of environmental data and ocean current transport. Using a consistent approach across organismal sizes that provides unprecedented resolution to measure changes in genomic composition between communities, we report a pan-ocean, size-dependent plankton biogeography overlying regional heterogeneity. We found robust evidence for a basin-scale impact of transport by ocean currents on plankton biogeography, and on a characteristic timescale of community dynamics going beyond simple seasonality or life history transitions of plankton.

Article accessible en ligne.

<https://www.biorxiv.org/content/10.1101/867739v2.abstract>

Contribution.

Dans cette étude, j’ai analysé les données d’imagerie quantitative et participé à la rédaction du paragraphe *Imaging methods*

A.3 Danko et al., Cell, 2020

A global metagenomic map of urban microbiomes and antimicrobial resistance

David Danko^{1,2,68}, Daniela Bezdán^{1,2,38,39,68}, Evan E. Afshin¹², Sofia Ahsanuddin³, Chandrima Bhattacharya^{1,2}, Daniel J. Butler^{1,2}, Kern Rei Chng⁴, Daisy Donnellan^{1,2}, Jochen Hecht⁵, Katelyn Jackson^{1,2}, Katerina Kuchin^{1,2}, Mikhail Karasikov^{6,64,65}, Abigail Lyons^{1,2}, Lauren Mak^{1,2}, Dmitry Meleshko^{1,2}, Harun Mustafa^{6,64,65}, Beth Mutai^{5,8}, Russell Y. Neches⁷, Amanda Ng⁴, Olga Nikolayeva⁹, Tatyana Nikolayeva⁹, Eileen Png⁴, Krista A. Ryon^{1,2}, Jorge L. Sanchez^{1,2}, Heba Shaaban^{1,2}, Maria A. Sierra^{1,2}, Dominique Thomas^{1,2}, Ben Young^{1,2}, Omar O. Abudayyeh¹⁰, Josue Alicea^{1,2}, Malay Bhattacharyya^{11,12}, Ran Blehman¹³, Eduardo Castro-Nallar¹⁴, Ana M. Cañas^{1,2}, Aspasia D. Chatziefthimiou^{1,2}, Robert W. Crawford¹⁵, Francesca De Filippis^{16,17}, Youping Deng¹⁸, Christelle Desnues¹⁹, Emmanuel Dias-Neto²⁰, Marius Dybwad²¹, Eran Elhaik²², Danilo Ercolini^{16,17}, Alina Frolova^{23,62}, Dennis Gankin¹⁰, Jonathan S. Gootenberg¹⁰, Alexandra B. Graf²⁴, David C. Green²⁵, Iman Hajirasouliha^{1,2}, Jaden J.A. Hastings^{1,2}, Mark Hernandez²⁶, Gregorio Iraola^{27,28,29}, Soojin Jang³⁰, Andre Kahles^{6,62,63}, Frank J. Kelly²⁵, Kaymisha Knights^{1,2}, Nikos C. Kyrpides⁷, Paweł P. Łabaj^{46,59,60}, Patrick K.H. Lee³¹, Marcus H.Y. Leung³¹, Per O. Ljungdahl³², Gabriella Mason-Buck²⁵, Ken McGrath³³, Cem Meydan^{1,2}, Emmanuel F. Mongodin³⁴, Milton Ozorio Moraes³⁵, Niranjana Nagarajan⁴, Marina Nieto-Caballero²⁶, Houtan Noushmehr³⁶, Manuela Oliveira³⁷, Stephan Ossowski^{5,38,39}, Olayinka O. Osuolale⁴⁰, Orhan Özcan⁴⁴, David Paez-Espino⁷, Nicolás Rascovan⁴¹, Hugues Richard^{42;43}, Gunnar Rättsch^{6,64,65}, Lynn M. Schriml³⁴, Torsten Semmler⁴³, Osman U. Sezerman⁴⁴, Leming Shi^{45,46}, Tielu Shi⁴⁷, Rania Siam⁶⁶, Le Huu Song^{48,49}, Haruo Suzuki⁵⁰, Denise Syndercombe Court²⁵, Scott W. Tighe⁵¹, Xinzhao Tong³¹, Klas I. Udekwu^{32,61}, Juan A. Ugalde^{52,63}, Brandon Valentine¹², Dimitar I. Vassilev⁵³, Elena M. Vayndorf⁵⁴, Thirumalaisamy P. Velavan^{55,56}, Jun Wu⁴⁷, María M. Zambrano⁵⁷, Jifeng Zhu^{1,2}, Sibozhu^{46,58}, Christopher E. Mason^{1,2,67,69}

The International MetaSUB Consortium

1 - Weill Cornell Medicine, New York, NY, USA

2 - The Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, New York, NY, USA

3 - Icahn School of Medicine at Mount Sinai, New York, NY, USA

- 4 - Genome Institute of Singapore, A*STAR, Singapore, Singapore
- 5 - Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
- 6 - ETH Zurich, Department of Computer Science, Biomedical Informatics Group, Zurich, Switzerland
- 7 - Department of Energy, Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
- 8 - Kenya Medical Research Institute – Kisumu, Kisumu, Kenya
- 9 - Independent Researcher, Zurich, Switzerland
- 10 - Massachusetts Institute of Technology, McGovern Institute for Brain Research, Cambridge, MA, USA
- 11 - Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
- 12 - Centre for Artificial Intelligence and Machine Learning, Indian Statistical Institute, Kolkata, India
- 13 - University of Minnesota, Minneapolis, MN, USA
- 14 - Universidad Andres Bello, Center for Bioinformatics and Integrative Biology, Facultad de Ciencias de la Vida, Santiago, Chile
- 15 - California State University, Sacramento, Sacramento, CA, USA
- 16 - Department of Agricultural Sciences, Division of Microbiology, University of Naples Federico II, Naples, Italy
- 17 - Task Force on Microbiome Studies, University of Naples Federico II, Naples, Italy
- 18 - University of Hawaii John A. Burns School of Medicine, Honolulu, HI, USA
- 19 - Aix-Marseille Université, Mediterranean Institute of Oceanology, Université de Toulon, CNRS, IRD, UM 110, Marseille, France
- 20 - Medical Genomics group, A.C.Camargo Cancer Center, São Paulo – SP, Brazil
- 21 - Norwegian Defence Research Establishment FFI, Kjeller, Norway
- 22 - Department of Biology, Lund University, Lund, Sweden
- 23 - Institute of Molecular Biology and Genetics of National Academy of Sciences of Ukraine, Kyiv, Ukraine
- 24 - University of Applied Sciences Vienna, Vienna, Austria
- 25 - Department of Analytical, Environmental and Forensic Sciences, King's College London, London, UK
- 26 - University of Colorado at Boulder, Boulder, CO, USA
- 27 - Microbial Genomics Laboratory, Institut Pasteur de Montevideo, Montevideo, Uruguay
- 28 - Center for Integrative Biology, Universidad Mayor, Santiago de Chile, Santiago,

Chile

29 - Wellcome Sanger Institute, Hinxton, UK

30 - Institut Pasteur Korea, Seoul, South Korea

31 - School of Energy and Environment, City University of Hong Kong, Hong Kong SAR, China

32 - Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Stockholm, Sweden

33 - Microba, 388 Queen St, Brisbane City, QLD 4000, Australia

34 - University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD, USA

35 - Fundação Oswaldo Cruz, Rio de Janeiro – RJ, Brazil

36 - University of São Paulo, Ribeirão Preto Medical School, Ribeirão Preto – SP, Brazil

37 - Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal

38 - Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany

39 - NGS Competence Center Tübingen (NCCT), University of Tübingen, Tübingen, Germany

40 - Applied Environmental Metagenomics and Infectious Diseases Research (AEMIDR), Department of Biological Sciences, Elizade University, Ilara-Mokin, Nigeria

41 - Microbial Paleogenomics Unit, Institut Pasteur, CNRS UMR2000, Paris 75015, France

42 - Sorbonne University, Faculty of Science, Institute of Biology Paris-Seine, Laboratory of Computational and Quantitative Biology, Paris, France

43 - Robert Koch Institute, Berlin, Germany

44 - Acibadem Mehmet Ali Aydınlar University, Istanbul, Turkey

45 - Center for Pharmacogenomics, School of Life Sciences and Shanghai Cancer Center, Fudan University, Shanghai, China

46 - State Key Laboratory of Genetic Engineering (SKLGE) and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Human Phenome Institute, Fudan University, Shanghai, China

47 - The Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai, China

48 - 108 Military Central Hospital, Hanoi, Vietnam

49 - Vietnamese-German Center for Medical Research (VG-CARE), Hanoi, Vietnam

- 50 - Keio University, Tokyo, Japan
- 51 - University of Vermont, Burlington, VT, USA
- 52 - Millennium Initiative for Collaborative Research on Bacterial Resistance, Santiago, Chile
- 53 - Faculty of Mathematics and Informatics, Sofia University “St. Kliment Ohridski,” Sofia, Bulgaria
- 54 - Institute of Arctic Biology, University of Alaska, Fairbanks, Fairbanks, AK, USA
- 55 - Institute of Tropical Medicine, Univeristätsklinikum Tübingen, Tübingen, Germany
- 56 - Faculty of Medicine, Duy Tan University, Da Nang, Vietnam
- 57 - Corporación Corpogen-Research Center, Bogotá, Colombia
- 58 - Department of Epidemiology, School of Public Health, Fudan University, Shanghai, China
- 59 - Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland
- 60 - Boku University Vienna, Vienna, Austria
- 61 - SciLife EVP, Department of Aquatic Sciences Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden
- 62 - Kyiv Academic University, Kyiv, Ukraine
- 63 - C+, Research Center in Technologies for Society, School of Engineering, Universidad del Desarrollo, Santiago, Chile
- 64 - University Hospital Zurich, Biomedical Informatics Research, Zurich, Switzerland
- 65 - Swiss Institute of Bioinformatics, Lausanne, Switzerland
- 66 - University of Medicine and Health Sciences, St. Kitts, West Indies and American University in Cairo, Cairo, Egypt
- 67 - The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA

Résumé.

We present a global atlas of 4,728 metagenomic samples from mass-transit systems in 60 cities over 3 years, representing the first systematic, worldwide catalog of the urban microbial ecosystem. This atlas provides an annotated, geospatial profile of microbial strains, functional characteristics, antimicrobial resistance (AMR) markers, and genetic elements, including 10,928 viruses, 1,302 bacteria, 2 archaea, and 838,532 CRISPR arrays not found in reference databases. We identified 4,246 known species of urban microorganisms and a consistent set of 31 species found in 97% of samples that were distinct from human commensal organisms. Profiles of AMR genes varied widely in type and density across cities. Cities showed distinct microbial taxonomic signatures that were

driven by climate and geographic differences. These results constitute a high-resolution global metagenomic atlas that enables discovery of organisms and genes, highlights potential public health and forensic applications, and provides a culture-independent view of AMR burden in cities.

Mots clés.

built Environment, metagenome, global health, antimicrobial resistance, microbiome, NGS, shotgun sequencing, de novo assembly, BGC, AMR

Article accessible en ligne.

<https://www.sciencedirect.com/science/article/pii/S0092867421005857>

Contribution.

Dans cette étude, j'ai participé à la collecte des données.

**Supplementary Information - Are we ready
for population metagenomics ?**

B.1 Methods for preliminary investigations

Preliminary investigations were conducted to determine if there were protist species known as present in the Mediterranean Sea and for which a reference (genome or transcriptome) would be available. In addition, we assumed that if a species was present in the metagenomic data, it should be detectable in the metabarcoding data. Therefore, an exploration of the abundances of the eukaryotic metabarcodes in the Tara Oceans Mediterranean samples was carried out.

B.1.1 Step 1. Bibliographic study of ecologically relevant protist species in the Mediterranean Sea

Evidence for the presence of protist species in the Mediterranean were searched in the scientific literature. Only planktonic (i.e. benthic removed) protists already observed in the Mediterranean were targeted. Available references were searched in the NCBI database. At the end of this first step, about twenty species supposed to be abundantly present in the Mediterranean Sea and having at least one reference genome or transcriptome available had been selected. In the next step, the abundance of their metabarcodes was estimated.

B.1.2 Step 2. Abundance analysis of the V9 eukaryotic metabarcodes in the TO Mediterranean samples

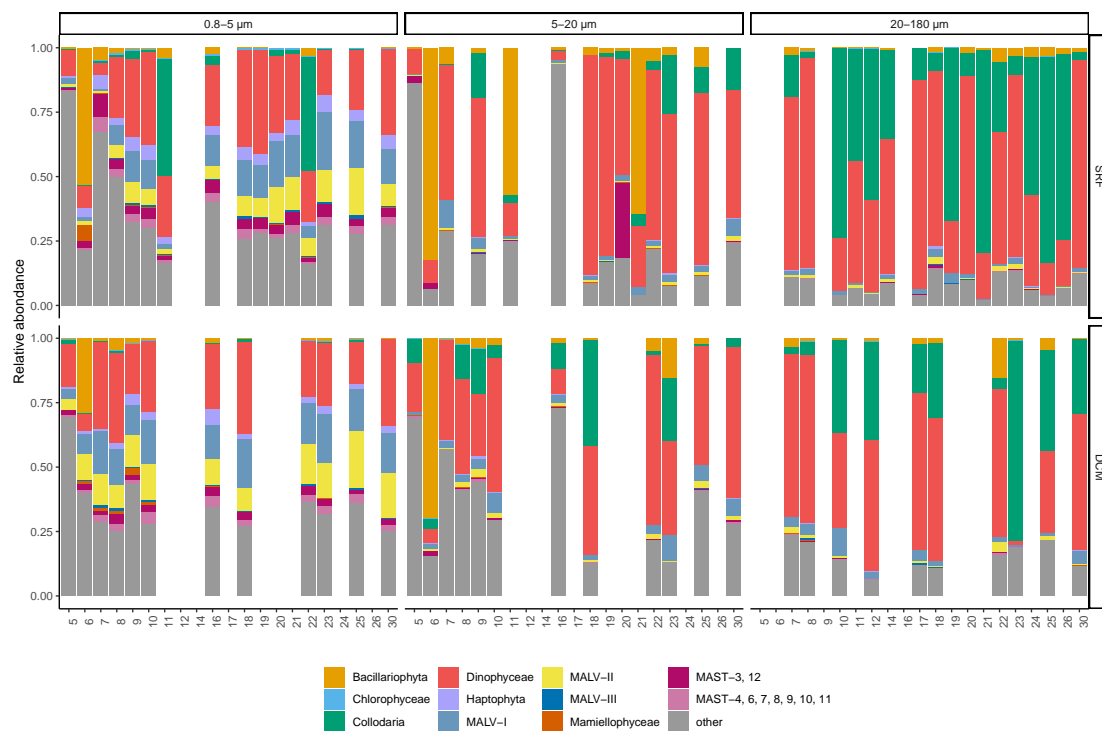


FIGURE B.1 – Distribution of the metabarcode abundances in the Mediterranean Sea (data from Ibarbalz et al., 2019). Exploration of the abundances of the eukaryotic metabarcodes in the Tara Ocean Mediterranean samples was carried out in order to determine if some species or taxa could be found in the metagenomic samples

B.1.3 Step 3. Availability of protist reference genome(s) or transcriptome(s)

Species	<i>Bathycoccus prasinus</i>	<i>Pelagomonas calceolata</i>	<i>Phaeocystis coralata</i>	<i>Rhaphidozoum sp.</i>	<i>Collozoum sp.</i>
Strain	RCC1105	RCC969	RCC1383	Personal communication	Personal communication
Reference data	Genome assembly (Moreau et al., 2012) (GCA_002220235.1) 15 Mb	Transcriptome assembly (MMETSP; Keeling et al., 2014; Johnson et al., 2019) 21 Mb	Transcriptome assembly (MMETSP; Keeling et al., 2014; Johnson et al., 2019) 9.4 Mb	Personal communication	Personal communication
Taxonomy (based on (Adl et al., 2019))	Chloroplastida Chlorophyta Mamiellophyceae Bathycoccus	Stramenopiles Gyrista Ochrophyta Diatomista Pelagophyceae Pelagomonadales Pelagomonas	Haptista Haptophyta Prymnesiophyceae Phaeocystales Phaeocystis	Rhizaria Retaria Radiolaria Polycystinea Collodaria Rhaphidozoum	Rhizaria Retaria Radiolaria Polycystinea Collodaria Collozoum
Estimated size from morphological studies	1-2 μm (Eikrem and Thronsdén, 1990; Moreau et al., 2012)	1.3 x 3 μm (Andersen et al., 1993)	3-3.5 μm long; 3-4 μm wide (Medlin and Zigone, 2007; Zigone et al., 1999)		

TABLE B.1 – Information about protists targeted in the study

References

Adl, S.M., Bass, D., Lane, C.E., Lukeš, J., Schoch, C.L., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A.A., Hoppenrath, M., James, T.Y., Karnkowska, A., Karpov, S., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, D.J., Lara, E., Le Gall, L., Lynn, D.H., Mann, D.G., Massana, R., Mitchell, E.A., Morrow, C., Park, J.S., Pawlowski, J.W., Powell, M.J., Richter, D.J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F.W., Torruella, G., Youssef, N., Zlatogursky, V. and Zhang, Q. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *J. Eukaryot. Microbiol.*, 66 : 4-119. <https://doi.org/10.1111/jeu.12691>

Andersen, R. A., Saunders, G. W., Paskind, M. P., Sexton, J. P. (1993). Ultrastructure and 18S rRNA gene sequence for *Pelagomonas calceolata* gen. et sp. nov. and the description of a new algal class, the pelagophyceae classis nov. 1. *Journal of Phycology*, 29, 5. 701–715.

Eikrem, W., Thronsdén, J. (1990). The ultrastructure of *Bathycoccus* gen. nov. and *B. prasinus* sp. nov., a non-motile picoplanktonic alga (Chlorophyta, Prasinophyceae) from the Mediterranean and Atlantic. *Phycologia*, 29, 3. 344–350.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., Coelho, L. P., Endo, H., Gasol, J.M., Gregory, A.C., et al. (2019). Global trends in marine plankton diversity across kingdoms of life. *Cell*, 179(5), 1084-1097.

Johnson, L. K., Alexander, H., Brown, C. T. (2019). Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience*, 8, 4. giy158.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Arm-

brust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., *et al.* (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) : illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS biology*, 12, 6. e1001889.

Medlin L., Zingone A. (2007) A taxonomic review of the genus *Phaeocystis*. In : van Leeuwe M.A., Stefels J., Belviso S., Lancelot C., Verity P.G., Gieskes W.W.C. (eds) *Phaeocystis*, major link in the biogeochemical cycling of climate-relevant elements. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6214-8_2

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., Van Bel, M., Poulain, J., Katinka, M., Hohmann-Marriott, M. F., *et al.* (2012). Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome biology*, 13, 8. 1–16.

Zingone, A., Chrétiennot-Dinet, M-J, Lange, M., Medlin, L. (1999). Morphological and genetic characterization of *Phaeocystis cordata* and *P. jahnii* (Prymnesiophyceae), two new species from the Mediterranean Sea // *Journal of Phycology*, 35, 6. 1322–1337.

B.2 Metagenomics analysis

STATION	FRACTION	SAMPLE	#READS	SIZE file.gz (Go)
TARA_004	0.8-5	ERR868369	181437081	14
TARA_005	0.8-5	ERR868497	215288687	17
TARA_006	0.8-5	ERR868431	193698390	15
TARA_007	0.8-5	ERR315802	163300930	19-20
TARA_009	0.8-5	ERR868407	185406539	16
TARA_016	0.8-5	ERR868412	163007547	13
TARA_018	0.8-5	ERR868393	168978860	14
TARA_020	0.8-5	ERR868504	167093109	13-14
TARA_022	0.8-5	ERR868403	205419622	17
TARA_023	0.8-5	ERR538173	198522116	17
TARA_024	0.8-5	ERR868456	160573608	13
TARA_025	0.8-5	ERR868356	185404639	16
TARA_026	0.8-5	ERR868402	190669700	16
TARA_030	0.8-5	ERR538186	215014987	18
TARA_004	5-20	ERR1726542	186521704	15
TARA_018	5-20	ERR1726555	182836931	16
TARA_022	5-20	ERR1726540	152747480	12
TARA_023	5-20	ERR538184	204887834	17
TARA_025	5-20	ERR1726722	175924502	15
TARA_030	5-20	ERR538178	155356395	13
TARA_007	20-180	ERR538183	223830672	18
TARA_011	20-180	ERR1726791	155977419	13
TARA_014	20-180	ERR1726870	158312226	13
TARA_018	20-180	ERR1726956	170491399	15
TARA_020	20-180	ERR1726660	164881714	13
TARA_022	20-180	ERR1726546	81559499	6.3-6.6
TARA_023	20-180	ERR538171	215234081	18
TARA_024	20-180	ERR1726675	135761345	11
TARA_025	20-180	ERR1726976	158887852	13
TARA_026	20-180	ERR1726673	165777892	13
TARA_030	20-180	ERR538189	149724616	12

TABLE B.2 – Metagenomic samples used for the analyses (from Carradec et al., 2018). When several samples were available, the largest one was systematically chosen. While in the exploration step, all the main eukaryotic size fractions (0.8-5 μm , 5-20 μm and 20-180 μm) were considered, the final species choice presented in the study led us to use only the samples corresponding to the size fractions 0.8-5 μm .

	Description	Bioinformatic tool	Settings
Step 1. Preprocessing	Remove bad quality regions	Trimmomatic (v0.33)	-threads 30; -phred33; LEADING :3; TRAILING :3; SLIDING-WINDOW :4 :15
Step 2. Mapping	Map metagenomic reads onto reference assembly	bwa mem (v0.7.5)	-M -t 30
Step 3. Aligned reads filtering	Filter reads aligned with less than 95% identity	samtools fillmd (v1.8-3) Python (v3.7.3)	>95% identity
	Remove poorly informative regions	samtools fasta PRINSEQ-lite(v0.20.4)	-lc_methoddust -lc_threshold 8
Step 4. Variant calling	Generate pileup file	samtools mpileup (v1.8-3)	-aa -uf
	Compute coverages	bash	
	Call variants	bcftools call (v1.8-8)	-ploidy 1 -v -m
Step 5. Variants filtering	Filter variants	bcftools filter (v1.8-8) bcftools view (v1.8-8)	-exclude 'QUAL <30' -g ^miss
	Extract SNPs	bcftools view (v1.8-8)	-v snps

TABLE B.3 – SNPs pipeline. Steps summary and associated tools, scripts and settings.

Species	<i>B. prasinos</i>	<i>P. calceolata</i>	<i>P. cordata</i>
SNP _{Total in all samples}	51,176	108,295	4,627
SNP _{Min - Max per sample}	13 - 44,467	195 - 79,740	604 - 1,867
SNP _{Mean per sample}	4,739.00	10,471.42	867.15
SNP _{Median per sample}	53	492.5	790
pSNP _{Min - Max per sample pair}	1 - 4,987	15 - 7,525	118 - 323
pSNP _{Mean per sample pair}	134.24	237.41	203.79
pSNP _{Median per sample pair}	9	108	201

TABLE B.4 – Distribution of the number of SNPs in the stations (SNP) obtained after filtration steps (i.e. removing SNPs with less than 4X) and the resulting number of SNPs used to compute genetic distances (FST) between pairs of stations (pSNP).

B.3 Geographic distances

From the geographic coordinates of the stations, pairwise geographic distances were estimated. They correspond to the shortest path between each pair of stations that do not cross through the land. An example of the pairwise paths are given in the Figure Supp. XXX. (station 12). The geographic distances are provided in the Table Supp. XXX.

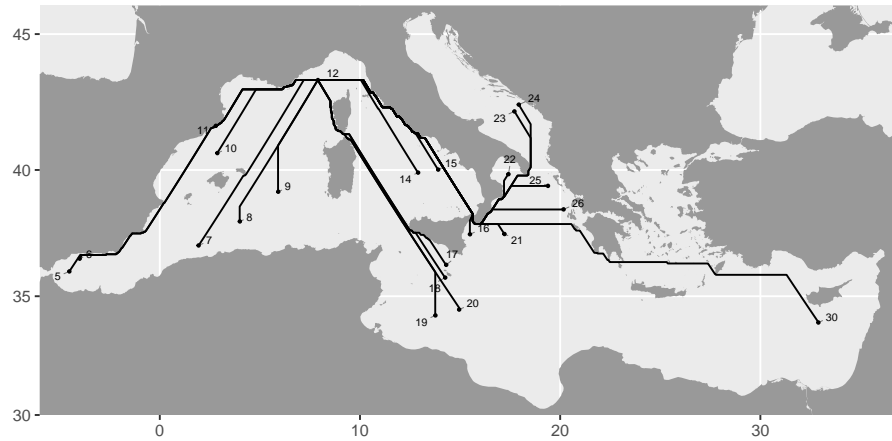


FIGURE B.2 – Representation of the geographic distances between station 12 and all others. The calculated paths correspond to the shortest land.

	5	6	7	8	9	10	11	12	14	15	16	17	18	19	20	21	22	23	24	25	26	30	
5	0																						
6	77	0																					
7	636	560	0																				
8	872	795	239	0																			
9	1102	1025	473	236	0																		
10	927	850	423	319	349	0																	
11	984	907	535	435	417	116	0																
12	1506	1429	892	687	512	581	522	0															
14	1752	1675	1122	884	685	965	953	627	0														
15	1844	1767	1214	976	778	1042	1030	702	93	0													
16	1969	1892	1334	1132	989	1334	1376	1049	426	347	0												
17	1815	1738	1180	1011	906	1250	1305	1019	506	544	199	0											
18	1842	1765	1207	1039	933	1278	1332	1047	533	571	241	61	0										
19	1893	1816	1258	1091	987	1331	1396	1186	678	716	412	243	185	0									
20	1984	1908	1349	1181	1076	1421	1479	1194	686	701	354	221	165	123	0								
21	2103	2026	1468	1242	1099	1444	1486	1159	536	457	153	333	369	506	401	0							
22	2253	2176	1622	1383	1240	1584	1627	1300	677	598	341	521	557	722	663	265	0						
23	2595	2518	1964	1724	1581	1926	1969	1641	1019	940	683	863	899	1038	975	576	380	0					
24	2615	2538	1983	1744	1601	1946	1989	1661	1038	959	703	883	919	1058	995	400	34	0					
25	2365	2288	1734	1495	1351	1696	1739	1412	789	710	453	633	669	808	690	303	198	354	0				
26	2379	2302	1748	1509	1366	1711	1753	1426	803	724	467	647	683	821	703	316	321	475	494	122	0		
30	3624	3548	2989	2821	2715	3060	3103	2776	2153	2074	1771	1815	1798	1781	1680	1618	1737	1882	1901	1541	1420	0	

FIGURE B.3 – Geographic distances between pairwise stations

B.4 Oceanographic distances

5	0	55	105	149	193	219	262	245	249	247	239	240	269	285	267	276	311	312	301	296	544
6	46	0	100	144	192	207	259	243	242	238	244	233	232	260	283	257	264	316	299	288	546
7	158	154	0	61	132	188	223	199	210	201	202	192	200	241	256	221	250	285	273	256	412
8	194	179	90	0	102	193	214	176	209	201	190	178	192	223	238	207	247	285	264	238	331
9	264	257	194	84	0	184	195	155	210	202	191	178	199	216	226	210	229	270	269	254	304
10	168	155	173	181	167	0	141	183	263	256	257	242	246	281	286	266	283	295	290	289	523
11	144	145	162	193	195	74	0	202	266	261	266	266	265	296	301	279	283	296	291	302	612
12	180	186	198	220	207	113	81	0	180	184	282	266	270	296	293	296	261	318	306	310	581
13	315	310	286	303	253	226	223	183	0	94	140	130	156	217	210	186	224	244	245	223	306
14	486	486	242	263	232	179	157	131	80	124	148	148	188	218	214	195	210	196	179	176	310
15	486	491	381	476	404	271	260	248	199	198	0	88	115	188	179	86	149	160	158	148	126
16	573	553	544	526	440	409	301	353	244	248	171	0	40	149	151	132	179	238	235	205	277
17	594	547	538	536	485	423	369	403	266	285	208	41	0	121	139	186	208	262	263	193	226
18	662	666	570	635	537	494	473	549	291	306	241	120	94	0	75	218	331	302	263	278	281
19	751	735	616	689	649	502	513	530	344	370	226	164	120	111	0	220	356	284	278	240	268
20	410	401	335	408	300	248	228	250	196	167	73	140	165	224	213	0	150	188	187	139	269
21	562	511	445	542	427	285	256	412	358	316	138	152	246	205	212	107	0	152	149	115	291
22	956	888	851	907	782	470	278	807	453	362	234	297	280	284	289	199	171	0	118	186	199
23	970	892	898	896	817	499	275	802	407	268	237	287	281	297	297	202	176	50	0	191	203
24	703	695	645	691	652	305	289	545	261	246	215	280	267	233	249	181	151	85	81	0	66
25	779	796	606	660	594	382	447	575	410	320	231	307	273	221	240	189	183	98	92	56	272
26	1292	1363	1368	1272	1305	1047	1066	1103	950	1030	531	818	773	473	493	531	445	304	299	273	259
30																					0

TABLE B.5 – Oceanographic distances computed as minimum connection time. The rows correspond to the departure stations and the columns to the arrival stations.

	5	6	7	9	16	18	20	22	23	24	25	26	30
5	0	6.76.10 ⁻⁴	5.11.10 ⁻⁴	8.33.10 ⁻⁴	0	0	0	0	0	0	0	0	0
6		0	5.63.10 ⁻⁴	8.08.10 ⁻⁴	0	0	0	0	0	0	0	0	0
7			0	1.55.10 ⁻³	8.83.10 ⁻⁵	3.66.10 ⁻⁴	2.59.10 ⁻⁴	2.25.10 ⁻⁵	0	0	0	0	0
9				0	1.16.10 ⁻⁴	3.50.10 ⁻⁴	3.57.10 ⁻⁴	4.81.10 ⁻⁵	0	6.60.10 ⁻⁵	0	0	0
16					0	2.89.10 ⁻⁴	2.85.10 ⁻⁴	1.69.10 ⁻³	1.02.10 ⁻⁴	1.62.10 ⁻⁴	1.32.10 ⁻³	1.42.10 ⁻³	4.20.10 ⁻⁵
18						0	2.33.10 ⁻³	1.52.10 ⁻⁴	1.38.10 ⁻⁵	2.20.10 ⁻⁵	2.18.10 ⁻⁴	1.72.10 ⁻⁴	5.46.10 ⁻⁵
20							0	8.32.10 ⁻⁵	1.03.10 ⁻⁵	1.20.10 ⁻⁵	1.27.10 ⁻⁴	1.20.10 ⁻⁴	4.83.10 ⁻⁵
22								0	5.38.10 ⁻⁴	5.60.10 ⁻⁴	2.38.10 ⁻³	1.20.10 ⁻³	8.67.10 ⁻⁶
23									0	9.07.10 ⁻³	1.14.10 ⁻⁴	3.83.10 ⁻⁴	2.75.10 ⁻⁶
24										0	1.58.10 ⁻³	1.02.10 ⁻³	3.18.10 ⁻⁶
25											0	5.93.10 ⁻³	3.27.10 ⁻⁵
26												0	6.76.10 ⁻⁵
30													0

TABLE B.6 – Oceanographic distances computed as probabilities of connection. The rows correspond to the departure stations and the columns to the arrival stations.

B.5 Overview of the study

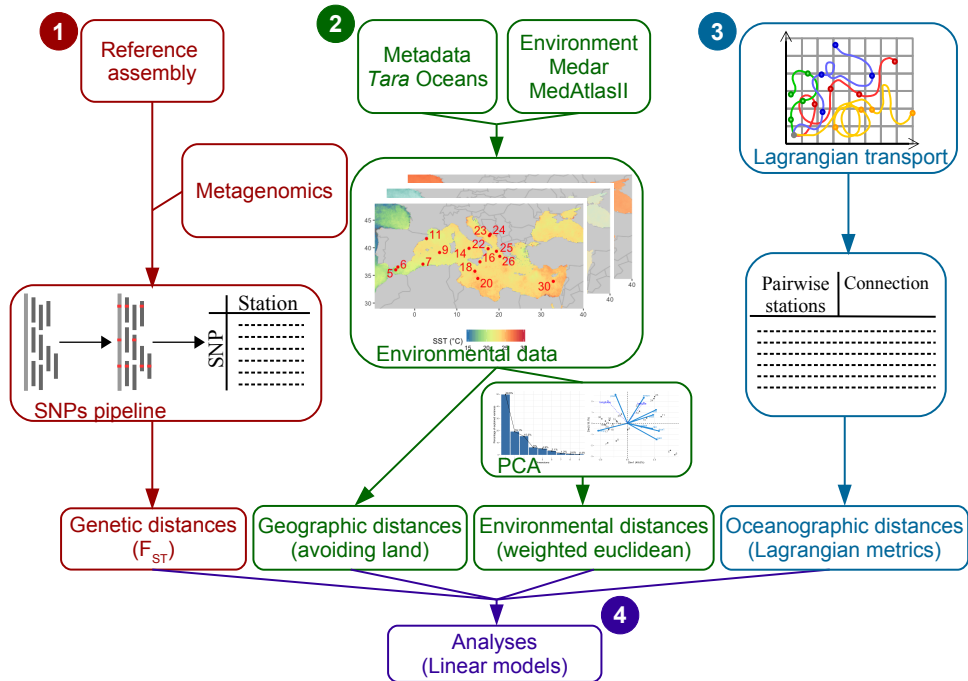


FIGURE B.4 – Analysis strategy for population metagenomics of marine protists detailed in 4 parts. The colours of the boxes indicate the genomic (red), geographic and environmental (green) or oceanographic (blue) parts of the study. PCA : Principal Component Analysis. SNP : Single Nucleotide Polymorphisms.

**Supplementary Information - From trait to
molecular basis of the transparency**

C.1 Morphological descriptors

name	description
area	Surface area of the object in square pixels
perim	The length of the outside boundary of the object
major	Primary axis of the best fitting ellipse for the object
area_exc	Surface area of the object excluding holes, in square pixels (=Area*(1-(%area/100))
esd	Equivalent Spherical Diameter
circ	Circularity = $(4 * \text{Pi} * \text{Area}) / \text{Perim}^2$; a value of 1 indicates a perfect circle, a value approaching 0 indicates an increasingly elongated polygon
feret	Maximum feret diameter, i.e., the longest distance between any two points along the object boundary
fractal	
skelarea	Surface area of skeleton in pixels. In a binary image, skeleton is obtained by repeatedly removing pixels from the edges of objects until they are reduced to the width of a single pixel.
symetrieh	Bilateral horizontal symmetry index
symetriev	Bilateral vertical symmetry index
symetriehc	Symmetry of the largest remaining object in relation to the horizontal axis after thresholding at the grey level Histcum1 value
symetrievc	Symmetry of the largest remaining object in relation to the vertical axis after thresholding at the grey level Histcum1 value
convperim	The perimeter of the smallest polygon within which all points in the object fit
convarea	The area of the smallest polygon within which all points in the object fit
elongation	major-minor
perimareaexc	$\text{perim}/(\sqrt{\text{area_exc}})$
feretareaexc	$\text{feret}/(\sqrt{\text{area_exc}})$
perimmajor	$\text{perim}/\text{major}$
perimferet	$\text{perim}/\text{feret}$
mean	Average grey value within the object; sum of the grey values of all pixels in the object divided by the number of pixels
stddev	Standard deviation of the grey value used to generate the mean grey value
mode	Modal grey value within the object
min	Minimum grey value within the object (0 = black)
max	Maximum grey value within the object (255 = white)
intden	Integrated density. This is the sum of the grey values of the pixels in the object (i.e. = Area*Mean)
median	Median grey value within the object
skew	Skewness of the histogram of grey level values
kurt	Kurtosis of the histogram of grey level values
slope	Slope of the grey level normalized cumulative histogram
histcum1	Grey level value at the first quartile of the normalized cumulative histogram of grey levels
histcum2	Grey level value at the second quartile of the normalized cumulative histogram of grey levels
histcum3	Grey level value at the third quartile of the normalized cumulative histogram of grey levels
fcons	Measure of contrast based in the texture feature descriptor (Amadasun and King, 1989)
thickr	Thickness ratio : relation between the maximum thickness of an object and the averag thickness of the object excluding the maximum
range	max - min
cv	$100 * (\text{stddev}/\text{mean})$
sr	$100 * (\text{stddev}/(\text{max}-\text{min}))$

TABLE C.1 – List of the 39 imaging descriptors used in the analyses

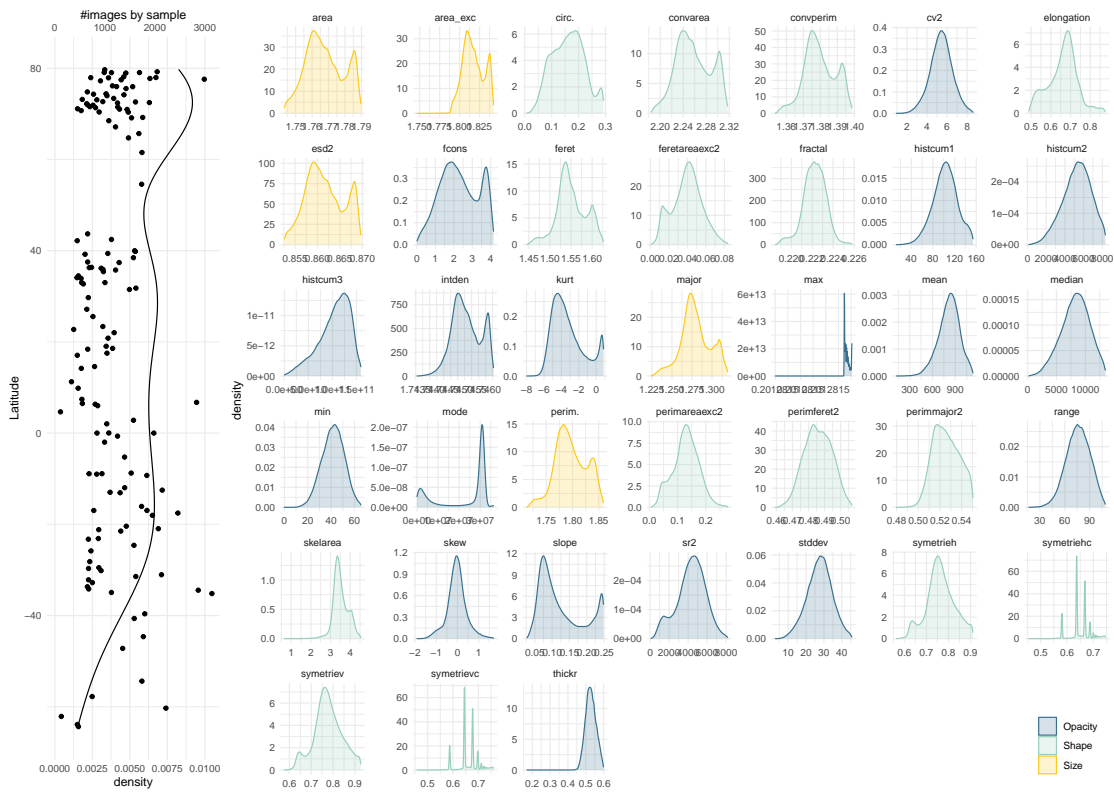


FIGURE C.1 – Distribution of the morphological descriptors

C.2 Environmental organization

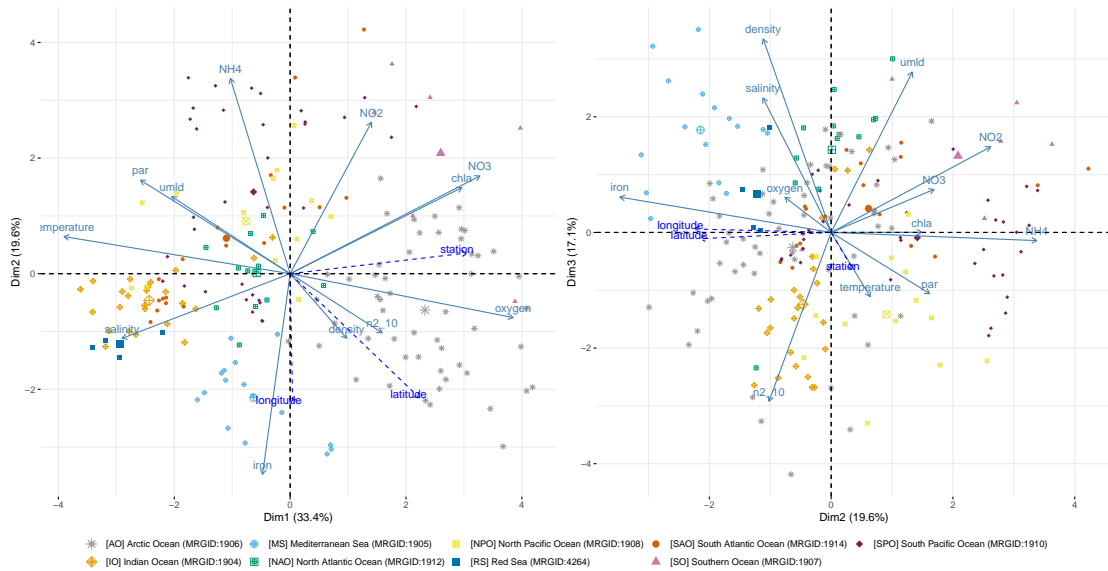


FIGURE C.2 – (A) Geographical positions of the stations and their projections in the plans composed by (B) the two first principal components (PC1 and PC2) and (C) the second and third principal components (PC2 and PC3) of the Principal Component Analysis performed on the environmental variables. Colors of the points represent biomes to which stations belong. Arrows represent projection of environmental variables in the reduced space composed by the three significant principal components (Kaiser-Guttman criterion).

C.3 Morphological organization

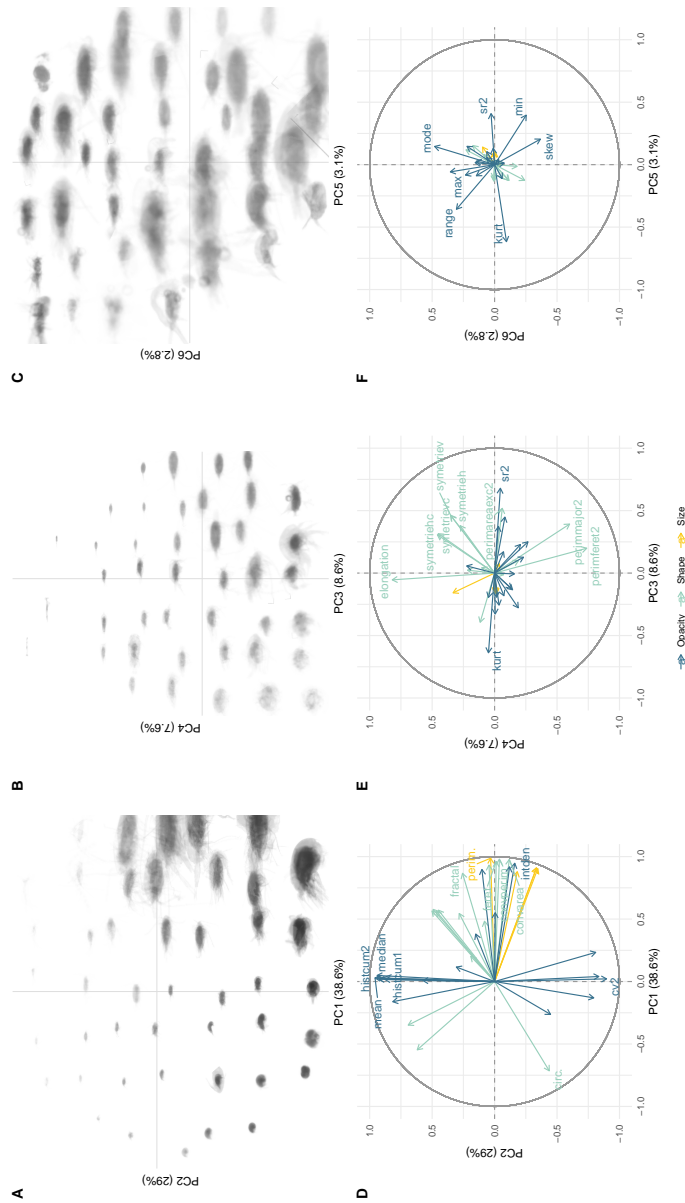


FIGURE C.3 – (A) Morphospace composed by the two first morphological principal components (PC1 and PC2) (B) Morphospace composed by the third and fourth morphological principal components (PC3 and PC4) and (C) Morphospace composed by the fifth and sixth morphological principal components (PC5 and PC6). Projections of the imaging descriptors in the (D) PC1/PC2 plan (E) PC3/PC4 plan and (F) PC5/PC6 plan

**Supplementary Information - Multi-scale
exploration of the planktonic ecosystem in
the North Atlantic Ocean**

D.1 Extended Material and Methods - Mathematical formalism

D.1.1 Datasets

The study is based on several datasets providing information on the different aspects of the Atlantic ecosystem (ecosystem components hereafter), namely environment, hydrodynamics, biological composition and community morphology.

Environment

The environmental dataset derived from several sources, including in situ data collected on board, data from modeling (Copernicus; Le Traon et al. 2019) and satellite data. Each sample collected during the Tara Pacific expedition is associated with environmental parameters such as physicochemical characteristics, nutrient and metal concentrations, and optical properties of the water bodies. These data are publicly accessible on zenodo (id : 4068293). They have been summarized in order to obtain a single value of each parameter per station. Briefly, after keeping the Atlantic samples (stations 1 to 27), the values of each environmental parameter were obtained by averaging per station the values obtained for each procedure and then for each sample device. Variables with more than 10% missing values were removed from the dataset. As some remaining variables were redundant, a manual selection procedure based on variable quality has been applied (by decreasing quality : in situ data, batos, a satellite of resolution $1/24^\circ$, a satellite of resolution $1/12^\circ$, modeling data extracted from Copernicus). Three stations (namely stations 1, 7 and 8) had many missing values and were therefore discarded from the environmental analysis. The remaining variables with missing values were completed by linear inference with the redundant variable of directly lower quality or, when this was not possible, by doing the local average over the two surrounding stations. In the end, the environmental matrix X_{env} resulting from this preprocessing is composed of 24 rows and 24 columns. All details about the procedure are available on GitHub.

Hydrodynamics

Hydrodynamic conditions of the stations were characterized by Eulerian and Lagrangian diagnostics computed from satellite-derived velocity fields with geostrophic and Ekman components.

Velocity field and trajectories computation.

The velocity field employed is the MULTI_OBS_GLO_PHY_REP_015_004 product downloaded from E.U. Copernicus Marine Environment Monitoring Service (CMEMS, <http://marine.copernicus.eu/>). This velocity field is obtained by combining the altimetry derived geostrophic velocities and modeled Ekman surface currents. The spatial resolution of the velocity field is 0.25° and the temporal resolution is one day. The trajectories of the virtual water particles were computed using a 4th order Runge-Kutta method with a time step of 3 hours.

D.1. Extended Material and Methods - Mathematical formalism

	In situ water sampling	Inline system	Satellite (1/24°)	Satellite (1/12°)	Copernicus assimilative model
pH					x
Oxygen					x
Wind underway		x			
Photosynthetically active radiation			x		
Kd 490			x		
Mixed layer depth					x
Temperature		x			
Nitrate	x				
Nitrite	x				
Phosphate	x				
Silicate	x				
Iron	x				
Cadmium	x				
Cobalt	x				
Copper	x				
Lead	x				
Manganese	x				
Nickel	x				
Zinc	x				
Chlorophyll		x			
Particulate Organic Carbon		x			
Salinity		x			For missing value interpolation
Particulate Inorganic Carbon			x		
Net primary production					x

TABLE D.1 – Environmental variables used for the analysis. Source of the data are indicated with crosses.

Lagrangian and Eulerian diagnostic.

Finite Time Lyapunov Exponents were computed as in Shadden et al. (2005). Values of 0.05, 0.1, 0.15 and 0.2° were used as initial separation. The advective times used ranged between 15 and 120 days (with steps of 15 days) backward in time. Lagrangian betweenness is a diagnostic based on Lagrangian flow network theory (Ser-Giacomi et al., 2015) and identifies regions of the ocean where water parcels passing by have originated from several different locations and that then spread over several final destinations. These regions can be identified as bottlenecks of the circulation. For this diagnostic, initial separation values were set to 0.05, 0.1, 0.15 and 0.2°. The temporal window over which the Lagrangian betweenness was computed was set to 30, 60, 90 and 120 days backward in time.

Two dimensional horizontal divergence is defined as $div = \frac{du}{dx} + \frac{dv}{dy}$ in which u and v are respectively the zonal and meridional components of the velocity field. Eulerian divergence was computed for each station by extracting the mean divergence value at the station’s location and the observation. Conversely, following the methodology in Hernandez-Carrasco et al. (2018), the Lagrangian divergence was obtained by computing the divergence for each position and instant of a given trajectory and then averaging in time.

In a two-dimensional velocity field, the vorticity is defined as $vort = \frac{dv}{dx} - \frac{du}{dy}$. Conversely, the Okubo-Weiss metric is obtained from $OW = s^2 - vort^2$, where s is the strain : $s = \left(\frac{du}{dx} - \frac{dv}{dy}\right)^2 + \left(\frac{dv}{dx} - \frac{du}{dy}\right)^2$. Negative Okubo-Weiss values identify regions where the vorticity intensity prevails

over the strain and thus are associated with eddy cores. As for the divergence, the vorticity and the Okubo-Weiss were computed in both Eulerian and Lagrangian frameworks. The retention time was computed as in d'Ovidio et al. (2013). This Lagrangian diagnostics identifies how many days a given water parcel has passed inside an eddy. This information is obtained by following the backward parcel trajectory, and by observing if the corresponding Okubo-Weiss values. If, for instance, the parcel, along its backward trajectory, displayed negative Okubo-Weiss values for 30 days, it means that it was retained inside an eddy for that amount of time. Thus, its retention time is 30 days. Finally, the absolute speed and kinetic energy were computed respectively as $abs\ speed = \sqrt{u^2 + v^2}$ and $kinenergy = \frac{1}{2}(u^2 + v^2)$.

Extraction of diagnostic in correspondence the station transects : the stadium shape.

During sampling, the boat was moving. A station thus corresponds to a trajectory between the sampling start position and the sampling end position. In order to best represent the water mass sampled, a stadium, with two semi-circles centered respectively on the start and end sampling positions, was defined for each station. The stadium was then filled with virtual particles equally separated by 0.03° and, for each of them, the Eulerian and Lagrangian diagnostics described in the previous section were computed. Finally, the value of a given diagnostic assigned to the station was obtained from the average values of stadium's particles. In this way, the stadium defines a homogeneous environment for the transect and has one parameter : the radius. Radius values ranged between 0.05° and 0.6° (with steps of 0.05°). The best correlations were obtained with a radius of 0.1° .

Morphology

Characterization and quantification of biodiversity are complex. Historical approaches have been based on community composition in which taxonomic signals are central and assuming that in similar connected environments, similar communities will be found (Chase, Leibold, 2009). However, non-related species may share functional traits (Violle et al., 2007), which could be either morphological traits (e.g. size, elongation), physiological or behavioral (Martini et al., 2020), that have a profound influence on their fitness. Thus, the use of functional traits responding to environmental pressure (*i.e.*, response traits) can also be relevant information to characterize the community structuring (Porcel et al., 2020). Thus, we also investigated morphological features based on imaging data.

Plankton image acquisition protocol depended on organism size. 20-200 μm organisms were collected using Dolphin device and analyzed with Flowcam, whereas 300-2000 μm organisms were collected with High Speed Net and analyzed with Zooscan. The imaging data are stored on the Ecotaxa imaging platform and accessible under project identifiers 1737 and 1744 for Flowcam and 1700 for Zooscan. These imaging data were acquired for 17 and 23 stations respectively. Flowcam datasets are extensive with many small organisms (*i.e.*, < 500 pixels). For these small organisms, only 30% of the data has been analyzed. This 30% threshold represents a compromise between manual preprocessing time and the representativeness of the complete dataset. Image

Metrics	Diagnostics	Type	Radius	Delta	Tau	Summarise
btw	Betweenness	Lagrangian	0.1	From 0.05 to 0.2 by 0.05	From -15 to -120 by -15	Mean for each station
ftle	Finite time Lyapunov exponents	Lagrangian	0.1	From 0.05 to 0.2 by 0.05	From -30 to -120 by -30	Mean for each station
div	Divergence	Eulerian	0.1	-	-	-
div_lag	Divergence	Lagrangian	0.1	-	From 0 to -120 by -5	Mean for each station
vort	Vorticity	Eulerian	0.1	-	-	-
ow	Okubo-Weiss	Eulerian	0.1	-	-	-
retime	Retention time	cell6	0.1	-	-	-
U	Absolute speed	Eulerian	0.1	-	-	-
ke	Kinetic energy	Eulerian	0.1	-	-	-

TABLE D.2 – Metrics characterizing mixing and transport properties of the water masses. The diagnosis and its type are specified, as well as the set of parameters. Radius : radius of the two semi-circles defining the stadium shape. Delta : initial separation values. Tau : number of days of the backward integration in time. Where there were multiple datasets, the synthetic metric was calculated as the average.

features have been automatically computed in EcoTaxa and 12 have been kept for downstream analyses (Table 2). Automatic image classification was carried out to determine the finest taxonomic assignment. Each assignment was then manually checked by an expert and either confirmed or corrected. Based on the taxonomic assignments, non living (*i.e.*, artefacts, louzy focus, bubble, detritus and fiber) and seaweed objects have been discarded. In total, 22027 images have been analyzed. From the sampling information (e.g. the volume of sampled water), abundances of organisms were normalized leading to comparable data between samples.

Taxonomic composition

Raw fastq files were merged and then trimmed. Each sample was dereplicated (i.e. strictly identical reads were merged) and converted into fasta format. Clustering was performed with Swarm (Mahé et al., 2015). Each cluster is characterized by a representative sequence. Representative sequences were assigned to taxa, using the 18S database EukRibo (*i.e.*, manually curated subset of sequences from PR2). In order to ensure good data quality, clusters and representative sequences were filtered (*e.g.*, only clusters containing at least 3 reads in 2 samples were retained).

The final occurrence table contains 369,451 OTUs and 3,955 samples.

D.1.2 Normalization and standardization

As several variables did not show normal distribution, normalizations have been performed to correct their distribution and homogenize their variance. In addition, to remove the influence of magnitude differences of variables in the multivariate analysis, all variables have been standardized.

Box-Cox transformation for biogeochemistry

Let X be the biogeochemical matrix

$$x_i^{(\lambda)} = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln x_i & \text{otherwise} \end{cases} \quad (\text{D.1})$$

with λ estimated using the profile likelihood function.

Log-modulus transformation for hydrodynamics

Let X be the hydrodynamics matrix

$$L(x) = \text{sign}(x) * \log(|x| + 1) \quad (\text{D.2})$$

$$\text{with } \forall a \in \mathbb{R}, \text{sign}(a) = \begin{cases} 1 & \text{if } a \geq 0 \\ -1 & \text{otherwise} \end{cases} \quad (\text{D.3})$$

Yeo-Johnson transformation for morphological traits

Two morphological datasets are corresponding to the 20-200 μm and 300-2000 μm size-fractions.

$$\psi(\lambda, x) = \begin{cases} ((x + 1)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, x \geq 0 \\ \log(x + 1) & \text{if } \lambda = 0, x \geq 0 \\ -[(-x + 1)^{2-\lambda} - 1]/(2 - \lambda) & \text{if } \lambda \neq 2, x < 0 \\ -\log(-x + 1) & \text{if } \lambda = 2, x < 0 \end{cases} \quad (\text{D.4})$$

Centered log ratio transformation for OTU composition

Lastly, OTUs abundances are in four matrices that refers to the 0.22-3 μm , 3-20 μm , 20-200 μm and 300-2000 μm size-fractions.

$$clr(x) = \left[\log \frac{x_1}{g(x)} \cdots \log \frac{x_D}{g(x_D)} \right] \quad (\text{D.5})$$

Standardization

Let X be one of the matrices to be standardized, such as :

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma(x_j)} \quad (\text{D.6})$$

D.1.3 Community weighted mean (CWM) of morphological traits

$$cwm_t = \sum_{i=1}^m p_i trait_t \quad (\text{D.7})$$

D.1.4 Principal component analysis (PCA)

The Principal Component Analysis (PCA, Legendre, Legendre 2012) is a dimensional reduction method that transforms initial variables into new variables, called principal components (PC), which are linear combinations of the initial variables. Each successive PC maximizes the remaining variance of the dataset (*i.e.*, variance not supported by previous PCs). Thus, the new multidimensional space is a reduced orthogonal base space. A PCA has been performed on each matrix to describe the variability of datasets and summarize the information carried by variables. Let Z be one of the studied standardized matrices with n lines (or stations) and p random variables $[Z_1, \dots, Z_p]$ with n joint realizations.

$$Z = \begin{pmatrix} z_{11} & \cdots & z_{1p} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{np} \end{pmatrix} \quad (\text{D.8})$$

$$\arg \max_{v_1 \in \mathbb{R}} \sum_{k=1}^K cov^2(x_{.k}, v_1) \quad (\text{D.9})$$

D.1.5 Multiple Factorial Analysis (MFA)

The principle is similar to that of the PCA with a successive calculation of the eigenvalues.

Calculate the 1st eigenvalue λ_1^j of the jth group of variables ($j=1,\dots,J$).
 Perform an overall PCA on the weighted table :

$$\left[\frac{X_1}{\sqrt{\lambda_1^1}}; \frac{X_2}{\sqrt{\lambda_1^2}}; \dots; \frac{X_J}{\sqrt{\lambda_1^J}} \right] \quad (\text{D.10})$$

X_j corresponds to the jth normalized or standardized tables.
 Here $j \in [bpc, hyd, tax, morph]$.

$$\operatorname{argmax}_{v_1 \in \mathbb{R}} \sum_{j=1}^J \sum_{k \in K_j} \operatorname{cov}^2\left(\frac{x.k}{\sqrt{\lambda_1^j}}, v_1\right) \quad (\text{D.11})$$

$$\operatorname{argmax}_{v_1 \in \mathbb{R}} \sum_{j=1}^J \mathcal{L}_g(K_j, v_1) \quad (\text{D.12})$$

where

$$\mathcal{L}_g(K_j, v_1) = \frac{1}{\lambda_1^j} \sum_{k \in K_j} \operatorname{cov}^2(x.k, v_1) \quad (\text{D.13})$$

The \mathcal{L}_g coefficient measures the connection between groups of variables :

$$\mathcal{L}_g(K_j, K_m) = \sum_{k \in K_j} \sum_{l \in K_m} \operatorname{cov}^2\left(\frac{x.k}{\lambda_1^j}, \frac{x.l}{\lambda_1^m}\right) \quad (\text{D.14})$$

The \mathcal{L}_g coefficient as an indicator of a group's dimensionality :

$$\mathcal{L}_g(K_j, K_j) = \frac{\sum_{k=1}^{K_j} (\lambda_k^j)^2}{(\lambda_1^j)^2} = 1 + \frac{\sum_{k=2}^{K_j} (\lambda_k^j)^2}{(\lambda_1^j)^2} \quad (\text{D.15})$$

The RV index as an indicator of pairwise relationships in the \mathcal{L}_g sense :

$$RV(K_j, K_m) = \frac{\mathcal{L}_g(K_j, K_m)}{\sqrt{\mathcal{L}_g(K_j, K_j)} \sqrt{\mathcal{L}_g(K_m, K_m)}} \text{ with } 0 \leq RV \leq 1 \quad (\text{D.16})$$

$RV = 0$: all variables in K_j and K_m are uncorrelated

$RV = 1$: the two-point clouds are homothetic

D.2 Environmental structuring



FIGURE D.1 – Environment structuring along the North Atlantic transect. (A) Spatial projections of the biogeochemical PCA coordinates of the stations. (B) Correlation between biogeochemical variables and biogeochemical PCs in the PCA space. (C) Projection of the stations in the hydrodynamic plan composed of the PC2 and PC3. Colors represent the hydrodynamic clusters.

D.3 Multiple factor analysis

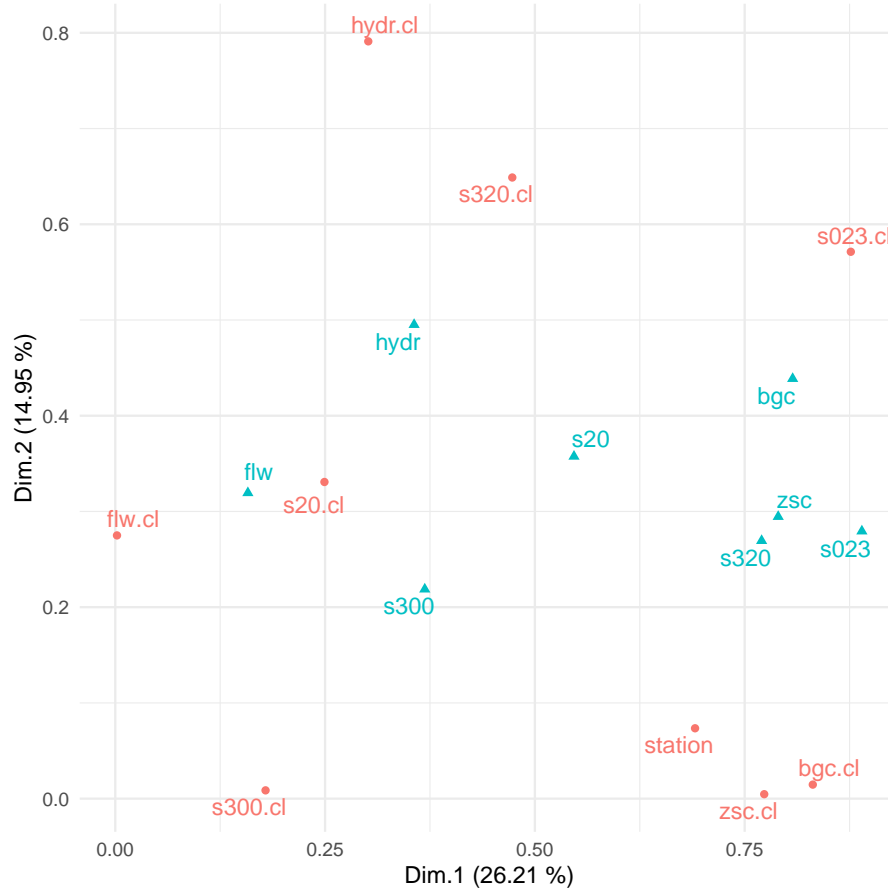


FIGURE D.2 – Graph of variables obtained after proceeding MFA. Each point represents a dataset. The point's color indicates if the dataset contributed to the analysis (pink : no, blue : yes).



FIGURE D.3 – Projection of the groups in the MFA space. Contribution of the significant principal components of each group of data in the construction of the first plane of the consensus space resulting from the MFA

Curriculum Vitae
