Dynamique de la structure des génomes et de leur biogéographie dans l'océan : analyses comparatives des données métagénomiques du projet Tara Oceans pour l'étude de la microalgue Bathycoccus et des communautés planctoniques globales

Thomas Vannier

THESE DE DOCTORAT
DE L'UNIVERSITE PARIS-SACLAY,
préparée à l'Université d'Evry Val D'Essonne

ECOLE DOCTORALE N° 577
Structure et dynamique des systèmes vivants

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

Par

Thomas Vannier

Dynamique de la structure des génomes et de leur biogéographie dans l'océan : analyses comparatives des données métagénomiques du projet *Tara* Oceans pour l'étude de la microalgue *Bathycoccus* et des communautés planctoniques globales.

Numéro national de thèse : 2017SACLE002

Thèse présentée et soutenue publiquement à Évry, le 21 Mars 2017 :

Composition du Jury :

M. Not Fabrice	CR1 (UPMC, Roscoff)	Rapporteur
Mme. Piganeau Gwenaël	DR (UPMC, Banyuls-sur-Mer)	Rapporteur
M. Dubow Michaël	PR (Université Paris-Sud, Gif-sur-Yvette)	Président
M. Hingamp Pascal	MC (Université Aix-Marseille II, Marseille)	Examinateur
M. Peterlongo Pierre	CR1 (INRIA, Rennes)	Examinateur
M. Jaillon Olivier	DR (CEA, Evry)	Co-directeur de thèse
M. Wincker Patrick	DR (CEA, Evry)	Directeur de thèse

Remerciements

Travailler en tant que jeune chercheur au sein du Genoscope a été une expérience enrichissante aussi bien d'un point de vue intellectuel que personnel. J'ai intégré le projet *Tara* Oceans en 2012. La goëlette *Tara* était sur la fin de sa mission mais ce n'était alors que le début de l'étude d'une ressource inestimable de données récoltées au cours de cette expédition. De nombreuses personnes se sont attelées au traitement et à l'analyse de ces données et cela a été propice à de nombreuses rencontres et collaborations.

Je tiens tout d'abord à remercier Patrick Wincker pour m'avoir donné l'opportunité de réaliser cette thèse sous sa direction au sein du Genoscope. Merci également à Olivier Jaillon pour son encadrement et son aide tout au long de ma thèse. Vos conseils m'ont aidé à acquerir la rigueur nécessaire à un bon raisonnement scientifique afin de mener à bien mon projet de recherche. Cela me sera utile dans mes futurs projets. C'est un privilège d'avoir pu intégrer cette grande communauté scientifique qu'est le consortium *Tara* Oceans. Cela m'a permis d'interagir avec de nombreux chercheurs et de me créer un réseau pour la suite de mon projet professionnel.
Merci à Jean Weissenbach, Pierre Leber et Marcel Salanoubat pour m'avoir accueilli pour la réalisation de cette thèse dans l'Institut de Génomique du CEA au sein de l'UMR8030 Génomique Métabolique.
Catherine Sarlande, Monique Meugnier, Nancy Delpech et Catherine Contrepois, merci pour votre aide lorsque j'en avais besoin et pour avoir participé au bon déroulement de ma thèse.

Le Genoscope est un gigantesque voilier sur lequel travaillent de concert de nombreux scientifiques, partageant leurs connaissances et leurs compétences afin de contribuer à l'amélioration de notre compréhension du monde qui nous entoure.
Je remercie les collègues de mon laboratoire, le LAGE, mais également ceux des autres laboratoires travaillant sur *Tara*, qui ont pu m'éclairer et m'aider sur mon travail, notamment au cours des discussions constructives lors des réunions *Tara* du lundi. Il y a eu un léger turnover de la communauté *Tara* au Genoscope ces 5 dernières années et pour cela j'espère n'oublier personne : Jean-Marc, Eric, Julie, Adriana, Valérie, Corinne, Benjamin, Gaëlle, Amine, Yoann, Jade, Samuel, Elise, Quentin, Marc, Kevin, Betina, Sarah, Marion, Amos et Mario. Merci également à toutes les personnes ayant participé à la production des séquences du projet.
Merci à Jean-Marc Aury et à l'équipe R&D bioseq dont j'ai fait partie un temps, pour la mise à disposition des outils permettant le traitement des données de séquençage haut débit. Vous êtes nombreux à m'avoir aidé dans mes travaux de recherche. Je n'oublierai pas les bons moments

passés à la cafétéria à déguster les plats de chez Picard mais également les discussions autour d'une (plusieurs) bière(s) au New Village.
Je remercie Claude Scarpelli qui m'a permis d'utiliser les clusters de calculs du TGCC. Merci également à l'équipe système pour la mise à disposition, l'évolution et le maintien de l'environnement et des ressources informatiques du Genoscope.
Merci à Maria, qui m'a précédé en tant que thésarde au Genoscope, pour ses conseils sur le bon déroulement d'une thèse. Merci à Jon-Jon avec qui j'ai partagé l'expérience des moments de rushs, du stress et des petits soucis que procure parfois une thèse. Et bien sûr, merci à tous les autres thésards du Genoscope avec qui j'ai passé de bons moments lors de nos réunions « journal club ».
Cyril, Elise, Alexis, Manu et Fabien, merci d'avoir partagé mon bureau et d'y avoir créé une excellente ambiance. C'était un plaisir de venir travailler chaque jour avec vous et vos aides m'ont été précieuses.

J'exprime ma gratitude à l'ensemble du consortium *Tara* Oceans. J'ai eu la chance de pouvoir collaborer étroitement avec certains de ses laboratoires. Daniele Iudicone, Maurizio Ribera et Romain Watteaux, merci pour votre accueil chaleureux dans cette ville fantastique qu'est Naples. Colomban de Vargas, Daniel Richter (Roscoff), Pascal Hingamp et Emilie Villar (Marseille), cela a été formateur et un plaisir de travailler avec vous tous.
Merci également aux membres de l'ANR MAPPI et HydroGen dont j'ai fait partie pour les discussions et collaborations fructueuses : Pierre Peterlongo, Dominique Lavenier, Nicolas Maillet, Gaëtan Benoit, Claire Lemaitre, Guillaume Collet, Guillaume Rizk (INRIA, Rennes), Mathieu Raffinot (LIAFA, Paris), Stéphane Robin, Julie Aubert (AgroParisTech, Paris), Sophie Schbath, Mahendra Mariadassou (Inra, Jouy-en-Josas), Hélène Touzet, Evguenia Kopylova et Samuel Blanquart (INRIA, Lille).

Merci aux membres de mon comité de thèse, Daniel Vaulot et Laurent Duret, pour les discussions, les conseils et les critiques toujours instructifs pour mes résultats et le déroulement de ma thèse.

Je remercie particulièrement mes rapporteurs, Gwenaël Piganeau et Fabrice Not ainsi que les autres membres de mon jury de thèse, Pascal Hingamp, Pierre Peterlongo et Michaël Dubow pour avoir accepté d'évaluer mon travail et de m'avoir accordé de leur temps.

Enfin, merci à mes proches qui ont été présents et m'ont soutenu au cours de ces dernières années. Merci à ma mère qui a participé à la relecture de ce manuscrit et qui, avec mon père, m'a toujours encouragé et aidé dans tout ce que j'ai entrepris.

The ocean's dying. Plankton's dying.
It's people. Soylent Green is made out of people.
They're making our food out of people.
Next thing, they'll be breeding us like cattle for food.
You've gotta tell them. You've gotta tell them!

Richard Fleischer, « Soylent Green » (1973)

Table des matières

TABLE DES MATIERES 1

ABREVIATIONS 5

INTRODUCTION 7
La démarche suivie dans cette thèse 13

CONTEXTE BIOLOGIQUE ET METHODOLOGIQUE 16

I. LE PLANCTON ET L'ECOSYSTEME PLANCTONIQUE 17
I.1 Le plancton : un rôle clef dans les équilibres nécessaires à la vie sur Terre 17
I.1.1 Qu'est-ce que le plancton? 17
I.1.2 Importance des micro-organismes planctoniques 19
I.2 Classification phylogénétique des espèces planctoniques 19
I.2.1 La séparation des branches du vivant 20
I.2.2 Classification phylogénétique des eucaryotes 20
I.2.3 Les algues vertes, composantes importantes du phytoplancton 21
I.2.4 Les Prasinophyceae 22
I.2.5 Les Mamiellales 24
Ostreococcus 24
Micromonas 25
Bathycoccus 26
I.3 L'écosystème planctonique 28
I.3.1 Qu'est-ce que l'écosystème planctonique ? 28
I.3.2 Etudes de la diversité des micro-organismes planctoniques 28
L'imagerie par microscope 28
I.3.3 Impact de l'environnement sur l'organisation spatiale du plancton 29
Les Courants marins 30
Les paramètres physico-chimiques 34
I.3.4 Mesures des données environnementales 35
I.3.5 La biogéographie des micro-organismes planctoniques 37
I.4 Les grandes expéditions océaniques 39
I.4.1 L'océanographie moderne et la collecte du plancton 39
I.4.2 Le H.M.S Challenger 39
I.4.3 L'expédition Global Ocean Sampling 42
I.4.4 L'expédition Tara Oceans 43
La Goélette *Tara* 43
Prélèvements des échantillons 45
Un projet multidisciplinaire 47

II. ÉTUDE GENOMIQUE ET METAGENOMIQUE DES MICRO-ORGANISMES PLANCTONIQUES 50
II.1. Le séquençage de l'ADN 50
II.1.1. Les débuts du séquençage 51
La méthode de Sanger 52
Les séquenceurs à capillaires 53

II.1.2 Les techniques de séquençage à haut débit 54
La technologie 454 54
La technologie « SOLiD » (Applied Biosystems) 55
La technologie Solexa/illumina 55
II.1.3 Technologies de séquençage de « 3e génération » 57
Technologie HeliScope 57
Technologie SMART 57
Technologie nanopores 58
Autres technologies de 3e génération 58
II.1.4 Performance des techniques de séquençage à haut débit 59
Temps d'un processus de séquençage 59
Quantité de lectures générées 59
Taille des lectures générées 60
Taux d'erreurs de séquençage 60
Le prix du séquençage 61
II.1.5 Comparaison des techniques de séquençage à haut débit 62
II.2. La génomique 63
II.2.1. La reconstruction des génomes 63
II.2.2. Le séquençage de novo 65
II.2.3. Le séquençage de génome à cellule unique 66
Méthode DOP-PCR 67
Méthode MDA 67
Méthode MALBAC 67
Co-assemblage des séquences de cellules uniques 67
II.2.4. Séquençage RNA-Seq 68
II.2.5. Séquençage de codes-barres à ADN 70
II.3. La métagénomique 71
II.3.1. Qu'est-ce que la métagénomique? 71
II.3.2. Différentes familles en métagénomique 72
La métagénomique quantitative 72
La métagénomique ciblée 72
La métagénomique fonctionnelle 73
La métagénomique comparative 73
II.3.3. La métatranscriptomique 74
Exemples d'applications en métatranscriptomique 74
Les limites de la métatranscriptomique 75
II.4 Les projets en métagénomique 76
II.4.1 Le projet metaHit 76
II.4.2 Le projet METASOIL 76
II.4.3 Global Ocean Sampling 77
II.4.4 Le projet Tara Oceans 78
II.4.6. Evolution au cours du temps du nombre de projets de métagénomique 81
II.4.7. La multidisciplinarité des projets de métagénomique 82

III. ANALYSE COMPARATIVE DES DONNEES METAGENOMIQUES DU PROJET *TARA* OCEANS 84
III.1 Comparaison du contenu génomique d'échantillons par homologie aux bases de données de références 84
III.1.1 MEGAN 85
III.1.2 MG-RAST 86
III.1.3 IMG/M 86

III.1.4 Outils optimisés pour les courtes séquences 86
III.2 Comparaison du contenu génomique d'échantillons via les biais compositionnels de l'ADN 88
III.2.1 Kraken 88
III.2.2 CLARK 89
III.3 Comparaison du contenu génomique d'échantillons avec les marqueurs génétiques 89
III.3.1 MetaPhlAn 90
III.3.2 MetaPhyler 90
III.3.3 mOTU 90
III.4 Comparaison du contenu génomique d'échantillons à partir des barcodes environnementaux 91
III.4.1 Le séquençage de barcodes environnementaux 91
III.4.2 Les métabarcodes pour étudier la diversité du plancton eucaryote 91
III.4.3 Calcul de distances de similarité avec les métabarcodes 93
La diversité beta 93
Notion de dissimilarité 93
Distance de Whittaker 93
Distance de Sørensen 94
Distance de Bray-Curtis 94
III.5 Comparaison du contenu génomique d'échantillons à partir des séquences *de novo* 95
III.5.1 crAss 96
III.5.2 TriageTools 96
III.5.3 Compareads 97
III.5.4 DSM 101
III.5.5 MetaFast 101
III.5.6 Mash 102
III.5.7 Simka 103
III.6 Utilisation de clusters de calculs pour la comparaison des échantillons métagénomiques de *Tara* Oceans 106
III.6.1 Les clusters de calculs 106
III.6.2 Ressource informatique au Genoscope 107
III.6.3 Centre de calcul du CEA 108
III.6.4 Parallélisations des processus pour la comparaison des métagénomes 108

CHAPITRE I 111
CARACTERISATION ET COMPARAISON DE LA BIOGEOGRAPHIE ET DE LA STRUCTURE DU REPERTOIRE DE GENES D'ALGUES VERTES PHOTOSYNTHETIQUES DANS LES EAUX OCEANIQUES DE SURFACE 111

I.1. Article 1: Survey of the green picoalga *Bathycoccus* genomes in the global ocean 111
I.2. Étude d'un gène dispensable chez *Bathycoccus prasinos* : La flavodoxine 125
I.2.1 Introduction sur la flavodoxine et la ferrédoxine 125
I.2.2 Récupération du gène de la flavodoxine chez TOSAG39-1 126
I.2.3 Le gène de la flavodoxine est-il un gène dispensable chez *Bathycoccus prasinos* ? 127
I.2.4 Tentative de récupération du gène de la flavodoxine de *Bathycoccus prasinos* dans les échantillons métagénomiques 128
I.3. Caractérisation de « gènes inconnus » chez *Bathycoccus prasinos* par une méthode de phylostratigraphie 129
I.3.1 La phylostratigraphie 129
I.3.2 Détection des gènes orphelins de *Bathycoccus prasinos* 130
I.3.3 Caractérisation structurale des gènes orphelins de *Bathycoccus prasinos* 132
I.4. Étude de génomique comparative des mamiellales 134

I.4.1 Phylostratigraphie des mamiellales 134
I.4.2 Biogéographie des mamiellales. 135
Conclusion du chapitre I **137**

CHAPITRE II **139**
ÉVALUATION ET AMELIORATION D'OUTILS PERMETTANT D'ETUDIER L'ORGANISATION GENOMIQUE DES COMMUNAUTES VIRALES, BACTERIENNES ET EUCARYOTES A L'ECHELLE DE LA PLANETE A PARTIR DES DONNEES METAGENOMIQUES **139**

II.1 Article 2: Environmental characteristics of Agulhas rings affect interocean plankton transport **139**
II.2 Article 3: *COMMET*: comparing and combining multiple metagenomic datasets. **153**
Conclusion du chapitre II **161**

CHAPITRE III **162**
ÉTUDE DE L'IMPACT DES COURANTS OCEANIQUES ET DES PARAMETRES PHYSICO-CHIMIQUES SUR L'ORGANISATION GENOMIQUE DES COMMUNAUTES DE MICROPLANCTONS **162**

III. Article 4 : Global plankton biogeography is shaped via ocean circulation dynamics **162**
Conclusion du chapitre III **211**

CONCLUSION GENERALE **212**
I. Dynamique de la structure des génomes et de leur biogéographie à l'échelle d'une espèce **212**
II. Dynamique de l'organisation biogéographique des communautés planctoniques globales **214**
III. Discussions et perspectives **216**
III.1. Impact de l'environnement sur la variabilité de la structure en gènes et de la biogéographie chez la micro-algue *Bathycoccus*. 216
III.2. Impact de l'environnement sur la biogéographie des communautés micro-planctoniques globales 218

REFERENCES **220**

ANNEXES **237**
Annexe I. Informations supplémentaires de l'article : Survey of the green picoalga *Bathycoccus* genomes in the global ocean. **237**
Annexe II. Informations supplémentaires de l'article : Environmental characteristics of Agulhas rings affect interocean plankton transport. **275**
Annexe III. Informations supplémentaires de l'article : Global plankton biogeography is shaped via ocean circulation dynamics. **276**

Abréviations

ADN : Acide Désoxyribonucléique

ADNc : Acide Désoxyribonucléique complémentaire

ARN : Acide Ribonucléique

ARNr : Acide Ribonucléique ribosomique

DCM : (*angl. Deep Chlorophyll Maximum*) Profondeur maximum en chlorophylle

Fld : Flavodoxine

ITS : (*angl. Internal Transcribed Spacer*) Espaceur interne transcript

LSU : (*angl. Large subunit*) Grande sous unité

Mb : Méga bases = 100 000 paires de bases nucléotidiques

mOTU : (*angl.metagenomic operational taxonomic units*) Unité taxonomique opérationnelle métagénomique

NGS : (*angl. Next-Generation Sequencing*) Technologie de Séquençage Nouvelle Génération

OTU : (*angl. Operational Taxonomic Unit*) Unité taxonomique opérationnelle

PCR : (*angl. Polymerase Chain Reaction)* Réaction en chaîne par polymérase

QC : (*angl. Quality control*) Score de qualité d'un séquençage

RCC : (*angl. Roscoff Culture Collection*) Collection de cultures de Roscoff

RNA-Seq : (*angl. RNA sequencing*) Séqençage de l'Acide Ribonucléique

RPKM : (*angl. Reads Per Kilobase per Million*) Méthode de normalisation pour la comparaison entre les gènes inter ou intra-échantillon.

SAG : (*angl. Single-cell amplified Genome*) Amplification de genome à cellule unique

SSU : (*angl. Small Subunit*) Petite sous unité

TRG : (*angl. Taxonomicaly Restricted Genes*) Gène taxonomiquement restreint

Introduction

La diversité de l'écosystème planctonique joue un rôle clef dans les équilibres nécessaires à la vie sur Terre. Elle représente la diversité en gènes, en organismes et en communautés planctoniques dans l'espace et dans le temps. Le plancton regroupe un large éventail d'êtres vivants qui dérivent le long des courants marins. Il est constitué d'organismes de tailles variables allant de plusieurs mètres de long, comme certaines méduses, à moins d'un pico mètre comme c'est le cas pour certains virus. Plusieurs types de planctons coexistent et interagissent dans les océans. Des eucaryotes multicellulaires comme le zooplancton, aux protistes unicellulaires composés de micro-algues vertes, en passant par les bactéries, les archées jusqu'au virus. Le zooplancton est un plancton animal qui se nourrit de matières vivantes et notamment de phytoplanctons. Ce dernier regroupe des micro-algues unicellulaires ainsi que des bactéries photosynthétiques. La photosynthèse, est un processus bioénergétique qui permet de synthétiser, en utilisant la lumière du soleil, de la matière organique, dont l'oxygène. 71% de la surface du globe est recouverte d'eau et 98% de la biomasse des océans est composée de plancton. On retrouve plusieurs milliards d'organismes phytoplanctoniques dans 1 litre d'eau de mer. Ainsi, des centaines de milliers d'espèces flottent à la surface des océans pour capter l'énergie du soleil et produisent autant d'oxygène que toutes les forêts et plantes terrestres. Le phytoplancton est donc le producteur de plus de 50% de l'oxygène que nous respirons. La matière organique qu'il fabrique constitue de la nourriture pour le zooplancton qui est lui-même mangé par de plus grosses espèces allant jusqu'aux baleines. Le plancton est donc à la base de la chaîne alimentaire marine. De plus, les cadavres de cette biomasse planctonique s'accumulent depuis des milliards d'années sur le fond des océans. Après un processus de plusieurs dizaines de millions d'années, cette couche de sédiments organiques se retrouve sous forme de charbon, de roche, de pétrole et de gaz naturel. Le plancton est donc intimement lié à nos activités humaines. Il est à travers le cycle du carbone un important régulateur de la machine climatique ainsi que de l'acidité des océans puisqu'il absorbe environ 70% du CO2 que nous produisons[1]. L'exploitation de ces combustibles fossiles comme source d'énergie non renouvelable depuis plus de cent ans perturbe le cycle du carbone et a donc un impact direct sur l'équilibre climatique planétaire.

L'écosystème planctonique est un système complexe qui regroupe des organismes planctoniques interagissant entre eux via le parasitisme, la prédation, la symbiose ou la compétition, mais également avec leur environnement. La circulation de l'eau de mer dans l'ensemble des océans est régie par les effets combinés du vent, des différences de températures,

de densité, de salinité mais également des interactions au sein des courants marins. Le plancton, transporté par ces mouvements d'eau, est intimement lié à ces différents facteurs. Combiné avec les différents modes d'interaction qui s'opèrent entre les organismes, les variations de ces paramètres environnementaux ont un impact sur l'organisation des communautés planctoniques, sur les variations de diversité et d'abondance d'une espèce ainsi que sur la malléabilité des génomes. L'ensemble de ces interactions a donc un rôle primordial sur l'écosystème planctonique tel que l'on peut l'observer aujourd'hui. Cependant, l'écosystème océanique est l'écosystème le moins connu de notre planète. De nombreuses espèces planctoniques sont encore à découvrir, la majorité des gènes qui les composent sont absents des bases de données publiques. De plus, l'impact des conditions environnementales sur la composition, la diversité et la biogéographie des communautés, des espèces et des gènes planctoniques à l'échelle de la planète reste encore peu connu. Ces lacunes limitent notre compréhension sur la façon qu'ont ces organismes à se regrouper en communautés complexes évoluant et interagissant entre elles et avec leur l'environnement pour former l'écosystème océanique. Alors que le climat de la planète ne cesse d'évoluer, mieux connaître cette écosystème, sa dynamique et ses capacités d'adaptation face aux modifications physico-chimiques des masses d'eau océanique s'avère indispensable pour appréhender l'impact des activités humaines sur celui-ci et donc sur les équilibres nécessaires à la vie sur Terre.

Pour étudier cet écosystème planctonique, de nombreuses expéditions océanographiques se sont succédées depuis le XIXème siècle. L'invention du filet pélagique en 1845 par Müller à permis d'effectuer les premières pêches de ces micro-organismes. Ainsi, le naturaliste Charles Darwin a pu décrire dans son ouvrage le Voyage du *Beagle* la diversité de ces micro-organismes marins. C'est ensuite, que le navire HMS *Challenger* réalise la première campagne d'exploration des fonds océaniques de 1872 à 1876. Un protocole d'échantillonnage est mis en place pour récolter les micro-organismes planctoniques tout en incluant les relevés des paramètres physiques, chimiques et biologiques. On assiste au début de l'océanographie moderne. Pendant cette expédition, plusieurs organismes planctoniques ont été décrits et dessinés en détail par le naturaliste Allemand Ernst Haeckel. Les mesures physico-chimiques ont permis de montrer la diversité des environnements océaniques. De plus, cette expédition a révélé qu'il était possible d'étudier la vie océanique dans son environnement. Cependant, le manque de moyens matériels et technologiques pour observer et étudier l'ensemble de la diversité des organismes micro-planctoniques ainsi que les lacunes sur nos connaissances des courants marins et des régions océaniques, ne permettaient pas d'apprécier l'impact des facteurs environnementaux sur l'organisation de la vie marine.

Au cours du XXème siècle, les avancées technologiques ont permis de combler ces lacunes. Deux domaines ont particulièrement contribué à améliorer nos connaissances sur l'écosystème marin : l'imagerie et la génomique. Deux méthodes en imagerie sont généralement utilisées pour étudier la biologie marine. La première méthode, utilise l'imagerie satellitaire pour suivre en temps réel la circulation océanique en surface et en profondeur dans l'ensemble des océans du globe. Il est par exemple possible de suivre le parcours de gigantesques tourbillons qui piègent et transportent parfois, pendant plusieurs années le plancton marin. Des modèles de simulation sont développés afin de prédire le temps que met une particule en suspension dans l'eau pour être transportée à différents points du globe. De plus, l'observation d'image satellite nous renseigne sur la quantité de chlorophylle *a* et donc sur l'abondance de phytoplancton présent dans l'eau. Ces images permettent de visualiser les efflorescences de phytoplancton lorsque les conditions environnementales sont réunies pour sa prolifération. La biogéographie est l'étude de la distribution géographique des espèces ou des communautés d'espèces. Elle permet d'identifier des écosystèmes en étudiant la distribution spatiale des organismes en relation avec les conditions environnementales. Ainsi, l'imagerie satellitaire a contribué aux propositions de découpage biogéochimique des océans initiées par Longhurst dans les années 2000[2–4]. Cependant, ce découpage reflète surtout la distribution spatiale des organismes phytoplanctoniques puisqu'elle intègre la mesure de données chimiques telles que la quantité en chlorophylle *a*. La deuxième méthode d'imagerie, a émergé via l'évolution des technologies d'imagerie microscopique. Celle-ci permet de décrire la morphologie de micro-organismes invisibles à l'œil nu. Elle a contribué à la découverte de nombreuses espèces planctoniques. De plus, il est possible de connaitre l'abondance et la composition taxonomique en plancton présent dans un milieu marin sans échantillonnage préalable. Cependant, il est parfois difficile d'identifier certains taxons de part les limites de résolutions et de détériorations lors du processus de préparation des échantillons.

Le développement des technologies de séquençage a grandement contribué à approfondir nos connaissances sur l'écosystème marin. Le séquençage de l'ADN consiste à déterminer l'ordre d'enchaînements des nucléotides d'un fragment d'ADN. Il permet d'obtenir, à partir du génome d'un individu, l'ensemble des informations issues des séquences d'ADN qui composent son matériel génétique. Cette méthode a été inventée dans la deuxième moitié des années 1970 dans un besoin de connaissance des gènes, étape indispensable à la compréhension des phénomènes biologiques au niveau moléculaire et cellulaire. Ainsi, de nombreuses espèces ont été séquencées depuis, dont le premier génome bactérien, *Haemophilius influenzae* en 1995[5] et le premier génome eucaryote, *Saccharomyces cerevisiae* en 1996[6]. Le séquençage de ces organismes nécessite néanmoins leurs mises en culture. Cependant, il est montré à cette époque que de nombreux organismes viables sont non cultivables. Le début du 21^{ième} siècle est marqué

par l'avènement de la nouvelle génération des techniques de séquençage qui visent à réduire le coût et le temps nécessaire pour séquencer un génome complet. Le séquençage du génome humain en 2001[7] puis 2004[8] et les découvertes qui s'ensuivent ont accéléré les progrès initiés dans ce domaine. Les séquenceurs à très haut débit sont maintenant capables de séquencer plusieurs dizaines de milliards de bases nucléiques par jour. L'augmentation de la quantité de données produites s'inscrit dans le développement du *big data* permettant de gérer et traiter de grandes quantités de données. Il est maintenant possible de séquencer une partie significative des fragments d'ADN d'un échantillon issu d'un environnement complexe et contenant le matériel génétique d'une multitude d'individus sans avoir à passer par une étape de culture. La métagénomique consiste en l'étude de l'ensemble de ces séquences d'ADN appartenant aux micro-organismes prélevés directement dans leur environnement[9]. C'est avec l'avènement des séquenceurs à haut débit et la gestion de *big data* que plusieurs projets de métagénomique ont émergé. Le séquençage du microbiome intestinal humain avec le projet metaHit [10] a été un des premiers projets de métagénomique sur l'humain. Il a permis de révéler un écosystème de micro-organismes jusque-là méconnus, notamment dût au fait que la plupart des espèces hébergées par notre système digestif ne sont pas cultivables in vitro. Notre intestin recèle près de 100 000 milliards de micro-organismes, soit dix fois plus que nos propres cellules. Ainsi, l'homme possède son propre écosystème puisque les microbes qui le composent vivent en symbiose avec lui et dépendent de facteurs génétiques et environnementaux. Cet écosystème longtemps occulté joue un rôle important dans de nombreux processus biologiques qui gouvernent notre existence. Plusieurs études de cette communauté de micro-organismes ont permis de montrer l'existence d'une situation d'interaction mutualiste entre le microbiote intestinal et l'hôte. Connaître cet écosystème est donc important pour comprendre les mécanismes qui nous régissent. Il en est de même pour l'écosystème océanique. Ainsi, de nombreux projets de recherche fondamentale en métagénomique océanique ont suivi. Ces projets avaient pour but de répertorier les espèces et les gènes qui n'avaient pu être révélés auparavant afin d'améliorer notre compréhension sur l'organisation des communautés microbiennes ainsi que le rôle écologique des micro-organismes qu'elles contiennent. Ainsi, les premières tentatives de reconstruction de génomes de micro-organismes non cultivés, ont été réalisées à partir de banques métagénomiques provenant de communautés virales issues d'océans et de fèces humains[11,12]. Craig Venter, qui a participé au séquençage du génome humain dans les années 2000, a également étudié l'écosystème microbien dans les océans. Il a tout d'abord mené un projet de métagénomique de la mer des Sargasses, au large des Bermudes[13]. Le but était de réaliser l'inventaire des gènes et des génomes microbiens présents dans l'océan de surface. Plus de deux millions de séquences ont été obtenues et annotées. Ce projet a révélé plus d'un million de séquences non répertoriées dans les banques de données ainsi que la grande

diversité de certaines familles de gènes. C'est ensuite que le *Venter Institue* a entrepris une campagne d'échantillonnages à l'échelle du globe. L'expédition *Global Ocean Sampling* (GOS) sur le *Sorcerer II* entre Février 2003 et Mai 2004[14] avait pour but de réaliser la comparaison des informations génomiques à grande échelle afin d'explorer la diversité microbienne des océans dans un contexte biogéographique. Des échantillons ont été collectés à la surface de l'eau dans une quarantaine de sites différents séparés de plus de 2 miles les un des autres afin d'obtenir des prélèvements d'environnement différents. Cette fois, 6.1 millions de protéines bactériennes et virales ont été annotées. C'est ainsi le premier projet à démontrer que le séquençage à haut débit peut permettre d'étudier les bactéries présentes dans les océans et ainsi de révéler la complexité en gènes de ces micro-organismes. Cependant, en ne prenant pas compte les organismes plus complexes que sont les eucaryotes, ce projet n'explore qu'une partie de l'écosystème marin. Les technologies ont continué d'évoluer en une dizaine d'années. L'accroissement de la profondeur de séquençage permet notamment de séquencer une partie plus importante des organismes présents dans un échantillon et permet d'observer une plus grande variété de planctons. De nombreuses expéditions ont suivi avec le déploiement de flottes de navires océanographiques mais aucun projet n'a refait d'échantillonnage sur l'ensemble des océans de la planète.

140 ans après l'expédition du *HMS Challenger*, l'expédition *Tara* Oceans (2009-2012) a pour but de réaliser une étude de l'écosystème du plancton des océans de surface à l'échelle de la planète afin de mieux connaitre et comprendre les mécanismes évolutifs et d'organisation qui régissent cet écosystème. Des prélèvements à bord de la goélette *Tara* connue pour ses expéditions dans les pôles, ont été réalisés avec les mêmes principes d'échantillonnages et une couverture mondiale à peu prêt similaire que le *Challenger* et que le *Sorcerer II*. Cependant, des outils et des méthodes plus complexes ont été utilisés pour la récolte des échantillons et pour les mesures physico-chimiques de l'environnement. Les sites d'échantillonages ont été choisis pour leurs complémentarités océanographiques. Ces échantillonnages ont été réalisés à différentes profondeurs, notamment en surface ainsi qu'au niveau où la concentration en chlorophylle est à son maximum et donc où le phytoplancton est abondant. Différents processus de filtration ont été utilisés afin de séparer les organismes en fonction de leurs tailles. Ainsi, les virus, les bactéries ainsi que les eucaryotes ont pu être séparés et ont été séquencés séparément. Parmi les eucaryotes, 4 fractions de tailles ont été obtenues pour étudier les planctons allant des protistes unicellulaires aux zooplanctons multicellulaires. Afin de pouvoir réaliser des études de génomique comparative, c'est-à-dire de comparer les génomes contenus dans différents échantillons, le même protocole d'échantillonnage a été utilisé sur l'ensemble des prélèvements. C'est ainsi plus de 35 000 échantillons qui ont été récoltés avec l'aide des avancées technologiques dans les domaines cités précédemment. Via les images satellites, la goélette a

effectuée des prélèvements dans des zones précises. Des échantillonnages ont été réalisés dans des stations ayant des phénomènes physico-chimiques particuliers. C'est le cas des *upwellings*, où les vents font remonter l'eau profonde, froide, riche en nutriments et en planctons à la surface. Des prélèvements ont été également effectués dans le cœur de tourbillons océaniques qui sont connus pour emprisonner et transporter le plancton sur de longues distances, parfois, entre deux océans. La Goélette a suivi, quand cela était possible, le trajet des grands courants marins, comme les gyres océaniques, afin de pouvoir étudier l'impact des déplacements de masses d'eau sur l'organisation et l'évolution des communautés de planctons. Comme pour les autres projets océanographiques, les paramètres physico-chimiques ont été mesurés dans chaque zone de prélèvement. Enfin, de nouveaux protocoles ont été mis en place lors de l'expédition. C'est le cas de la reconstruction partielle de génomes d'organismes non cultivables par le prélèvement et le séquençage de cellules uniques. Le séquençage de l'ensemble de ces échantillons a été réalisé au Genoscope qui est le Centre National de Séquençage. L'utilisation des séquenceurs à haut débit de dernière génération a permis d'obtenir une profondeur de séquençage au delà de ce qui avait pu être réalisé auparavant. Ainsi, une partie plus importante des séquences génomiques des micro-organismes prélevés dans leur environnement, et ce, dans l'ensemble des océans du globe est maintenant à disposition. C'est la première fois qu'il est possible de combiner cette quantité massive de données génomiques avec les mesures physico-chimiques et les modèles d'océanographie physique pour une étude de l'écosystème planctonique à l'échelle de la planète. C'est donc par une approche multidisciplinaire que ce projet d'envergure veut tenter de résoudre les questions que nous nous posons sur l'écosystème océanique. Toutefois, cette étude interdisciplinaire est un réel défi pour les scientifiques qui doivent développer des méthodes et des outils pour analyser les données mesurées et générées.

Ainsi, les expéditions océanographiques qui se sont succédées aux cours de ces deux derniers siècles et les progrès technologiques et méthodologiques qui ont été mis place ont permis d'améliorer nos connaissances sur l'écosystème planctonique et de prendre conscience de l'importance de celui-ci sur l'équilibre de notre planète. En effet, les observations de Darwin et l'expédition *HMS Chalenger* ont mis en évidence la diversité des micro-organismes planctoniques ainsi que de l'environnement océanique. Ensuite, l'avènement de la métagénomique avec l'évolution des technologies de séquençage à haut débit ont permis lors de l'expédition *Global Ocean Sampling* (GOS) de réaliser l'étude des micro-organismes planctoniques dans leur environnement. Une proportion importante d'espèces et de gènes non répertoriés a été observée sans pour autant l'estimer. De plus, ce projet a révélé la complexité en gènes de ces micro-organismes qui sont en étroite interaction avec l'environnement biologique, physique et chimique. Le projet *Tara* Oceans est la première expédition à avoir réalisé une collecte et un séquençage exhaustif de l'ensemble des micro-organismes planctoniques des

océans de surface à l'échelle de la planète tout en intégrant les mesures des différents facteurs environnementaux et des modèles pour l'étude du transport de particules le long des courants marins. Cependant, les progrès du séquençage à haut débit ont permis d'augmenter la profondeur de séquençage pour obtenir la quasi-totalité des séquences génomiques des organismes présents dans un échantillon. Avec plus de 35 000 échantillons séquencés, la quantité de données à analyser est de l'ordre du péta-octet. Les outils actuellement développés pour les analyses en métagénomique comparative ne sont pas adaptés à cette masse de données. De plus, l'utilisation de la métagénomique pour étudier la diversité d'un écosystème est récente et demande encore à être évaluée par rapport aux autres méthodes utilisées actuellement. Enfin, cette approche multidisciplinaire qui intègre l'étude des courants océaniques, des paramètres physico-chimique et des données génomiques correspondant aux différents organismes micro planctoniques à l'échelle de la planète n'a encore jamais été réalisée. Cela demande d'adapter les outils et les méthodes existants voir même d'utiliser de nouvelles méthodes pour explorer cet écosystème océanique.

Cette thèse consiste à étudier l'écosystème planctonique des eaux de surface océanique en intégrant dans les analyses les données génomiques, métagénomiques, les paramètres physico-chimiques correspondant aux échantillons récoltés lors de l'expédition *Tara* Oceans ainsi que les courants marins. L'objectif étant de mieux comprendre dans quelle mesure les facteurs océaniques impactent la dynamique de la structure des génomes et de leur biogéographie dans l'océan.

La démarche suivie dans cette thèse

Thomas Kuhn, dans son livre *La Structure des révolutions scientifiques* publié en 1962, a montré que la science progresse, non pas dans une ascension continue, mais de façon révolutionnaire, par transformation des principes d'explication ou paradigme. Face à cet océan d'ignorance que représente l'écosystème planctonique, les scientifiques issus de disciplines variées travaillent de concert pour amorcer cette révolution. Ces interactions ont fait naître un système complexe nécessitant la compréhension du rapport entre différentes échelles (des gènes aux micro-organismes, d'une micro-algue isolée à une communauté planctonique), des rapports entre systèmes différents (de l'environnement physico-chimique aux courants marins, de régions biogéochimiques aux *genocenoses*) ou encore entre des temps différents (d'un *turn-over* d'une communauté au temps de déplacement d'une particule dans un courant). Cette thèse de Bio-informatique s'inscrit dans un champ de recherche multidisciplinaire et c'est donc par

une approche regroupant des échelles, des systèmes et des temps différents que j'ai travaillé sur la problématique qui m'a été posée.

Dans ce manuscrit, je présenterai dans un premier temps le contexte biologique et méthodologique dans lequel s'inscrit cette thèse. Trois parties présenteront les avancées dans les domaines et les disciplines ainsi que les outils et les méthodes qui m'ont permis de mener à bien mon projet de recherche. Dans un deuxième temps, les résultats obtenus, pour la plupart valorisés durant cette thèse sous forme d'articles scientifiques, seront présentés en trois chapitres. Un premier chapitre présente des analyses qui consistent à étudier la diversité des espèces ainsi que de leurs gènes via les données métagénomiques. Un second chapitre présente l'évaluation d'une méthode et l'amélioration d'un outil permettant de réaliser un regroupement d'échantillons possédant des similarités génomiques afin de permettre de passer à l'échelle de l'étude des communautés d'organismes dans l'analyse de la diversité planctonique. Un troisième chapitre présente une analyse sur l'étude de l'impact des courants océaniques et des paramètres physico-chimiques sur l'organisation des communautés de microplanctons. La discussion des résultats obtenus, ainsi que les perspectives envisageables concluent ce manuscrit.

Contexte biologique et méthodologique

Ce chapitre a pour objectif de présenter un état de l'art des domaines et des disciplines étudiés pendant cette thèse. Les outils, les données et les méthodes qui ont permis de répondre aux questions soulevées pendant cette thèse seront également présentés. Ce chapitre est constitué de trois parties. La première partie consiste à présenter le plancton et les différents acteurs qui composent l'écosystème planctonique ainsi que les moyens mis en place pour les étudier. La seconde partie présente l'évolution des technologies de séquençage et l'avènement de la métagénomique permettant d'étudier la génomique des communautés de micro-organismes planctoniques prélevés directement dans leur environnement. Enfin, dans la troisième partie, seront présentés les différents outils et les méthodes utilisés pour étudier l'organisation génomique de communauté de micro-organismes planctoniques dans les océans de surface.

I. Le plancton et l'écosystème planctonique

L'étendue de l'océan, son caractère dynamique et la difficulté technique d'échantillonner le milieu océanique a contribué au fait que celui-ci a été plus tardivement étudié et pris en compte comme un écosystème jouant un rôle clef, au même titre que l'écosystème terrestre, dans les équilibres nécessaires à la vie sur Terre. Je définirai dans cette partie le plancton et présenterai les différents acteurs qui composent l'écosystème planctonique ainsi que les techniques et les moyens mis en place pour les étudier.

I.1 Le plancton : un rôle clef dans les équilibres nécessaires à la vie sur Terre

I.1.1 Qu'est-ce que le plancton?

Le terme « plancton » vient du grec *planktós* signifiant flottant ou errant. Homère l'employait dans l'Odyssée pour désigner les animaux surnageant à la surface de la mer. Ce terme familier à Aristote et à d'autres auteurs de l'antiquité, a été repris dans l'Océanographie par Hensen en 1887[15] pour définir « tout ce qui flotte dans l'eau ». Cette définition trop vaste, a été précisée depuis, et on comprend actuellement sous le nom de plancton « l'ensemble des organismes vivants, de nature végétale ou animale, n'ayant pas d'attaches directes avec le sol, et passant leur vie, entièrement ou partiellement, dans le milieu liquide, dans lequel ils flottent plus ou moins passivement »[16]. Les organismes planctoniques sont généralement mobile, soit par la contraction de leurs corps, soit via leurs organes locomoteurs, les cils ou encore les flagelles dont ils sont pourvus. Cependant, ces mouvements, trop faibles pour lutter contre les courants, servent principalement à les maintenir en état de flottabilité. Ils sont parfois capables de déplacements verticaux pouvant atteindre plusieurs centaines de mètres d'amplitude[17]. Ceux-ci sont facilités par divers moyens, externes, morphologiques ou internes. Les formes extérieures des composants du plancton sont variées mais peuvent être représentées en cinq types principaux. La première comprend les formes sphériques, vésiculaires ; la deuxième les formes aplaties, discoïdales ; la troisième les formes bacillaires ; la quatrième les formes rubanées et la cinquième les formes échinoïdales, c'est-à-dire sphéroïdales munies à la périphérie d'épines, de piquants, de spicules ou de cornes. Pour les dispositifs internes, les planctons possèdent parfois des vacuoles remplies d'air ou de gaz, remplacées chez certains par des flotteurs contenant les mêmes éléments gazeux. A la fin des années 1970, le plancton a été catégorisé en différentes

classes de tailles[18] allant du pico-plancton au méga-plancton. Cependant, cette fourchette d'ordres de grandeur n'intègre pas les virus. Ces dernières années, de nombreuses découvertes sur la structure[19,20], les modes et la capacité des virus à interagir avec les différentes branches du vivant[21] et entre eux[22] supposent que ces derniers ont un rôle important dans l'organisation des communautés planctoniques. Le plancton est donc constitué d'organismes de tailles très variables allant de plusieurs mètres de long, comme certaines méduses, les *Siphonophores Apolemia* ou encore les chaînes de Salpes, à moins d'un pico-mètre comme c'est le cas pour certains virus (Figure I.1.1). Parmi le plancton, on retrouve les micro-organismes planctoniques qui peuvent être décrits en cinq groupes d'organismes : les virus, les bactéries, les archées, les protistes et les métazoaires. Dans le cadre de cette étude, ces différents groupes ont été séparés en fonction de leurs tailles de manière arbitraire en six fractions taille qui seront détaillées dans la partie I.4.4.

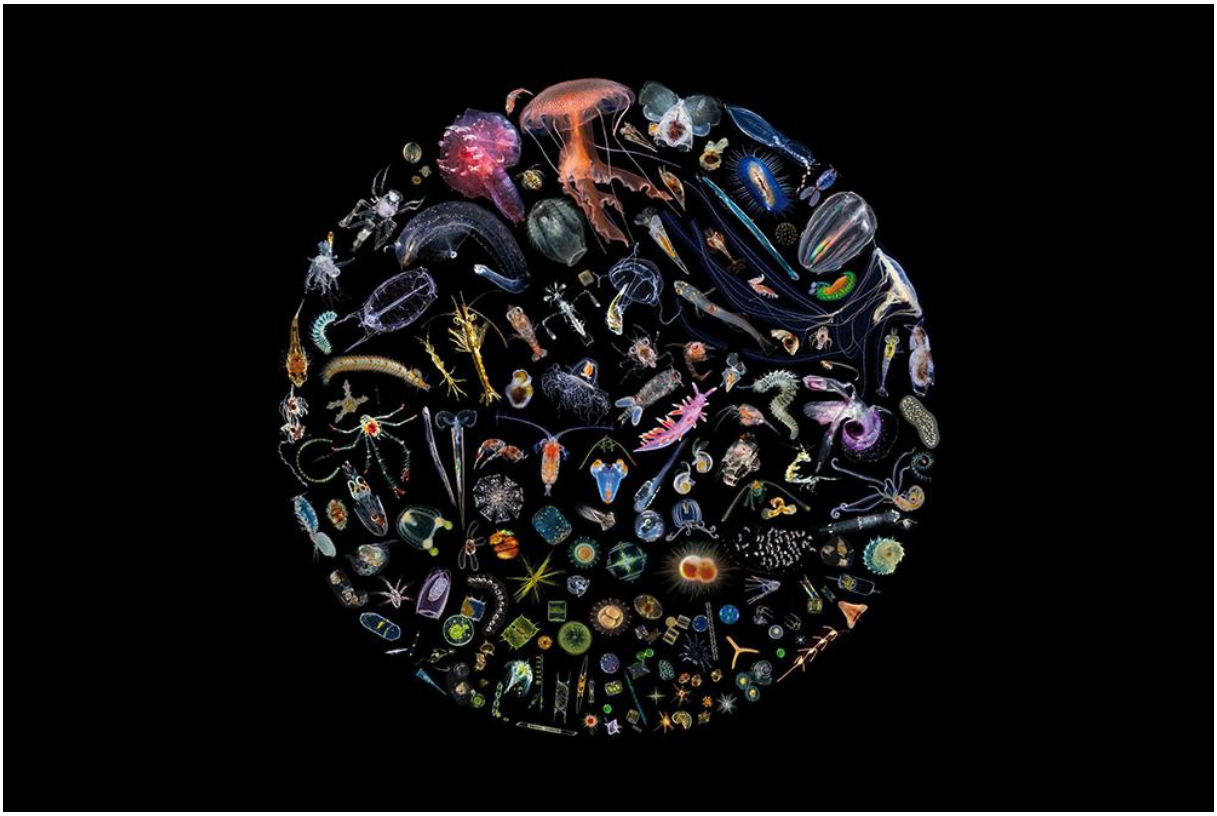


Figure I.1.1 | Mandala du plancton. Dans la partie supérieure figurent les organismes les plus volumineux (méduses, siphonophores, cténophores, salpes) faisant partie du zooplancton. Dans la partie centrale les organismes zooplanctoniques dont les tailles varient de quelques millimètres à plusieurs centimètres (chaetognathes, annélides, ptéropodes, copépodes). Dans la partie inférieure on retrouve les organismes microscopiques mesurant moins d'un millimètre (protistes, radiolaires, foraminifères, diatomées, dinoflagellés). Les virus et les bactéries ne sont pas représentés. (Figure extraite de Sardet 2013[23])

I.1.2 Importance des micro-organismes planctoniques

Nous pouvons distinguer deux sortes de micro-organismes planctoniques marins, le Zooplancton et le Phytoplancton. Ces deux types de planctons, par leur abondance dans les océans, ont un rôle majeur dans les cycles biogéochimiques de notre planète et dans notre alimentation. L'importance du plancton a été reconnue au Moyen Age par les pêcheurs italiens chez lesquels existait le dicton « che i pesci crede, che sia plancton », ce qui veut dire : qui dit poissons, dit plancton. En effet, la matière organique que le phytoplancton fabrique constitue de la nourriture pour le zooplancton herbivore puis pour les carnivores, pour finir en ressource nutritive pour les grands prédateurs parmi lesquels l'Homme constitue le haut de la chaîne trophique[24]. Le plancton est donc à la base de la chaîne alimentaire marine. De plus, le phytoplancton regroupe des micro-algues unicellulaires ainsi que des bactéries photosynthétiques. Ces cellules végétales vivent à la surface des océans, dans la couche euphotique. Même s'ils ne représentent que 1 à 2% de la biomasse végétale sur la Terre, ces organismes microscopiques produisent à eux seuls autant d'oxygène que toutes les forêts et plantes terrestres[25]. De plus, le plancton est à l'origine des énergies fossiles telles que le charbon, le pétrole et le gaz naturel. Il est donc intimement lié à nos activités humaines et est à travers le cycle du carbone[26] un important régulateur de la machine climatique ainsi que de l'acidité des océans puisqu'il absorbe environ 70% du CO2 que nous produisons[1]. L'écosystème planctonique est un système complexe qui regroupe des organismes planctoniques interagissant entre eux via le parasitisme, la prédation, la symbiose ou la compétition, mais également avec leur environnement. Les virus en agissant sur l'expression des gènes de leurs hôtes[21] ont un rôle sur les processus photosynthétiques. Le plancton peut être décrit comme une biocénose, dans laquelle les composants dépendent les uns des autres, et dont l'existence et l'abondance se trouvent en rapport étroit avec les caractères du milieu ambiant. La diversité planctonique et cette biocénose ont été révélées au moyen des pêches planctoniques et des diverses expéditions océanographiques qui seront pour les plus importantes décrites dans la partie I.4.

I.2 Classification phylogénétique des espèces planctoniques

Darwin est le premier à avoir illustré et popularisé le concept d'un arbre de la vie en 1859[27]. Encore maintenant, les biologistes de l'évolution utilisent des diagrammes en forme d'arbre pour décrire l'évolution des espèces. Je présenterai dans cette partie l'organisation phylogénétique des micro-organismes planctoniques en développant particulièrement le groupe des mamiellales dont fait partie *Bathycoccus* qui a été un organisme d'étude pendant cette thèse.

I.2.1 La séparation des branches du vivant

Linné au XVII^{e} siècle a établi les bases d'une classification du vivant en nommant les espèces. La notion d'espèce est basée sur les capacités de reproduction sexuée produisant des descendants fertiles. Aujourd'hui, de nouvelles espèces ne cessent d'être découvertes. D'autant plus qu'avec les progrès de la génétique, il est maintenant possible de différencier les espèces cryptiques, c'est-à-dire, les espèces qui ne sont pas distinguables d'un point de vue morphologique. La phylogénie est l'étude des liens de parenté entre les organismes. Parmi les groupes de micro-organismes qui constituent le plancton, les bactéries et les archées, les procaryotes, étaient déjà présentes lors de l'Éoarchéen, soit il y a plus de 3600 millions d'années. Ces cellules microscopiques ne possèdent ni noyau, ni membrane interne, sauf pour l'embranchement des cyanobactéries qui ont une invagination de la membrane externe. Les picocyanobactéries sont majoritairement composées des genres Prochlorococcus et Synechococcus. Par leur abondance, ils sont fortement représentés dans le phytoplancton et ont donc un rôle important dans la production primaire[28]. C'est ensuite que des cellules plus complexes contenant un noyau et d'autres organites délimités par des membranes ont fait leurs apparitions. Le réticulum endoplasmique et la membrane nucléaire auraient sans doute évolué à partir de l'invagination de la membrane externe. La théorie endosymbiotique[29] propose une évolution des mitochondries et des chloroplastes à partir des bactéries ingérées et restées presque intactes. Ainsi, les mitochondries et les chloroplastes ont leur propre ADN qui ressemble à celui des procaryotes. Les eucaryotes comprennent les protistes, les métazoaires et les végétaux pluricellulaires. Les protistes sont unicellulaires et comprennent entre autres des organismes autotrophes, c'est-à-dire qu'ils effectuent la production de matière organique par réduction de matière inorganique, comme c'est le cas pour les cyanobactéries qui sont des microalgues. D'autres sont hétérotrophes comme les protozoaires qui ingèrent leur nourriture par phagocytose. Enfin, une grande majorité des protistes sont mixotrophes. Ces derniers sont alors capables de se nourrir par les deux modes trophiques. Les protistes ne forment pas un groupe monophylétique. Les métazoaires désignent, par opposition aux protozoaires, les eucaryotes multicellulaires. La plupart sont mobiles et hétérotrophes.

I.2.2 Classification phylogénétique des eucaryotes

Dans la dernière classification phylogénétique suite aux découvertes de nouvelles espèces, on peut caractériser parmi les eucaryotes cinq grands groupes[30]. Les opisthochontes (Opisthokonta) regroupent des organismes très divers en apparence et comprennent les champignons et les métazoaires. Les Excavés (Excavata), qui signifie creusés, sont des protistes hétérotrophes, généralement flagellés et rassemblent une grande variété de cellules libres et symbiotiques dont certaines sont des parasites pour l'Homme. Les amibozoaires (Amoebozoa)

du grec amoibē signifiant *transformation*, constituent un grand groupe de protozoaires simples qui se déplacent pour la majorité par vagues cytoplasmiques internes. Le SAR (Harosa) est un groupe constitué des Stramenopiles, des Alveolata et des Rhizaria. La première lettre du nom de chaque groupe explique l'origine du nom SAR. Enfin, le groupe des Archaeplastida, ou lignée verte, regroupent des êtres vivants capables de faire de la photosynthèse (Figure I.2).

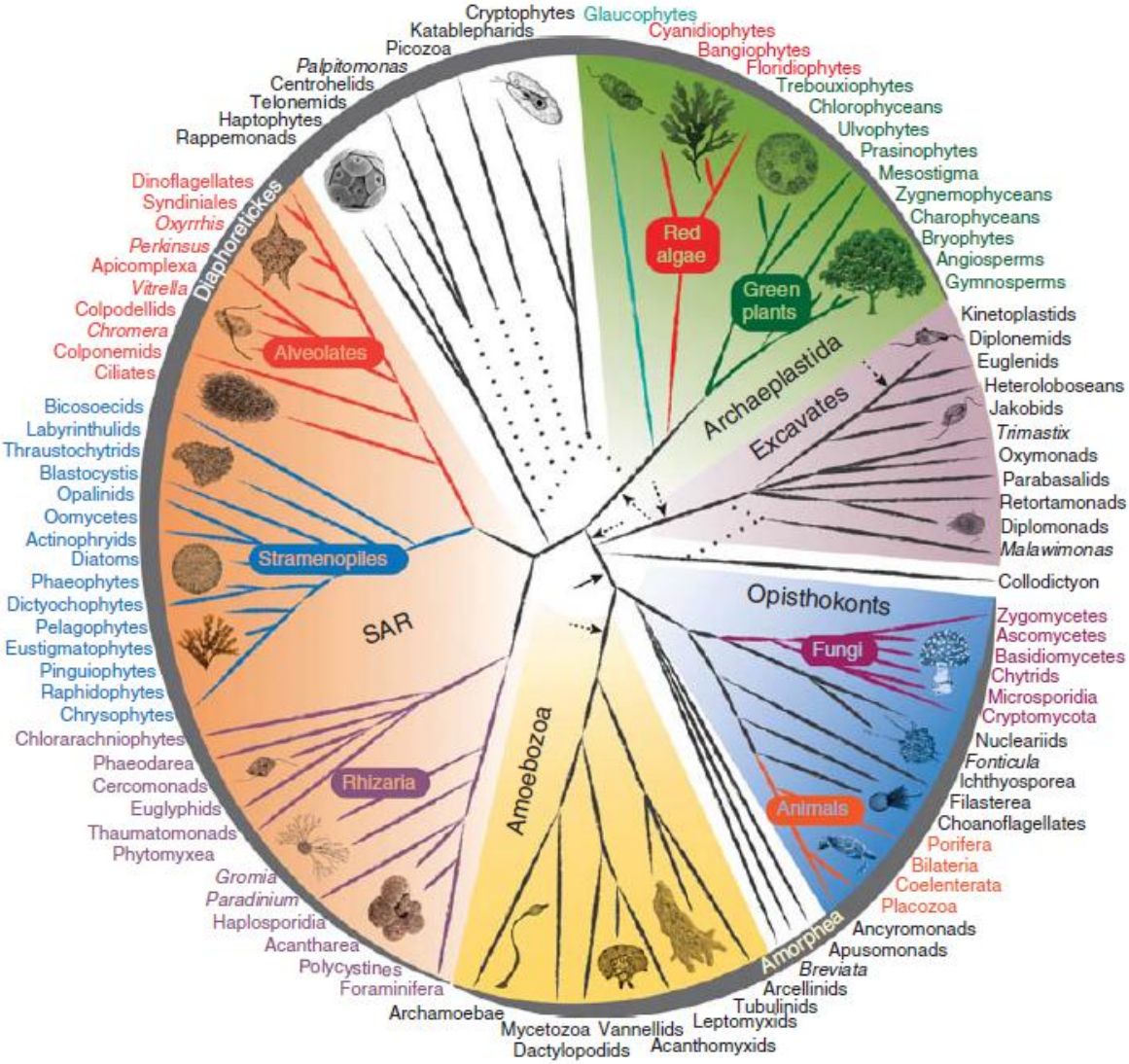


Figure I.2 | Phylogénie des eucaryotes. (Figure extraite de Burki 2014[30])

I.2.3 Les algues vertes, composantes importantes du phytoplancton

Tous les organismes eucaryotes photosynthétiques proviennent d'un évènement d'endosymbiose entre un eucaryote et une cyanobactérie[31]. Ce phénomène d'endosymbiose serait un évènement unique qui aurait eu lieu il y a environ 1,56 milliard d'années[32]. De cet évènement sont issues trois lignées anciennement regroupées sous le terme « Règne Végétal » et désormais désignées par le groupe des Archaeplastida. Ces lignées sont constituées par les Chloroplastida (anciennement chlorophytes comprenant les algues vertes et les plantes

terrestres), les Rhodophyceae (anciennement rhodophytes qui regroupent les algues rouges) et les Glaucophyta (anciennement glaucophytes ou glaucocystophytes). Les picophytoplanctons de moins de 2 à 3µm sont les contributeurs d'environ la moitié de la production primaire et de la biomasse globale[25]. Le groupe des Archaeplastida est un acteur important dans cette production, notamment avec les algues vertes.

I.2.4 Les Prasinophyceae

Parmi les algues vertes unicellulaires ou multicellulaires possédant des pigments chlorophylliens on dénombre actuellement quatre classes : les Prasinophyceae, les Chlorophyceae, les Trebouxiophyceae, et les Ulvophyceae. Les Pedinophyceae ont été récemment considérés comme une classe indépendante des algues vertes[33]. L'étymologie du nom Prasinophyceae vient du mot grec « prasinos » qui désigne la couleur « verte poireaux » de ces algues. La classe des Prasinophyceae regroupe des taxons très divers, appartenant à des lignées différentes, mais considérés comme les algues vertes ayant conservé le plus de caractères primitifs[34]. Leur phylogénie a tout d'abord été fondée sur des caractères ultrastructuraux comme la présence d'écailles, la présence et la position de flagelles. Les études de biologie moléculaire consistant à analyser les séquences du gène de l'ARNr 18S ont contribué à l'amélioration de cette classification en intégrant des organismes possédant d'autres caractéristiques morphologiques comme l'absence de flagelle chez *Bathycoccus*, l'absence d'écaille chez *Micromonas* ou encore l'absence d'écaille et de flagelle chez *Pycnococcus.* Cependant, la taxonomie de ce groupe reste en constante évolution par la diversité des nouvelles zones d'échantillonnages et l'évolution des méthodes utilisées pour d'écrire de nouvelles espèces. On peut tout de même considérer aujourd'hui une dizaine d'ordres[35–39] dans la classe des Prasinophyceae : l'ordre des Pyramimonadales (clade I), l'ordre des Pseudoscourfieldiales, l'ordre des Chlorodendrales (clade IV), les clades VII, VIII et IX, l'ordre des Prasinococcales (clade VI), et enfin l'ordre des Monomastigales, Dolichomastigales et des Mamiellales qui constituent la lignée importante des Mamiellophyceae[39–43]. L'étude des séquences de la petite sous unité 18S du gène de l'ADN ribosomal a mis en exergue notre manque de connaissance sur la distribution des Mamiellophyceae[44]. Pourtant, les phytoplanctons constituant cette lignée occupent un large panel d'environnements, des pôles jusqu'aux tropiques[41,45–48]. Leur temps de génération rapide et leur implication dans la production primaire font que ces algues vertes sont écologiquement importantes[49]. De plus, elles ont un rôle non négligeable sur l'équilibre de l'écosystème océanique et sont fortement impactées par les perturbations environnementales[50,51]. L'étude de la biogéographie et de la diversité des Mamiellophyceae est donc importante afin de comprendre comment la chaîne trophique et le cycle du carbone océanique peuvent être affectés par ces micro-organismes. Les données 18S ont permis

d'étudier leur distribution dans des environnements restreints[37]. Afin de connaitre leur distribution géographique globale et leur diversité phylogénétique il est nécessaire de réaliser un échantillonnage dans des environnements plus divers avec un protocole d'échantillonnage et de séquençage unique. Une étude récente a étudié la diversité et la biogéographie des Mamiellophyceae avec les données récoltées lors de l'expédition *Tara* Oceans[52]. Cette expédition sera décrite dans la partie I.4.4. Une classification phylogénétique des régions V9 de l'ARNr 18S des Mamiellophyceae les plus abondants a été réalisée et leur abondance relative dans les stations d'une partie du projet *Tara* Oceans calculée (Figure I.2.4). Plusieurs études ont cependant établi que ce marqueur phylogénétique n'était pas adéquat pour décrire la diversité de cette communauté picoplanctonique[53].

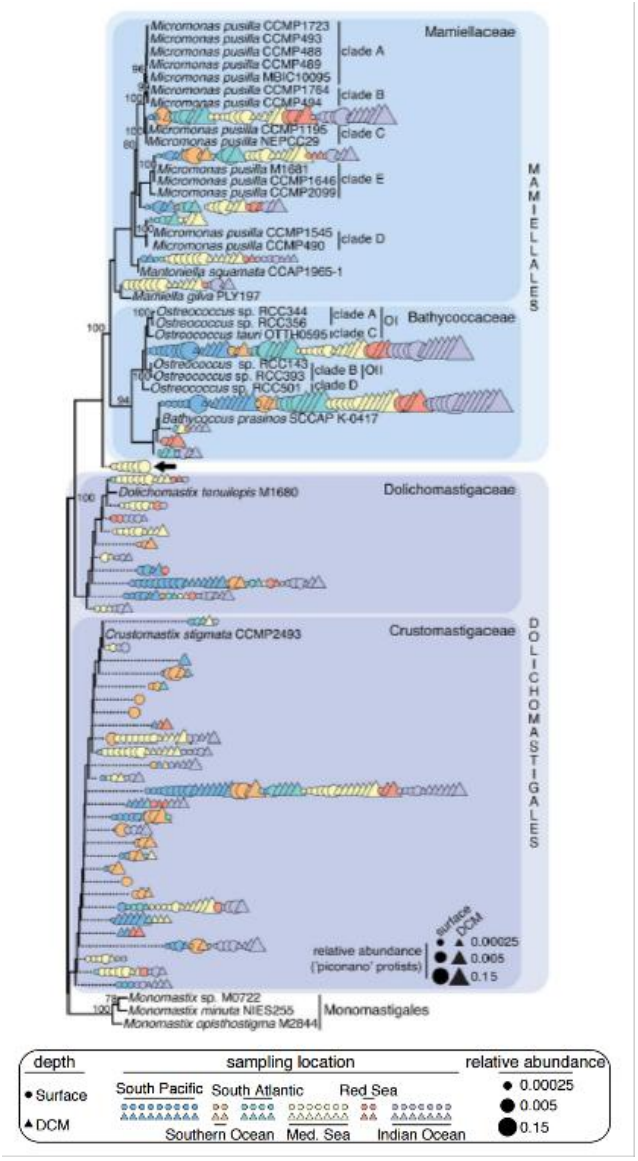


Figure I.2.4 | Classification phylogénétique et abondance relative des Mamiellophyceae. La classification à été obtenue à partir des séquences de la région V9 de l'ARNr 18S sur une partie des stations du projet *Tara* Oceans. (Figure extraite de Monier *et al.* 2016[52])

I.2.5 Les Mamiellales

L'ordre des Mamiellales est composé des genres *Ostreococcus*, *Bathycoccus*, *Micromonas*, *Mantoniella*, *Mamiella*, *Dolichomastix*, *Crustomastix* et *Monomastix*. Le genre Monomastix est le plus éloigné phylogénétiquement des autres et est le seul à contenir des espèces d'eau douce[39]. Les Prasinophyceae planctoniques marins comprennent trois genres principaux, *Ostreococcus sp.*[54], *Bathycoccus sp.*[55] et *Micromonas sp.*[56]. Ces algues vertes unicellulaires marines servent de modèles pour l'étude de l'évolution et de l'écologie de ce groupe. En effet, la plupart des séquences du gène de l'ARNr 18S assignées Mamiellales trouvées dans l'environnement marin appartiennent à ces algues vertes[50]. *Ostreococcus sp.*, *Bathycoccus sp.* et *Micromonas sp.* ont tous les trois un plastide et une mitochondrie mais ont des tailles et des morphologies différentes (Figure I.2.5.i).

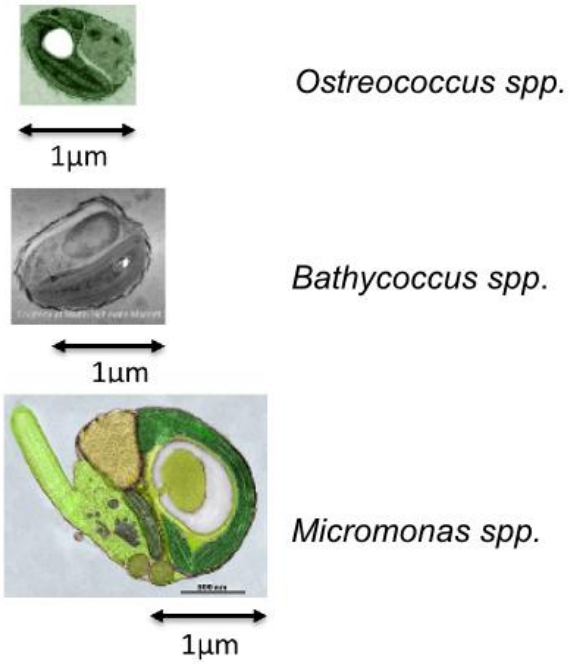


Figure I.2.5.i | Les trois principaux Prasinophyceae planctoniques marins.
(Image adaptée d'après http://genome.jgi.doe.gov, http://bioinformatics.psb.ugent.be et http://roscoff-culture-collection.org)

Ostreococcus

Ostreococcus est le plus petit organisme photosynthétique libre connu actuellement[57] avec une taille pour *Ostreococcus Tauri* de 0.8µm[54]. Il se différencie des deux autres genres par sa cellule nue alors que *Bathycoccus* est recouvert d'écailles polysaccharidiques et *Micromonas* est doté d'un flagelle. Ces trois genres de Mamiellales présentent une certaine diversité génétique avec différentes espèces au sein d'un groupe qui sont adaptées à différentes régions, profondeurs et environnements physico-chimiques. Ainsi, trois espèces d'*Ostreococcus* ont été décrites : *Ostreococcus tauri*[54], *Ostreococcus mediterraneus*[58] et *Ostreococcus lucimarinus*[59] et une

dizaine de souches ont été caractérisées[60]. Parmi ces clades, plusieurs souches semblent adaptées à des intensités lumineuses différentes mais ces préférences écologiques impliquent probablement d'autres paramètres environnementaux comme la température ou les nutriments[47]. *Ostreococcus* possède le plus petit génome eucaryote photosynthétique connu (environ 12 Mb). Les génomes de quatre espèces sont disponibles actuellement. *Ostreococcus tauri* possède le plus petit génome avec 20 chromosomes[61]. Il a été prélevé dans le lagon de Thau en France en 1994, son génome a été séquencé au laboratoire Arago à Banyuls. *Ostreococcus lucimarinus* a quant à lui 21 chromosomes, compte environ 13Mb et à été séquencé au JGI sous la dénomination CCE9901[59]. *Ostreococcus mediterraneus*[58], précédemment connu sous le nom d'*Ostreococcus* clade D possède 20 chromosomes. Des analyses de la sous unité de l'ARN ribosomal et des comparaisons de la structure secondaire de l'ITS2 (*internal transcribed spacer 2*) ont récemment permis de montrer qu'il s'agissait bien d'une espèce distincte. Enfin, un dernier génome est disponible à la Roscoff Culture Collection, Ostreococcus RCC809.

Micromonas

Le genre *Micromonas*, ubiquitiste et abondant dans l'écosystème marin, a été initialement décrit comme une seule espèce sous le nom de *Chromulina pusilla*[56]. La dénomination latine *pusilla* s'explique par sa petite taille. Le nom de *Micromonas pusilla* lui à été attribué ensuite et l'étude de nombreuses souches a montrée l'existence de plusieurs espèces[45]. En effet, l'analyse comparée des génomes de deux souches de *Micromonas pusilla* a permis de montrer que malgré leurs morphologies identiques, ces deux souches présentent des génomes avec plus de 10% de différences globales et représente donc deux espèces distinctes[62]. L'analyse de ARN ribosomal 18S a montré que différents clades de *Micromonas* sont présents dans des environnements variés[62,63]. Par exemple, un écotype de *Micromonas* est restreint aux eaux polaires[46,64]. Deux génomes, d'environ 21 Mb de *Micromonas* sont disponibles : *Micromonas pusilla* CCMP1545 avec 21 chromosomes et *Micromonas sp.* RCC299 avec 17 chromosomes. Des séquences de répétition d'intron fortement conservés ont été révélées chez CCMP1545 alors qu'ils sont absents dans RCC299.

Bathycoccus

Bathycoccus à été prélevé initialement à 100 mètres de profondeur au niveau DCM dans la mer méditerranéenne[55]. Des cellules recouvertes d'écailles polysaccharidiques et ayant une même morphologie ont été également observées dans l'Océan Atlantique[65]. *Bathycoccus* a ensuite été trouvé dans différents environnements océaniques, ce qui fait de lui un organisme cosmopolite. Il a été en particulier prélevé dans des eaux côtières[66,67]. Le génome de la souche côtière *Bathycoccus prasinos* RRC1105 est actuellement disponible[68]. Celle-ci est composée de 19 chromosomes et d'environ 15 Mb. Les chromosomes 14 et 19 sont dits « outlier » du fait de leurs différences structurales et fonctionnelles en comparaison des autres chromosomes (Figure I.2.5.ii). En effet, ces deux chromosomes sont de petite taille, la proportion de leur contenu en bases guanine-cytosine est faible et leur contenu en gènes de transfert horizontaux est élevé par rapport aux autres chromosomes. Le chromosome 19, appelé SOC pour « small outlier chromosome » est le plus petit des deux. Il ne possède pas de colinéarité avec le génome des autres Mamiellales et contient beaucoup de gènes sans homologues avec les autres espèces. Le chromosome 14, dit BOC pour « big outlier chromosome » a un contenu en introns et un niveau d'expression plus élevés que les autres chromosomes. Les données métagénomiques ont suggéré l'existence de deux écotypes de *Bathycoccus*[40,41,48], récemment nommé BI et BII. Ces deux écotypes ont les mêmes séquences de l'ARN ribosomal 18S et ne peuvent donc pas être différenciés en utilisant les marqueurs comme les régions V4 ou V9 des gènes de l'ARNr 18S[48]. Actuellement, un seul génome de *Bathycoccus* est disponible dans les bases de données publiques. Il n'est donc pas possible d'observer la répartition géographique et les préférences écologiques de ces deux écotypes ni de comparer leur structure génomique afin d'étudier les mécanismes d'évolution qui régissent ces algues vertes unicellulaires. Récemment, un second génome de *Bathycoccus* a été généré par la méthode d'amplification à cellule unique dans le cadre de l'expédition *Tara* Oceans. Cette méthode appelée Single-cell Amplified Genome (SAG) sera présentée dans la partie II.2.3. Ce génome partiel de *Bathycoccus* possède une longueur de 10.3 Mb et une couverture d'assemblage estimée à 64% du génome de référence. Les cellules ayant le même ARN 18S que la souche RCC1105 ont été récoltées dans la mer d'Arabie à la station 39 du projet *Tara* Oceans d'où sa nomination TOSAG-39-1 (premier SAG de la station 39 du projet *Tara* Oceans).

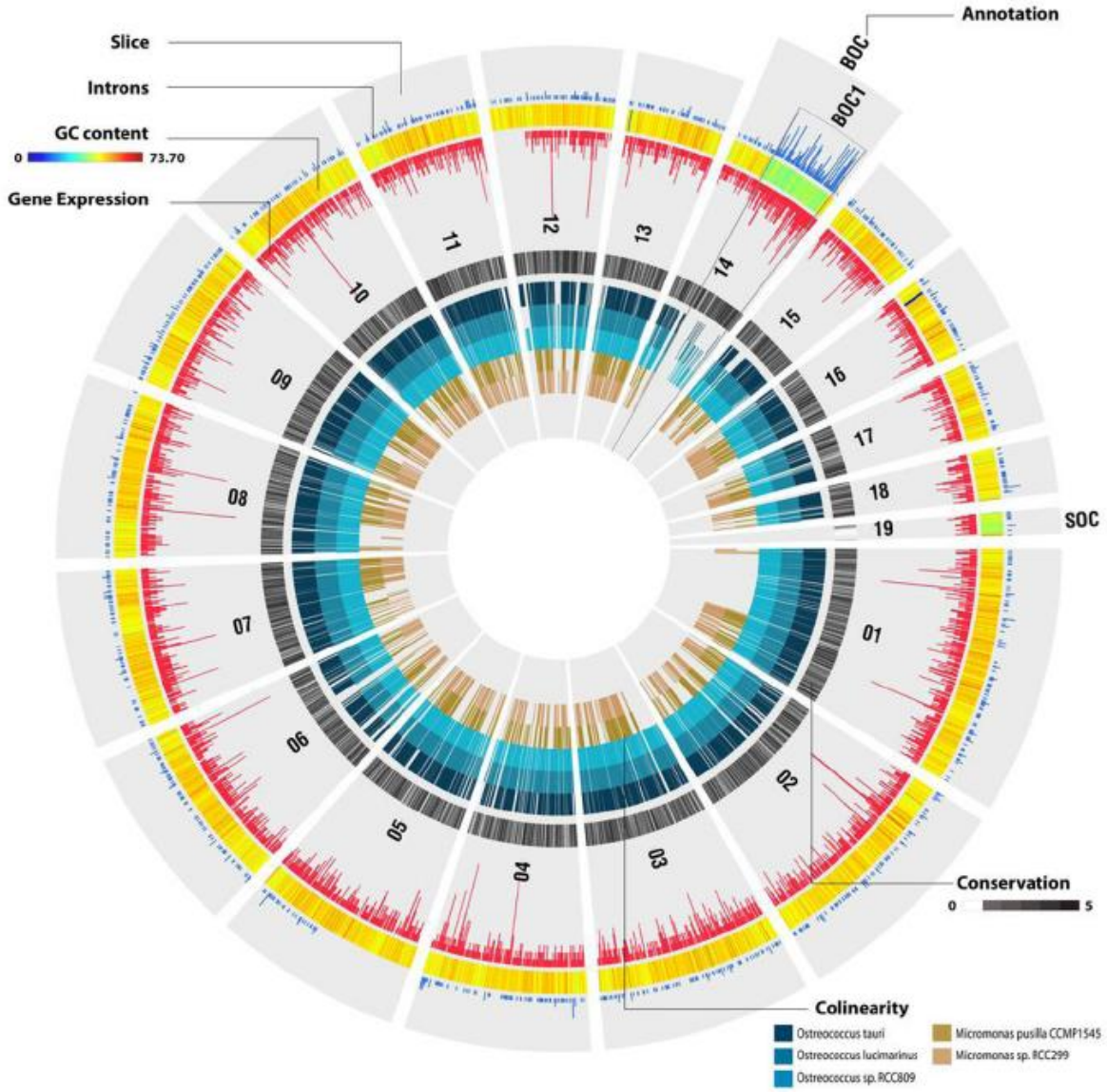


Figure I.2.5.ii | Représentation comparative des caractéristiques structurales et fonctionnelles du génome de la souche RCC1105 de *Bathycoccus* prasinos.(Figure extraite de Moreau *et al.* 2012[68])

Malgré l'importance écologique supposée de ces différents genres de Mamiellales, il reste encore certaines espèces non connues en raison de leur petite taille qui, souvent ne permet pas de repérer leurs différences morphologiques[39,45,58]. De plus, le manque de résolution des marqueurs phylogénétiques couramment utilisés pour les étudier[53] est également un obstacle à leur caractérisation.

I.3 L'écosystème planctonique

I.3.1 Qu'est-ce que l'écosystème planctonique ?

Le concept d'écosystème a été introduit en 1935 par Tansley[69]. L'écosystème peut être défini par trois concepts[70]. Le concept de population consiste à étudier comment chaque espèce se répartit dans un milieu donné. Le concept de biocénose consiste à examiner l'ensemble des interactions entre espèces vivantes dans un milieu donné et à décrire leur dynamique et leur évolution. Ce terme biocénose vient du grec bios (vie) et koinos (commun) et a été crée par le zoologiste allemand Karl Möbius qu'il conçoit donc comme étant une « *communauté de vie* »[71]. Enfin le concept d'écosystème au sens strict voit les vivants et leur milieu comme un système dans lequel circulent des flux de matière, d'énergie et d'information. Ainsi, un écosystème est l'ensemble formé par une association ou une communauté d'êtres vivants et son environnement et où les êtres vivants coexistent et interagissent entre eux, et avec l'environnement. Ces différents acteurs ont été pris en compte pour étudier la diversité et l'organisation de la vie microscopique marine.

I.3.2 Etudes de la diversité des micro-organismes planctoniques

Deux techniques sont utilisées pour explorer la diversité des espèces planctoniques : l'imagerie et la biologie moléculaire avec la génétique.

L'imagerie par microscope

La découverte du plancton est liée aux avancées technologiques en imagerie avec l'invention du microscope par Hooke en 1665. Cette technologie a tout d'abord permis de visualiser et d'identifier les micro-organismes planctoniques via la microscopie optique. Les premières descriptions scientifiques datent du XIX^{ème} siècle avec Victor Hensen. Au XX^{ème} siècle, le développement des techniques de fluorescence a permis d'identifier le phytoplancton. Avec la spectrophotométrie, il est possible de doser la chlorophylle *a* et donc de connaitre la concentration en phytoplancton. La chromatographie liquide à haute performance (HPLC) basée sur la séparation des pigments et la détermination de leur fluorescence[72–79] permet d'analyser les pigments phytoplanctoniques afin de quantifier l'abondance relative des différents groupes de phytoplancton. La fluorométrie permet quant à elle d'identifier et de caractériser les pigments photosynthétiques. Cette méthode consiste à exciter les pigments photosynthétiques par une source de lumière à différentes longueurs d'onde spécifique pour filtrer la fluorescence émise par l'échantillon afin de déterminer les pics correspondant aux différents pigments photosynthétiques[80,81]. La cytométrie en flux (CMF) est une technique utilisée pour l'analyse de l'évolution spatiale et temporelle du phytoplancton[82–87] mais également pour la découverte de

nouvelles espèces marines, comme cela a été le cas pour *Prochlorococcus* et *Ostreococcus*[57,88]. Enfin, les propriétés optiques des cellules et leur composition pigmentaire permet de détecter et discriminer des populations mixtes issues d'un même échantillon[88–91]. Ces approches, coûteuses en temps, évoluent vers des systèmes automatisés d'imagerie à haut-débit avec des appareils embarqués directement sur des instruments sous-marins ou sur les navires océanographiques[92]. De nombreux instruments ont été développés pour réaliser l'identification rapide d'un très grand nombre d'échantillons en un temps limité. Le FlowCAM[93] par exemple, est un procédé d'imagerie à haute résolution. La cytométrie de flux permet de numériser dans un flux d'eau les particules et les images sont ensuite stockées sur un ordinateur. Le ZooScan[94] est un système d'analyse d'images qui permet quant à lui d'énumérer, de mesurer et d'identifier le zooplancton. Le développement des techniques d'imagerie a donc permis de connaitre la concentration, la distribution, la biomasse et la composition en micro-organismes planctoniques, notamment en phytoplanctons d'un échantillon. Mais, parfois, la morphologie ne permet pas de distinguer et de caractériser certains groupes taxonomiques[95].

La génomique

L'avènement de la génomique avec le développement des technologies de séquençage a contribué à approfondir nos connaissances sur l'écosystème marin. Avec l'introduction en 2003 du concept de codes-barres ADN (« DNA barcoding »)[96] qui permet de distinguer et d'identifier des espèces proches morphologiquement. Cette technique récente de phylogénie moléculaire permet donc de réaliser la caractérisation génétique d'un individu ou d'un échantillon d'individus à partir d'un gène de son génome. En effet, cette technique a l'avantage d'avoir accès à des organismes non cultivables prélevés directement dans leur milieu. Il est donc possible via la technique dite de métabarcoding d'identifier les différentes espèces d'échantillons naturels[97]. Cette technique sera détaillée dans la partie III.4.1.

Ainsi, via l'utilisation des techniques d'imagerie et de la génétique, il est possible de caractériser taxonomiquement et quantitativement les micro-organismes présents dans un échantillon. Cela permet de décrire la répartition des micro-organismes planctoniques sur la surface des océans mais ne permet pas d'expliquer cette répartition. En effet, le contexte environnemental est une composante importante à intégrer pour pouvoir réaliser une étude dite, de biogéographie des communautés planctoniques.

I.3.3 Impact de l'environnement sur l'organisation spatiale du plancton

La circulation de l'eau de mer dans l'ensemble des océans est régie par les effets combinés du vent, des différences de températures, de densité, de salinité mais également des interactions au sein des courants marins. Le plancton, transporté par ces mouvements d'eau, est

intimement lié à ces différents facteurs. Les courants marins et les facteurs environnementaux ont un impact sur l'organisation du plancton.

Les Courants marins

Les grands courants marins sont d'immenses autoroutes se déplaçant dans les mers sur toute la planète. Ces déplacements d'eau, chaude ou froide, transportent le plancton sur tous les océans du globe. Le vent et la densité de l'eau sont les deux phénomènes à l'origine des courants marins. Le vent est responsable des courants superficiels qui peuvent atteindre jusqu'à 800 mètres de profondeur. Dans l'hémisphère Nord, la friction provoquée par la rencontre des vents dominants d'ouest et des alizés du nord-est met en mouvement les eaux de surface dans le sens des aiguilles d'une montre, et inversement dans le sud de l'équateur. Les courants marins se déplacent de façon rectiligne à la surface de la Terre en rotation et subissent la force de Coriolis qui est perpendiculaire à leur vitesse de déplacement. Ces courants sont donc déviés vers la droite dans l'hémisphère Nord et vers la gauche dans l'hémisphère Sud sous l'effet de la rotation de la Terre qui engendre cette force physique. Il existe environ 18 courants océaniques principaux de surface (Figure I.3.3.i). Certains sont chauds comme le Gulf Stream qui s'étale des côtes de la Floride jusqu'au Groenland. Un autre courant chaud est celui des Agulhas. Celui-ci parcourt l'océan Indien en longeant la côte Est de l'Afrique et en passant dans le canal du Mozambique pour se terminer au niveau du Cap de Bonne-Espérance. A ce niveau, ce courant chaud rencontre le courant froid du Benguela dans de l'Atlantique Sud qui longe la côte Ouest Africaine. Cette rencontre d'eau chaude et d'eau froide forme de gigantesques tourbillons pouvant mesurer de 100 à 400 km de diamètre[98]. Ces tourbillons, également appelés anneaux d'Agulhas, permettent une connexion entre l'Océan Indo-Pacifique et l'Océan Atlantique. En effet, ces anneaux formés à la sortie du canal du Mozambique passent dans l'Océan Atlantique et poursuivent leur route le long du gyre de l'Atlantique Sud. Le trajet de certains de ces anneaux peut se terminer jusqu'à la côte Est de l'Amérique latine après plusieurs années[98,99]. Ainsi, ces anneaux d'Agulhas jouent un rôle important dans la distribution de la chaleur et de la salinité dans les océans du globe. Ces transferts de l'Océan Indo-Pacifique vers l'Atlantique ont augmenté au cours de ces dernières décennies[100] et font de cette connexion interocéanique une zone importante pour l'étude des courants océaniques ainsi que celle des différents scénarios du changement climatique[101]. Enfin, l'étude de la biogéographie de fossiles de diatomées au niveau de cette connexion a permis de montrer que ces phénomènes ne sont pas des barrières pour le transfert du microplancton[102].

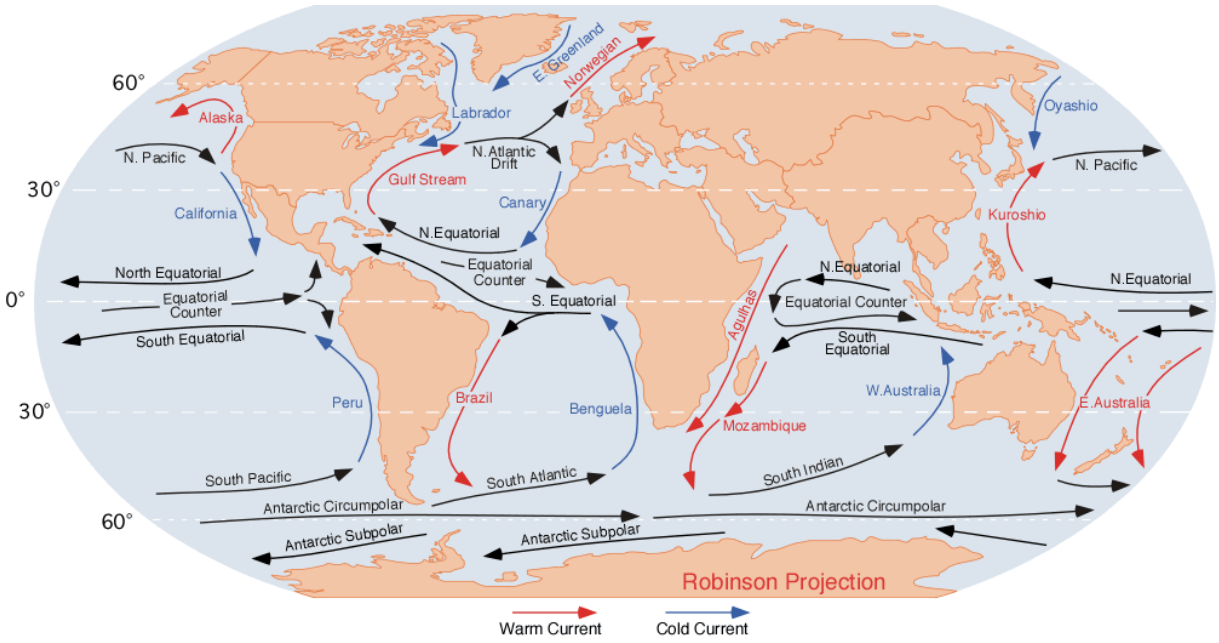


Figure I.3.3.i | Les principaux courants marins mondiaux. (Image issue de http://blue.utb.edu/paullgj/geog3333/lectures/physgeog.html)

Le second phénomène à l'origine des courants marins dépend de la densité de l'eau qui affecte les courants de surface et ceux de l'océan profond. La densité varie en fonction de la température et de la salinité. Ainsi, les eaux froides, plus denses, plongent sous les eaux chaudes. Les eaux chaudes de surface au niveau des tropiques, refroidissent près des pôles et voient donc leur densité augmenter et s'enfoncent vers les profondeurs pour former des courants profonds et froids. Les eaux salées sont plus denses et subissent donc le même phénomène de mouvements verticaux. Le Gulf Stream subit ce phénomène car il est constitué pour ses trois-quarts d'un courant chaud de surface et pour un quart d'un courant froid[103]. Ainsi les eaux de surface se refroidissent suffisamment pour se mélanger aux courants profonds existants. C'est également le même phénomène qui se déroule au large de Gibraltar, où les eaux très salées de la Méditerranée s'enfoncent sous celles de l'Atlantique permettant aux courants chauds de surface de passer de l'Atlantique vers la Méditerranée[104]. La vitesse de ces courants peut varier de quelques centimètres à 3 mètres par seconde pour le Gulf Stream[105]. Ainsi, la température (thermo) ainsi que la salinité (halin) permettent de faire circuler les courants dans toutes les mers du globe. C'est ce que l'on appelle la circulation thermohaline qui est la circulation océanique à grande échelle engendrée donc par les différences de densité de l'eau de mer (Figure I.3.3.ii). Il a été estimé qu'il faut près de mille ans à 1 m^3 d'eau de mer pour réaliser le parcours le long de la circulation thermohaline globale[106]. C'est donc un trajet de plus de trois millions de kilomètres dans tous les bassins océaniques du globe qu'effectue une particule d'eau. Si l'on suit le parcourt d'une goute d'eau dans ce courant, celle-ci passe dans les courants de

surface au large de l'Australie, traverse l'océan Indien, longe la côte de Madagascar et de la Tanzanie, remonte vers l'Oman, la pointe de l'Inde puis l'Indonésie avant de regagner l'Australie. Puis cette goutte repart vers les Comores, passe le cap de Bonne-Espérance pour s'engager dans l'océan Atlantique. Celle-ci longe la côte d'Amérique du Sud, du Brésil jusqu'au golfe du Mexique, suit le tropique du Cancer jusqu'à la Mauritanie et repart en mer des Caraïbes pour traverser l'océan Arctique. Elle plonge ensuite dans les profondeurs, rejoint l'Atlantique Sud pour aller jusque dans l'océan Austral. Cette goutte est ensuite entraînée vers l'est, en direction de l'océan Indien et du Pacifique où elle remonte en surface pour regagner l'Atlantique Sud puis pourra réitérer ce trajet dans ce circuit fermé. Ainsi, le planton, qui par définition dérive le long des courants marins, est transporté, telle une particule d'eau dans ces immenses autoroutes.

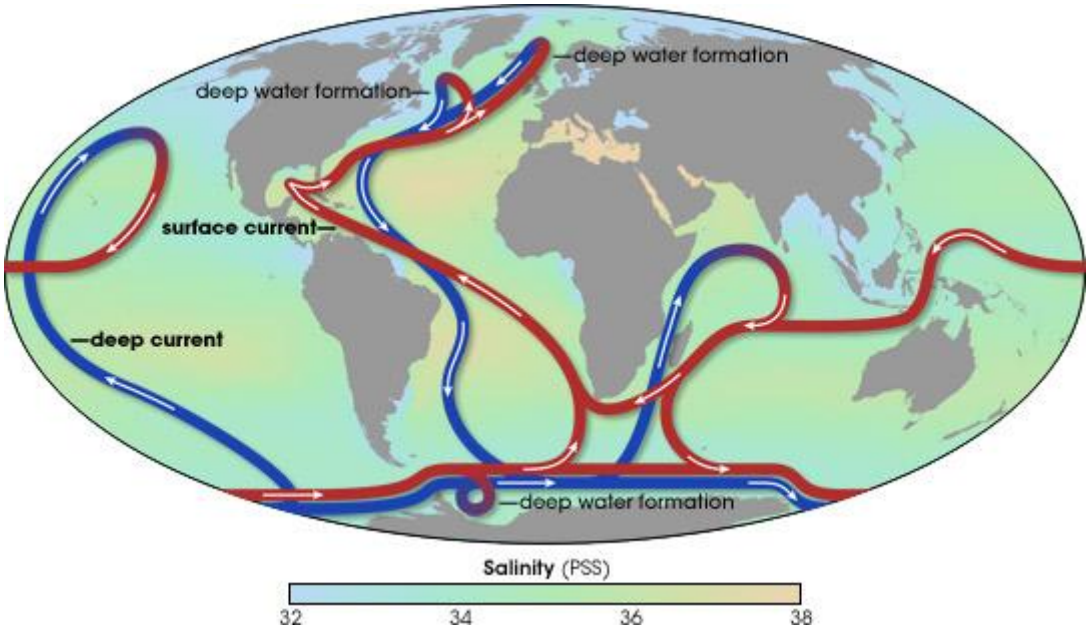
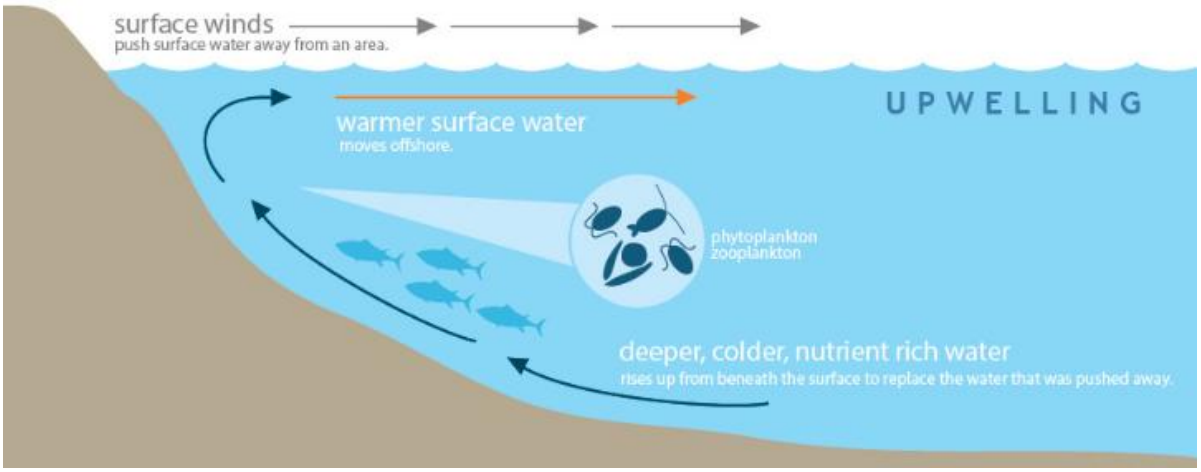


Figure I.3.3.ii | La circulation thermohaline des océans. Les courants chauds de surface sont représentés en rouge et les courants froids profonds en bleu. (Représentation de Robert Simmon issue de http://earthobservatory.nasa.gov/)

D'autres types de courants ont un impact sur l'organisation du plancton dans les océans. C'est le cas des *upwelling* (remontée d'eau en français) qui est un phénomène océanographique correspondant à la remontée vers la surface d'eaux profondes riches en minéraux, due à l'action des vents et des courants de surface. Il existe au moins cinq types d'upwelling dont le plus connu est le type côtier. Ce dernier se produit lorsque des vents forts poussent l'eau de surface vers le large. Il y a alors création d'un déficit près de la côte qui est aussitôt comblé par les eaux profondes qui remontent vers le littoral. Ces eaux chargées en sels minéraux, permettent le développement en surface de phytoplancton, puis de zooplancton et de toute la chaîne alimentaire marine (Figure I.3.3.iii). Les zones d'*upwelling* sont donc riches en phytoplancton et la mesure de la concentration en chlorophylle est un bon indicateur de la production primaire de

ces courants, elles peuvent ainsi être observées via les satellites d'observation terrestre. Il existe quatre zones principales d'*upwelling* côtier qui recouvrent à peine 0.1% de la surface globale océanique mais représentent 30% des captures de la pêche[107]. Deux *upwelling* se situent dans l'hémisphère Sud : L'*upwelling* du Benguela au sud de l'Angola, de la Namibie et de l'Afrique du Sud et l'*upwelling* de Humboldt au large du Pérou et du Chili. Deux autres *upwelling* se situent quant à eux dans l'hémisphère Nord : l'upwelling de Californie au large des Etats-Unis et au nord du Mexique ainsi que l'upwelling des Canaries au niveau des côtes du Maroc, de Mauritanie, du Sénégal et de la Gambie. Certaines expéditions ont pour objectif d'étudier la diversité planctonique s'y trouvant[40].

a



b

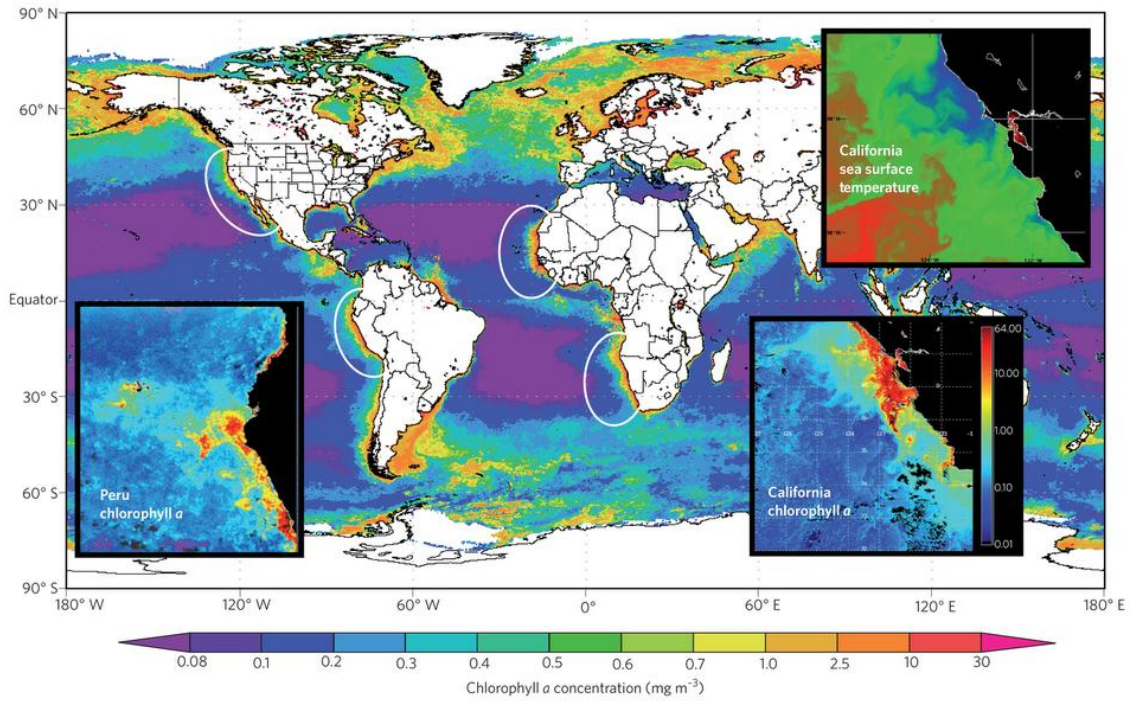


Figure I.3.3.iii | Les *upwelling* côtiers. a, Le mécanisme d'*upwelling* (Figure extraite de www.oceanservice.noaa.gov), **b,** les quatre principaux *upwelling* côtiers (Figure extraite de Capone et Hutchins 2013[108]).

Ainsi, la circulation thermohaline des océans et le phénomène de courants verticaux maintiennent la productivité primaire océanique en ramenant à la surface les nutriments et la matière organique nécessaires au développement des micro-organismes planctoniques. Ces micro-organismes sont eux mêmes transportés par ces courants marins. Le contexte physico-chimique dans lequel ils évoluent est donc un facteur important à prendre en compte puisqu'il contribue à son développement.

Les paramètres physico-chimiques

Les paramètres physico-chimiques ont un impact sur le développement des micro-organismes planctoniques et sont donc à considérer pour suivre leur organisation et leur évolution.

La température de l'eau est un descripteur de base pour la connaissance du milieu marin. Celle-ci conditionne la présence et la répartition des espèces selon leur thermorésistance. En effet, la température influe sur l'activité enzymatique et peut avoir ainsi un impact sur l'évolution physiologique des organismes.

La salinité est un autre paramètre à étudier. Elle représente la proportion de sels minéraux dissous dans l'eau de mer. Les espèces planctoniques se développeront plus ou moins bien en fonction de leur tolérance si la salinité varie. De plus, comme précisé précédemment, la salinité a une influence sur la densité de l'eau de mer et permet donc l'obtention d'informations sur la circulation océanique. Deux méthodes sont couramment utilisées pour mesurer la salinité : la méthode volumétrique et la méthode conductimétrique[109].

La concentration en nutriments, ou sels nutritifs regroupe différents composants. Les nitrates, les nitrites, l'ammonium, les phosphates et les silicates sont les composés généralement mesurés lors des prélèvements. En effet, ces éléments sont nécessaires à la nutrition du phytoplancton et sont donc essentiels à sa croissance. Ces sels nutritifs sont présents naturellement dans l'eau de mer et sont apportés par les eaux douces des fleuves et des rivières des bassins versants littoraux. Les nitrates et les phosphates peuvent également provenir des activités humaines. Les silicates ont une origine directement liée à l'érosion.

L'oxygène dissous est l'élément de base pour la survie des organismes vivants, hors bactéries anaérobiques. Il facilite également la dégradation des matières organiques et l'accomplissement des cycles biochimiques. L'oxygène dissous provient des échanges entre l'atmosphère et la surface de l'eau mais résulte également du processus de photosynthèse réalisé par le phytoplancton[110]. L'oxygénation des eaux est régulée par les conditions physiques et physico-chimiques du milieu. Une hausse de température ou de salinité limitera cette oxygénation.

Le pH est un paramètre qui nous permet de mesurer l'acidité, l'alcalinité ou la basicité de l'eau. La modification du CO2 par la photosynthèse entraîne une modification du pH et permet donc de connaître l'activité photosynthétique du phytoplancton dans un milieu donné.

La mesure des matières en suspension (MES) ou de la turbidité de l'eau est un paramètre qui permet de se rendre compte de la concentration en micro-planctons. Celle-ci a donc un rôle sur le développement du phytoplancton car elle peut être néfaste à un certain niveau en empêchant la lumière du soleil de pénétrer dans l'eau, et donc d'empêcher le phytoplancton à réaliser sa photosynthèse.

Enfin, la chlorophylle *a* est un paramètre indispensable pour la mesure de la présence planctonique. Il est en effet possible de quantifier la biomasse du phytoplancton à travers son activité chlorophyllienne. La chlorophylle *a* est un pigment photosynthétique qui permet de quantifier la biomasse totale de phytoplancton actif. Le phytoplancton étant le premier maillon de la chaîne alimentaire marine dont va dépendre le reste des organismes vivants supérieurs, celle-ci peut donner également un indice sur l'abondance des autres espèces de micro-organismes planctoniques.

I.3.4 Mesures des données environnementales

Les mesures réalisées en océanographie physique ont pour but d'améliorer nos connaissances fondamentales de l'océan ainsi que celles du fonctionnement de notre planète. Pour coordonner, superviser et assurer la qualité des mesures réalisées, la National Science Foundation des USA a lancé en 1982 un programme appelé WOCE (World Ocean Circulation Experiment) et a créer le WOCE Hydrographic Programme Office. Il est en effet important d'assurer la qualité des données collectées et d'homogénéiser les mesures réalisées pour exploiter les informations ainsi apportées. Quelques capteurs et instruments utilisés pour évaluer les paramètres utiles aux océanographes peuvent être référencés[111]. La température, la pression et la conductivité sont trois grandeurs dont la mesure est essentielle pour déterminer la masse volumique des océans. Celle-ci est calculée à partir d'équations empiriques appelées équations d'état de l'eau de mer datant de 1980[112]. Pour mesurer ces paramètres, des bathysondes scientifiques comme la sonde CTD (*Conductivity Temperature Depth*) permettent de numériser sous l'eau l'information issue des capteurs et la transmettent en temps réel au bateau par un câble électroporteur. La température et la salinité peuvent être également mesurées par un thermosalinographe.

La turbidité de l'eau vient de la présence de diverses matières en suspension telles que les matières organiques et inorganiques, le plancton et les autres micro-organismes. Elle est couramment définie comme la propriété optique qui fait que la lumière incidente est diffusée et

absorbée plutôt que transmise en ligne droite. Ces phénomènes sont mesurés à l'aide de transmissiomètres[113], néphélomètres[114] et fluorimètres[115] pouvant être intégrés sur les bathysondes. Ces capteurs bio-optiques peuvent ainsi mesurer la concentration en nitrates, le taux d'oxygène et la fluorescence à différentes longueurs d'ondes.

L'utilisation des courantomètres à rotor ou bien des courantomètres émettant des ondes acoustiques et qui utilisent l'effet Doppler généré par le déplacement des particules présentes dans les masses d'eau permet de mesurer la vitesse et la direction des courants marins. L'ancrage d'instruments appelés flotteurs dérivants dans les masses d'eaux ou bien le suivi de traceurs chimiques permettent également d'avoir des informations sur les courants marins.

Les navires océanographiques permettent de mettre en place le déploiement de ces instruments. La mise à l'eau des lignes instrumentées conçues pour se maintenir en position fixe en milieu marin permet de connaître l'évolution d'un ou plusieurs paramètres sur de longues périodes temporelles. Ces mouillages peuvent être de type « eulériens » par référence aux repères d'Euler qui sont fixes par rapport à la terre, ou bien de type « lagrangiens » par référence aux repères mobiles de Lagrange. Il est également possible d'utiliser des flotteurs dérivants pour étudier l'évolution des propriétés du milieu sur de grandes échelles de temps et d'espace[116]. Des planeurs sous-marins appelés *gliders* peuvent ainsi être mis en place pour mesurer les données environnementales entre différentes zones marines.

Toutes ces mesures ne peuvent être effectuées sans l'aide de systèmes satellitaires de positionnement et de datation qui permettent de définir les lieux de prélèvements des échantillons. En effet, les satellites couvrent la surface des océans en quelques jours, ils permettent donc de connaitre l'état instantané des océans. Les satellites peuvent nous donner des informations sur la température de surface des océans, la turbidité, la concentration en chlorophylle *a* des océans via différents types de capteurs analysant la couleur de l'eau. Le principe de base de la méthode consiste à mesurer le signal de radiance ou de luminance réémis par la couche de surface océanique après absorption et diffusion de la lumière solaire incidente. Les longueurs d'ondes considérées sont celles du visible (400 à 700nm). Le contenu en chlorophylle *a* qui est un pigment qui absorbe fortement le bleu et plus faiblement le rouge permet d'indiquer les eaux riches en phytoplanctons qui apparaissent vertes du fait que les rayonnements bleu et rouge ne ressortent que partiellement de l'eau. Il est ainsi possible de suivre les floraisons de phytoplanctons au cours des saisons qui sont liées à l'ensoleillement et à la concentration en nutriments. Les satellites permettent également de connaitre le niveau de la mer et la topographie des océans, ce qui nous informe sur la vitesse de la circulation océanique. Certains satellites ont été mis en orbite spécifiquement pour étudier les océans. C'est le cas des

satellites NOAA (National Oceanic and Atmospheric Administration). Ces satellites météorologiques américains à orbite polaire observent la Terre depuis une altitude d'environ 850km. Ils ont pour missions principales l'observation des phénomènes météorologiques, la cartographie de la structure thermique superficielle des océans, l'agro-météorologie et l'étude de l'évolution de l'environnement marin et côtier. Les mesures des satellites une fois traités sont ensuite diffusées et utilisées lors des campagnes océanographiques pour l'étude, par exemple du plancton marin.

Ainsi, via les mesures de données environnementales et la description des micro-organismes présents dans différents environnements marins, il est possible de décrire et de d'expliquer la distribution géographique de ces micro-organismes. Cette discipline, appelée biogéographie fait partie intégrante des analyses pour la compréhension de l'organisation de l'écosystème planctonique.

I.3.5 La biogéographie des micro-organismes planctoniques

La biogéographie s'intéresse à l'étude de la distribution des organismes vivants sur Terre et cherche à expliquer les raisons de leur répartition géographique. Cette discipline permet donc d'identifier l'écosystème marin par l'étude de la distribution spatiale des organismes planctoniques avec les conditions environnementales[117]. Les trois quarts de la planète apparaissent comme une immensité bleue, homogène qui ne se différencie que par des variations en surface de couleurs ou d'agitation[118]. L'étude des conditions environnementales ainsi que de la distribution des micro-organismes planctoniques montre pourtant des variations dans leurs distributions à l'échelle mondiale. C'est pourquoi, une première division de l'océan a été proposée en 1872 par Mary Somerville[119] en se basant sur des études réalisées en milieu terrestre. Plus tard, les multiples expéditions océaniques ont permis le prélèvement d'échantillons marins et la mesure de données environnementales. Cela a permis de réaliser le découpage des océans fondé soit sur la distribution spatiale des organismes[120,121] soit sur les conditions biogéochimiques marines[122]. Le lancement de satellites à partir des années 1970 pour étudier l'océan à l'échelle globale a permis de réaliser des mesures en temps réel des différents paramètres environnementaux de surface. Les images satellitaires ont permis de retrouver les zones biogéographiques décrites auparavant en étudiant directement les paramètres physico-chimiques qui participent au développement du phytoplancton[123]. Des unités géographiques ont ainsi été identifiées tout d'abord en Atlantique Nord[124,125] puis dans l'océan global avec le découpage biogéochimique des océans proposé par Longhurst dans les années 2000[2–4] . Ce découpage repose sur l'identification de 4 biomes océaniques représentant les grands types de végétation océanique (biome polaire, biome des vents d'ouest, biome des vents d'est et les biomes côtiers). Puis la subdivision de chaque biome, a permis de former des provinces

océaniques qui délimitent les spécificités régionales biogéochimiques, physiques et écologiques (Figure I.3.5). De récentes études basées sur la distribution globale de paramètres physico-chimiques ainsi que d'organismes trophiques ont confirmé que chacune des 56 provinces pouvait être considérée comme une unité écologique unique[126]. Ces provinces ont depuis été réévaluées par l'intégration d'autres paramètres environnementaux[4] mais cette division océanique reste un référentiel géographique des grands écosystèmes mondiaux pour le suivi des populations marines exploitées, pour le changement climatique ou encore pour la mise en place de campagne océanographique.

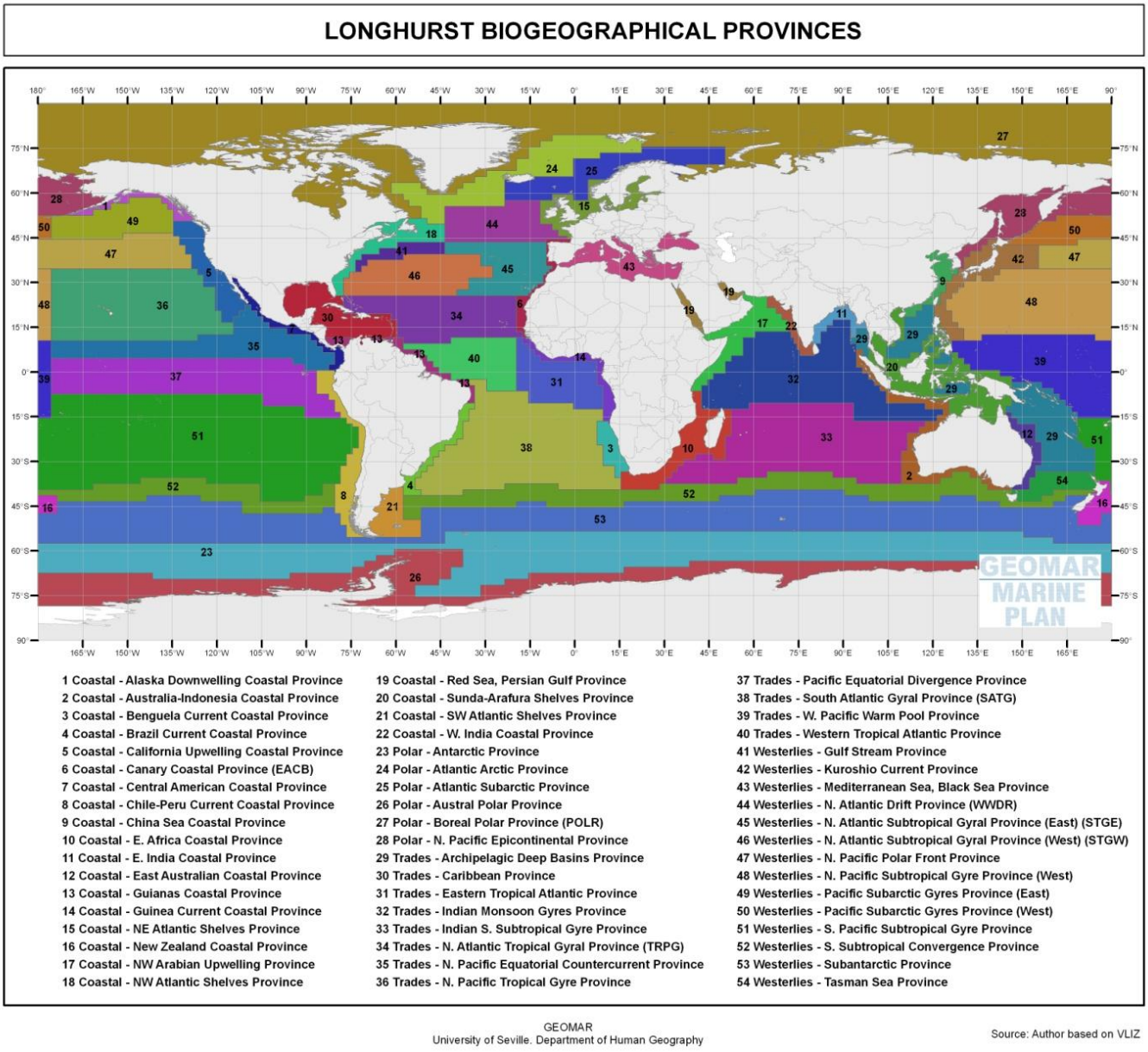


Figure I.3.5 | Provinces biogéographiques proposées par Longhurst en 2007. (Figure extraite de www.marineplan.es)

L'ensemble de ces connaissances sur l'écosystème planctonique ainsi que la conception et l'évolution des technologies et des méthodes permettant son étude n'aurait été possibles sans

la mise en place d'expéditions océaniques d'échantillonnage du plancton et les mesures physico-chimiques de l'environnement marin.

I.4 Les grandes expéditions océaniques

I.4.1 L'océanographie moderne et la collecte du plancton

L'échantillonnage du zooplancton remonte au XIXème siècle lorsque le docteur Thomson inventa un filet pour prélever les larves de crustacés en 1968[127,128]. Les premières études qualitatives sur le zooplancton ont eu pour objectif de découvrir et de répertorier de nouvelles espèces. C'est ensuite que Victor Hensen a développé le filet Hensen pour réaliser une étude quantitative du plancton. Le prélèvement de phytoplancton à été initié par Teodor Cleve avec l'utilisation de filets en soie[129]. De nombreux autres types de filet on été développés depuis. Les différents types de filet se distinguent notamment par des modifications au niveau de la forme, de l'ouverture, de la longueur et de la maille du filet pour récolter différents types et tailles de plancton. Les filets à nappes ou filets multiples permettent d'échantillonner sur la colonne d'eau à différentes profondeurs. Pour prélever des échantillons sur différentes colonnes d'eau, il est également possible d'utiliser des bouteilles de prélèvements. La bouteille de Nansen a été la première à être développée au début de XX^{ème} siècle. La bouteille de Niskin a suivi en 1966 et est encore couramment utilisée pendant les expéditions océanographiques. Ces bouteilles servent généralement à réaliser des échantillonnages en profondeur, pouvant aller jusqu'à plusieurs milliers de mètres. Le prélèvement d'échantillons dans la colonne d'eau se fait par ouverture puis fermeture de clapets lorsque la profondeur désirée est atteinte. Ces bouteilles peuvent être associées sur une structure en rosette où est souvent intégrée une sonde CTD ainsi que d'autres appareillages permettant d'enregistrer en continu les données environnementales pendant toute la durée de la descente. La découverte du nanoplancton a été possible par le développement des techniques de centrifugation[130]. Le développement de ces technologies a permis l'émergence de l'océanographie moderne avec de nombreuses expéditions océanographiques qui se sont succédées depuis le XIXème siècle afin d'étudier les micro-organismes planctoniques.

I.4.2 Le H.M.S Challenger

Une des plus marquantes expéditions scientifiques effectuées est celle du voyage du H.M.S (Her Majesty Ship) *Challenger*, de 1872 à 1876. Sous l'impulsion de la Royal Society, le gouvernement anglais organisa une expédition autour du monde, et réarma alors un bâtiment de guerre, le *Challenger*. Ce navire quitta l'Angleterre au printemps 1872 et y revint trois ans et

demi plus tard après avoir parcouru 70 000 miles et réalisé une exploration des fonds océaniques dans des profondeurs à plus de 2 000 mètres sur un total de 362 stations. Le sondage le plus profond est de 8 200 mètres dans la fosse des Mariannes en mer des Philippines. Ce point appelé *Challenger Deep* a été considéré pendant longtemps comme le point le plus profond des océans. Un protocole d'échantillonnage a été mis en place pour récolter les micro-organismes planctoniques tout en incluant les relevés des paramètres physiques, chimiques et biologiques. Les instruments de mesure et de collecte étaient encore très simples. Ce sont environ 13 000 échantillons qui ont été collectés pendant cette expédition à l'aide de filets. La mesure de la température est réalisée avec un thermomètre positionné dans une capsule métallique pour le protéger de la pression. Une fois ramenés à la surface, les échantillons sont triés et décrits par les scientifiques puis dessinés en détail. Le naturaliste Allemand Ernst Haeckel participa à la représentation des organismes planctoniques observés pendant l'expédition, comme cela a été le cas pour de nombreux radiolaires[131] (Figure I.4.2). Les échantillons ont ensuite été conservés dans des bocaux étiquetés. Ce sont environ 1 500 nouvelles espèces qui ont été décrites à la suite de cette expédition. Ce projet a révélé qu'il était possible d'étudier la vie océanique dans son environnement. C'est le début de l'océanographie moderne. Presque 150 ans après, cette expédition reste un modèle pour les scientifiques qui étudient cet écosystème. Cependant, le manque de moyens matériels et technologiques pour observer et étudier la diversité des micro-organismes planctoniques ne permettait pas encore d'apprécier l'impact des facteurs environnementaux sur l'organisation de la vie marine.

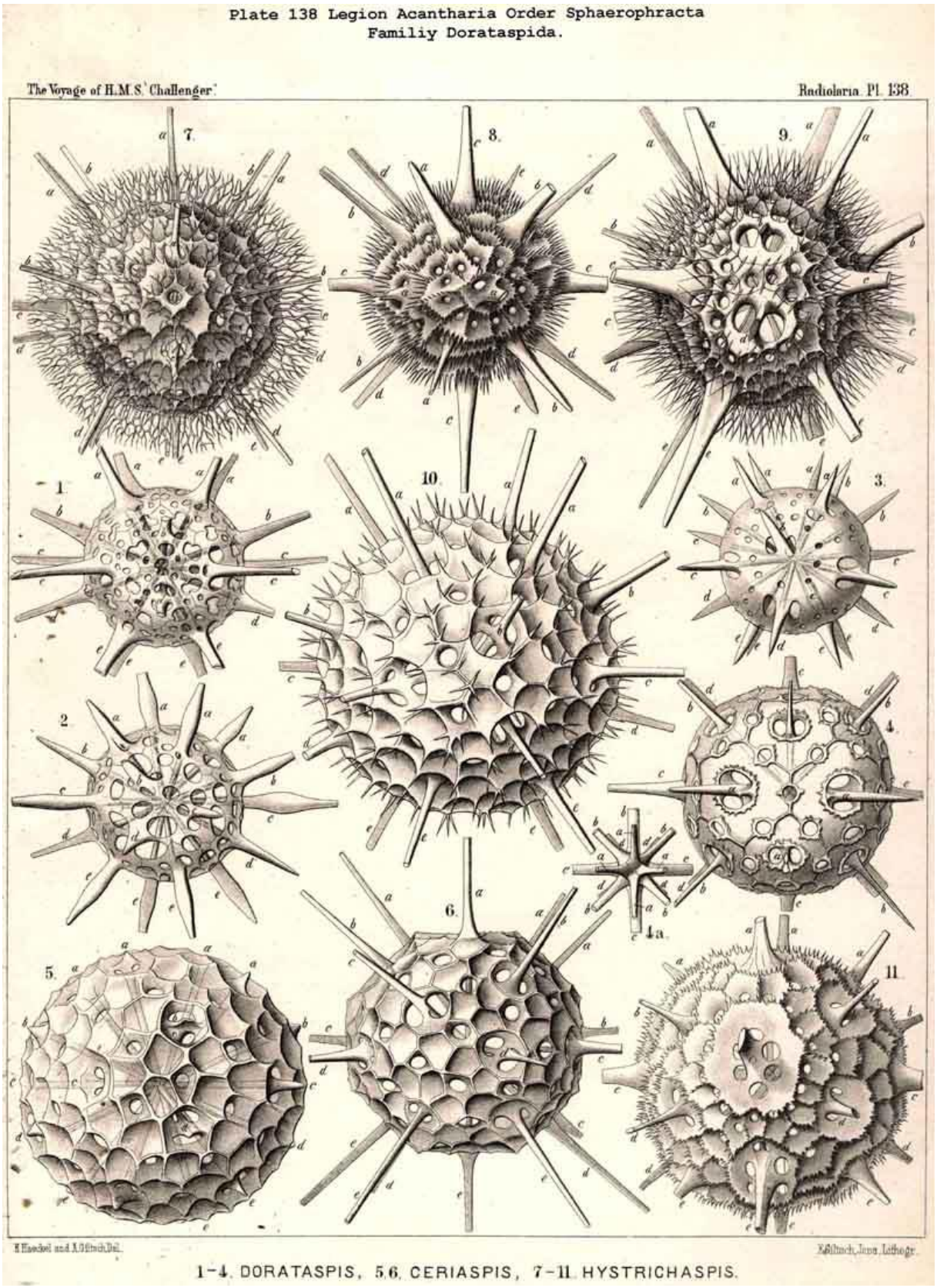


Figure I.4.2 | Illustration de *Radiolaires* de l'ordre *Sphaerophracta* et de la famille *Dorataspida* décrits et dessinés par Haeckel. (Figure issue de la planche 138 extraite de Heackel 1880[131])

I.4.3 L'expédition Global Ocean Sampling

Avec les avancées récentes en génomique et notamment avec le séquençage massif de communautés de micro-organismes issues directement de leur environnement (cette discipline correspond à la métagénomique décrite à la partie II.3) ainsi que les analyses bioinformatiques, il est maintenant possible d'avoir accès à l'information taxonomique et génomique des micro-organismes planctoniques des océans. Ces données permettent d'étudier la diversité de ces micro-organismes, les gènes qui les composent ainsi que leur profil d'expression[132,133]. C'est ainsi que le *Venter Institue* a entrepris une campagne d'échantillonnages des micro-organismes bactériens. L'expédition *Global Ocean Sampling* (GOS) sur le *Sorcerer II*[14] avait pour but de réaliser la comparaison des informations génomiques à grande échelle afin d'explorer la diversité microbienne des océans dans un contexte biogéographique. C'est sur le yacht de Craig Venter, transformé en laboratoire flottant, que la collecte des échantillons marins débuta en Février 2003 à partir de la Nouvelle-Ecosse jusqu'à la Polynésie Française. En tout, 41 échantillonnages ont été réalisés dans les eaux de surface en longeant la côte Est de l'Amérique du Nord, en passant par la mer des Caraïbes, le canal du Panama puis dans l'Océan Pacifique. Les prélèvements ont été séparés les uns des autres de plus de 200 miles afin d'obtenir des échantillons provenant d'environnement différents. Ces échantillons, une fois filtrés ont été séquencés via la technique de séquençage haut débit qui est décrite dans la partie II.1.2. 7,7 millions de lectures ont été générées recouvrant au total 6,3 gigabases d'ADN. Ces données, après analyse, ont montré qu'une grande majorité des micro-organismes de l'océan global reste inconnu[14,134]. Il a notamment été relevé qu'au sein des espèces abondantes des différences génétiques entre les échantillons existaient. Ces différences ont suggéré des adaptations des micro-organismes en relation avec leur environnement. De plus, des populations génétiquement isolées ont mis en évidence des préférences d'environnements qui diffèrent selon les organismes. Des échantillons similaires au niveau génétique peuvent se trouver dans des localisations géographiques très différentes. Ainsi, la comparaison des informations génomiques à grande échelle entre différents milieux a été réalisée afin d'explorer la diversité microbienne des océans dans un contexte biogéographique. Le projet GOS a pu fournir aux scientifiques une base de données métagénomique des communautés planctoniques dans ces régions. D'autres analyses ont montré que ce type de données pouvait permettre d'identifier les relations entre la composition fonctionnelle en gènes et les facteurs environnementaux[135,136], ce qui n'avait pas été possible lors du projet H.M.S *Challenger*. Ainsi, un échantillonnage permettant d'étudier l'étendue de l'écosystème des micro-organismes planctoniques en réalisant des prélèvements à différentes profondeurs, en regardant les différents domaines de la vie et les différentes tailles d'organismes combinés avec les données environnementales dans l'ensemble des océans du

globe permet d'avoir une vision globale de l'écosystème planctonique. C'est dans cette vision que le projet *Tara* Oceans a été mis en place.

I.4.4 L'expédition Tara Oceans

L'expédition *Tara* Oceans a pour objectif d'explorer la diversité de l'écosystème planctonique des océans de surface de la manière la plus complète et approfondie possible. Celle-ci doit son nom au voilier *Tara* sur lequel ont été prélevés les échantillons dans l'ensemble des océans du globe.

La Goélette *Tara*

C'est sur la goélette *Tara*, connue pour ses expéditions dans les régions polaires, que les prélèvements d'échantillons marins ont été réalisés (Figure I.4.4.i). Ce navire avec sa coque en aluminium et en forme de noyau d'olive de 36 m de long et de 10m de large a été initialement conçu pour résister à la compression des glaces en mouvement et aux basses températures afin de se laisser dériver sur la banquise. D'abord baptisée *Antarctica* par le médecin explorateur Jean-Louis Etienne en 1989, elle fut rachetée par le célèbre navigateur Peter Blake qui la renomma *Seamaster*. Ce dernier entreprit des missions d'explorations pour son programme de défense de l'environnement. C'est en 2003 que le bateau est rebaptisé *Tara* par Etienne Bourgois et que le lancement du projet *Tara* Expédition est réalisé. Ce projet est à l'origine de plusieurs expéditions qui ont eu lieu depuis. Parmi ces expéditions figure l'expédition *Tara* Arctic qui commença en septembre 2006 et pendant laquelle *Tara* dériva sur près de 1 800 km sur la banquise de l'océan arctique pendant 2 ans afin d'étudier et *comprendre* les phénomènes de changement climatique des hautes latitudes. C'est ensuite que l'expédition *Tara* Oceans fut lancée en Septembre 2009 pour se terminer en Mai 2012 puis dans la continuité de l'étude de la diversité océanique, l'expédition *Tara Polar Circle* a été réalisée. Cette dernière débuta en Mai 2013, le bateau réalisa une circumnavigation de l'océan arctique de 25 000 km en 6 mois. D'autres expéditions ont eu lieu depuis comme l'expédition *Tara* Méditerranée en 2014 pour étudier l'impact des micro-plastiques en mer Méditerranée et enfin, à l'heure où j'écris ce mémoire de thèse, se déroule l'expédition *Tara* Pacific qui est prévue jusqu'en 2018 pour étudier le corail marin ainsi que les organismes en interaction avec ce dernier.

Figure I.4.4.i | Le voilier *Tara*. (© F. Latreille, Tara Expeditions)

Trajet de l'expédition *Tara* Oceans

Pendant cette expédition, la goélette *Tara* a réalisé un parcours traversant les principales grandes zones océaniques qui présente des propriétés océanographiques contrastées tout en tentant de parcourir la plus large couverture océanique possible pour essayer d'établir une corrélation entre la structure des écosystèmes planctoniques et l'environnement. L'expédition a débuté à Lorient, ville de port d'attache de la goélette. C'est un quasi tour du monde qui a été effectué (Figure I.4.4.ii). Le bateau, après avoir traversé le canal de Gibraltar, a échantillonné en mer Méditerranéenne puis en mer Rouge, au Nord de l'Océan Indien connu pour être une région d'acidification importante avec en profondeur des eaux pauvres en oxygène, au sud de l'Océan Indien pour ensuite traverser le canal du Mozambique puis passer le Cap de Bonne Espérance pour suivre le gyre de l'Atlantique Sud. Pendant la traversée de cet océan, des prélèvements ont été effectués au niveau de l'*upwelling* du Chili et également au niveau de deux anneaux provenant du courant des Agulhas. Ces deux anneaux ont la particularité d'avoir été formés à des moments différents afin d'échantillonner et d'étudier l'évolution des communautés planctoniques ainsi que des données environnementales. Ainsi, des échantillonnages d'un anneau au niveau du début du gyre de l'Atlantique Sud 9 mois après sa formation puis d'un autre au niveau des côtes Brésiliennes 3 ans après sa formation ont été réalisés. Quelques prélèvements ont été effectués dans l'océan Austral puis, après le passage du cap Horn, le voilier est remonté dans l'océan Pacifique Sud connu comme étant un désert océanique. Des

prélèvements ont été réalisés dans la région des Galápagos puis après un passage dans les régions équatoriales, dans le Pacifique Nord avec notamment un prélèvement dans l'*upwelling* de Californie. Des échantillonnages ont ensuite été effectués dans le Golf du Mexique, une partie du Gulf Stream pour finir par des prélèvements le long du gyre de l'Atlantique Nord.

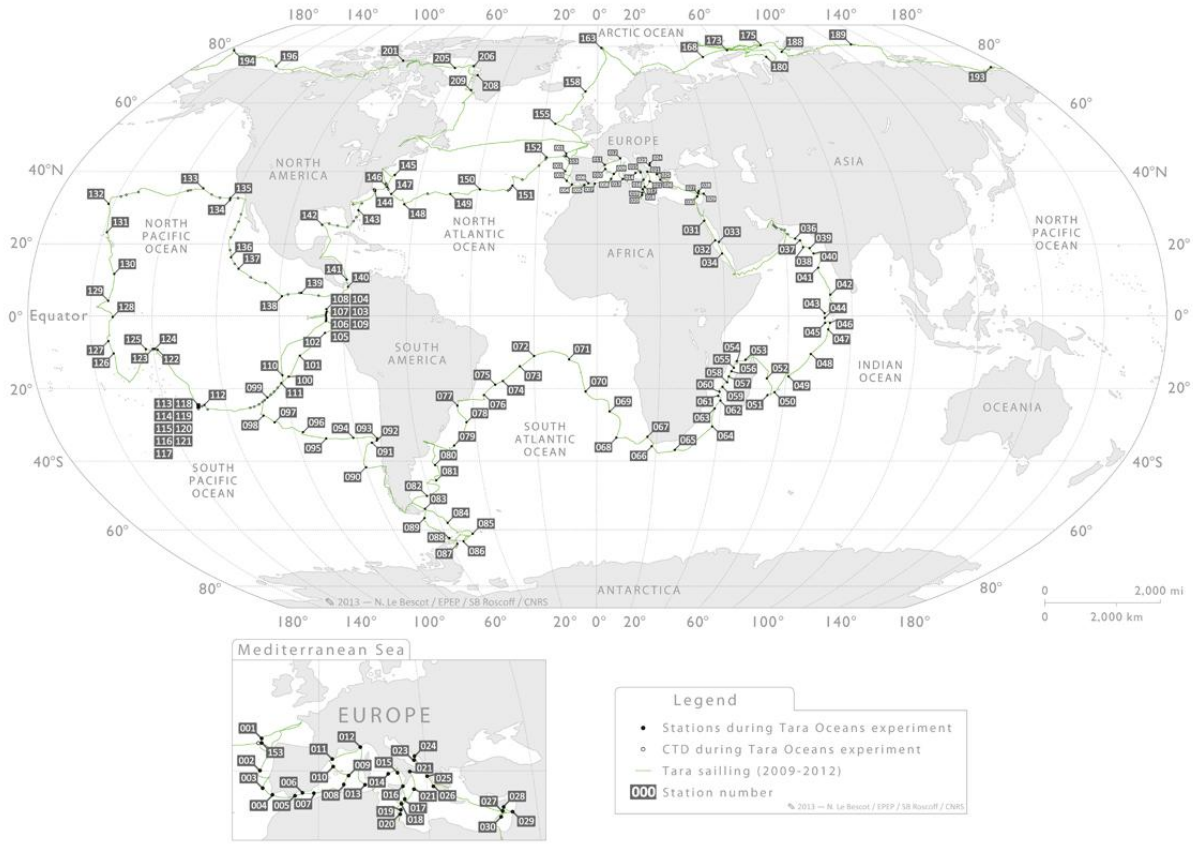


Figure I.4.4.ii | Parcours de la goélette *Tara* lors de l'expédition *Tara* Oceans et *Tara Polar Circle*. Le trajet est représenté par le tracé vert et chaque point correspond à une station de prélèvement (Image exraite de Noan Le Bescot 2013).

Prélèvements des échantillons

C'est donc pendant deux ans et demi de traversée que des prélèvements sur 153 zones, appelées stations ont été effectués. Les satellites nous renseignent sur la localisation des courants, des tourbillons, sur la température de l'eau en surface et sur la présence de différents types de bactéries et de protistes photosynthétiques. Ainsi, le choix de chaque station a été réalisé par rapport aux caractéristiques connues de l'espace à étudier associées aux mesures satellites et des différents instruments de mesures embarqués sur le bateau. Pour chaque station, un protocole d'échantillonnage a été mis en place pour collecter les organismes ayant des tailles très diverses. Les prélèvements ont été réalisés avec des bouteilles Niskin placées sur la rosette et des filets avec différentes taille de mailles. Une pompe péristaltique permettant de

récolter l'eau qui est ensuite filtrée dans des tamis de plus en plus petits afin de séparer toutes sortes d'organismes de tailles à été utilisée. Ainsi, les processus de filtration ont permis d'obtenir des échantillons correspondants à différents filtres de tailles qui représentent un type majoritaire d'organismes. Il y a par exemple des échantillons correspondants aux communautés virales (0-0.2µm), bactériennes (0.22-3µm), protistes (0.8-5µm), et les métazoaires répartis sur 3 filtres (5-20µm, 20-180µm et 180-2000µm) (Figure I.4.4.iii). Les outils de mesures physiques pour relever de manière continue des paramètres environnementaux ont permis d'identifier et de caractériser les zones de prélèvements. L'échantillonnage a été réalisé à trois niveaux de profondeurs. En surface au niveau de la zone photique correspondant à la zone éclairée. Entre 20 et 100 mètres de profondeur dans la zone appelée « *Deep Chlorophyll Maximum* » (DCM) où il y a un maximum de concentration en chlorophylle et enfin dans la zone méso-pélagique se trouvant entre 300 et 400 mètres de profondeur, où la lumière est absente. Plus de 27 000 échantillons ont été collectés, la moitié pour effectuer des analyses génomiques et l'autre moitié pour réaliser de l'imagerie quantitative. La mesure des paramètres environnementaux des différentes colonnes d'eau échantillonnées a été réalisée avec la rosette CTD ainsi que d'autres instruments pour permettre l'acquisition de profils décrivant une dizaine de paramètres tels que les paramètres physiques comme la pression, la température et la conductivité ; Les paramètres chimiques comme l'oxygène et les nitrates ; Les paramètres géochimiques avec la distribution des petites et des grandes particules; Les paramètres bio-optiques avec la mesure des propriétés optiques de l'eau ou encore les paramètres biologiques comme la fluorescence. Des laboratoires situés sur le pont et à l'intérieur du bateau ont permis de réaliser les manipulations pour récupérer et conserver les différents types de micro-organismes, mais également de les visualiser et d'effectuer des analyses à l'aide de microscopes, d'un FlowCam ou encore d'un zooscan pour décider de la suite des opérations d'échantillonnage. Les échantillons ont été congelés dans de l'azote liquide et stockés à bord du bateau puis ont été transférés toutes les 6 à 8 semaines vers les laboratoires sans rompre la chaîne du froid. En tout, 13 envois d'échantillons vers les laboratoires ont été effectués avec environ 225 litres d'échantillons par déchargement. Parmi ces laboratoires, le Genoscope a été chargé de réaliser le séquençage de ces échantillons afin d'obtenir l'information génétique des micro-organismes s'y trouvant.

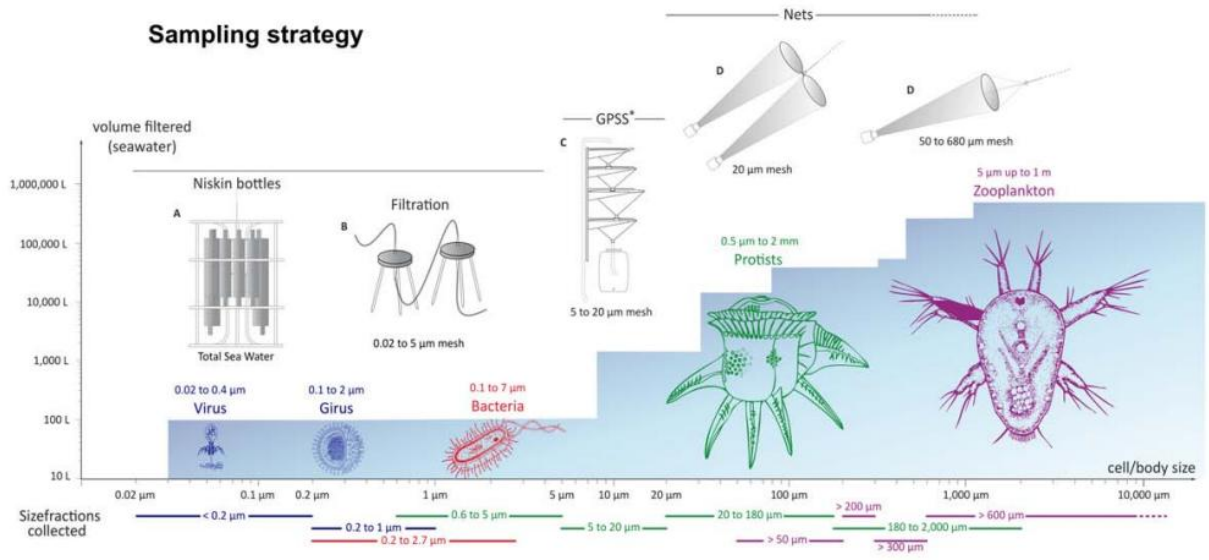


Figure I.4.4.iii | Protocole d'échantillonnage des organismes par classe de tailles et d'abondance. Le fond bleu indique le volume nécessaire pour obtenir un nombre suffisant d'organismes pour les analyses. (Figure extraite de Noan Le Bescot 2013)

Un projet multidisciplinaire

Cette expédition est la première à permettre de combiner les données génomiques, les mesures physico-chimiques et les modèles d'océanographie physique pour étudier l'écosystème planctonique, et ce, à l'échelle de la planète. C'est donc par une approche multidisciplinaire que ce projet veut tenter de résoudre les questions que nous nous posons sur l'écosystème océanique. Ainsi, de nombreuses disciplines sont impliquées dans ce projet. On retrouve notamment l'océanographie, la biologie marine, la biologie moléculaire, l'écologie, la microbiologie (bactériologie et virologie), la modélisation, la génomique, la bioinformatique ou encore l'imagerie par microscope. Ce projet multidisciplinaire fait intervenir des spécialistes de 21 laboratoires dans 7 pays. Chaque laboratoire est responsable d'un domaine d'étude ou d'un type de micro-organisme. Pour en citer quelques uns, le choix des zones de prélèvements a été fait par des spécialistes en imagerie spatiale à la station zoologique Anton Dohrn à Naples, au Department of Earth, Atmospheric and Planetary Sciences *Department of Earth, Atmospheric and Planetary Sciences* du MIT ainsi qu'à l'Institut Universitaire Européen de la Mer à Brest. L'imagerie est analysée à Villefranche sur Mer, Roscoff, Dublin ou encore à l'EMBL à Heidelberg en Allemagne. L'étude des protistes est coordonnée à la station océanographique de Roscoff, des girus à l'Institut de microbiologie de la Méditerranée à Marseille ou encore des virus à l'université d'Arizona. Les données océanographiques sont analysées et conservées à l'Observatoire de Villefranche-sur-Mer. Le Genoscope est quant à lui responsable du séquençage

et des analyses Bioinformatiques des échantillons récoltés pendant l'expédition. Les analyses génomiques de ces échantillons ont en partie pour but de mieux comprendre comment s'organise l'écosystème planctonique dans l'environnement global. Pour cela, les milliers d'échantillons prélevés dans des environnements complexes lors de l'expédition ont été séquencés et l'ensemble des séquences d'ADN générées ont été analysés avec les outils et les méthodes adéquates.

II. Étude génomique et métagénomique des micro-organismes planctoniques

Le séquençage a été inventé dans la deuxième moitié des années 1970 dans un besoin de connaissance des gènes, étape indispensable à la compréhension des phénomènes biologiques au niveau moléculaire et cellulaire. Le développement des technologies de séquençage a grandement contribué à approfondir nos connaissances sur l'écosystème marin. En effet, les progrès réalisés ces dernières années dans ce domaine ont permis le séquençage et l'étude de nombreux génomes d'organismes pris individuellement mais également l'obtention de l'information génétique de communautés de micro-organismes prélevés directement dans leur environnement. Je présenterai dans cette partie les avancées technologiques et méthodologiques dans le domaine du séquençage de l'ADN qui ont permis à de nouvelles méthodes d'analyses d'émerger, comme cela est le cas de la métagénomique.

II.1. Le séquençage de l'ADN

Suite à la découverte de la structure de l'ADN en 1953 par Jim Watson et Francis Crick, différentes méthodes de séquençage de l'ADN se sont développées à partir de la fin des années 1970. Le séquençage de l'ADN consiste à déterminer l'ordre d'enchaînements des nucléotides d'un fragment d'ADN. Il permet ainsi d'obtenir, à partir du génome d'un individu, l'ensemble des informations issues des séquences d'ADN qui composent son matériel génétique. Toutes les techniques de séquençages existantes ont un point commun : elles se basent sur les connaissances qui ont été acquises depuis une trentaine d'années sur les mécanismes de réplication de l'ADN avec l'utilisation de l'ADN polymérase. Cette enzyme est capable de synthétiser un brin complémentaire d'ADN à partir d'un brin matrice. Elle a l'avantage d'avoir une activité corrective, c'est à dire qu'à chaque incorporation d'une nouvelle base, la polymérase vérifie que cette base est la bonne, ce qui limite le taux d'erreurs. On peut estimer ce dernier à une erreur toutes les 3 milliards de bases. Depuis une dizaine d'années, des nouvelles technologies dites de séquençage « haut débit » ou « nouvelle génération » (*Next-Generation Sequencing :* NGS) ont été développées et commercialisées. Ces nouvelles technologies améliorent considérablement la rapidité et le coût du séquençage entraînant une augmentation exponentielle du nombre de séquences produites[137] (Figure II.1).

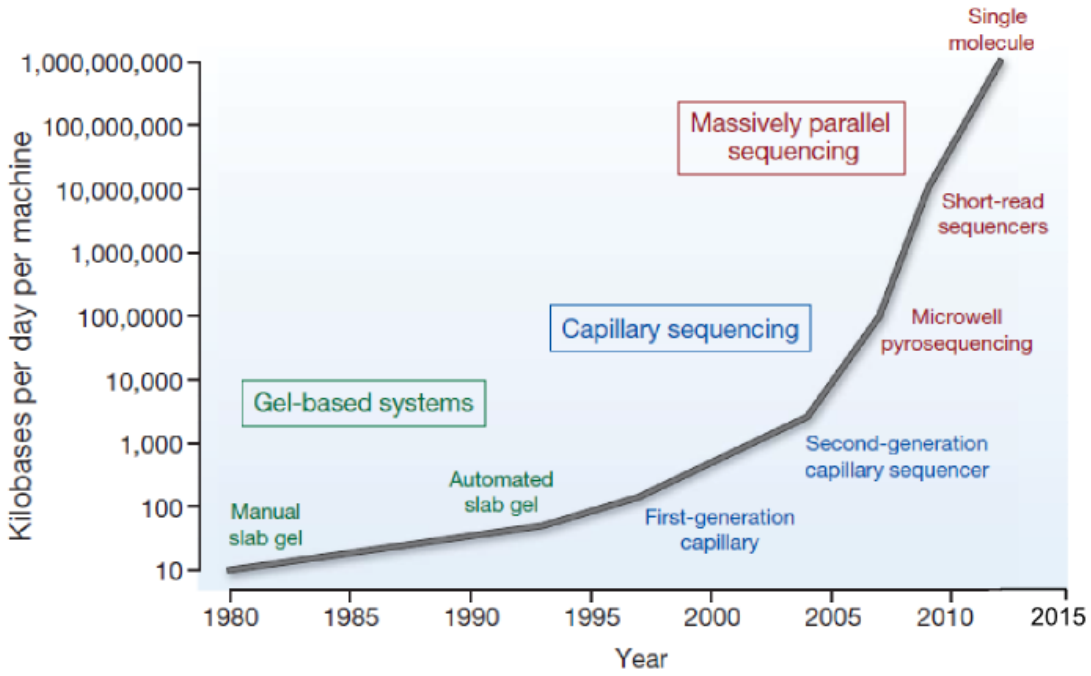


Figure II.1 | Evolution au cours du temps du nombre de kilobases produites par jour par les technologies de séquençage. (Image adaptée d'après Stratton et al. 2009[137])

II.1.1. Les débuts du séquençage

En 1977, Frederick Sanger et Alan Coulson ont publié en Grande Bretagne une méthode de séquençage de l'ADN qui a révolutionné le champ de la biologie moléculaire[138]. Cette technique dite « méthode de Sanger » a nettement amélioré la technique développée aux Etats Unis par Maxman et Glibert, publiée la même année[139] ainsi que la méthode « plus and minus » publiée par Sanger et Coulson deux ans auparavant[140] . Pour cette découverte, Gilbert et Sanger ont été récompensés par le prix Nobel de chimie en 1980. Cependant, c'est à partir de la méthode de Sanger qu'ont été développées les techniques qui suivirent. En effet, la méthode de Maxam et Gilbert nécessite des réactifs chimiques toxiques et reste limitée quant à la taille des fragments d'ADN qu'elle permet d'analyser (inférieure à 250 nucléotides). De plus, la stratégie utilisée par Sanger bénéficie de l'invention de la technique de PCR élaborée par Kary Mullis[141] et du développement de l'électrophorèse à capillaire, permettant de simplifier la partie séparative et analytique afin d'augmenter la vitesse de séquençage.

La méthode de Sanger

Initialement, la méthode de Sanger nécessitait de disposer d'un ADN simple brin qui servait de matrice pour la synthèse enzymatique du brin complémentaire. Pour cette raison, le premier organisme biologique dont le génome a été séquencé en 1977 est le virus bactériophage ΦX174. Ce dernier a son génome constitué d'ADN simple brin encapsulé dans la particule virale[142]. Le principe de cette méthode consiste à initier la polymérisation de l'ADN à l'aide d'une amorce complémentaire à une partie du fragment d'ADN à séquencer. L'élongation de l'amorce est réalisée par des ADN polymérases thermostables. Les quatre désoxyribonucléotides sont ajoutés ainsi qu'une faible concentration de l'un des quatre didésoxynucléotides (ddNTP). Ces ddNTP une fois incorporés dans le nouveau brin synthétisé, empêchent la poursuite de l'élongation. La terminaison se fait de manière statistique sur toutes les positions possibles. On obtient ainsi un mélange de fragments d'ADN de tailles croissantes qui se terminent tous au niveau d'une des bases dans la séquence. Ces fragments sont séparés par la méthode d'électrophorèse sur gel de polyacrylamide et leur détection se fait en incorporant un traceur dans l'ADN synthétisé. Initialement, ce traceur était radioactif, attaché soit à l'oligonucléotide, soit au didésoxyribonucléotide (Figure II.1.1). On obtient avec cette méthode environ 1 Kb d'ADN en 6 à 8 heures et une seule lecture par échantillon. Cependant, lors de la survenue d'un homopolymère, c'est-à-dire de la répétition d'une même base, il est difficile de connaitre le nombre de bases présentes ce qui peut induire une insertion-délétion dans la séquence. Malgré cet inconvénient, la méthode de Sanger a été l'unique méthode de séquençage utilisée pendant près de 30 ans. L'automatisation des laboratoires et la parallélisation des processus ont conduit à l'établissement de centres de séquençage hébergeant des centaines de séquenceurs. L'utilisation de cette technique est à la base du projet de séquençage du génome humain qui a débuté en 1990 pour s'achever en 2001 avec la publication de la première ébauche de ce génome[7]. Le Genoscope, Centre National de Séquençage, a apporté sa contribution à cet effort international en publiant la séquence complète et l'analyse du chromosome 14[143].

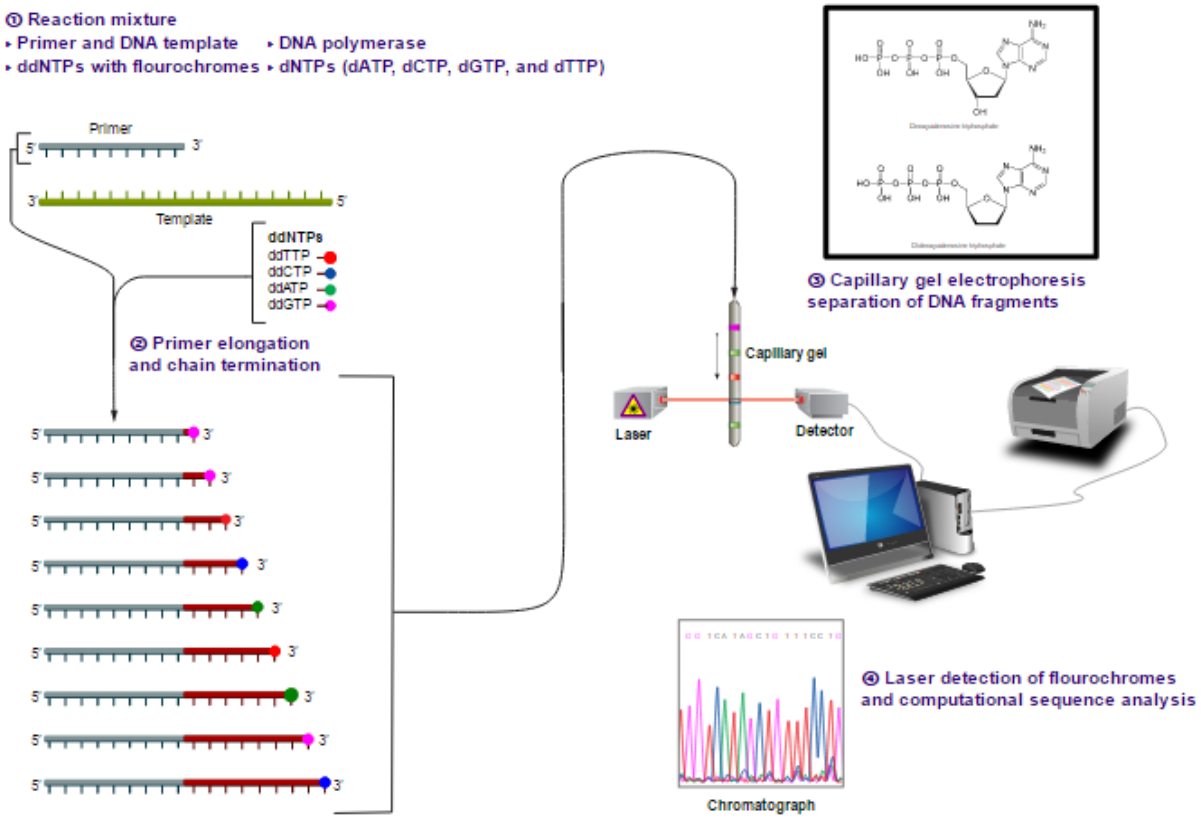


Figure II.1.1 | Principe de la méthode de Sanger et Nicken. Représentation des différentes étapes du séquençage par la méthode de Sanger**.**
(Figure extraite de https://en.wikipedia.org/wiki/Sanger_sequencing/)

Les séquenceurs à capillaires

Cette technique est basée sur l'utilisation de l'électrophorèse à capillaire. Elle est apparue dans les années 1990 suite au remplacement du marqueur radioactif par un marqueur fluorescent. On utilise donc des tubes capillaires de verre de seulement quelques microns de diamètre, sur plusieurs dizaines de centimètres de longueur pour séparer l'ADN durant l'électrophorèse. Les quatre nucléotides passent dans le même tube capillaire et on utilise quatre marqueurs fluorescents différents pour caractériser les quatre nucléotides du brin d'ADN séquencé. Avec la parallélisation des échantillons qui permet de lire plus de séquences en même temps, on obtient avec cette technique 300 kb d'ADN par lecture en 3 heures.

C'est principalement à partir de ces deux méthodes qu'ont été mises en place depuis 2005 de nouvelles méthodes de séquençage ayant en commun le clonage et l'amplification moléculaire.

II.1.2 Les techniques de séquençage à haut débit

Le lancement du concours « Archon prize for genomics » en octobre 2006 avait pour but de stimuler l'élaboration de ces techniques. Ce projet consistait à séquencer 100 génomes en 10 jours, pour moins de 10 000 dollars chacun. Trois modèles de synthèse sont sortis de ce concours basés sur le principe même du séquençage Sanger, sans le révolutionner mais en permettant son automatisation et l'augmentation du nombre de créations de séquences analysables par jour. Ces méthodes sont appelées « techniques de séquençage à haut débit de 2e génération » ou encore « séquençage de nouvelle génération » (*Next Generation Sequencing* : NGS). Elles permettent d'amplifier spécifiquement un fragment d'ADN isolé. Ainsi, les étapes de clonage bactérien particulièrement longues sont évitées. Les trois plateformes commerciales qui ont dominé le marché lors de l'apparition des NGS sont le « 454 Genome Sequencer » de Roche, le « Genome Analyser » de Illumina et le système « SOLiD » de Applied Biosystems. Pour ces trois méthodes le principe est basé sur l'obtention d'un grand nombre de séquences courtes.

La technologie 454

Le système « 454 », développé par la société Life Sciences et racheté par Roche en 2007, a été commercialisé en 2005 et constitue la première plateforme de séquençage haut débit disponible sur le marché[144]. La technologie 454 utilise la méthode de pyroséquençage qui consiste à la luminescence de fragments d'ADN par libération de pyrophosphate. Ces fragments sont isolés dans des micro-gouttes faisant office de micro-réacteurs de PCR au sein d'une émulsion. Ce protocole a pour avantage de ne pas contaminer les autres réactions de PCR. La méthode de séquençage 454 se déroule en deux étapes (Figure II.1.2.a). Dans un premier temps, l'ADN est fractionné aléatoirement en morceaux de 300 à 800 paires de bases. Des adaptateurs de quelques nucléotides spécifiques des extrémités 3' et 5' sont attachés. Chaque fragment est fixé à une bille et chaque bille est amplifiée dans une gouttelette d'une PCR en émulsion. Il y a alors la génération de copies multiples d'un même fragment d'ADN sur chaque bille. Dans un second temps, les billes sont capturées sur une plaque avec des puits d'un volume d'un picolitre et le pyroséquençage est réalisé en parallèle sur chaque fragment d'ADN. L'incorporation des nucléotides est détectée par le largage d'un pyrophosphate inorganique (PPi), ce qui conduit à la génération enzymatique de photons enregistrés par une caméra *charge coupled device* (CCD) qui assure la conversion d'un signal lumineux en un signal électrique. Ce cycle est itérativement répété pour les quatre bases adénine (A), thymine (T), guanine (G) et cytosine (C). Les erreurs majeures de séquences proviennent avec cette méthode, comme la méthode de Sanger, des homopolymères qui induisent des insertions délétions dans la séquence.

La technologie « SOLiD » (Applied Biosystems)

La société Agencourt, rachetée par Applied Biosystem en 2006, a quant à elle basé son système de détection sur le principe de l'amplification par émulsion et hybridation-ligation avec des oligonucléotides marqués[145–147]. La technologie SOLiD a une procédure d'amplification similaire au « 454 », mais la stratégie de séquençage est différente (Figure II.1.2.b). Les billes sont déposées sur une lame de verre et la séquence est déterminée par une hybridation ainsi qu'une ligation séquentielle d'oligonucléotides quasi aléatoire, avec une paire de bases identifiables par un fluorophore. Après que la couleur ait été enregistrée et l'oligonucléotide ligué enlevé, ce processus est alors répété six à sept fois afin d'obtenir une longueur de séquence d'environ 35 paires de bases. La lecture des séquences est donc effectuée dans un espace de couleur : le codage des résultats est effectué sur deux bases dans un espace de 4 couleurs ce qui permet une très grande fidélité de la lecture des résultats. On peut faire ainsi la différence entre les erreurs de séquençage et les variants réels comme les insertions, les délétions ou le polymorphisme d'un seul nucléotide (SNP). Cela représente un débit de 30 Gb/jour. Cependant, cela a un impact sur le temps de séquençage qui est plus important comparé aux autres techniques de séquençage à haut débit. En effet, le système de codage dans l'espace de couleur rend l'analyse informatique complexe.

La technologie Solexa/illumina

La plateforme de séquençage Solexa a été commercialisée en 2006 par la société Illumina. C'est aujourd'hui la technologie la plus utilisée dans les laboratoires[148]. La technologie Solexa réalise le séquençage par synthèse chimique. Celle-ci utilise des méthodes d'amplification sur support solide (lame de verre) permettant l'incorporation de bases terminateurs de chaînes réversibles marquées par des fluorochromes[149]. La première étape du séquençage Solexa (Figure II.1.2.c) est basée sur l'amplification par PCR de fragment d'ADN d'environ 150 à 200 paires de bases avec des amorces ancrées aléatoirement sur une surface solide. De multiples cycles d'amplification sont ensuite réalisés pour créer un millier de copies simple brin de chaque fragment d'ADN. Le séquençage est effectué séquentiellement à l'aide d'amorces, de l'ADN polymérase et de quatre nucléotides labellisés par un fluorophore, bloquant réversiblement la PCR. Après l'incorporation d'un nucléotide, l'image est capturée par une caméra CDD et l'identité de la première base est enregistrée. Les fluorophores sont ensuite retirés et les étapes d'incorporation, de détection et d'identification sont répétées. Contrairement à la méthode de Sanger et 454, on reconstruit base après base la lecture ce qui évite les erreurs d'insertion-délétion dues aux homopolymères. Cependant, il peut y avoir des erreurs lors de la PCR qui induiront des substitutions dans la séquence. Cette technique est efficace avec un taux d'erreurs

assez faible. De plus, avec la technologie Solexa, il est possible de séquencer les fragments par leurs deux extrémités. Après avoir terminé de séquencer les fragments dans un sens, les séquences amplifiées sont éliminées et les matrices peuvent être régénérées *in situ,* pour être séquencées dans l'autre direction. Cette méthode est appelée « séquençage pairé » (*Paired-end*). Il est également possible de réaliser du séquençage *Mate Pair.* Cette technique, réalisée au moment de la préparation des échantillons, permet de marquer deux fragments ayant une distance connue entre eux dans le génome. Pour cela, des fragments d'ADN génomique de tailles connues sont biotinylés à leurs extrémités. L'ADN est ensuite circularisé et à nouveau fragmenté pour générer des petits fragments dont certains contiennent les deux extrémités du fragment originel, donc les deux étiquettes de biotine. Ces fragments, dits *Mate Pair,* sont enrichis en utilisant les étiquettes et séquencés.

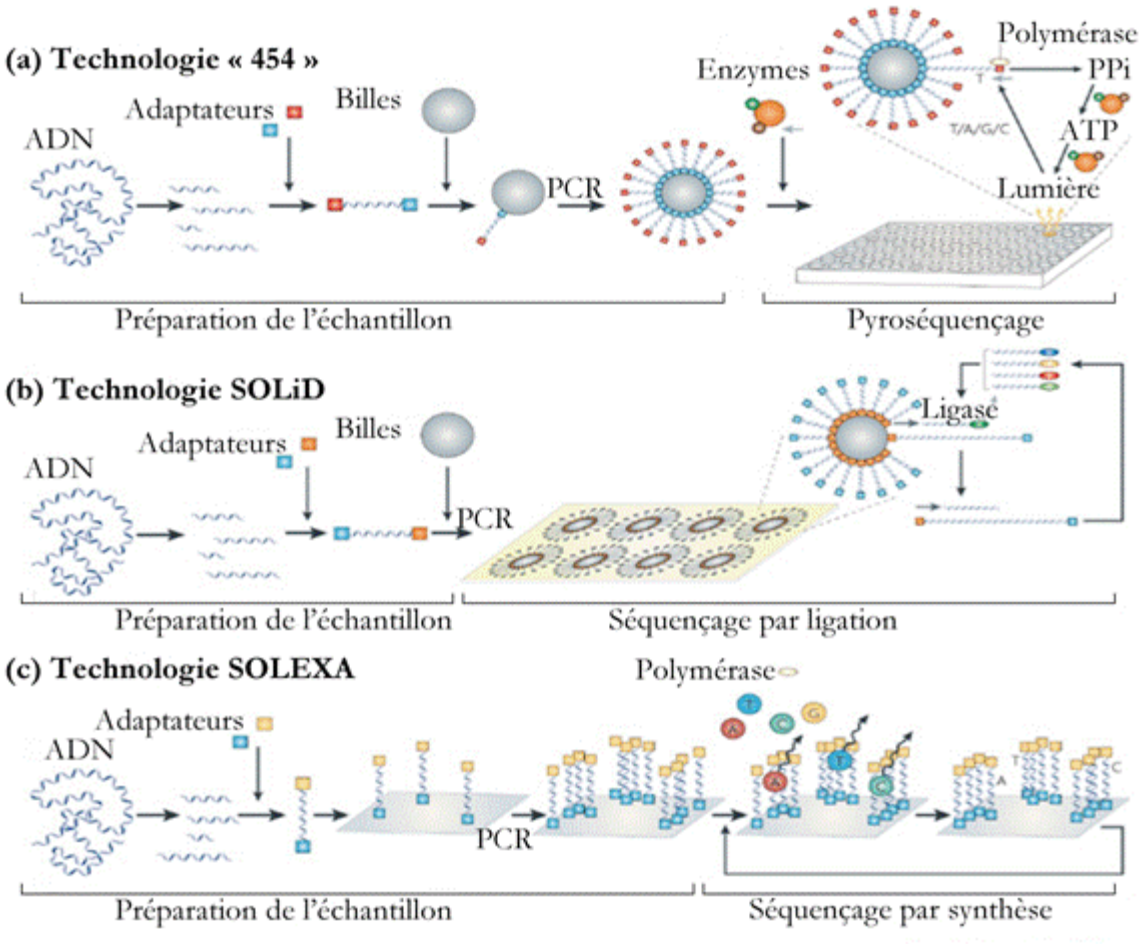


Figure II.1.2 | Illustration des différentes techniques de séquençage à haut débit. a, La méthode de séquençage « 454 », **b**, La technologie SOLiD et **c**, La technologie SOLEXA (Figure extraite de Medini *et al.* 2008[150])

II.1.3 Technologies de séquençage de « 3e génération »

Suite à l'arrivée des séquenceurs à haut débit de 2e génération, de nouvelles technologies de séquençage sont encore en train d'émerger. Ces technologies de séquençage, dites de « 3e génération » permettent le séquençage en temps réel à partir d'une molécule d'ADN unique. Cela permet d'augmenter considérablement la vitesse de séquençage et de s'affranchir des erreurs potentiellement introduites lors de l'amplification PCR. De plus, contrairement aux technologies de séquençage de 2e génération qui produisent de courtes séquences, les technologies de séquençage de 3e génération ont la capacité théorique de générer de plus longues séquences d'ADN.

Technologie HeliScope

La société Helicos Biosciences a développé la technologie HeliScope qui est le premier système de séquençage à partir d'une molécule unique[151]. Elle est basée sur la détection de fluorescence sur les nucléotides marqués avec le même fluorophore. Actuellement, Helicos ne vend plus d'appareils mais propose un service de séquençage.

Technologie SMART

Pacific Biosciences a développé le premier instrument capable de séquencer une molécule unique en temps réel. Le PacBio RS utilise la technologie « SMART » utilisant des cellules *SMRT* (*Single Molecule Real Time technology*)[152,153]. Cette technologie consiste donc à réaliser la mesure en temps réel de l'incorporation de chaque base associée à un fluorochrome avec une caméra CDD placée sous la plaque support. Cette plaque en verre est supportée par une plaque en métal perforée de trous d'une dizaine de nanomètres de diamètre formant une nano-structure (*Zero Mode Waveguide*). Ce système permet de détecter des quantités de fluorescence très faibles. L'ADN polymérase étant immobilisé au fond des puits, à chaque fois qu'un nucléotide marqué est incorporé par cette dernière, sa fluorescence est détectée. L'intervalle de temps entre chaque pic ainsi que la durée de chaque pic de fluorescence sont propres à chaque nucléotide et permettent ainsi leur identification. Ce système permet de générer des séquences de taille moyenne de 1 000 pb avec un temps de préparation de l'échantillon d'environ 30 minutes et un temps de processus de séquençage de quelques minutes seulement. Il est possible ainsi de séquencer 600 bases/minute, alors que le taux de polymérisation de l'ADN polymérase dans la cellule est d'environ 3 000 bases/minute.

Technologie nanopores

Le séquençage par nanopores[154] permet de déterminer une séquence d'ADN à la résolution du nucléotide sans aucune amplification. Les bases et leur statut de méthylation peuvent être déterminés en temps réel grâce au courant qui traverse le pore, avec une grande précision (99,8%). Une molécule unique d'ADN traverse un pore formé par une protéine ancrée dans une bicouche lipidique par application d'un potentiel. Une exonucléase clive chaque nucléotide à l'entrée du nanopore et celui-ci est alors détecté de façon électronique via une cyclodextrine[155]. Aucun traitement ni marquage préalables ne sont nécessaires et de faibles quantités d'ADN suffisent. De plus, aucun système optique n'est nécessaire ce qui réduit le coût d'un séquençage. Plusieurs compagnies travaillent sur cette technologie : Oxford Nanopore Technologie, Nabsys, Electronic BioSciences, BioNanomatrix, GE research, LingVitae, Complete Genomics, CrackerBio ou encore IBM. Oxford Nanopore a fait une avancée importante sur la miniaturisation de ce type de séquenceur et commercialise actuellement le MinION qui a la taille d'une clé USB et dont le coût ne dépasse pas les 1 000$ (Figure II.1.3).

Figure II.1.3 | Le séquenceur MinION. La miniaturisation des séquenceurs haut débit permet une utilisation dans divers environnements, comme sur un bateau par exemple (Figure extraite de https://nanoporetech.com/).

Autres technologies de 3e génération

D'autres technologies ont été développées en parallèle et essaient d'émerger pour arriver à occuper le marché du séquençage de 3e génération. La start-up française Depixus (anciennement PicoSeq) commercialise en 2016 une technologie qui utilise une approche biophysique pour extraire l'information génétique et épigénétique d'une molécule unique d'ADN. Cette technologie utilise une nouvelle approche SIMDEQ™[156] (*Single-molecule Magnetic DEtection and Quantification*). Une autre méthode consistant à la détection optique par

microscopie électronique[157] est également explorée par les sociétés LightSpeed, Halcyon et ZS Genetics. La résolution de l'image obtenue lors de l'observation de l'ADN n'est pas suffisamment élevée pour permettre le séquençage. Cependant, avec un marquage différentiel des bases de l'ADN avec des atomes lourds ou des métaux, il est possible d'obtenir un signal correspondant à chaque acide aminé. La société IBM a travaillé sur un séquençage à l'aide de transitors alors que la société Life Technologies a quant à elle développé un séquenceur appelé Starlight utilisant des particules nanométriques à base de semi-conducteurs. De nombreuses autres sociétés ont élaboré leur propre technique de séquençage dans le but de la commercialiser et dans l'espoir de l'utiliser dans de nombreux laboratoires. Il semblerait qu'à l'heure actuelle, les sociétés Oxford Nanopore Technologie et Pacific Biosciences sont bien parties pour dominer le marché du séquençage dans les prochaines années.

Ainsi, les séquenceurs de 3e génération ont l'avantage de générer de longues séquences en un temps record. Cependant, la faible qualité des séquences qui sont produites reste un obstacle pour l'assemblage de génomes complexes. L'utilisation conjointe des deux types de technologies de séquençage semble être une bonne alternative pour améliorer ces assemblages complexes[158]. De part leur maturité, les séquenceurs de deuxième génération restent actuellement les plus utilisés dans les laboratoires. Cependant, la miniaturisation et l'amélioration des performances des séquenceurs de 3e génération ne font aucun doute sur la place dominante que prendront ces derniers dans les laboratoires dans les années à venir.

II.1.4 Performance des techniques de séquençage à haut débit

Les performances des technologies de séquençage peuvent être évaluées sur cinq critères principaux[159,160].

Temps d'un processus de séquençage

Le temps d'un processus de séquençage ou *run* correspond au fait de réaliser un processus complet, préparation des échantillons puis séquençage. Celui-ci est un facteur important à prendre en compte lorsque de nombreux échantillons sont à séquencer. Cela peut-être le cas pour certains projets de métagénomique. Plusieurs autres critères sont à considérer pour évaluer ce temps comme notamment la quantité de lectures générées au terme du *run*.

Quantité de lectures générées

Le nombre de lectures générées pendant un *run* est lié à la profondeur de séquençage. La profondeur de séquençage est le rapport entre la longueur de l'ensemble des séquences lues mises bout à bout et la longueur du génome cible. Par exemple, si l'on séquence 100Mb pour un

génome de 10Mb, on a une profondeur de 10 équivalents génome, ce que l'on note 10X. Plus la profondeur de séquençage est grande et plus le nombre de lectures chevauchantes que l'on peut assembler et donc la fraction de génome couverte sera importante.

Taille des lectures générées

Plus la taille des séquences est grande et plus on a accès à la diversité nucléotidique par séquence. Il est possible d'avoir des séquences de plus grande taille par des méthodes d'assemblages utilisées après l'étape de séquençage. La taille des lectures est un facteur à prendre en compte pour mener à bien un assemblage. Actuellement, ces méthodes permettent de reconstituer le génome d'un organisme séquencé individuellement. Cependant, l'assemblage des séquences issues du séquençage d'un échantillon contenant différents organismes ne permet pas d'obtenir l'ensemble des génomes des organismes présents dans cet échantillon. En effet, la couverture des différentes espèces dans un échantillon peut être très inégale, et certains organismes peuvent avoir une faible couverture. Les logiciels d'assemblage ne fonctionnent pas correctement sur ces données. De plus, des organismes distincts peuvent posséder dans leur génome des régions d'ADN similaires. Il peut y avoir alors une fusion des séquences provenant de ces différents organismes créant alors des séquences chimériques[161]. Cependant, certains logiciels d'assemblages permettent, non pas d'essayer de reconstituer des génomes complets, mais de créer de plus grandes séquences que les séquences initiales. Les séquences ainsi générées sont appelées des contigs[162]. Un des défis des technologies à haut débit est donc de séquencer directement des séquences de grande taille sans pour autant perdre en qualité de séquençage.

Taux d'erreurs de séquençage

Il est possible de calculer le score de qualité d'un séquençage (*Quality Control* : QC) en se basant sur la probabilité d'erreurs du nucléotide. Le premier programme créé pour développer un système précis et puissant de score assigné à chaque base a été le programme Phred[163]. Les scores de qualité Phred Q ont pour propriété d'être reliés de façon logarithmique à la probabilité d'erreurs d'identification d'une base P^2.

$$Q = -10\, log_{10} P$$

Ainsi, si le score qualité Phred est de 40, la probabilité d'identification d'une base incorrecte est de 1 pour 10 000. Ce score est devenu un standard pour caractériser la qualité d'une séquence d'ADN et est utilisé pour comparer l'efficacité des différentes techniques de séquençage.

Le prix du séquençage

Le coût d'un séquençage varie entre les différentes techniques. Cela s'explique notamment par le type et le volume de réactifs utilisés pour le séquençage, de la miniaturisation du processus de séquençage, de la fixation des fragments sur support solide ou encore de l'utilisation d'un système optique. Depuis l'avènement des technologies de séquençage à haut débit, le prix du séquençage ne cesse de baisser et suit une décroissance exponentielle (Figure II.1.4). La croissance exponentielle de la puissance de calcul des ordinateurs peut-être représentée comme suivant la « loi de Moore ». La baisse du prix d'un séquençage de l'ADN est plus rapide que l'augmentation de la puissance des ordinateurs depuis 2003. La différence entre les deux exponentielles est elle-même une exponentielle. C'est-à-dire que chaque année, il est possible de séquencer pour le même prix encore plus d'ADN par minute de calculs possibles. Ainsi, la capacité à séquencer de l'ADN croit plus vite que la puissance informatique. Ce coût de séquençage continue encore à décroitre avec notamment une rupture en 2015 de la baisse constante du prix de séquençage observée depuis 2012. Ce faible coût de séquençage implique avant tout une quantité importante de données générées par les séquenceurs qui devront être traitées et analysées avec des moyens personnels et des matériels conséquents. Notamment avec l'utilisation de puissance de calcul performante et des outils développés spécialement pour analyser ces « big data ». Cette notion sera étudiée en partie III.6.

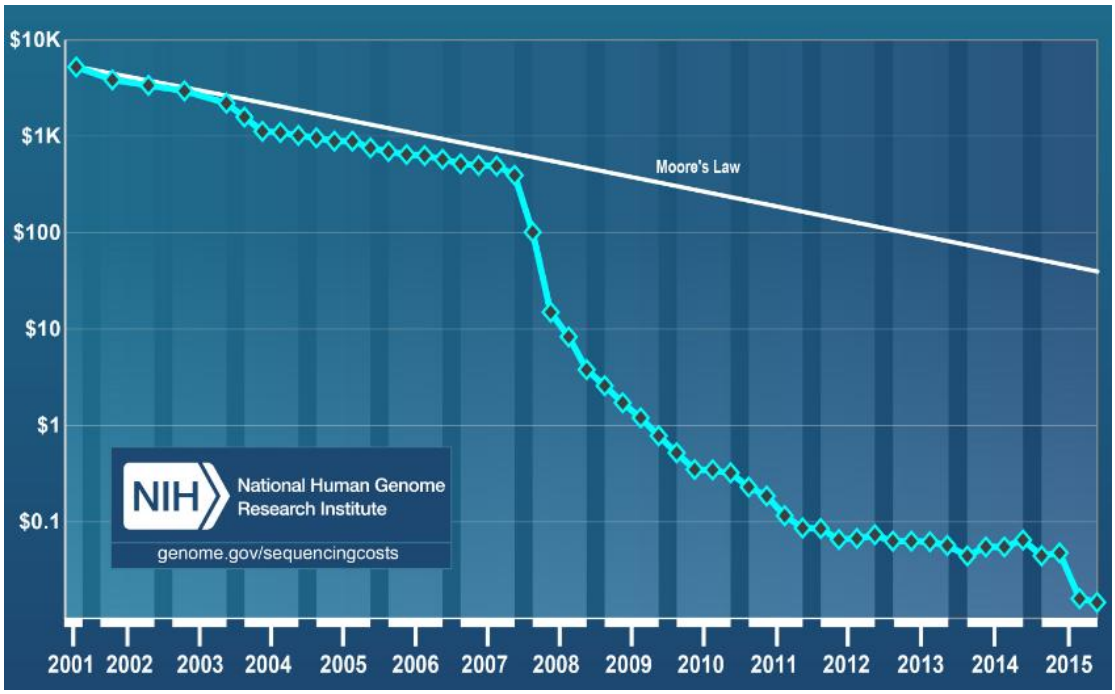


Figure II.1.4 | Représentation de l'évolution dans le temps du coût de séquençage d'une megabase d'ADN (échelle logarithmique). La loi de Moore montre l'évolution parallèle de la puissance de calcul disponible pour traiter ces données.
(Figure extraite de www.genome.gov/sequencingcostsdata).

II.1.5 Comparaison des techniques de séquençage à haut débit

Depuis leurs apparitions, ces techniques ne cessent d'évoluer en termes de performance. Je me focaliserai dans cette partie, sur les caractéristiques des technologies disponibles en 2009, date à laquelle le projet *Tara* Oceans a été mis en place, et donc principalement sur les séquenceurs de 2e génération. Cela permettra d'expliquer le choix de la technologie utilisée pour le séquençage des échantillons marins ainsi que de décrire les caractéristiques des séquences qui ont été analysées pendant cette thèse. Chacune de ces techniques de séquençage a ses avantages et ses inconvénients (Figure II.1.5). Le type de séquenceur à utiliser dépendra du type d'analyse à effectuer ainsi que de la quantité de données à séquencer. La méthode Sanger a l'avantage de générer de grandes lectures de bonne qualité. Cette technique a permis de fournir des informations sur la diversité microbienne de différents sites anatomiques du corps humain[164,165]. Cependant, le coût élevé et la faible quantité de données générées pour réaliser des analyses sur un grand nombre d'échantillons font que cette technique a été supplantée par les séquenceurs à haut débit. La technologie 454 a la particularité de produire des lectures de plus grande taille en un temps plus rapide que les technologies SOLiD et Illumina. Cependant, le nombre d'erreurs de séquençage est plus élevé avec la technologie 454. En effet, les principales erreurs de séquences détectées sont des insertions/délétions dues à la présence de régions homopolymères. Leur identification repose sur l'intensité du signal lumineux produit par la réaction de pyroséquençage. Des signaux d'intensité trop élevée ou trop faible entraînent une sous ou surestimation du nombre de nucléotides[166]. De plus, le coût de séquençage 454 est presque dix fois plus élevé qu'avec les autres techniques de nouvelle génération. La technologie Illumina présente l'avantage d'avoir un débit important et un coût/base le plus faible. Malgré le fait que de courtes séquences soient générées, cette technologie a été adoptée par la communauté scientifique. De ce fait, de nombreux outils bio-informatiques en libre accès ont été développés pour l'analyse des séquences générées par les séquenceurs Illumina. C'est pour ces raisons que la technologie Illumina a été utilisée pour le séquençage de la majorité des échantillons *Tara* Oceans. Le séquençage réalisé pour ce projet sera décrit en partie II.4.4.

(a)

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400 ~ 900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

(b)

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Instrument price	Instrument $500,000, $7000 per run	Instrument $690,000, $6000/(30x) human genome	Instrument $495,000, $15,000/100 Gb	Instrument $95,000, about $4 per 800 bp reaction
CPU	2* Intel Xeon X5675	2* Intel Xeon X5560	8* processor 2.0 GHz	Pentium IV 3.0 GHz
Memory	48 GB	48 GB	16 GB	1 GB
Hard disk	1.1 TB	3 TB	10 TB	280 GB
Automation in library preparation	Yes	Yes	Yes	No
Other required device	REM e system	cBot system	EZ beads system	No
Cost/million bases	$10	$0.07	$0.13	$2400

Figure II.1.5 | Tableau comparatif des NGS en 2010. a, Avantage et type de séquençage. **b**, Coût de séquençage et caractéristiques des machines (Figure extraite de Wooley 2010[160]).

II.2. La génomique

L'évolution des technologies de séquençage a permis le séquençage et l'étude de nombreux génomes issus de divers organismes pris individuellement. Le génome est l'ensemble du matériel génétique d'un individu ou d'une espèce codé dans son ADN. La génomique consiste à l'étude du vivant à l'échelle du génome. La connaissance des gènes est une étape indispensable à la compréhension des phénomènes biologiques d'un organisme. Aujourd'hui, les applications sont de plus en plus nombreuses dans les domaines de la médecine et des industries pharmaceutiques, biotechnologiques, agro-alimentaires, ainsi que dans d'autres domaines en prise directe avec les processus biologiques comme l'étude de l'environnement. Pour toutes ces applications, la séquence est le point de départ.

II.2.1. La reconstruction des génomes.

Les séquenceurs actuels ne permettent pas d'obtenir des séquences assez longues pour avoir directement accès au génome complet d'un organisme. Pour cela, il est indispensable de passer par une étape d'assemblage. L'assemblage de séquences consiste à aligner et fusionner les lectures issues du séquençage de l'ADN afin de reconstruire le génome ou la séquence de

départ. Un assemblage est généralement plus difficile si le génome étudié est grand et riche en séquences répétées. C'est pour cela que les virus qui possèdent de petits génomes sans séquences répétées font partie des premiers génomes séquencés et assemblés. *Haemophilus influenzae* est le premier génome bactérien à avoir été séquencé en 1995[5]. De nombreux autres procaryotes ont depuis été séquencés dans leur intégralité. La taille des génomes de procaryotes est de l'ordre de quelques millions de paires de bases, on parlera alors de mégabases (Mb). Pour les organismes eucaryotes, la taille de leur génome pouvant dépasser le milliard de paires de bases nécessite un effort de séquençage et de traitement en aval plus important. *Saccharomyces cerevisiae* en 1996[6] est le premier génome eucaryote séquencé. Le nombre de génomes séquencés ne cesse d'augmenter depuis, notamment en 2013 la venue des séquenceurs de 2e génération (Figure II.2.1). Le traitement de l'ensemble des séquences générées pour l'obtention de l'assemblage de génomes complets est impossible aujourd'hui avec les moyens matériels et humains disponibles. En effet, en 2016, « seulement » 1 546 génomes séquencés ont pu être assemblés dans leur intégralité alors que 19 220 génomes « brouillons » ont été générés.

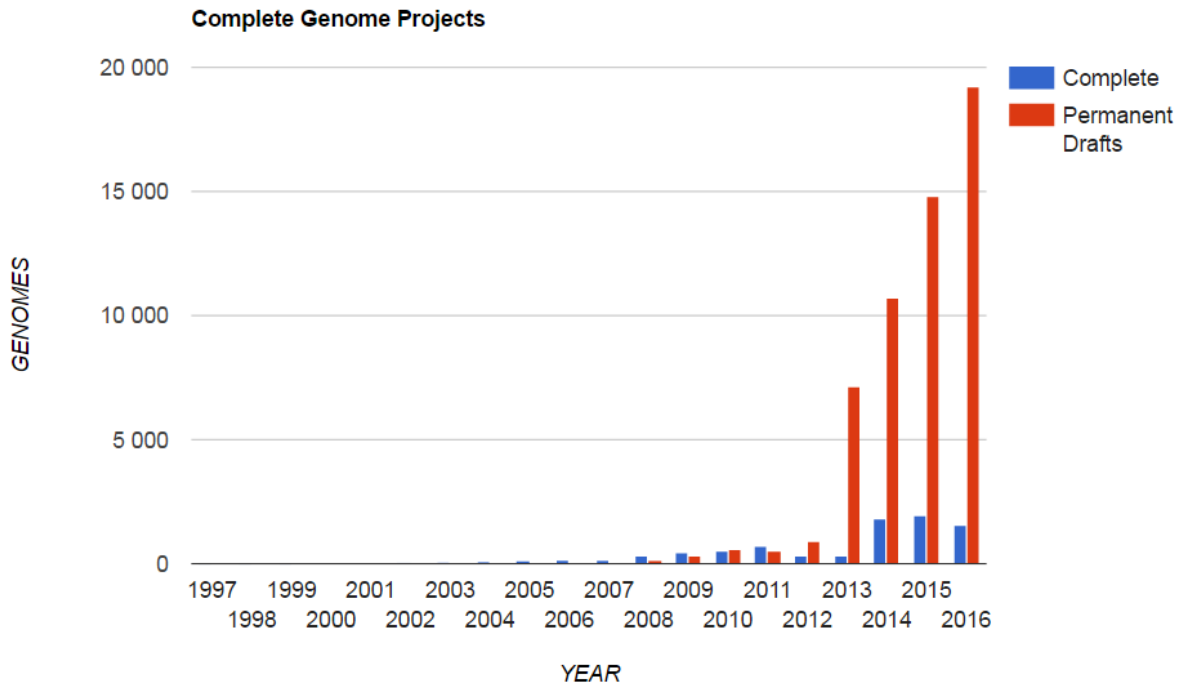


Figure II.2.1 | Nombre de génomes séquencés au cours du temps. On observe une très forte hausse de génomes séquencés à partir de 2013 avec la baisse du prix de séquençage et l'augmentation de séquences produites par les séquenceurs de 2e génération. Il y a presque 10 fois moins d'assemblages de génomes complets que d'assemblages de génomes « brouillons » en 2016 du fait de la proportion importante de séquences produites par rapport au traitement nécessaire à l'assemblage (Figure extraite de Genome Online Database[167,168]).

Lorsque les organismes ne sont pas connus, donc pas encore séquencés ni assemblés comme c'est le cas de la majorité des micro-organismes planctoniques, aucune référence n'est disponible. Il est donc nécessaire de reconstruire ces génomes à l'aide de différentes approches.

II.2.2. Le séquençage de novo

Le séquençage *de novo* consiste à obtenir pour la première fois la séquence génétique d'un organisme. Les technologies de séquençage haut débit permettent de séquencer des génomes *de novo* plus vite et à moindre coût qu'en utilisant la méthode Sanger. En effet, le séquençage dit *shotgun* va couper aléatoirement l'ADN en de nombreux petits segments qui seront amplifiés et séquencés pour obtenir des lectures. Ces lectures ainsi séquencées sont ensuite assemblées. L'assemblage est généralement réalisé en 2 étapes (figure II.2.2). Dans un premier temps, les lectures sont assemblées en utilisant leur chevauchement. Résultera de ces assemblages de plus grandes séquences appelées contigs. Plusieurs méthodes existent pour assembler les lectures en contigs mais la plus utilisée actuellement est la modélisation en graphes de *De Bruijn*. Celle-ci repose sur un graphe où les fragments séquences sont décomposés en *k-mers*, des petits fragments de taille k homogènes. Le graphe de *De Bruijn*, dont les nœuds sont les (k – 1)-mers, et les arêtes relient deux (k – 1)-mers apparaissant dans un k-mer séquencé. On peut alors reconstruire la séquence initiale à la recherche d'un chemin eulérien en passant par toutes les arêtes du graphe. Cette méthode est notamment utilisée par les assembleurs Velvet[169] et Minia[170]. Dans un second temps, une orientation est donnée aux contigs et la taille des espaces entre ces derniers est estimée. C'est la phase de scaffolding. Le séquençage *Mate Pair* est alors souvent utilisé dans le séquençage *de novo* puisque cette approche permet de réduire les zones non couvertes dans le génome et relier les contigs l'un à l'autre pour créer les scaffolds. En effet, cette technique réalisée au moment de la préparation des échantillons pour le séquençage, permet de marquer deux fragments d'une taille de plusieurs kilo-bases ayant une distance connue entre eux dans le génome. Le séquençage *paired end* permet également lors de cette étape de scaffolding de donner les informations sur l'orientation des lectures sur les contigs afin d'obtenir l'orientation relative des contigs 2 à 2. De plus, cela nous permet de définir le sens dans le génome afin d'en déduire l'ordre des contigs. Enfin, la connaissance de la distance séparant les deux lectures d'une paire donne approximativement la distance entre deux contigs. De nombreux outils pour le scaffolding existent tel que Sspace[171], Sopra[172] ou encore Opera[173] qui tentent de contourner les problèmes inhérents à celui-ci. Par exemple, les répétitions dans le génome sont difficilement repérables au moment du contigage. De plus, il peut y avoir de potentielles erreurs lors du séquençage. Ainsi, bien avant le scaffolding, les erreurs précédentes vont conduire à des erreurs de contigage créant ainsi des contigs dits chimériques, qui n'ont aucune existence dans la réalité biologique.

La résolution de ces problèmes de scaffolding demande une gestion du temps de calcul qui est d'autant plus long que le génome à assembler est grand. Suite à l'arrivée des séquenceurs de 3e génération qui produisent de longues séquences de faible qualité, de nouveaux assembleurs sont élaborés utilisant de nouvelles approches. C'est le cas de l'assembleur NaS[158] qui combine les longues lectures MinION avec les courtes lectures Solexa afin de générer des assemblages de génomes microbiens ou de petits eucaryotes de très bonne qualité en des temps plus rapides que les assembleurs classiques.

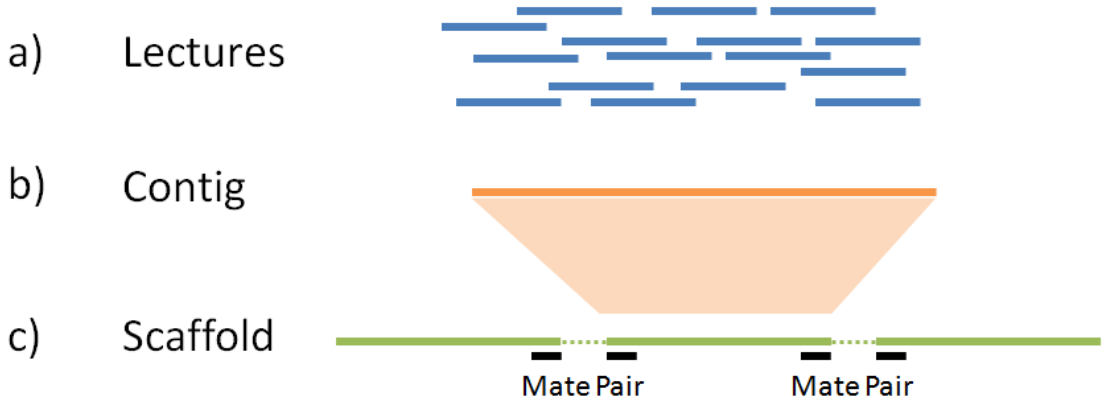


Figure II.2.2 | Les étapes d'assemblage d'un séquençage de novo. a, Lectures issues du séquençage haut débit ; **b,** assemblage des lectures en contigs ; **c**, assemblage des contigs en scaffolds avec l'utilisation du séquençage Mate Pair.

II.2.3. Le séquençage de génome à cellule unique

Il est maintenant possible de séquencer le génome d'un organisme à partir d'une cellule unique prélevée directement dans son environnement. Cela permet d'identifier et d'assembler des génomes d'organismes non cultivables[174]. De plus, en utilisant cette méthode de séquençage les fonctions cellulaires peuvent faire l'objet d'études approfondies[175] puisque l'on ne se base pas à l'échelle de l'organisme, dans le cas d'organisme multicellulaire, comme cela était réalisé auparavant[176,177]. Cette technique implique l'isolation d'une cellule unique puis la réalisation d'une amplification du génome dite *whole-genome-amplification* (WGA) avant de procéder au séquençage et à l'assemblage. L'étape d'amplification est donc cruciale puisqu'il est impossible de séquencer un génome avec une quantité d'ADN aussi faible que celle obtenue dans une seule cellule. Les techniques d'amplification ont le désavantage de copier en grand nombre certaines régions du génome au détriment d'autres régions qui seront alors indétectables au séquençage. De plus, celles-ci peuvent introduire des mutations ou encore des séquences chimèriques[178]. Ces

techniques permettent de couvrir en moyenne 40% du génome[179]. Différentes méthodes d'amplifications sont généralement utilisées pour ce type de séquençage (Figure II.2.3).

Méthode DOP-PCR

La méthode DOP-PCR[180] (Degenerate Oligonucleotide Primed-PCR) fait intervenir des amorces aléatoires suivies d'une amplification par PCR qui amplifie préférentiellement des sites spécifiques dans le génome. Il en résulte une faible couverture de séquençage mais une bonne uniformité dans l'amplification.

Méthode MDA

La méthode MDA[181] (Multiple Displacement Amplification) consiste quant à elle à réaliser une amplification iso-thermique en faisant intervenir des amorces aléatoires de type hexamères et une enzyme, la phi29. Cette enzyme parcourt le brin néo-synthétisé et déplace un brin complémentaire pour continuer sa synthèse. Les brins générés par cette technique peuvent atteindre 100 kb avec un taux d'erreurs moins élevé que dans la technique précédente mais une moins bonne uniformité de l'amplification. Une nouvelle méthode hybride d'amplification de l'ADN a été développée pour améliorer la couverture de génome séquencé ainsi que l'uniformité de l'amplification.

Méthode MALBAC

La méthode MALBAC[182] (*multiple annealing and looping-based amplification cycles*) consiste donc à initier l'amplification iso-thermique de l'ADN isolé d'une cellule à l'aide d'amorces aléatoires associées à 27 nucléotides dont la séquence est identique pour chaque amorce. Un cycle de PCR est généré produisant des amplicons partiels. Au cours du cycle suivant, ces amplicons partiels vont servir de support aux amorces et permettre la formation d'amplicons complets qui possèdent à chacune de leurs extrémités des séquences complémentaires. Ils vont alors former une boucle et ne serviront plus de support aux amorces lors des prochains cycles, limitant ainsi le biais d'amplification. Cependant, avec cette technique, un tiers des variations nucléotidiques n'est toujours pas détecté.

Co-assemblage des séquences de cellules uniques

D'autres alternatives pour améliorer la couverture de séquençage ainsi que la détection des variations nucléiques existent. Il est possible par exemple de séquencer plusieurs cellules appartenant au même organisme. On réalise ensuite un co-assemblage avec les séquences issues du séquençage de chaque cellule unique. Chaque cellule participe à l'assemblage d'un contig

permettant d'augmenter la taille des génomes séquencés. Il faut cependant faire attention d'utiliser des cellules appartenant à la même espèce sans variation intra-population. En effet, cela pourrait fragmenter l'assemblage voire générer des séquences chimériques. Cette méthode a été utilisée pour des cellules récoltées lors de l'expédition *Tara* Oceans[183]. Certains génomes, comme *Bathycoccus prasinos* ont pu être reconstitués avec une couverture pouvant atteindre 62%. Une étude de génomique comparative entre ce génome partiel et le génome de la souche de référence a été réalisée et est présentée dans le chapitre I.1.

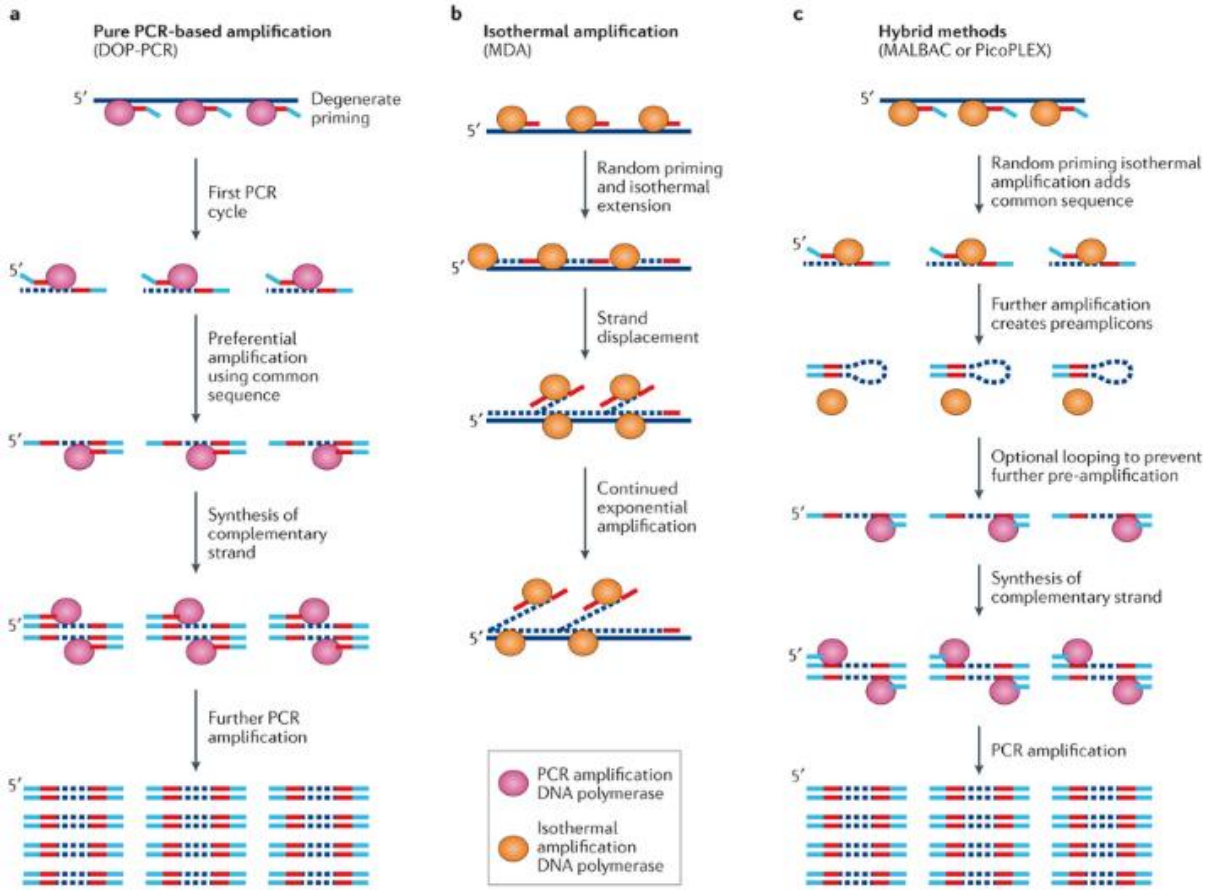


Figure II.2.3 | Les trois méthodes principales d'amplification utilisées pour un séquençage de génome à cellule unique. a, Amplification par PCR : DOP-PCR ; **b**, Amplification isothermique: MDA ; **c**, Méthodes hybrides : MALBAC ou PicoPLEX (Figure extraite de Gawad *et al.* 2016[178]).

II.2.4. Séquençage RNA-Seq

Le séquençage de l'ARN ou RNA-Seq (*RNA sequencing*) utilise également le séquençage haut débit afin d'identifier et de quantifier l'ARN issu de la transcription d'un génome à un instant donné. Ici, on n'étudie plus le génome mais le transcriptome. Le transcriptome représente l'ensemble des transcrits d'une cellule. L'analyse qualitative et quantitative de ces

transcrits permet d'obtenir des informations fondamentales pour comprendre l'expression des gènes. En effet, il est possible d'assembler des séquences d'ARN messagers à partir des données RNA-seq afin d'obtenir les transcrits différents d'un même gène. L'étude de ces transcrits consiste à faire de la génomique fonctionnelle. L'approche de RNA-Seq[184] (Figure II.2.4) consiste dans un premier temps à extraire l'ARN total des micro-organismes du milieu. Les ARN polyadénylés permettent de sélectionner les ARNm stables en étant isolés par l'utilisation de billes magnétiques liées à des olygonucléotides poly(T). Il est également possible de séparer l'ARN ribosomal qui ne participe pas à la traduction en protéine par une étape de déplétion ribosomique. Différents protocoles permettent également d'isoler l'ARN nucléaire, l'ARN cytoplasmique ou encore d'autres types d'ARN. L'étape de transcription inverse permet ensuite de convertir les ARN en ADN complémentaire (ADNc). Les librairies de séquençage sont ensuite élaborées en liguant les adaptateurs sur les extrémités de l'ADNc puis, après fragmentation, ces séquences sont amplifiées par PCR. Les librairies sont enfin séquencées via un séquençage haut débit. Comme pour le génome, il est possible d'étudier le transcriptome à partir d'une cellule unique[185].

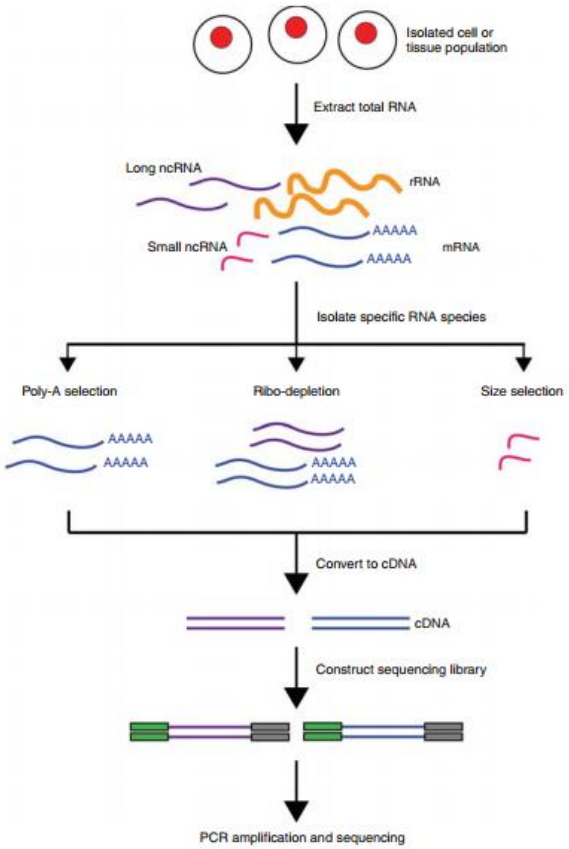


Figure II.2.4 | La méthode RNA-Seq (Kukurba *et al.* 2015[184]).

II.2.5. Séquençage de codes-barres à ADN

Les études de génomique ces dernières années ont permis d'approfondir nos connaissances sur les micro-organismes marins, qui sont pour la plupart inconnus, notamment pour les protistes. En effet, les analyses en imagerie par microscopie ne donnent que peu d'informations sur l'identité de ces cellules de nano- et pico-planctons du fait de leur petite taille et de leurs morphologies souvent proches voir identiques dans le cas d'espèces cryptiques. L'utilisation depuis 2003 des codes-barres ADN ou « *DNA barcoding* » en phylogénie moléculaire a permis d'identifier des organismes non cultivables à partir d'un gène[186]. Cette méthode consiste à extraire l'ADN d'un organisme puis à amplifier par PCR le fragment cible correspondant aux codes-barres à l'aide d'un couple d'amorces préalablement défini. La quantification de la biodiversité taxonomique d'un échantillon est alors possible en utilisant ces codes-barres phylogénétiquement informatifs. Cela permet d'assigner des espèces inconnues à un taxon d'ordre supérieur. On peut utiliser les différentes séquences obtenues pour définir des unités opérationnelles, ou OTUs. Le DNA barcoding peut permettre d'identifier des espèces cryptiques directement prélevées dans leur milieu. Chez les protistes, le groupe de marqueurs le plus couramment utilisé pour une analyse phylogénétique est situé sur l'ADN ribosomal car celui-ci est présent partout dans le monde vivant en de nombreuses copies et répété dans le génome. De plus, il évolue relativement lentement. Ces marqueurs correspondent à l'ADNr 18S, l'ADNr 28S, les espaceurs transcrits internes 1 (ITS1) et 2 (ITS2) ainsi que l'ADNr 5,8S (Figure II.2.5).

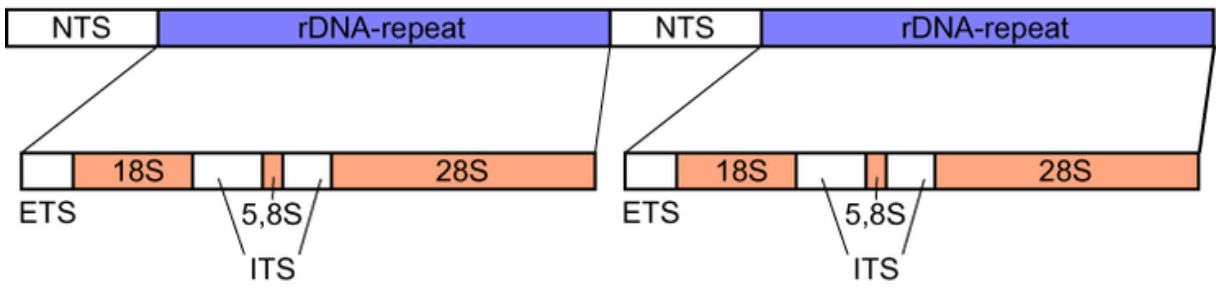


Figure II.2.5 | Structure générique d'un opéron d'ADN ribosomique eucaryote. Répétition en tandem des domaines NTS, ETS, 18S, ITS1, 5,8S, ITS2 et 28S (Figure extraite de http://www.wikiwand.com/fr/ADN_ribosomique)

Ainsi, des études génomiques utilisant le séquençage du gène de l'ARNr 18S ont permis d'obtenir un catalogue des gènes de micro-organismes eucaryotes marins et de découvrir de nouveaux groupes de protistes non cultivables[43,44,187–196]. Les régions hypervariables V9 du gène de l'ARNr 18S ont notamment été utilisées pour constituer une base de données de séquences de référence chez les dinoflagellées[197] à l'aide des données séquencées lors de l'expédition *Tara*

Oceans. Cependant, la nature conservée de ces régions fait que l'on sous estime la diversité d'espèces dans une communauté, comme cela est le cas pour le zooplancton[198]. De plus, la résolution phylogénétique obtenue avec l'ARNr 18S est parfois inadéquate pour distinguer les organismes micro-planctoniques eucaryotes au niveau de l'espèce[53]. Enfin, il existe des biais de séquençage et de clonage avec cette méthode[199]. Il peut y avoir un risque d'amplification de pseudogène qui peut affecter la fiabilité des résultats.

Le séquençage de l'ADN du génome d'un organisme permet de découvrir de nombreux micro-organismes, de mieux comprendre leur organisation spatiale ou encore d'étudier les fonctions qui les régissent. Cependant, les approches génomiques se basent à l'échelle de l'organisme et non des communautés ce qui peut-être restrictif lorsque l'on souhaite étudier l'écosystème marin. Avec les avancées du séquençage haut débit, une autre discipline a émergé et permet d'étudier non pas le génome d'un seul micro-organisme, mais les génomes d'une communauté de micro-organismes directement prélevés dans leur environnement.

II.3. La métagénomique

Moins de 1% des micro-organismes peuvent être cultivés en laboratoire suivant les méthodes nécessaires à un séquençage[200]. Ainsi, avec les méthodes traditionnelles de séquençages, seule une infime partie des génomes des organismes nous est accessible. De plus, l'étude d'un génome issu d'un organisme mis en culture ne permet pas d'avoir l'information de l'impact de la variation de son environnement d'origine. Les récentes avancées réalisées avec le séquençage de cellules uniques commencent à donner accès aux génomes partiels de micro-organismes non cultivables. Cependant, on peut trouver jusqu'à plusieurs centaines de milliers d'espèces différentes dans le même échantillon. Il est donc impossible actuellement de séquencer les cellules de chaque organisme d'une communauté. Il est donc devenu nécessaire de pouvoir étudier les micro-organismes comme des mélanges de génomes prélevés dans leur environnement.

II.3.1. Qu'est-ce que la métagénomique?

Un métagénome correspond à l'ensemble des séquences d'ADN appartenant aux organismes prélevés directement dans leur environnement sans même avoir besoin de passer par des méthodes d'isolation et de culture[9]. Le terme métagénomique a été introduit pour la première fois par Handelsman en 1998[201] et a été rapidement employé par la communauté scientifique. Le préfixe méta, vient du grec « μετά » qui fait référence à « ce qui vient après »,

« au-delà » la génomique[202]. Ainsi, la métagénomique consiste à étudier le contenu génétique d'un métagénome et non plus d'un seul génome.

II.3.2. Différentes familles en métagénomique

Différentes analyses génomiques peuvent être réalisées, on peut donc distinguer différentes familles en métagénomique[203].

La métagénomique quantitative

La métagénomique quantitative consiste à obtenir le catalogue des gènes connus présents dans un métagénome. On peut par la suite estimer l'abondance de ces gènes ou encore des espèces présentes dans les échantillons[204]. Il sera alors possible de comparer les différents échantillons en fonction des espèces et des gènes partagés. Pour qualifier les organismes présents dans un échantillon, les régions hypervariables de l'ARNr sont généralement utilisées car elles ont une signature taxonomique bien définie comme cela a été décrit dans la partie II.2.5. L'utilisation de ces codes barres génétiques obtenus suite au séquençage d'un environnement complexe est ce que l'on appelle le métabarcoding. La métagénomique quantitative permet d'observer par exemple la diversité d'une classe d'organismes dans différents échantillons[197,205]. Monier *et al.* ont réalisé une étude sur la diversité et la biogéographie des Mamiellophyceae en utilisant les données issues du séquençage des barcodes présents dans les échantillons du projet *Tara* Oceans[52]. Avec cette approche il est donc nécessaire d'avoir des séquences de référence connues. L'absence de ces références peut être un obstacle pour estimer la diversité génétique d'un échantillon.

La métagénomique ciblée

Une autre famille utilisant des séquences de références est la métagénomique ciblée. Celle-ci permet d'identifier un organisme dans un métagénome. Il est donc possible de savoir si cet organisme est présent ou non dans un échantillon ou encore de s'intéresser à un gène en particulier. Ainsi, il est possible par cette approche d'étudier les organismes ou les gènes faiblement représentés dans un métagénome[206]. De plus, le fait de ne pas avoir à réaliser de culture en laboratoire permet d'étudier la quasi totalité des organismes. Il est par exemple possible d'observer la présence ou non de certains gènes dans des milieux aux conditions physico-chimiques variées afin d'étudier les adaptations au sein du génome d'organismes difficilement cultivables[207]. Cependant, étant donné qu'il est nécessaire d'utiliser un génome ou un marqueur de référence préalablement séquencé, assemblé et annoté, la métagénomique ciblée ne s'intéressera qu'aux taxons connus. Dans le premier chapitre de cette thèse sera présentée la biogéographie des deux génomes de *Bathycoccus* ainsi que celles des autres

mamiellales obtenue avec la métagénomique ciblée. L'étude de la présence ou non de certains gènes de *Bathycoccus prasinos* dans les échantillons métagénomiques sera également présentée.

La métagénomique fonctionnelle

La métagénomique fonctionnelle permet d'étudier les interactions entre les différents intervenants d'un écosystème en observant les fonctions exprimées dans un métagénome. Pour cela, les séquences métagénomiques sont comparées à des bases de données référençant des séquences annotées fonctionnellement. La base de données Uniprot[208] constituée en 2002 est un entrepôt pour les séquences protéiques qui peuvent être déposées par les équipes scientifiques du monde entier. UniProt propose diverses annotations qui peuvent y être associées, telles que les fonctions, les ontologies, les références bibliographiques liées à la séquence, le découpage de la protéine en domaines ou encore les liens vers d'autres séquences ou des bases de données plus spécialisées (*cross-references*). Il est également possible d'améliorer la prédiction de fonction des protéines en étudiant leur composition en domaines structuraux et fonctionnels. Ainsi, des méthodes comme Pfam[209] et PRIAM[210] ont été développées pour découper les protéines en domaines, trouver comment les identifier et y associer une activité biologique. Enfin, la ressource KEGG[211] rassemble un grand nombre de méthodes présentant des informations sur les génomes et sur le métabolisme. La métagénomique fonctionnelle peut permettre de participer à l'annotation fonctionnelle de gènes en étudiant les gènes différentiellement exprimés dans des environnements contrastés. Elle peut permettre également la découverte de nouveaux antibiotiques ou de nouvelles enzymes sécrétées dans l'environnement[212,213], de protéines impliquées dans la résistance aux antibiotiques[214] ou encore des processus impliqués dans la dégradation de polluants[215].

Aucune des familles de métagénomique présentées précédemment ne permettent de réaliser de la métagénomique *de novo*, c'est-à-dire sur l'ensemble des séquences contenues dans différents métagénomes et donc sur des séquences issues de génomes ou de gènes inconnus. La dernière famille permet de réaliser de la métagénomique *de novo*.

La métagénomique comparative

La métagénomique comparative a pour objectif de comparer les métagénomes entre eux. Cette comparaison peut être réalisée de différentes manières. Une première façon est d'utiliser des séquences de références pour identifier des données connues dans les différents métagénomes et de comparer les échantillons sur la base de ces références. Cependant, la quantité importante de séquences inconnues dans les milieux complexes fait que ces méthodes n'utilisent pas l'ensemble des informations génétiques présentes dans les métagénomes. Ainsi, la

seconde manière de réaliser de la métagénomique comparative ne nécessite pas d'utiliser des séquences connues. Il est aujourd'hui possible de comparer l'ensemble du matériel génétique généré par le séquençage d'un métagénome. Une telle comparaison peut permettre d'obtenir le nombre de séquences métagénomiques similaires entre différents échantillons. Ainsi, on peut connaître le degré de similarité génomique entre ces échantillons et les regrouper sur la base de leur contenu génétique. Pour réaliser cette comparaison, les algorithmes d'alignements locaux de séquence de type blast[216] peuvent être utilisés pour de petits jeux de données. En effet, malgré les efforts réalisés afin d'optimiser le temps d'alignement dans la conception de nouveaux outils comme BLAT[217], LAST[218], USEARCH[219] ou encore KLAST[220], cela reste insuffisant pour aligner l'ensemble des séquences de nombreux et volumineux jeux de données dans des temps raisonnables. Pour cela, des outils capables de comparer *de novo* un grand nombre de métagénomes contenant plusieurs centaines de millions de séquences ont été développés et seront détaillés dans la partie III.5.

II.3.3. La métatranscriptomique

Les technologies de séquençage haut débit ont également permis de faire émerger d'autres sous domaines de la métagénomique. C'est le cas de la métatranscriptomique. Ce domaine consiste à réaliser l'analyse d'une partie des gènes exprimés par une communauté de micro-organismes dans un milieu complexe à un moment donné[221,222].

Les études de métatranscriptomique impliquaient au départ l'utilisation des puces à ADN. Cependant, cette méthode ne donne des informations que sur les séquences connues. La métatranscriptomique consiste à analyser les ARN de l'ensemble des organismes contenus dans un échantillon. Pour cela, le séquençage haut débit permet de séquencer aléatoirement le transcriptome d'une communauté de micro-organismes de la même manière que pour un organisme par l'approche de RNA-Seq[184] présenté en partie II.2.4. Comme pour la métagénomique, la métatranscriptomique peut être divisée en sous familles avec différentes applications.

Exemples d'applications en métatranscriptomique

Le premier aperçu d'un transcriptome environnemental en 2005 a permis de construire des bibliothèques d'ARNm procaryotes prélevées dans deux sites aquatiques et d'avoir un aperçu du profil d'expression génétique de ces environnements[223]. Ensuite, une étude sur des données de métatranscriptomique d'échantillons marins générées par la technologie de pyroséquençage 454 a permis d'obtenir un enrichissement important d'ARNm comparé aux ARNr. Cette approche a montré que les études de métatranscriptomique de communautés

microbiennes sont non seulement possibles, mais que si les analyses métatranscriptomiques sont associées à la métagénomique, cela permet de mieux explorer la structure et la fonction de communautés microbiennes[224]. En 2009, le premier métatranscriptome de plancton eucaryote a été obtenu[221] puis de nombreuses études sur des métatranscriptomes issues de différents environnements comme le sol forestier[225], les sources d'eau chaude[226], l'intestin de l'homme[227] ont été réalisées et continuent encore aujourd'hui. L'analyse de ces gènes exprimés peut permettre d'étudier les fonctions réalisées in situ par les micro-organismes dans différentes conditions environnementales. Par exemple, l'utilisation de la métatranscriptomique a permis de comprendre les différences jour/nuit dans l'expression des gènes dans les eaux de surface du Gyre subropical du Pacific Nord[228]. Cette analyse a fourni des informations sur les processus métaboliques qui dominent au sein de plancton bactérien et a révélé des changements pertinents dans les patterns d'expression de certains processus biogéochimiques.

Les limites de la métatranscriptomique

La métatranscriptomique peut impliquer différentes sources de biais techniques. Le choix du protocole d'extraction de l'ARNm est important pour avoir suffisamment d'ADN de qualité tout en éliminant les contaminants. Celui-ci demande un travail minutieux pour sa mise en place. En effet, il existe de nombreux kits d'extraction et de purification de l'ADN qui doivent être adaptés en fonction de l'origine, du type de micro-organismes étudiés, de la quantité et de la méthode de préparation des échantillons. Par exemple, les protocoles d'extraction d'ARNm procaryotes sont particulièrement difficiles. En effet, ceux-ci sont dépourvus de queues poly(A) et les ARNm ont un temps de demi vie relativement court ce qui rend l'isolement des ARNm plus compliqué[229]. Les évènements de transferts horizontaux et le manque de références de génomes complets dans les bases de données contribuent également à ces biais techniques[230]. De plus, les ARNm sont moins abondants que les ARNr après une extraction totale des ARN ce qui induit des bruits de fond qui compliquent la récupération des ARNm[228].

Ainsi, malgré ces limites techniques, la métatranscriptomique permet d'inventoriser les gènes, et donc les fonctions sans passer par les limites de la métagénomique. Elle peut donc permettre de valider l'annotation de gènes non présents dans les bases de données mais détectés dans les données métagénomiques comme cela sera présenté dans le chapitre I.3. De plus, elle est en mesure de révéler le pattern d'expressions des gènes d'une communauté de micro-organismes dans un environnement complexe. Ce domaine peut-être utilisé pour identifier les conditions environnementales qui ont un impact majeur sur l'expression des gènes.

Les projets de métagénomiques qui ont vu le jour ces dernières années, comme le projet *Tara* Oceans utilisent dans leurs analyses ces différentes familles de métagénomique et de métatranscriptomique.

II.4 Les projets en métagénomique

La métagénomique est un domaine récent. En effet, les premières études utilisant du matériel génétique prélevé dans un milieu naturel datent de 1986[231,232]. Celles-ci se focalisaient sur l'étude de l'ARN ribosomique. Les premières analyses en métagénomique ont consisté à étudier la variété de nouvelles espèces ainsi que des communautés formées par celles-ci. Des projets de métagénomique d'envergure ont été montés et ont nécessité la constitution de différents consortiums regroupant des laboratoires de disciplines variées afin de pouvoir étudier les micro-organismes dans leur environnement.

II.4.1 Le projet metaHit

Le séquençage du microbiome intestinal humain avec le projet metaHit[10] (META*genomics of the* H*uman* I*ntestinal* T*ract*) a été l'un des premiers projets d'envergure de métagénomique sur l'humain. Ce projet est organisé au sein d'un consortium rassemblant 15 instituts de 8 pays différents et a pour but d'explorer le microbiome intestinal humain afin de comprendre les relations entre ces micro-organismes et la santé humaine. Ce projet regroupe des analyses de métagénomiques ciblées, fonctionnelle, quantitative et comparative. Les objectifs étant dans un premier temps d'obtenir un catalogue de gènes et de génomes microbiens présents dans l'intestin humain[10], puis d'observer la présence et la fréquence de certains gènes chez différents individus[233]. Une étude comparative de personnes saines ou atteintes de certaines pathologies à été réalisée afin d'établir des associations entre le microbiome intestinal et la santé de l'individu. Ce projet a notamment permis de révéler un écosystème composé de micro-organismes jusque-là méconnus étant donné que la plupart des espèces hébergées par notre système digestif ne sont pas cultivables in vitro.

II.4.2 Le projet METASOIL

Le consortium Terragenome (*International Soil Metagenome Sequencing Consortium*) s'atelle quant à lui à l'étude d'échantillons du sol. Ce consortium travaille sur six projets de métagénomique dont le projet METASOIL *(Metagenomic discovery and exploitation of the soil microbial community*). Ce projet, débuté en 2009, s'intéresse à l'étude d'échantillons de sols qui n'ont pas été utilisés par l'homme depuis plus de 150 ans. La comparaison des différents métagénomes prélevés dans ces sols permet une meilleure compréhension de l'adaptation des

micro-organismes en fonction de trois paramètres : la profondeur d'échantillonnage sous la terre, les saisons et différentes techniques d'extraction et de purification de l'ADN[234]. La métagénomique quantitative et fonctionnelle a été utilisée ici pour comparer les métagénomes issus de différentes conditions sur la base de leur composition en espèces et des informations fonctionnelles[235].

II.4.3 Global Ocean Sampling

D'autres projets de métagénomique ont pour objectif l'étude de l'environnement marin. L'expédition *Global Ocean Sampling* (GOS) précédemment décrite dans la partie I.4.3 a permis de séquencer et d'étudier 44 métagénomes représentatifs d'une partie des océans du globe. Cette expédition a enrichi nos connaissances sur les micro-organismes ainsi que sur les fonctions ou sur les protéines qu'ils expriment. Un total de 6,12 millions de protéines bactériennes et virales ont été annotées à partir de l'assemblage de 7,7 millions de séquences[134]. Une comparaison des métagénomes a également été réalisée à partir des données brutes issues du séquençage haut débit[14]. Pour cela, la métagénomique comparative *de novo* a été utilisée. Ainsi, la similarité génétique entre deux échantillons a été calculée à partir de la comparaison des lectures métagénomiques et non de l'ARNr 16*S* traditionnellement utilisé. La similarité génétique est considérée ici comme l'estimation de la fraction de séquences partagées par deux échantillons. Pour cela, les séquences ont été alignées afin de construire une base de données de séquences chevauchantes. Le calcul de la similarité entre deux échantillons a été réalisé à partir de cette base de données de chevauchement. On est ainsi capable d'obtenir une distance de similarité entre deux échantillons. Par une méthode de clusterisation hiérarchique il est ensuite possible de regrouper les échantillons génétiquement similaires. Ces comparaisons ont montré que les métagénomes issus de localisations géographiquement proches ou partageant des facteurs environnementaux communs ont tendance à se regrouper dans les mêmes clusters. Cependant, cette méthode ne permet pas d'associer une distance de similarité par échantillon qui indiquerait le nombre de séquences d'un échantillon retrouvées dans le second et inversement. Cela peut être problématique lorsqu'un échantillon possède des lectures redondantes et ne permet donc pas d'avoir l'information sur la complexité de ces échantillons. De plus, il est impossible de récupérer les séquences similaires entre deux métagénomes par cette approche ce qui ne permet pas d'effectuer d'analyse taxonomique ou fonctionnelle en aval. Enfin, la méthode d'alignement utilisée ne permet pas de réaliser les comparaisons sur de nombreux métagénomes contenant plusieurs 100ènes de millions de séquences comme cela peut-être le cas dans le cadre du projet *Tara* Oceans. L'amélioration d'un outil de comparaison de séquences déjà existant pour résoudre ces problèmes sera présentée dans le chapitre II.2.

II.4.4 Le projet *Tara* Oceans

Le projet *Tara* Oceans est un projet de métagénomique à plus large échelle que les précédents projets. En effet, jamais autant d'échantillons marins n'avaient été prélevés auparavant au cours d'une même expédition. De plus, ce projet réunit une étude de métagénomique, de métatrancriptomique, de métabarcoding, et de séquençage de génomes à cellule unique. La quantité d'information génétique des micro-organismes présents dans ces échantillons n'aurait été possible sans les progrès réalisés ces dernières années dans le domaine du séquençage de l'ADN. Au cours de cette expédition, les échantillons marins récoltés ont été envoyés au Genoscope qui était en charge de réaliser l'ensemble du séquençage de ce projet. Différentes stratégies de séquençage ont été mises en place en fonction du type de donnée génomique et des organismes à étudier. Ainsi, la mise en place de protocoles de séquençage métagénomique, métatranscriptomique, métabarcoding et de génome à cellule unique a été réalisée[236]. Le séquençage des échantillons a été effectué pour chacune des fractions de taille de micro-organismes (Figure II.4.5). Les échantillons avec une majorité de virus, de procaryotes, de protistes unicellulaires et enfin de métazoaires ont été séquencés avec des protocoles spécifiques à chaque taille de filtres. Le modèle de séquenceur HiSeq2000 commercialisé en 2010 permet de générer jusqu'à 600 Gb de séquences correspondant à 3 milliards de lectures de taille 2 x 101 pb en 8 jours. Cela représente un débit de 75 Gb/jour. Il y a génération d'environ dix fois plus d'informations par *run* que pour la méthode 454 et donc la couverture de séquençage est plus importante avec un séquençage Illumina. Les fragments générés avec la technologie Illumina sont plus petits mais de meilleure qualité qu'avec un séquenceur 454. C'est pourquoi la technologie Illumina a été privilégiée pour réaliser le séquençage de ces échantillons. Le séquenceur HiSeq2500, plus rapide que son prédécesseur a été commercialisé depuis. L'expédition *Tara* Oceans ayant débuté en 2009, c'est avec le séquenceur Hiseq2000 que les échantillons métagénomiques, métatranscriptomiques et les SAGs ont été séquencés (Figure 12 B). Pour le projet *Tara* Oceans, sans prendre en compte l'expédition *Tara* Polar Circle, un total de 644 échantillons métagénomiques de surface et DCM ont été séquencés. Pour cela, 955 runs métagénomiques ont été réalisés et environ 240 milliards de lectures ont été générées, ce qui représente plus de 24 trillions de base ($24*10^{12}$ pb) séquencées au Genoscope. Cela constitue à ma connaissance le plus gros jeu de donnée métagénomique généré dans le cadre d'un projet d'étude d'un biome.

Concernant le séquençage métatranscriptomique, 441 échantillons ont été pour l'instant séquencés et exploités. Un total de 1441 runs métatranscriptomiques ont été réalisés et environ de 340 milliards de lectures ont été générées ce qui représente plus de 51 trillions de paires de

base ($51*10^{12}$ pb). Un catalogue de 116 849 350 gènes eucaryotes a été constitué et annoté à partir d'une méthode de clusterisation des échantillons métatranscriptomiques.

Le séquençage des métabarcodes 16S et 18S a été réalisé avec les séquenceurs Hiseq2500, GAIIx et Miseq. Des unités taxonomiques opérationnelles (OTU) ont été générées afin d'étudier la composition en micro-organismes des échantillons. 15 222 OTUs viraux ont ainsi été produits à partir d'une méthode de regroupement du contenu protéique[237,238]. Cela a permis d'estimer le nombre total d'organismes viraux dans les échantillons *Tara* Oceans à environ un million. De plus, à l'aide d'un marqueur ribosomique de l'ADN, 7 250 miTAG OTUs procaryotes ont été construits[239]. Enfin pour les eucaryotes, l'utilisation des séquences marqueurs correspondant à la région V9 de l'ADNr 18S a permis d'obtenir 1,3 billions de séquences ($1,3\ 10^9$) représentant 474 303 OTUs au total. Ce nouveau jeu de données eucaryotes vient d'être obtenu avec l'addition d'un précédent jeu généré dans le cadre des premières analyses de *Tara* Oceans[205]. La construction de ces banques de données de gènes sur les communautés microbiennes, virales, et eucaryotes a permis de montrer que la plupart de ces gènes demeure inconnue, c'est-à-dire qu'ils n'ont pas de référence dans les bases de données publiques. Cela s'explique par le fait que l'on travaille sur des micro-organismes souvent difficilement cultivables et souvent phylogénétiquement éloignés des organismes séquencés. De plus, jusqu'alors aucune autre expédition n'avait récolté autant d'échantillons provenant d'endroits, de profondeurs et de types d'organismes si variés et ce, dans les différents océans du globe.

a

Size fractions (µm)	Genomic analysis	Mainly targeted organisms	Sample storage laboratory	Sequencing laboratory
< 0.2 µM	Metagenomics	Viruses	M. Sullivan lab (University of Arizona, AZ, US)	CEA, Genoscope, France
0.2-1.6, 0.1-0.2, 0.45-0.8, 0.2-0.45	Metagenomics	Giruses	N. Grimsley lab (CNRS, Banyuls-sur -Mer, France)	
0.2-1.6, 0.2-3	16S metabarcoding, metagenomics, metatranscriptomics by random priming	Viruses, Giruses, Prokaryotes	S. Gonzales-Acinas lab (ICM-CSIC, Barcelona, Spain)	
0.8-inf, 3-inf, 0.8-5 (0.8-3), 5-20 (3-20), 20-180, 180-2000	18S and 16S metabarcoding, metagenomics, metatranscriptomics on polyA$^+$ RNA	Protists and metazoa	C. De Vargas lab (CNRS/UPMC, Roscoff, France P. Wincker lab (CEA, Genoscope, France)	
Isolated samples for SAGs	*De novo* sequencing	Protists	M. Sieracki lab (Bigelow lab, ME, US)	

b

Samples		Library insert size (pb)	Sequencing instrument	Read length (PE mode)	Mean number of reads per sample (millions of paired reads)
Metagenomic libraries		180	HS2000	101	160
Viral samples	TO	150-900	HS2000	101	50
	TPC				200
SAGs		150-900	HS2000	101	20
Metatranscriptomic libraries		100-600	HS2000	101	160
18S metabarcode libraries	TO	160	GAIIx	151	1.5
	TPC		HS2500	151	3
16S metabarcode libraries		400	Miseq	301	0.1
			HS2500	251	1
TO: *Tara* Oceans expedition					
TPC: *Tara* Oceans Polar Circle expedition					

Figure II.4.5 | Tableaux récapitulatifs des données séquencées lors du projet *Tara* Oceans. **a**, Fractions de tailles et types de données séquencées. **b**, Librairies générées et séquenceurs utilisés lors du séquençage des échantillons *Tara* Oceans. (Figure extraite de Alberti *et al.*[236])

D'autres projets de métagénomique ont été montés en parallèle. Ces projets s'intéressent à divers milieux et écosystèmes. Comme par exemple de la métagénomique des stations de métro à New York[240], d'os de mammouth[241] ou encore des sédiments à environ 2,5 km sous le plancher océanique[242].

II.4.6. Evolution au cours du temps du nombre de projets de métagénomique

Il est possible de suivre l'évolution du nombre de projets de génomique et de métagénomique en observant les données de séquençage déposées dans les bases de données publiques (Figure II.4.6). Ainsi, avec la venue du séquençage haut débit, les projets de séquençage se sont multipliés depuis le séquençage du premier génome bactérien en 1995. Les projets de séquençage de génomes bactériens ont eu une évolution presque exponentielle au cours du temps et sont aujourd'hui six fois plus nombreux que dans les autres domaines. Un total de 979 études de métagénomique ont été référencées en 2016 contre 56 en 2006. De plus, on compte presque autant de projets de séquençage métagénomique que de projets de séquençage de génomes eucaryotes en 2016. Si cette évolution suit l'allure actuelle, les projets de séquençage métagénomique dépasseront ceux de génomes eucaryotes l'an prochain. Cela illustre bien l'importance d'étudier les génomes des micro-organismes prélevés dans leur environnement afin d'améliorer nos connaissances sur les différents écosystèmes existants.

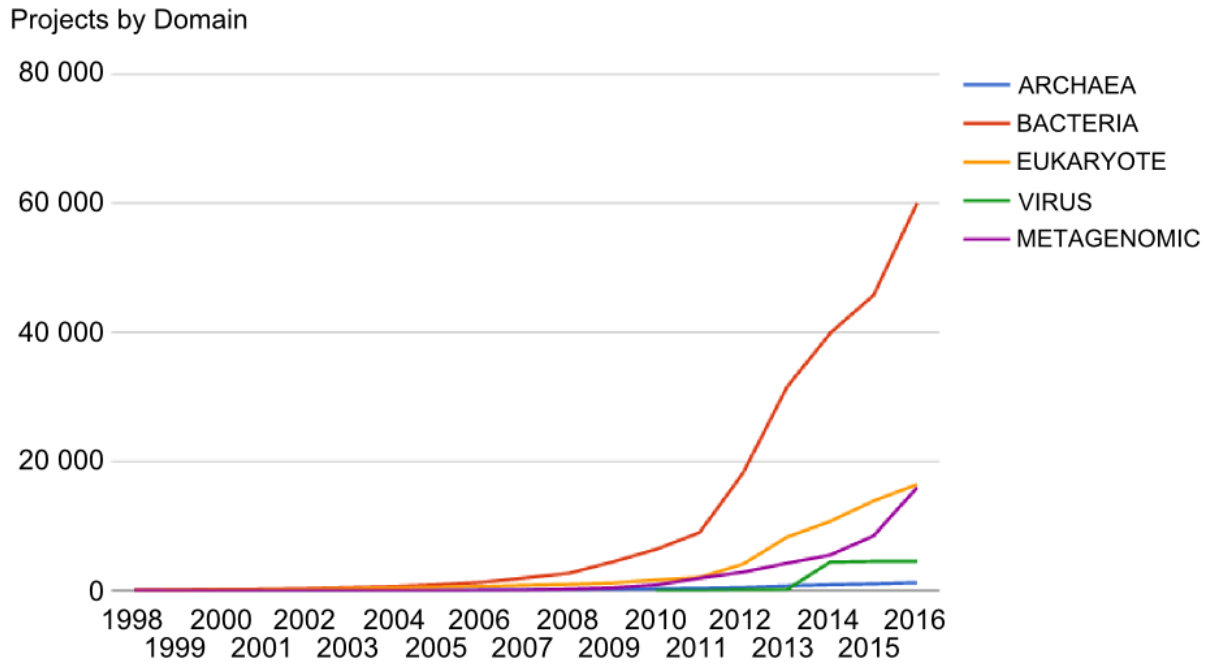


Figure II.4.6 | Comparaison de l'évolution du nombre de projets de séquençage dans différents domaines. La courbe bleue représente les projets de séquençage d'archée, en rouge les bactéries, en orange les eucaryotes, en vert les virus et en violet la métagénomique (Figure extraite de *Genome Online Database*[167,168]).

II.4.7. La multidisciplinarité des projets de métagénomique

Les projets de métagénomique ont comme point commun la possibilité d'avoir accès aux gènes, aux organismes et à l'information sur les communautés d'organismes et sur l'environnement du prélèvement. Ces trois éléments peuvent être analysés par paires ce qui permet l'étude de trois disciplines : la métagénomique, l'écologie et la génomique (Figure II.4.7). L'étude du vivant à l'échelle du génome consiste à faire de la génomique, étape indispensable à la compréhension des phénomènes biologiques d'un organisme. L'étude des relations des êtres vivants avec leur habitat et leur environnement, ainsi qu'avec les autres êtres vivants désigne l'écologie. Ce terme fut inventé en 1866 par Ernst Haeckel qu'il désigne comme étant « la science des relations des organismes avec le monde environnant, c'est-à-dire, dans un sens large, la science des conditions d'existence »[243]. L'étude de ce domaine consiste à analyser deux grands ensembles : La biocénose et le biotope. Le terme biocénose vient du grec βίος, bios (« vie ») et κοινός, koinós (« commun »). Il a été introduit en 1877 par le zoologiste allemand Karl Möbius qui le conçoit donc comme étant une communauté de vie[244]. Ce terme permet de décrire l'organisation, non pas d'un seul organisme vivant, mais de l'ensemble des organismes vivants qui cohabitent dans un espace déterminé. Le biotope correspond au type de lieu de vie défini par les caractéristiques physico-chimiques. Enfin, l'étude de l'ensemble des génomes des organismes prélevés dans leur environnement correspond à la métagénomique. Comme en écologie, il est possible avec la métagénomique de décrire l'organisation des communautés d'organismes vivants basée sur leur contenu génétique dans un espace donné. Dans cette thèse j'introduirai un nouveau terme, *genocenose* en référence au terme *biocenose* pour illustrer cette notion qui sera décrite dans le chapitre III. Même si la majorité des projets de métagénomique ont accès à ces trois types de données, peu de projets essayent d'étudier conjointement les trois disciples qui en résultent. Pourtant, une telle étude permettrait une analyse plus exhaustive d'un écosystème.

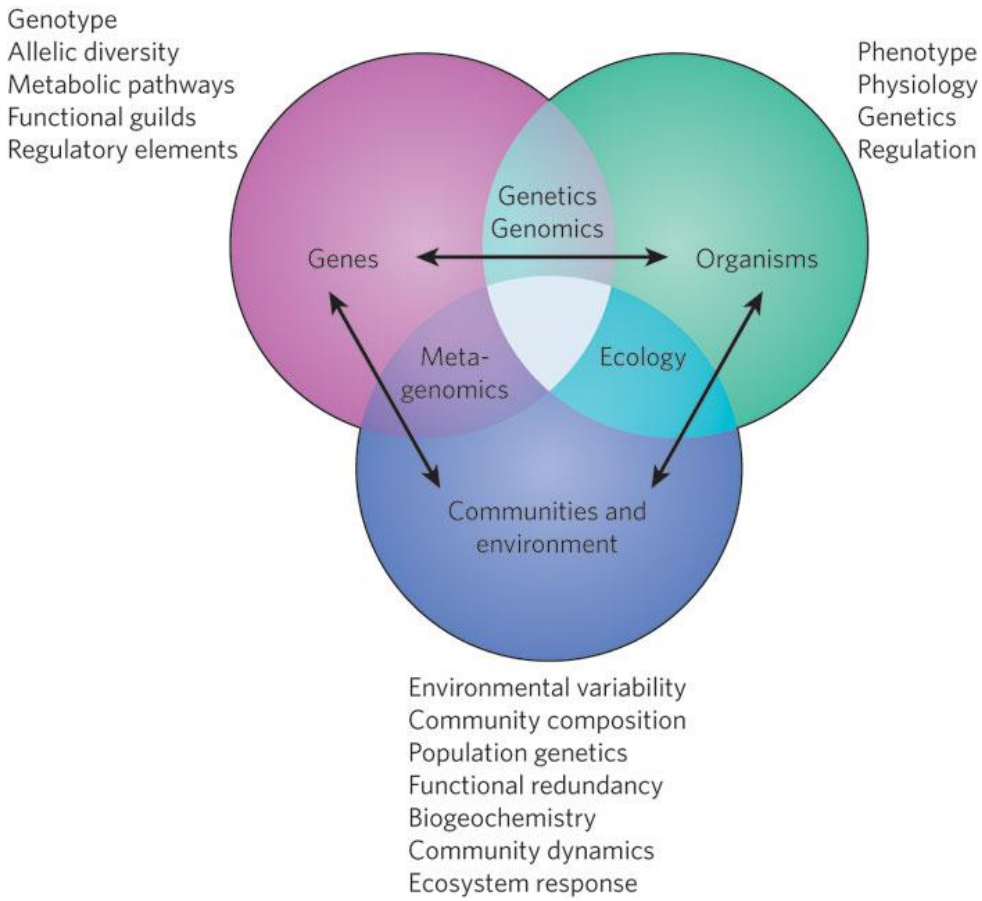


Figure II.4.7 | Intersection de la génomique, de la métagénomique et de l'écologie. Chaque discipline se croise pour étudier les gènes, les organismes ainsi que les communautés et leur environnement. (Figure extraite de DeLong 2009[245])

Tara Oceans est un des premiers projets qui permet d'avoir accès à ces différents éléments représentatifs de l'écosystème océanique global. L'étude conjointe des domaines qui en résulte demande la conception de nouvelles approches ainsi que l'élaboration de nouvelles méthodes permettant de relever les défis que pose une telle analyse multidisciplinaire. De plus, le volume de données génomiques à analyser est actuellement un des plus important obtenu dans le cadre d'un projet. Cette quantité de données générées par les séquenceurs haut débit pose des défis en bio-informatique. En effet, la capacité à séquencer de l'ADN croit plus vite que la puissance informatique. Il est donc devenu nécessaire de concevoir de nouveaux outils et des algorithmes toujours plus performants pour pouvoir traiter l'ensemble des informations produites par les séquenceurs haut débit dans des temps raisonnables.

III. Analyse comparative des données métagénomiques du projet *Tara* Oceans

Avec l'avènement des séquenceurs haut débit, il a été possible de séquencer l'ensemble des échantillons prélevés lors du projet *Tara* Oceans. La quantité de données métagénomiques ainsi générée est actuellement la plus importante pour un projet d'étude d'un biome. Avec l'augmentation ces dernières années du nombre de projets en métagénomique, des outils ont été spécifiquement développés pour analyser ce type de données. Bien que ces outils prennent en compte le volume massif de données générées par ces projets, il n'en reste pas moins qu'ils ne sont parfois pas adaptés pour mener à bien l'analyse de certains d'entre eux, tels que le projet *Tara* Oceans. De plus, la génomique du plancton océanique étant majoritairement inconnue des bases de données publiques, des stratégies doivent être mises en place pour pouvoir explorer la diversité globale des micro-organismes planctoniques. Enfin, la multidisciplinarité d'un tel projet demande la conception de nouvelles approches ainsi que l'élaboration de nouvelles méthodes pour pouvoir étudier l'organisation de la génomique des micro-organismes planctoniques dans leur environnement à l'échelle de la planète. Je présenterai dans cette partie les outils qui ont été développés ainsi que les ressources de calculs nécessaires pour effectuer une étude comparative du matériel génétique généré par le séquençage d'un métagénome.

Il est utile de comparer la génomique des micro-organismes planctoniques présents dans différentes localisations pour mieux comprendre comment les communautés planctoniques sont organisées et affectées par les facteurs environnementaux. Avec la multiplication du nombre de projets en métagénomique ces dernières années, des méthodes et des outils ont été développés pour analyser et comparer le contenu génétique de différents métagénomes. Une telle analyse entre dans le domaine de la métagénomique comparative présentée en partie II.3.2.

III.1 Comparaison du contenu génomique d'échantillons par homologie aux bases de données de références

L'utilisation de séquences de références pour assigner taxonomiquement ou fonctionnellement des séquences métagénomiques peut être réalisée pour comparer le contenu génomique de différents échantillons. Il est possible d'utiliser comme références des marqueurs génétiques tels que l'ADN ribosomique ou encore des banques de gènes de références annotés

taxonomiquement ou fonctionnellement. Différents outils ont été développés dans cette optique. Je présenterai ici les outils les plus utilisés par la communauté scientifique.

III.1.1 MEGAN

Le logiciel MEGAN[246] (*MEtaGenome ANalyzer*) est couramment utilisé en bioinformatique pour explorer le contenu taxonomique d'un échantillon métagénomique. Cet outil utilise le logiciel d'alignement Blast[216] pour comparer les séquences métagénomiques contre des séquences issues d'une base de référence. Ainsi, si une séquence s'aligne sur une référence, son assignation taxonomique est obtenue en utilisant la taxonomie du NCBI (*National Center for Biotechnology Information*, États-Unis). Lorsqu'une séquence du métagénome s'aligne sur plusieurs séquences avec des assignations sur différents taxons, une étape dite de *Lowest Common Ancestor* (LCA) est réalisée. Le LCA est une entité théorique qui correspond au dernier nœud de l'arbre phylogénétique à partir duquel divergent les branches de chacune des lignées en question (Figure III.1). Cette étape permet donc d'assigner la séquence au dernier ancêtre commun partagé par les autres séquences. Pour le métagénome étudié, MEGAN produit un arbre phylogénétique avec l'information des taxons présents et leurs abondances respectives. Il est donc possible par la suite, de comparer l'abondance de différents taxons dans les échantillons métagénomiques. Cependant, l'utilisation d'une méthode d'alignement de type blast ne permet pas d'obtenir des résultats en des temps raisonnables pour de grands jeux de données métagénomiques. Plusieurs versions de ce logiciel ont été développées afin de s'adapter à la quantité de séquences à analyser et afin d'améliorer ainsi le temps d'alignement. La version 6 de MEGAN est actuellement disponible[247].

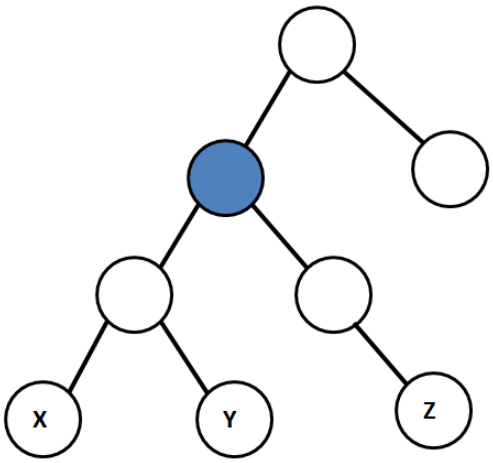


Figure III.1 | Schéma explicatif de la recherche du *Lowest Common Ancestor.* Dans cet arbre, le plus petit ancêtre commun de X, Y et Z est le nœud bleu.

III.1.2 MG-RAST

MG-RAST[248] (*Metagenomic Rapid Annotations using Subsystems Technology*) est un serveur utilisé pour les analyses phylogénétiques et fonctionnelles de métagénome. Ce serveur aligne les séquences sur différentes bases de données protéiques avec blastx. En parallèle, un alignement blastn sur des bases d'ARNr est réalisé pour leur détection. Avec les informations sur les ARNr et les alignements protéiques, il est possible de reconstruire l'arbre phylogénétique correspondant au contenu génétique de l'échantillon. Il est également possible d'obtenir une classification fonctionnelle des séquences étudiées. Cet outil peut donc être utilisé en métagénomique fonctionnelle. Cependant, le fait d'avoir recours à un serveur en ligne requiert de transmettre l'ensemble des séquences métagénomiques pour pouvoir exécuter les requêtes d'alignement. Cette étape est difficilement réalisable lorsque les échantillons à étudier sont trop volumineux. De plus, comme avec MEGAN, l'utilisation d'outils d'alignement de type blast pose des problèmes de temps de calculs pour les échantillons métagénomiques volumineux en séquences.

III.1.3 IMG/M

IMG/M[249] (*Integrated Microbial Genomes with Microbiome samples*) est un logiciel hébergé au *Joint Genome Institure* (JGI). Ce serveur est une ressource publique qui permet d'analyser et d'annoter les métagénomes via une approche comparative. Il intègre donc des jeux de données métagénomiques publiques pour pouvoir réaliser les analyses. En 2016, 7091 métagénomes et métatranscriptomes sont intégrés au serveur. De nombreux outils ont été implémentés pour explorer les métagénomes afin de rechercher un organisme, un gène ou encore une fonction donnée. L'information phylogénétique des micro-organismes présents dans un métagénome peut être obtenue pour réaliser des analyses comparatives entre différents métagénomes. Il est alors possible de comparer les profils d'abondances des métagénomes ou encore la distribution phylogénétique de gènes dans différents métagénomes. Les métagénomes étant présents dans les bases de données du JGI, certains calculs coûteux en temps sont déjà réalisés. Cependant, certains métagénomes nouvellement séquencés ne pourront être analysés.

III.1.4 Outils optimisés pour les courtes séquences

D'autres outils plus récents utilisent également l'alignement des lectures métagénomiques sur des bases de références afin de réaliser une assignation taxonomique des échantillons métagénomiques. Ces outils utilisent des algorithmes d'alignement optimisés pour les courtes séquences et permettent donc d'obtenir des résultats dans des temps plus courts que ceux d'un alignement de type blast. Par exemple, Genometa[250] est une interface graphique qui

utilise l'outil d'alignement Bowtie[251]. Cet outil d'alignement a été spécialement conçu pour aligner les courtes séquences issues de séquençage haut débit contre de grand génome. BWA[252] est un autre outil d'alignement également optimisé pour les alignements de courtes séquences. Ce dernier est notamment utilisé par l'outil GOTTCHA[253]. Cet outil analyse les régions génomiques uniques à chaque référence afin d'assigner les lectures métagénomiques en limitant le nombre de faux positifs. Taxator-tk[254] quant à lui utilise l'aligneur LAST[255] qui permet d'optimiser les paramètres d'alignements selon les données utilisées. D'autres outils utilisent des algorithmes d'alignement optimisés pour les courtes séquences métagénomiques. Cependant, dans le cadre des projets actuels de séquençage métagénomique, le séquençage d'un échantillon produit des centaines de millions de séquences et plusieurs centaines d'échantillons sont séquencés. L'alignement de la totalité des séquences sur des bases de données volumineuses implique des temps de calculs souvent trop importants. D'autres outils ont été développés et utilisent des alignements sur la base de *k-mers*, des petits fragments de tailles k homogènes, afin d'accélérer les temps de calculs.

III.2 Comparaison du contenu génomique d'échantillons via les biais compositionnels de l'ADN

L'utilisation des biais compositionnels de l'ADN permet d'exploiter les caractéristiques intrinsèques des reads. Les outils présentés ci-dessous s'intéressent à la fréquence des oligonucléotides en se servant des *k-mers*.

III.2.1 Kraken

L'outil Kraken utilise les *k-mers* pour assigner une lecture à un taxon (Figure III.2.1). Les lectures sont d'abord fractionnées en *k-mers* chevauchants. Puis chacun de ces *k-mers* est assigné au LCA des génomes contenant ce *k-mer*. Pour cela, une base de données contenant cette information est préalablement calculée. Pour chacune des lectures, un arbre de classification est construit en retenant seulement les taxons associés aux *k-mers* de cette lecture. Cet arbre est utilisé pour avoir la classification taxonomique de la lecture. Avec cet outil, 4,1 millions de lectures métagénomiques seraient assignées par minute.

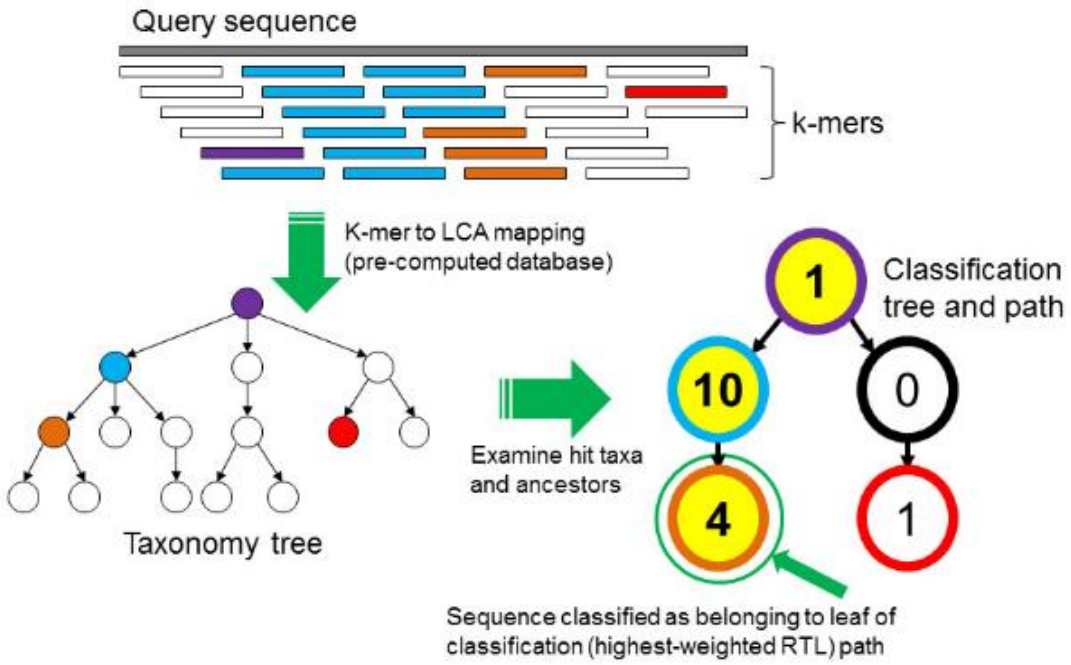


Figure III.2.1 | Algorithme de classification des séquences de l'outil Kraken. (Figure extraite de Wood et Salzberg 2014[256])

III.2.2 CLARK

CLARK[257] (*CLAssifier based on reduced K-mers*), comme Kraken, est un outil permettant d'assigner taxonomiquement les lectures de grands jeux de données métagénomiques. L'indexation des *k-mers* appartenant aux séquences de la base de données est réalisée en éliminant les *k-mers* communs. On parle de *k-mers* discriminants car ceux-ci représentent des régions génomiques caractérisant uniquement chacune des séquences cibles. Une lecture est assignée à une séquence cible lorsqu'elle partage le plus grand nombre de *k-mers*. Un score de confiance est indiqué pour chaque assignement. Cet outil permettrait d'assigner 32 millions de lectures métagénomiques par minute. Cependant, CLARK a été conçu pour l'analyse de métagénomes bactériens et cela ne permet pas d'étudier d'autres métagénomes.

LMAT[258] ou encore OneCodex[259] sont deux autres outils utilisant les *k-mers* pour l'alignement sur les bases de références pour permettre de réaliser des analyses comparatives sur la distribution taxonomique d'échantillons métagénomiques. Ces deux outils semblent être moins performants que CLARK et Kraken en terme de temps d'exécution et sur la sensibilité de l'assignation taxonomique[260]. Selon les données utilisées et selon l'analyse à effectuer, chaque outil a ses avantages et ses inconvénients. C'est à l'utilisateur de choisir l'outil le plus approprié pour ces analyses. Ainsi, l'utilisation des *k-mers* permet d'accélérer le processus d'assignation taxonomique des lectures métagénomiques contre des banques de références. Cependant les références de ces banques sont souvent redondantes et leur assignation n'est pas forcément résolutive. De plus une erreur d'assignation peut advenir s'il manque un clade dans la banque de références. Pour éviter ces inconvénients, l'utilisation de marqueurs génétiques peut être une bonne alternative.

III.3 Comparaison du contenu génomique d'échantillons avec les marqueurs génétiques

Le barcoding ADN décrit en partie II.2.5 peut servir pour définir des clusters génétiques qui seront utilisés comme des unités taxonomiques moléculaires opérationnelles (OTU). L'utilisation de ces unités permet d'estimer la richesse en espèces de communautés de micro-organismes parfois méconnues. Certains outils utilisent ces marqueurs génétiques pour décrire la diversité et la structure de communautés génétiques.

III.3.1 MetaPhlAn

MetaPhlAn[261] (Metagenomic Phylogenetic Analysis) utilise un set de gènes marqueurs identifiés à partir de 3 000 génomes de références bactériens et d'archées. Ce set de gènes est composé d'environ 1 million de séquences pour assigner taxonomiquement les échantillons métagénomiques bactériens et d'archées. Les métagénomes eucaryotes ne peuvent donc pas être pris en compte dans l'analyse. Les gènes marqueurs sont fortement conservés au sein d'un clade et ne doivent pas être similaires à des séquences en dehors de ce clade. La résolution des assignations peut atteindre le niveau de l'espèce. Cet outil aligne les lectures métagénomiques sur le set de gènes marqueurs avec Bowtie2[262]. Il permet ainsi d'avoir l'estimation de l'abondance relative des micro-organismes dans un métagénome par comptage des lectures qui ont un alignement sur ces gènes marqueurs.

III.3.2 MetaPhyler

Metaphyler[263] comme MetalAn utilise un set de gènes marqueurs construit à partir de séquences de l'ARNr 16S. Il utilise l'aligneur blastx pour assigner les lectures métagénomiques aux gènes marqueurs. L'assignation taxonomique peut atteindre le niveau du genre.

III.3.3 mOTU

mOTU[264] est un outil qui permet de quantifier les micro-organismes connus mais également ceux non répertoriés dans les bases de données publiques. Pour cela, la base de données est construite à partir de deux sets de gènes marqueurs. Un set de 10 gènes marqueurs extraits à partir de 3 496 génomes procaryotes de références et un autre set de gènes marqueurs métagénomiques obtenu depuis les contigs issus de 263 échantillons métagénomiques des projets MetaHIT et HMP. Une étape de regroupement (*clustering*) a été réalisée en prenant ces deux sets de gènes pour former une banque dite mOTUs pour unité métagénomique taxonomique opérationnelle (*metagenomic operational taxonomic units*). Les lectures métagénomiques sont donc alignées sur cette banque de mOTUs qui n'ont pas forcément d'assignation taxonomique connues mais représente un micro-organisme dont la résolution peut représenter une espèce.

Ocean Microbial Reference Gene Catalog[239] (OM-RGC) est une banque de données de gènes marqueurs de micro-organismes planctoniques récemment constituée à partir d'une méthode similaire à mOTU. Cette base de données comprend plus de 40 millions de séquences non redondantes appartenant à des micro-organismes allant des virus, aux procaryotes et aux pico-eucaryotes. Pour cela des contigs issus de l'assemblage d'échantillons métagénomiques du projet *Tara* Oceans ont été utilisés. Ces séquences ont été combinées avec d'autres séquences

métagénomiques issues de différents projets dont l'expédition GOS ainsi que des génomes de références.

Ces différents outils utilisant des marqueurs génétiques sont souvent limités à l'étude de métagénomes procaryotes. Si l'objectif est de comparer la diversité génétique entre différents échantillons métagénomiques, ces méthodes ne sont pas appropriées lorsque l'on étudie un milieu complexe tel que l'océan. En effet, la majorité des micro-organismes planctoniques n'est pas répertoriée dans les bases de données[14,134] et de ce fait, une grande partie de l'information génétique ne sera pas prise en compte dans les analyses. Une alternative serait d'utiliser l'information des métabarcodes.

III.4 Comparaison du contenu génomique d'échantillons à partir des barcodes environnementaux

III.4.1 Le séquençage de barcodes environnementaux

Les méthodes de séquençage haut débit permettent d'obtenir des centaines de millions de métabarcodes à partir d'un environnement complexe en des temps records. Ce *barcoding* environnemental permet de mesurer la diversité globale des micro-organismes planctoniques. En effet, avec cette méthode il est actuellement possible d'avoir accès à l'information de pratiquement la totalité des organismes présents dans un échantillon issus d'un environnement complexe[205]. Pour cela, un séquençage de barcodes environnementaux est réalisé. Celui-ci consiste dans un premier temps à utiliser des amorces universelles pour amplifier massivement par PCR les régions génomiques conservées chez un groupe d'organismes. Cela peut, par exemple, être une région du gène de l'ARN 16S pour les bactéries ou encore de l'ARN 18S pour les eucaryotes. Le produit de l'amplification de ces barcodes est ensuite séquencé pour générer les courtes séquences d'ADN spécifiques aux différents micro-organismes présents dans l'échantillon. Couplée à l'utilisation de librairies de références de codes-barres à ADN, cette méthode permet d'obtenir l'identité d'une partie des espèces présentes dans les échantillons. En effet, selon l'écosystème étudié, une plus ou moins grande fraction des micro-organismes n'est pas référencée dans les bases de données publiques. Ainsi, certains métabarcodes n'auront pas d'assignation taxonomique.

III.4.2 Les métabarcodes pour étudier la diversité du plancton eucaryote

Dans le cadre du projet *Tara* Oceans, un premier jeu d'environ 766 millions de métabarcodes obtenus à partir de la région V9 de la sous unité de l'ARNr eucaryote a été généré

depuis 47 stations et de 334 échantillons[205]. Ces échantillons comprennent les fractions de tailles d'organismes eucaryotes allant des protistes unicellulaires aux plus gros mésoplanctons (0.8 to 5 µm, 5 to 20 µm, 20 to 180 µm et 180 to 2000 µm). Les analyses ont montré que les quantités de métabarcodes obtenues saturent la biodiversité eucaryote aussi bien localement que globalement à environ 150 000 OTUs (Figure III.4.2). Près d'un tiers de la diversité ribosomale n'a pu être assigné taxonomiquement. De plus, plus de 85% de la diversité planctonique eucaryote correspond à des protistes marins d'où l'importance d'étudier cette catégorie de micro-organismes pour la plupart méconnus[265] alors que celle-ci est un acteur majeur dans la production primaire océanique[49,266]. Cette banque de métabarcodes vient d'être complétée avec l'addition des métabarcodes obtenus en utilisant la totalité des échantillons du projet *Tara* Oceans aux deux profondeurs, surface et DCM. Au total, 474 303 OTUs sont disponibles pour étudier la diversité planctonique eucaryote dans les océans de surface.

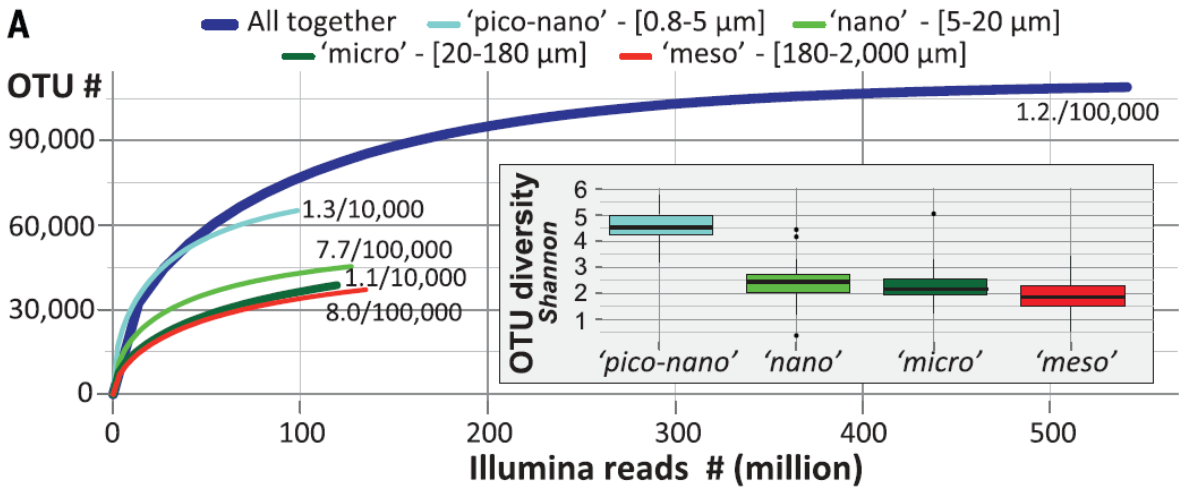


Figure III.4.2 | Diversité ribosomique du plancton eucaryote de la zone photique. Les courbes de raréfaction et de diversité des OTUs issues de la région V9 de l'ARNr pour chaque fraction de taille planctonique eucaryote sont représentées. La saturation est indiquée ici par une pente qui cesse de s'accentuer à la fin de chaque courbe de raréfaction. 1,2/100,000 veut dire qu'au niveau du quasi plateau il y a 1,2 nouveaux métabarcodes obtenus tous les 100 000 reads d'ARNr séquencé (Figure extraite de de Vargas 2015[205]).

Avec le séquençage haut débit, l'utilisation des métabarcodes est bien plus rapide que l'utilisation des autres méthodes décrites précédemment. Les métabarcodes permettent d'envisager une étude de la biodiversité d'un écosystème[267,268]. Il est ainsi possible de comparer la diversité taxonomique de différents échantillons prélevés dans des environnements physico-chimiques différents. Cependant, comme pour les marqueurs génétiques, l'utilisation de la PCR peut induire des biais d'amplification[199]. De plus, dans le cadre d'un écosystème peu connu comme l'écosystème planctonique, beaucoup de métabarcodes n'auront pas d'assignation

taxonomique et ne pourront donc pas être pris en compte dans la comparaison du contenu génomique de deux échantillons. Pour cela, il est possible de calculer à partir de l'information du métabarcoding la distance génomique séparant différents échantillons pour comparer la diversité beta de ces échantillons.

III.4.3 Calcul de distances de similarité avec les métabarcodes

La diversité beta

La comparaison de la diversité des espèces entre les écosystèmes le long de gradients environnementaux est ce qu'on appelle la diversité beta. Ce terme a été introduit par Whittaker en 1960[269]. La diversité beta mesure donc à quel point les systèmes locaux sont différents. Cette définition est assez vague et fait toujours l'objet de débats[270]. La diversité beta peut être mesurée de différentes façons[271]. Dans le cadre d'une étude sur la variation de la diversité en micro-organismes dans différents échantillons on quantifie les différences de communautés d'un point de vue spatial ou temporel. On peut également étudier le roulement, ou *turnover,* d'une communauté. On quantifie alors le changement des communautés le long d'un gradient spatial, temporal ou environnemental[272]. Avec ces deux approches, on s'intéresse à la dissimilarité en espèces entre différents échantillons.

Notion de dissimilarité

Une similarité ou dissimilarité est toute application à valeurs numériques qui permet de mesurer le lien entre les individus d'un même ensemble ou entre les variables. Pour une similarité, le lien est d'autant plus fort que sa valeur est grande. Avec cette notion, on peut estimer la distance de dissimilarité entre différents échantillons.

Distance de Whittaker

Whittaker a introduit une formule pour mesurer la diversité beta entre deux communautés[273] :

$$\beta = \frac{S}{\bar{\alpha}} - 1$$

S correspond au nombre total d'espèces communes aux deux communautés et $\bar{\alpha}$ la moyenne du nombre d'espèces trouvées au sein des communautés. D'autres distances écologiques sont utilisées pour estimer la dissimilarité en espèces entre deux échantillons.

Distance de Sørensen

Une distance couramment utilisée est la distance de dissimilarité de Sørensen initialement appliquée sur l'écosystème végétal[274] :

$$\beta = 1 - \frac{2c}{S1 + S2}$$

S1 et S2 correspondent au nombre total d'espèces présentes dans le premier et le second échantillon, c au nombre d'espèces communes aux deux échantillons. Ainsi, cette distance varie entre 0 et 1. Lorsqu'elle est égale à 0, les deux échantillons partagent les mêmes espèces. Lorsqu'elle est égale à 1, la distance est élevée et les deux échantillons ne partagent pas d'espèces en commun. Cette distance se calcule à partir de la présence/absence des espèces. Il est possible de calculer des distances écologiques en regardant l'abondance des espèces.

Distance de Bray-Curtis

Une distance très similaire à celle de Sørensen est la distance de Bray-Curtis. Celle-ci se base sur l'abondance brute des espèces présentes dans l'échantillon. Il faut donc considérer des échantillons ayant la même taille. Cette distance comme son nom l'indique a été établie par Bray et Curtis pour l'étude de l'écosystème forestier[275] :

$$BC_{jk} = 1 - \frac{2\sum_{i=1}^{p} \min(N_{ij}, N_{ik})}{\sum_{i=1}^{p}(N_{ij} + N_{ik})}$$

N_{ij} représente l'abondance d'une espèce i dans l'échantillon j et N_{ik} l'abondance de la même espèce i dans l'échantillon k. La nomenclature min() correspond au minimum obtenu pour deux comptes sur les mêmes échantillons. Cette distance est généralement utilisée pour réaliser des regroupements d'échantillons basés sur leurs similarités en espèces pour observer la biogéographie de communautés d'organismes[276–280]. Il est possible de connaitre le nombre de métabarcodes communs entre les échantillons. La distance de Bray-Curtis calculée à partir de ce contenu en métabarcodes d'échantillons planctoniques a permis d'établir la biogéographie des communautés de planctons bactériens dans l'océan Atlantique[281]. Ainsi, cette distance reflète la composition en espèces de deux échantillons. Plus les échantillons partagent de métabarcodes et plus les espèces qui les composent sont similaires. Cependant, en utilisant les métabarcodes pour le calcul des distances entre les échantillons, la fraction virale ne sera pas prise en compte car il n'y a pas de marqueur universel connu pour les virus. De plus, l'assignation taxonomique obtenue avec ces marqueurs n'est pas forcément résolutive pour assigner un micro-organisme au bon niveau phylogénétique[53]. En effet, deux espèces peuvent partager la même séquence

nucléotidique correspondant au marqueur utilisé. C'est le cas pour les deux espèces d'haptophytes *Emiliania huxleyi* et *Gephyrocapsa oceanica* qui possèdent une séquence identique du gène de l'ARNr 18S mais des différences morphologiques[282]. Un autre exemple sera présenté dans le chapitre I.1 avec le protiste *Bathycoccus prasinos* composé de deux espèces cryptiques ayant une séquence identique sur l'ARNr 18S.

Ainsi, l'utilisation de cette méthode ne permettra pas de différencier deux espèces possédant le même métabarcode environnemental. Il est fort probable que de nombreuses autres espèces micro-planctoniques partagent un même marqueur génétique. L'analyse des distances à partir des métabarcodes entre différents échantillons ne prendra pas en compte ce biais résolutif. L'étude de la variation de la diversité en micro-organismes ou encore du *turnover* de communautés sera impactée par ce biais en occultant la variabilité génétique d'organismes phylogénétiquements proches. L'utilisation des données métagénomiques s'abstrait de ce biais résolutif et se trouve être une méthode adéquate pour comparer la diversité génétique d'échantillons planctoniques.

III.5 Comparaison du contenu génomique d'échantillons à partir des séquences *de novo*

Sans passer par une étape de culture en laboratoire, la métagénomique permet d'apprécier la diversité génétique des échantillons marins. De plus, le fait de travailler sur les séquences génomiques représentatives des micro-organismes présents dans un échantillon a l'avantage de s'abstraire des biais résolutifs du métabarcoding ainsi que des biais d'amplification. Enfin, étudier les génomes de micro-organismes prélevés directement dans leur environnement permet de pouvoir analyser comment ce dernier à un impact sur l'organisation des communautés planctoniques. Il a été présenté dans la partie II.3.2 que la métagénomique comparative a pour objectif de comparer les métagénomes entre eux et ce, de deux façons. La première méthode utilise les bases de données de références et ne permet donc pas d'explorer la totalité de l'information génétique d'un échantillon puisqu'une partie importante de micro-organismes planctoniques est inconnue. La seconde méthode consiste à comparer l'ensemble du matériel génétique généré par le séquençage de métagénomes. Pour cela, l'utilisation d'algorithmes d'alignement de type blast n'est pas envisageable lorsque l'on souhaite comparer des centaines d'échantillons métagénomiques contenant plusieurs millions de lectures. Des outils capables de comparer *de novo* des métagénomes ont été développés.

III.5.1 crAss

crAss[283] est un des premier logiciel développé pour comparer des métagénomes *de novo*. Il est nécessaire avec celui-ci de passer par une étape d'assemblage de l'ensemble des séquences métagénomiques issues des différents échantillons. Les contigs générés proviennent de différents échantillons. Il est possible de retrouver le nombre de séquences de chaque métagénome qui a construit le contig. Cela permet de déterminer si des contigs sont partagés par différents échantillons. Des distances peuvent être calculées avec ces informations permettant de comparer deux à deux les métagénomes et de construire l'arbre phylogénétique des métagénomes analysés. Ainsi, cette méthode n'utilise pas les banques de références pour éviter d'analyser uniquement les organismes répertoriés dans les bases de données. Cependant, l'assemblage d'un métagénome issu d'un milieu complexe tel que l'océan n'est pas forcément informatif ni résolutif. En effet, les assembleurs utilisés en génomique sont conçus pour l'assemblage de génomes uniques avec une couverture uniforme. En métagénomique, on a un mélange de séquences pouvant provenir de génomes d'espèces proches qui partagent une similarité sur une partie de leur ADN. Les logiciels d'assemblage peuvent fusionner deux séquences provenant d'espèces différentes et former des séquences dites chimériques[161]. De plus, selon le logiciel d'assemblage utilisé, les résultats ne sont pas homogènes et seulement une partie des séquences métagénomiques est représentée dans les contigs. On ne regarde donc plus l'ensemble de l'information génétique d'un échantillon. Enfin, un assemblage est fortement coûteux en temps de calculs et il est donc difficile d'assembler l'ensemble des échantillons d'un projet métagénomique tel que *Tara* Oceans.

III.5.2 TriageTools

TriageTools[284] est un outil qui compare directement les séquences métagénomiques en comptant le nombre de *k-mers* qu'elles partagent. Le nombre de *k-mers* partagés entre les jeux de données est utilisé pour estimer la distance entre différents échantillons métagénomiques. La méthode consiste dans un premier temps à récupérer les k-mer des séquences métagénomiques. Chaque k-mer est transformé en séquence binaire en codant chaque nucléotide sur deux bits. Ensuite, chaque k-mer est associé à une case d'un tableau comportant l'ensemble des *k-mers* pouvant être retrouvés dans les jeux métagénomiques. Si une case est associée à un k-mer elle prend la valeur 1, sinon, elle est nulle. Enfin, Les séquences cibles sont découpées en *k-mers* et converties en séquences binaires pour les faire pointer sur une case du tableau associatif. Si la case est à 1, le k-mer est présent dans le métagénome. Chaque séquence cible à un score correspondant au nombre de *k-mers* partagés avec les séquences métagénomiques. Cet outil a l'avantage d'être rapide mais se trouve être peu sensible pour l'identification de nombreux faux

positifs. Un faux positif peut apparaître lorsqu'une séquence cible a ses *k-mers* présents dans l'index et qui sont sur des séquences métagénomiques différentes. De plus, cet outil ne permet pas de récupérer les séquences similaires pour pouvoir réaliser des analyses approfondies sur les lectures partagées entre les différents échantillons.

III.5.3 *Compareads*

Compareads[285] a été développé dans l'équipe GenScale à INRIA de Rennes dans le cadre du projet ANR *MAPPI* auquel j'ai participé. Cet outil a été conçu pour des analyses de métagénomique comparative *de novo*. *Compareads* permet de détecter les lectures similaires entre deux échantillons métagénomiques. Pour cela, les lectures des échantillons sont comparées deux à deux en se basant sur les *k-mers* partagés. Un *k-mers* partagé correspond à un mot de taille k dont la version *forward* et/ou *reverse complement* existe dans les deux échantillons à comparer. Ainsi, les séquences similaires sont définies avec cet outil comme étant deux séquences qui partagent avec une identité stricte au moins *n k-mers* non chevauchants (figure III.5.3.i). Le nombre ou la taille des l-mers sont fixés par l'utilisateur en fonction du type de données et du niveau de stringence recherché. Un score de similarité entre les échantillons métagénomiques est calculé. La similarité ici ne représente pas la vraie similarité entre deux séquences. En effet, celle-ci est une heuristique puisque des lectures peuvent être considérées à tort comme similaires entre les échantillons. Cela peut être le cas lorsqu'une séquence de 100pb partage 2 *k-mers* de 31 mots avec une autre séquence et être différente sur tout le reste de la séquence. Dans d'autres cas, l'ordre des *k-mers* partagés peut être différent entre les séquences considérées comme étant similaires. C'est pourquoi un symbole a été proposé pour définir cette similarité. Ainsi, la similarité entre deux échantillons A et B, normalement notée $(A \vec{\cap} B)$ est notée ici $(A \widetilde{\cap} B)$ comme étant les lectures de l'échantillon A « similaires » aux lectures de l'échantillon B.

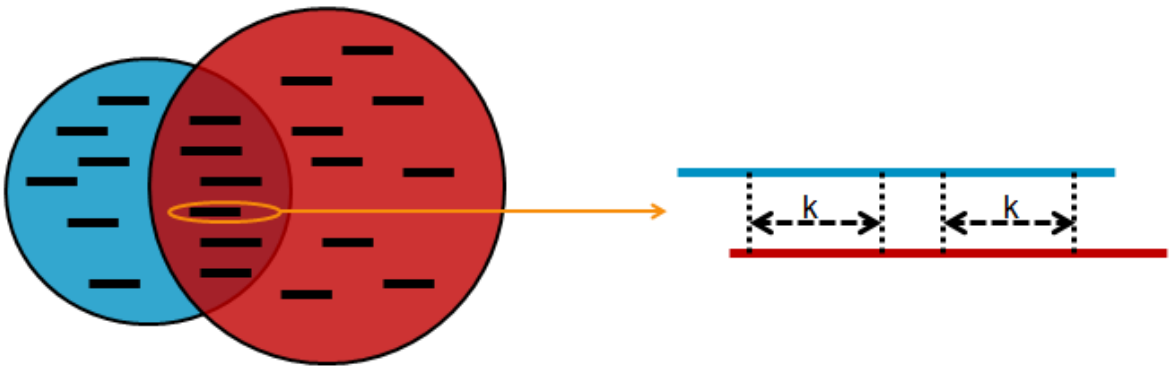


Figure III.5.3.i | Définition des séquences similaires par *Compareads*. Dans cet exemple, deux séquences sont dites similaires entre deux jeux de données métagénomiques lorsqu'elles partagent au moins 2 *k-mers* de tailles n non chevauchant avec une identité stricte.

Pour calculer la similarité de $(A \widetilde{\cap} B)$ le programme procède à une première étape d'indexation des *k-mers* de l'échantillon *B*. Une représentation de chaque k-mer des séquences de *B* est stockée dans un index. Afin de gagner en temps de calcul et en mémoire, l'indexation utilisée est basée sur un filtre de Bloom[286] dont la structure a été améliorée par les auteurs et appelée BDS (*Bloom Data Structure index*)[285]. Le filtre de Bloom est une structure de données probabiliste ou chaque mot à indexer est associé à un entier obtenu à partir d'une fonction de hachage. Dans un second temps, on regarde successivement pour chaque k-mer qui compose une lecture de l'échantillon *A* s'il existe dans *B*. Si le k-mer existe il y a incrémentation du nombre de *k-mers* partagés pour cette lecture et le prochain k-mer non chevauchant est testé. Sinon, on passe au k-mer de la position suivante sur la lecture. Enfin, si le nombre de *k-mers* partagés est égal ou dépasse celui indiqué pour que deux séquences soient similaires, la lecture étudiée est ajoutée à l'ensemble $(A \widetilde{\cap} B)$ et on passe à l'étude de la lecture suivante. Ainsi le score de similarité de lectures de l'échantillon A avec l'échantillon B est calculé. Cependant l'utilisation du filtre de Bloom peut induire des faux positifs (figure III.5.3.ii).

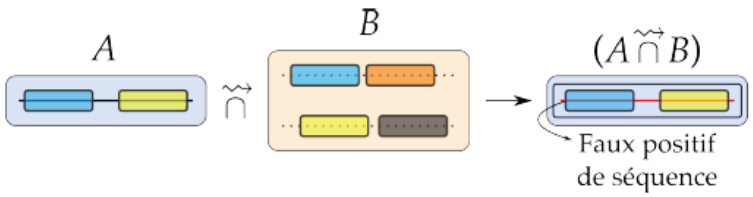


Figure III.5.3.ii | Représentation d'un faux positif de séquence. La lecture représentée dans le jeu *A* n'a aucune lecture similaire dans le jeu *B* avec laquelle elle partage au moins 2 *k-mers*. (Figure extraite de Maillet 2013[203])

Pour réduire ce nombre de faux positifs, les auteurs ont développé une méthode à trois étapes pour réaliser la comparaison complète de deux échantillons *A* et *B.* En effet, pour obtenir le nombre de lectures de *A* similaires au nombre de lectures de *B :* $(A \widetilde{\cap} B)$ puis le nombre de lectures de *B* similaires au nombre de lecture de *A :* $(B \widetilde{\cap} A)$ une troisième étape est effectuée (Figure III.5.3.iii). On a ainsi, une première étape avec le calcul de$(A \widetilde{\cap} B)$, puis une deuxième étape avec le calcul du set de lectures de *B* qui sont retrouvées dans *A* : $(B \widetilde{\cap} (A \widetilde{\cap} B))$ et enfin la troisième étape avec le set de lectures de *A* qui sont retrouvées dans B : $(A \widetilde{\cap} B) \widetilde{\cap} (B \widetilde{\cap} (A \widetilde{\cap} B))$. Ainsi, l'ensemble $(A \widetilde{\cap} (B \widetilde{\cap} A))$ est plus proche que $(A \widetilde{\cap} B)$ du résultat attendu$(A \vec{\cap} B)$.

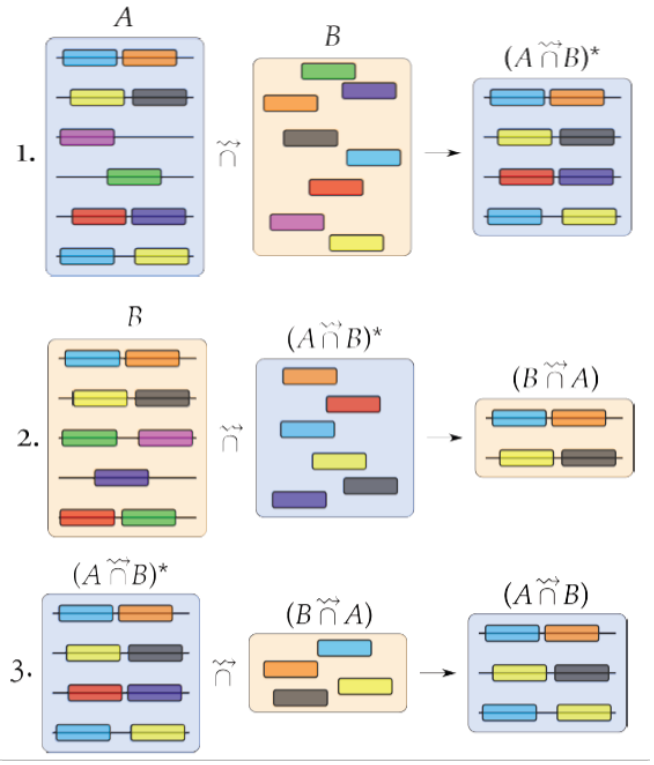


Figure III.5.3.iii | Représentation des trois étapes permettant de comparer les deux échantillons *A* et *B.* Un faux positif de séquence présent dans la première étape ($A \tilde{\cap} B$) a été éliminé lors de la troisième étape. (Figure extraite de Maillet 2013[203])

Deux scores de similarité peuvent être calculés avec cette méthode. Le premier est le score de similarité d'un jeu de données. Celui-ci correspond au nombre de séquences d'un premier échantillon similaire aux séquences d'un second jeu divisé par le nombre total de séquences du premier jeu :

$$S(A_B) = \frac{|(A \tilde{\cap} B)|}{|A|} \text{ x } 100$$

Ce score permet donc d'avoir l'information du nombre de lectures partagées pour les deux sens de comparaisons. Cette dissymétrie nous informe sur la complexité en séquence génomique d'un échantillon sur un autre. Le second score correspond au score de similarité global. Ce score, contrairement au premier est symétrique entre les deux échantillons comparés. Il correspond à la division du nombre total de séquences similaires trouvées, par le nombre total de séquences :

$$S(A,B) = \frac{|(A \tilde{\cap} B)| + |(B \tilde{\cap} A)|}{|A| + |B|} \text{ x } 100$$

Le calcul de ce dernier est proche de celui effectué pour obtenir la distance de Jaccard[287]. Cette dernière est utilisée pour mesurer la diversité et la similarité de deux ensembles. La seule différence avec le score de similarité global de *Compareads* est que le calcul de celui-ci se fait sur

des multi-ensembles pouvant contenir plusieurs fois la même lecture alors que le score de Jaccard se fait sur de simples ensembles :

$$JS(A,B) = \frac{|A \cap B|}{|A \cup \mathrm{B}|} \mathrm{x}\, 100$$

Ainsi, l'outil *Compareads* permet de calculer la similarité de deux jeux métagénomiques de plusieurs centaines de millions de lectures en quelques heures sur un ordinateur de bureau. En effet, le calcul de cette similarité a l'avantage d'être simple et la structure d'indexation utilisée ne nécessite pas de réaliser d'alignement entre les séquences. De plus, le fait de ne pas charger les séquences en mémoire pour les comparer permet à cet outil d'être très peu gourmand en mémoire vive. Enfin, cet outil a l'avantage de pouvoir fournir les séquences similaires communes aux différents échantillons métagénomiques. Dans le cadre de ma thèse, j'ai pu évaluer cet outil et effectuer des analyses sur une partie des échantillons *Tara* Oceans. Ces échantillons correspondent aux prélèvements réalisés dans une partie de l'Océan Indien ainsi que dans l'Océan Atlantique Sud afin d'étudier la connexion génétique qui s'opère entre ces deux océans via les anneaux d'Agulhas. L'analyse qui en résulte sera présentée dans le chapitre II.1.

Cependant, malgré les progrès en terme de gain de temps apportés par cette approche, le passage à plus grande échelle se trouve être compromis lorsque la quantité d'échantillons à comparer est élevée. En effet, dans le cadre du projet *Tara* Oceans, comparer l'ensemble des échantillons des 6 filtres de tailles d'organismes pris individuellement correspond à réaliser près de 60 900 comparaisons deux à deux. Étant donné qu'une comparaison de deux échantillons métagénomiques de 100 millions de lectures s'effectue en une dizaine d'heures, il faudrait 69 ans de temps CPU pour terminer la totalité des comparaisons. De plus, la récupération des lectures similaires entre les différents échantillons est utile mais prend beaucoup d'espace disque lorsque l'on passe à cette échelle. Ainsi, il a été nécessaire d'apporter des modifications à l'outil *Compareads* pour pouvoir comparer les échantillons à plus grande échelle. L'outil *COMMET* a été développé afin d'optimiser le temps de calcul et de pouvoir récupérer les lectures similaires aux échantillons sans pour autant nécessiter d'un espace de stockage conséquent. Celui-ci utilise le même algorithme que *Compareads* mais propose des améliorations sur plusieurs niveaux. Lors de ma thèse, j'ai participé à la conception de l'outil *COMMET* qui sera présenté dans le chapitre II.3.

D'autres outils ont été développés ensuite pour réaliser les comparaisons entre échantillons métagénomiques sur la base des *k-mers* avec la même optique de gain de temps tout en gardant une bonne sensibilité dans la qualité des résultats.

III.5.4 DSM

DSM[288] (*Distributed String Mining*) est un outil qui permet de calculer simultanément la similarité génomique entre des paires d'échantillons pour plusieurs jeux de données à partir d'un comptage de *k-mers*. La méthode utilisée estime dans un premier temps la fréquence des *k-mers* issus des séquences de l'ensemble des échantillons métagénomiques à analyser. Pour cela une normalisation est appliquée pour prendre en compte la différence de couverture et de profondeur de séquençage qu'il peut exister entre les échantillons. Une étape de filtrage des *k-mers* qui sont peu informatifs est également réalisée. Ces *k-mers* correspondent par exemple à ceux retrouvés dans tous les échantillons avec la même abondance. Ceux-ci représentent également les *k-mers* présents dans seulement un échantillon et pouvant donc être une erreur de séquençage. Dans un deuxième temps, l'algorithme décide si un k-mer est informatif ou non en regardant sa fréquence dans l'ensemble des métagénomes. Jellyfish[289] et DSK[290] sont deux outils permettant de réaliser un comptage de *k-mers* en stockant l'information dans des tables de hachage. Cependant le calcul est réalisé pour un seul échantillon et il est difficile de paralléliser le processus pour effectuer ce calcul sur plusieurs échantillons simultanément. Pour cela, les auteurs utilisent un *framework* DSM[291] qui permet de réaliser ce comptage sur l'ensemble des échantillons via l'utilisation d'un cluster de calculs. La troisième étape consiste à obtenir les distances de dissimilarité entre les échantillons métagénomiques en utilisant l'information de la fréquence des *k-mers* précédemment filtrés. Une distance de Jaccard est calculée donnant l'information sur la présence/absence des kmers dans deux échantillons. Deux autres métriques basées sur les distances euclidiennes sont utilisés pour avoir l'information sur l'abondance de *k-mers* partagés entre les échantillons. Ainsi, cette méthode permet de retrouver rapidement les *k-mers* similaires entre plusieurs échantillons avec une bonne sensibilité dans les résultats obtenus. Cependant, lorsque de nombreux échantillons sont à analyser, l'étape de comptage de *k-mers* demande un stockage intermédiaire important sur la mémoire vive. De plus, contrairement à *Compareads* et *COMMET*, il est impossible de récupérer les séquences des lectures similaires entre les échantillons.

III.5.5 MetaFast

L'algorithme de MetaFast[292] s'effectue en quatre étapes. Premièrement un assemblage *de novo* de chaque échantillon métagénomique est réalisé en se basant sur le graphe de Bruijn. Les séquences générées sont similaires aux contigs mais plus courtes. Ensuite, la construction d'un graphe de Bruijn à partir de toutes les séquences assemblées pour tous les métagénomes permet de rechercher des composants qui sont connectés entre ces séquences. Puis, un calcul d'un vecteur caractéristique de chaque métagénome de longueur égale au nombre de composants

connectés est effectué. Chaque élément vectoriel représente le nombre de *k-mers* d'un composant connecté qui sont présents dans les lectures du métagénome. Ainsi, les métagénomes sont comparés entre eux par la construction d'une matrice de dissimilarité avec des distances de Bray-Curtis basées sur les vecteurs calculés dans l'étape précédente. Ce programme a l'avantage d'être rapide et d'utiliser peu de mémoire. Cependant, l'utilisation des assemblages initiaux pour le calcul des dissimilarités restreint l'analyse des séquences qui participent à ces assemblages et n'évite pas des éventuelles erreurs d'assemblage.

III.5.6 Mash

Mash[293] est un outil très récent qui utilise un sous échantillonnage de *k-mers*. Pour réaliser cet échantillonnage, l'approche MinHash[294] est réalisée. MinHash a été utilisé à l'origine pour la détection de page web et d'images dupliquées. Son utilisation en génomique a débuté il y a une dizaine d'années[295] et a servi à la résolution de problèmes d'assemblage[296] dans le regroupement de gènes de l'ARNr 16S[297] ou encore dans le regroupement de séquences métagénomiques[298]. En effet, cette approche a l'avantage d'utiliser très peu de mémoire vive et de CPU ce qui la rend très performante pour étudier d'importantes quantités de données. Ainsi, Mash propose deux fonctions pour comparer deux échantillons métagénomiques. La première consiste à convertir les séquences métagénomiques en un « *MinHash sketch* ». Les séquences des deux échantillons sont découpées en *k-mers* et chaque k-mer est enregistré dans une fonction de hachage de 32 ou 64 bit correspondant à la taille des *k-mers*. Cette méthode conserve par défaut 1 000 *k-mers* par échantillon. La deuxième fonction consiste à calculer la distance de Jaccard à partir du nombre de *k-mers* partagés et spécifiques à chaque *sketch*. Celle-ci est approximative puisque l'on n'utilise qu'une partie des *k-mers* issues de l'union des deux *sketch* choisis aléatoirement. Cet outil permet par cette approche de récupérer l'information sur la présence/absence des *k-mers* entre deux échantillons en un temps extrêmement rapide par rapport aux outils présentés précédemment et avec une utilisation mémoire très faible. Mash est donc très intéressant pour manipuler de nombreux jeux de données métagénomiques tels que ceux du projet *Tara* Oceans. Cependant, il faudrait évaluer l'impact de la sélection d'une fraction des *k-mers* sur le score de similarité. En effet, dans le cas d'échantillons métagénomiques, la sélection d'une partie de l'information génétique peut ne pas être représentative de l'ensemble de la diversité génétique intra-échantillons. De plus, cette méthode ne donne pas l'information sur l'abondance des *k-mers*. Enfin, il est impossible de récupérer les lectures similaires entre deux échantillons.

III.5.7 *Simka*

Simka[299] est un outil développé dans l'équipe GenScale à INRIA de Rennes dans le cadre du projet ANR Hydrogen auquel je participe. Cet outil permet de réaliser une comparaison multiple des métagénomes via l'utilisation des *k-mers*. Les auteurs exploitent ici l'abondance de l'ensemble des *k-mers* de la totalité des échantillons métagénomiques à comparer. Pour cela, une matrice composée de l'ensemble des *k-mers* distincts issus des échantillons métagénomiques doit être générée. De plus, pour chaque k-mer, leur abondance dans chacun des échantillons permet d'estimer les distances entre ces échantillons. Pour cela, lorsqu'une centaine d'échantillons est à considérer, la construction de telles matrices d'abondances contenant des milliards de *k-mers* est impossible à réaliser car la mémoire requise est trop importante pour les ordinateurs actuels. Pour contourner ce problème, l'algorithme traite la matrice ligne par ligne, ou k-mer par k-mer pour ne pas stocker l'ensemble de la matrice. Pour cela, une première étape de comptage de *k-mers* de l'ensemble des jeux de données est effectuée. Comme précisé précédemment, l'outil DSK[290] permet de réaliser ce comptage échantillon par échantillon. Pour effectuer ce comptage multi-échantillons cet outil a dû être modifié (Figure III.5.7.i). La nouvelle méthode GATB–DSK implémentée à *Simka* consiste à lire les *k-mers* de chaque échantillon et à les stocker dans un fichier. Les *k-mers* sont ensuite classés dans chacun des fichiers pour permettre de repérer les *k-mers* identiques qui seront comptés pour connaitre l'abondance de chaque k-mer. Cette information sera fusionnée à l'ensemble des échantillons avec la construction de vecteurs d'abondance. On peut donc ici paralléliser les calculs sur plusieurs cœurs pour la construction de ces vecteurs. Cette notion de parallélisation des tâches pour accélérer le temps de calcul sera développée dans la partie III.6.4.

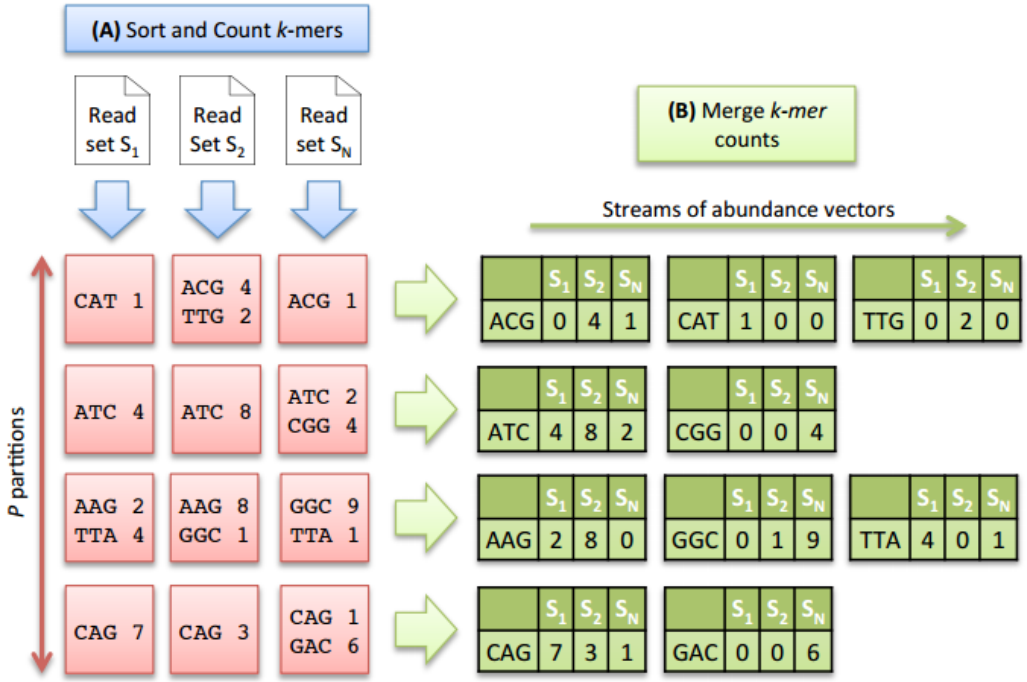


Figure III.5.7.i | Etapes pour le comptage de *k-mers* d'échantillons multiples avec la méthode GATB – DSK. a, La première étape consiste à enregistrer les *k-mers* de chaque échantillon dans des fichiers qui seront triés pour connaître le nombre des *k-mers* de chacun des échantillons. **b**, La deuxième étape fusionne le comptage de *k-mers* et crée des vecteurs d'abondance (Figure extraite de Benoit *et al.* 2016[299]).

Un filtre pour éliminer les *k-mers* pouvant résulter d'erreurs de séquençage est réalisée pendant l'étape de comptage. Ces *k-mers* correspondent à ceux dont l'abondance est inférieure à un certain seuil. Ce filtre permet de réduire le nombre de *k-mers* distincts et donc d'accélérer le temps d'exécution du programme. Les distances de similarité entre paires d'échantillons sont directement calculées à partir des vecteurs d'abondance de *k-mers* ce qui permet de ne pas stocker ces vecteurs en mémoire. *Simka* génère donc différentes mesures de distances écologiques en présence/absence ou en abondance de *k-mers* entre chaque paire d'échantillons. Ainsi, on peut par exemple obtenir les distances de Bray-Curtis ou encore de Jaccard représentées sous forme de matrice de distances symétriques. En effet, contrairement à *COMMET*, la comparaison ne se fait pas dans les deux sens. C'est-à-dire que l'on ne regarde pas les séquences d'un échantillon présentes dans un autre échantillon et inversement. De plus, *COMMET* calcule une distance de similarité basée sur les lectures qui partagent des *k-mers*. *Simka* ne se réfère qu'aux *k-mers* pris individuellement. Pour évaluer l'heuristique de *Simka*, les distances de similarité de ces deux outils ont été comparées. Une évaluation a également été effectuée à partir des alignements réalisés avec le logiciel Blat[217] (figure III.5.7.ii). Les mesures de similarité obtenues avec *Simka* et *COMMET* sont fortement corrélées (coefficient de

corrélation de Spearman r = 0.989). Une corrélation (r > 0.89) est observée entre le pourcentage de *k-mers* similaires et le pourcentage de lectures similaires définies par l'alignement Blat. L'outil *Simka* permet donc d'obtenir en des temps rapides des distances de similarité entre échantillons adéquates.

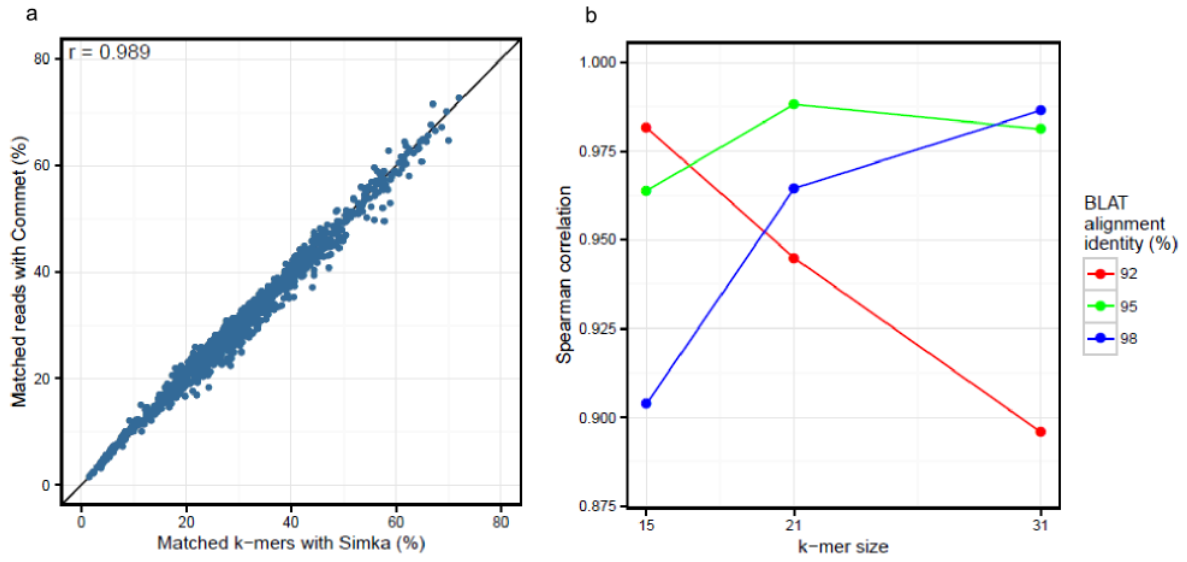


Figure III.5.7.ii | Evaluation de l'heuristique de l'outil *Simka*. a, Comparaison des mesures de similarité *Simka* et *COMMET*. Les deux outils ont été utilisés avec une taille de *k-mers* de 33pb. Chaque point représente une paire d'échantillons. **b**, Comparaison des distances *Simka* et BLAT avec différentes valeurs de tailles de *k-mers* et de seuils d'identité BLAT. (Figure extraite de Benoit *et al.* 2016[299])

Malgré le fait que *Simka* ne permette pas d'avoir l'information sur la différence de complexité de séquences entre les échantillons et qu'il est encore une fois impossible de récupérer les lectures similaires entre échantillons, l'utilisation de cet outil sur de nombreux jeux de données composés de centaines de millions de lectures est possible. Cependant, celui-ci doit alors être associé à l'utilisation de clusters de calcul composés de plusieurs cœurs et d'une mémoire conséquente. En effet, la procédure GATB – DSK pour le comptage des *k-mers* est optimisée pour être utilisée via la parallélisation des processus sur plusieurs cœurs. De plus, le stockage temporaire du comptage des *k-mers* de chaque échantillon dans des fichiers indépendants demande une mémoire de stockage conséquente. Un tel algorithme proposant en des temps raisonnables les différentes distances écologiques entre les échantillons métagénomiques donne la possibilité d'étudier les différences de diversités microplanctoniques à l'échelle planétaire. Ces distances de dissimilarité génomique prises deux à deux, combinées à des distances océanographiques, ou encore à des mesures physico-chimiques peuvent aider à

mieux comprendre comment un changement d'environnement influe sur la composition de la génomique de communautés planctoniques. Une telle étude sera présentée dans le chapitre III.

Avec l'augmentation exponentielle des projets en métagénomique ainsi que la quantité de données toujours plus volumineuses générées par les séquenceurs haut débit, de nombreux outils ont été pensés afin de pouvoir réaliser des analyses de métagénomique comparative de cette quantité massive de données. Ainsi, il est maintenant possible de comparer l'ensemble du contenu génétique d'un grand nombre de métagénomes. Cependant, ces outils sont dépendants de l'utilisation de clusters de calculs lorsque la quantité de données à comparer devient très importante, comme cela est le cas dans le cadre du projet *Tara* Oceans.

III.6 Utilisation de clusters de calculs pour la comparaison des échantillons métagénomiques de *Tara* Oceans

Pour comparer l'ensemble des échantillons *Tara* Oceans de surface et DCM des 6 filtres de tailles d'organismes pris individuellement il faut réaliser près de 60 900 comparaisons deux à deux. Les algorithmes des outils *COMMET* et *Simka* sont appropriés pour effectuer ces comparaisons. L'utilisation d'un nombre important de lectures métagénomiques doit être prise en compte afin de tenter d'obtenir l'information génétique de l'ensemble des micro-organismes planctoniques présents dans chaque échantillon. De plus, une normalisation du nombre d'échantillons doit être effectuée pour respecter les critères pour le calcul des distances. Pour cela, l'utilisation de 100 millions de lectures par échantillon pour les comparaisons correspond à un seuil adéquat. Malgré les performances de ces deux outils, il est nécessaire pour comparer l'ensemble des échantillons avec cette quantité de données d'utiliser des clusters de calculs appropriés.

III.6.1 Les clusters de calculs

Un *cluster* ou grappe en français, est une architecture composée de plusieurs ordinateurs formant des nœuds et où chacun des nœuds est capable de fonctionner indépendamment les uns des autres (figure III.6.1). Un cluster de calculs permet donc de répartir une charge de travail parmi un grand nombre de serveurs afin d'utiliser la performance cumulée de chacun des nœuds. Les nœuds peuvent être constitués d'un processeur multi-cœur. Ce processeur possède plusieurs cœurs physiques fonctionnant simultanément. Un cœur physique est un ensemble de circuits capables d'exécuter des programmes de façon autonome. Il est donc possible de répartir des tâches sur un ou plusieurs cœurs pour permettre de traiter des informations de manière

simultanée. On réalise alors une parallélisation pour réaliser un grand nombre d'opérations afin d'accélérer les temps de calculs.

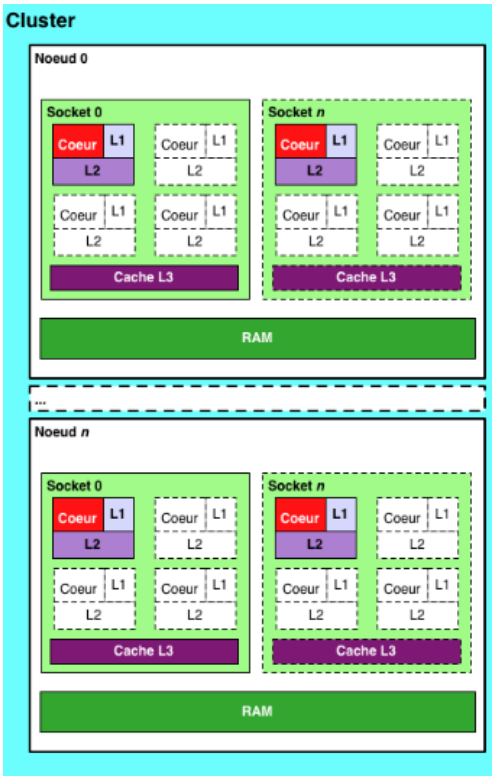


Figure III.6.1 | Schéma synthétique d'un cluster de calculs. Les cœurs partagent une mémoire cache (cache L3), mais possèdent également chacun leur propre mémoire cache (cache L1 et L2). (Figure extraite de http://cs-ljk.imag.fr/)

III.6.2 Ressource informatique au Genoscope

Le Genoscope est pourvu d'un réseau informatique composé d'une cinquantaine de serveurs de calculs représentant un total de 430 cœurs. Ces serveurs de calculs sont exploités sous deux gestionnaires de ressources et ordonnanceurs de tâches. Un *Load Sharing Facility* (LSF) et un *Simple Linux Utility for Ressource Management* (*SLURM*). Ces *batch managers* sont utilisés pour distribuer les différentes tâches en fonction de la charge des machines. Certaines tâches seront suspendues en fonction de leur priorité ainsi que du nombre de processus que l'utilisateur doit exécuter. En effet, ces ressources informatiques sont partagées avec l'ensemble des laboratoires du Genoscope. Leurs utilisations doivent donc respecter des règles d'usages pour ne pas perturber les calculs des autres utilisateurs. Dans le cadre des comparaisons métagénomiques des échantillons de *Tara* Oceans, l'utilisation des ressources informatiques du Genoscope a été possible lorsqu'il s'agissait d'étudier les stations d'une partie du projet. A l'échelle globale, trop de cœurs sont à allouer pour effectuer l'ensemble des comparaisons sur ces machines. Il a fallu alors utiliser une autre ressource de calculs.

III.6.3 Centre de calcul du CEA

Dans le cadre du projet France Génomique, le Genoscope a accès à des supercalculateurs du Très Grand Centre de Calcul du CEA (TGCC). Celui-ci est pourvu entre autre du supercalculateur *Airain* fourni par l'entreprise *BULL* en 2012. Il est composé d'une puissance de calcul de 420 Tflops avec de nombreux nœuds et cœurs de calculs à disposition (Figure III.6.3). Ces nœuds sont interconnectés par un réseau haute performance *InfiniBand QDR* permettant une gestion efficace de la parallélisation des calculs. La capacité de stockage local des données sur le supercalculateur *Airain* est de 2,3 Péta-octet ce qui donne la possibilité d'utiliser des outils, tel que *Simka*, qui stocke temporairement de grosses quantités de données. *Airain* utilise le système d'ordonnancement des tâches *SLURM* qui permet de vérifier si les ressources nécessaires pour les comparaisons sont disponibles ainsi que de savoir si les conditions spécifiées sont respectées. Ainsi, avec la quantité de métagénomes à comparer, il est possible d'utiliser ce supercalculateur pour paralléliser les comparaisons sur les différents nœuds et cœurs de calculs.

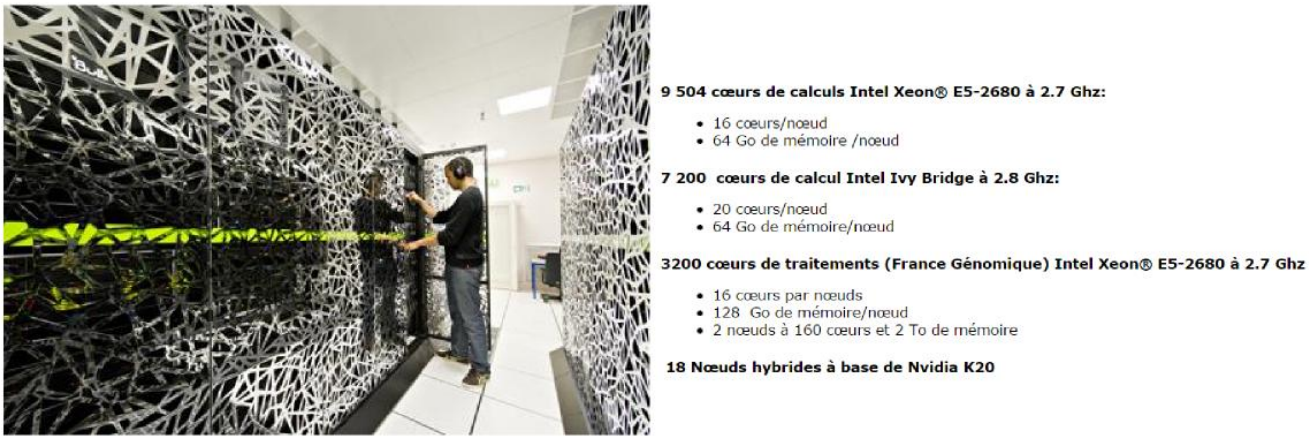


Figure III.6.3 | Architecture du supercalculateur *Airain*. (Figure extraite de http://www-ccrt.cea.fr/)

III.6.4 Parallélisations des processus pour la comparaison des métagénomes

L'outil *COMMET* compare les échantillons deux à deux. La réalisation de l'ensemble des comparaisons des échantillons *Tara* Oceans nécessite d'exécuter 60 900 comparaisons. Il est possible d'utiliser un micro-ordonnanceur de tâches pour organiser la répartition des processus sur les différents cœurs. Pour cela *GLOST* est un micro-ordonnanceur de tâches développé par le support applicatif du Centre de Calcul Recherche et Technologie (CCRT) du CEA utilisant le *Message Passing Interface* (MPI) utilisable avec *SLURM*. Il permet de regrouper un nombre important de tâches similaires et de faible durée dans une même tâche *SLURM*. De plus, *GLOST* assure l'utilisation optimale des ressources qui lui sont allouées pour l'ensemble des tâches qu'il

gère. Ainsi, afin d'optimiser l'enchainement des comparaisons des échantillons métagénomiques avec *COMMET*, l'ordonnanceur *GLOST* peut être utilisé pour permettre la répartition et l'exécution de ces comparaisons sur les différents cœurs de calculs d'*Airain*. Cela a permis d'exécuter automatiquement l'ensemble des comparaisons. Cependant, des stratégies ont dû être mises en place pour éviter les conflits d'entrée/sortie pour la lecture des fichiers fastq et l'écriture dans les fichiers de sortie. De plus, même si la puissance de calcul que dispose *Airain* est très importante, il a fallu partitionner les 60 900 comparaisons pour respecter le temps de calcul maximum autorisé par processus. L'outil *Simka* possède son propre ordonnanceur de tâches pour paralléliser les processus de la phase de comptage. Il n'a pas été nécessaire avec *Simka* de fractionner les échantillons pour réaliser les multi-comparaisons des échantillons de chaque fraction de tailles d'organismes. Cependant, une quantité plus importante d'échantillons aurait certainement nécessité de modifier le code de l'outil ou d'utiliser une machine plus puissante.

Ainsi, l'utilisation de l'ensemble des lectures issues d'un séquençage métagénomique est une approche adéquate pour explorer la diversité globale des micro-organismes planctoniques à un niveau taxonomique résolutif. L'utilisation d'outils conçus spécialement pour réaliser rapidement des comparaisons métagénomiques complétés par des clusters de calculs puissants qui rend possible la réalisation de l'ensemble des comparaisons des échantillons du projet *Tara* Oceans. Ces outils génèrent différentes informations. Ainsi, l'outil *COMMET* permet notamment de récupérer les lectures similaires présentes dans les intersections des comparaisons métagénomiques. Des analyses taxonomiques peuvent par exemple être réalisées en aval. De plus, l'utilisation de l'outil *Simka* permet d'obtenir différentes distances écologiques entre les échantillons métagénomiques. Ces distances, corrélées aux données océanographiques et physico-chimiques permettent par exemple d'étudier l'impact de l'environnement sur l'organisation des communautés génomiques micro-planctoniques. Les résultats de telles analyses seront présentés dans le chapitre III.

Chapitre I

Caractérisation et comparaison de la biogéographie et de la structure du répertoire de gènes d'algues vertes photosynthétiques dans les eaux océaniques de surface

La génomique comparative vise à comparer les génomes d'organismes pour mettre en exergue les variations génomiques existantes entre ces organismes. Ces différences peuvent s'expliquer par des mécanismes évolutifs et d'adaptation à un environnement donné. Il est possible, via la métagénomique ciblée, de décrire la biogéographie d'une espèce ainsi que l'abondance de ces gènes dans différents environnements. Les analyses suivantes sont réalisées sur des micro-organismes planctoniques appartenant au groupe des *Mamiellales* présenté dans la partie I.2.5 du contexte biologique et méthodologique.

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

Bathycoccus prasinos est une algue verte photosynthétique unicellulaire décrite comme étant abondante dans les océans et un contributeur majeur de la production primaire. Sa distribution cosmopolite dans les océans soulève des questions sur sa diversité et ses adaptations en fonction des conditions environnementales.

L'exploration de la diversité génomique et spatiale de *Bathycoccus* a été réalisée à partir du génome de la souche côtière méditerranéenne (RCC1105) et d'un SAG (*Single Amplified Genome*) récolté dans l'Océan Indien lors de l'expédition *Tara* Oceans. Le Genoscope a séquencé et annoté ce dernier. L'assemblage du SAG estimé comme étant complet à 64% du génome est nommé TOSAG39-1. Une analyse de génomique comparative de ces deux génomes partageant la même séquence de l'ARNr 18S a révélé qu'ils sont génétiquement distants. En effet, ils partagent une identité protéique à 78% et ont une synténie incomplète. De plus, les deux génomes possèdent différents *Internal Splicing marker* (ITS). Ainsi, ces génomes correspondent à deux espèces de *Bathycoccus*. L'analyse de leur distribution géographique dans 122 échantillons métagénomiques a été réalisée. Les deux écotypes se trouvent dans des environnements riches

en nutriment, mais ont des niches écologiques distinctes. TOSAG39-1 est retrouvé dans des milieux plus chauds et au niveau de la profondeur DCM alors que RCC1105 est détecté dans des milieux plus froids, en surface et riche en oxygène. Enfin, les deux *Bathycoccus* sont généralement absents des stations possédants une faible concentration en fer.

L'étude de la variation génomique de la souche Méditerranéenne RCC1105 dans les différents échantillons a été réalisée. Les gènes dispensables, qui correspondent aux gènes présents ou absents dans certains échantillons, ont été détectés. À partir des données métatranscriptomiques, leur niveau d'expression a été quantifié. Un total de 108 gènes dispensables (représentant environ 1% du génome) a été observé. L'étude de leurs structures et de leurs positions sur les chromosomes a été réalisée. La moitié de ces gènes est localisée sur le chromosome 19 connu pour être un chromosome *outlier* chez *Bathycoccus*. L'autre moitié est positionnée aléatoirement sur le génome, souvent en cassettes de gènes successifs. Il a été vérifié, à l'aide des assemblages métagénomiques, que la continuité des régions génomiques autour de ces cassettes était bien présente. Une étude de synténie entre les génomes de RCC1105 et de TOSAG39-1 a également été effectuée pour comparer les régions contenant ces gènes dispensables positionnés en cassettes. Cette analyse a montré l'absence de ces cassettes dans le génome de TOSAG39-1. La couverture en lecture qui s'aligne sur ces cassettes de gènes diffère selon les bassins océaniques. Cela implique plusieurs types génomiques au sein d'une même espèce qui varient en fonction du milieu écologique.

Ces résultats ont fait l'objet d'une publication le 30 Novembre 2016 dans le journal *Scientific Report* et sont présentés dans l'article suivant. Les informations supplémentaires de cet article se trouvent dans l'annexe I.

SCIENTIFIC REPORTS

OPEN

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

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

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

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

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

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

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

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

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

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

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

Results

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

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

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

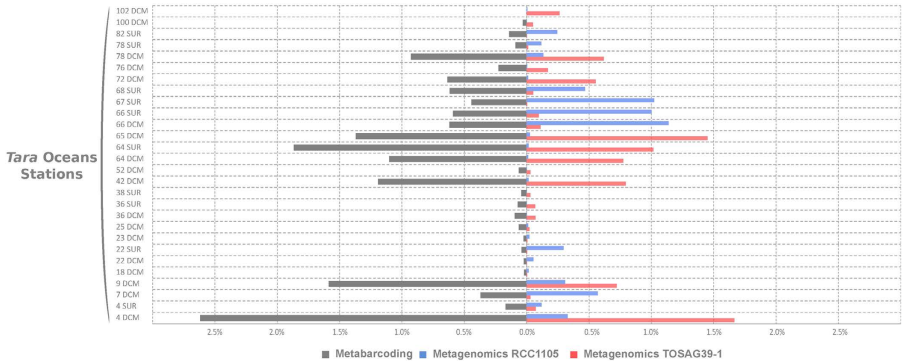


Figure 1. Comparisons of relative abundances of *Bathycoccus* in the 0.8–5 μm size fraction samples from *Tara* Oceans stations. Left: relative 18S rRNA V9 amplicons abundance (percent of reads). Right: relative metagenomic abundances (percent of metagenomic reads) from direct mapping of metagenomic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

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

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

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

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

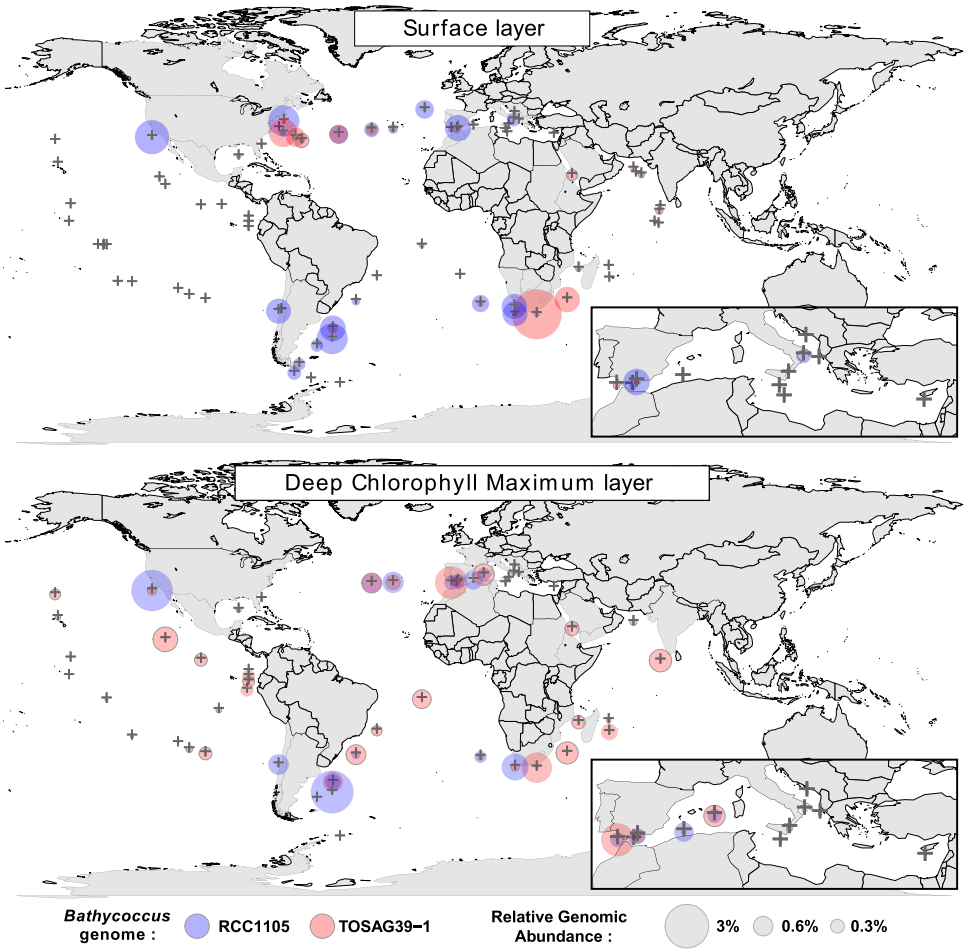


Figure 2. Geographical distribution of two *Bathycoccus* genomes, RCC1105 and TOSAG39-1, along *Tara* Oceans expedition stations from recruitments of metagenomic reads. Top and bottom maps correspond to the surface and deep chlorophyll maximum (DCM) samples respectively. Gray crosses indicate *Tara*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

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

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

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

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

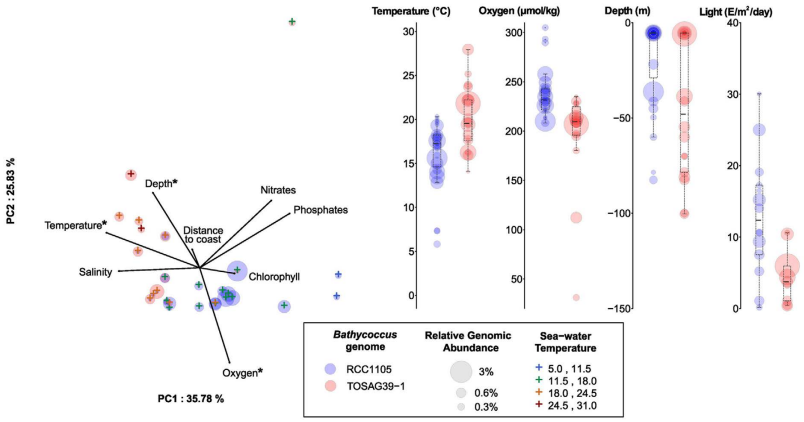


Figure 3. Relationships between environmental parameters and *Bathycoccus* genome abundance. Left: Principal component analysis. We only considered stations where we detected 98% of the genes for one *Bathycoccus* genome, and for which all environmental parameters were available (Oxygen, Nitrates, Phosphates,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

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

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

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

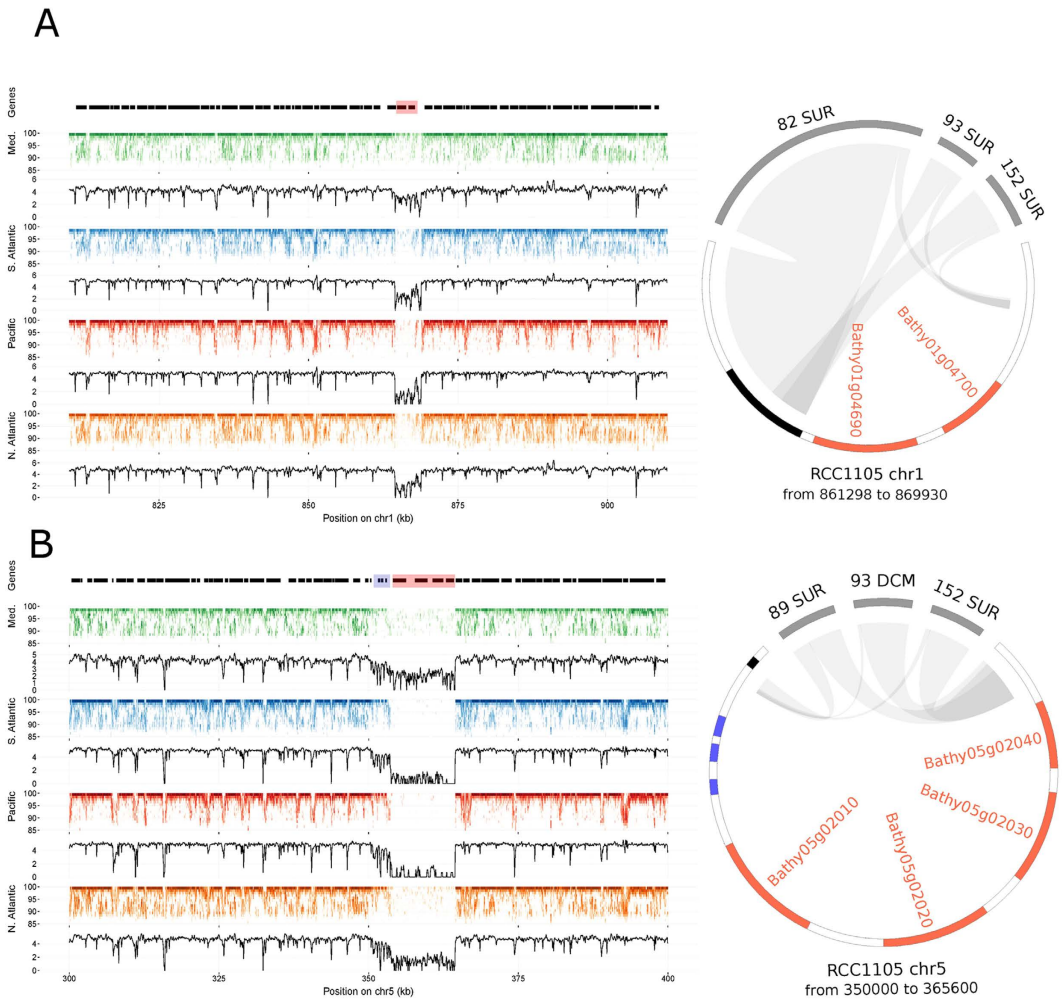


Figure 4. Evidence for cassettes of dispensable genes in *B. prasinos* RCC1105. Left and right sides of the figures represent fragment recruitment and genomic alignments of dispensable gene cassettes, respectively. Fragment recruitments plots are displayed by marine zones (left legend). Each dot corresponds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

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

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

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

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

Discussion

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

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

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

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

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

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

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

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

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

Material and Methods

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

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

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

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

References

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

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

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

Acknowledgements

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

Author Contributions

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

Additional Information

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

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

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

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

I.2. Étude d'un gène dispensable chez *Bathycoccus prasinos* : La flavodoxine

I.2.1 Introduction sur la flavodoxine et la ferrédoxine

La flavodoxine (Fld) est une flavoprotéine qui intervient dans le transport des électrons dans la chaîne respiratoire. La flavodoxine est donc impliquée dans les réactions d'oxydoréduction. La ferrédoxine est également une protéine réalisant des transferts d'électrons dans les réactions d'oxydoréduction. Cependant, cette dernière est particulièrement sensible au stress oxydatif et à la limitation en Fer. C'est pourquoi chez certain micro-organismes photosynthétiques la ferrédoxine a été remplacée par la flavodoxine qui est plus résistante à ces différents stress. La flavodoxine, généralement retrouvée chez certaines bactéries et algues eucaryotes localisées dans des régions pauvres en Fer, est absente dans les régions riches en Fer. Cela indique une pression de sélection en faveur de la ferrédoxine dans les régions costales et dans les eaux froides riche en fer. La flavodoxine a été perdue au cours de l'évolution et est absente chez les plantes terrestres. Des résultats récents ont montré un profil d'évolution très dynamique de la flavodoxine chez les lignées d'algues eucaryotes étroitement lié à la disponibilité en Fer[300] (Figure I.2). Cela est le cas pour les mamiellales où la flavodoxine est absente chez certaines espèces et présente dans d'autres. La flavodoxine est retrouvée chez l'écotype méditerranéen côtier de *Bathycoccus prasinos,* mais n'est pas présente dans les deux métagénomes prélevés dans l'*upwelling* du Chili[40] ainsi que dans le métagénome de l'océan atlantique tropical[301]. Cela indiquerait que le gène de la flavodoxine est un gène dispensable chez *Bathycoccus prasinos* qui aurait persisté ou serait réapparu chez certains génotypes au cours de l'évolution. Je présente ici une étude de métagénomique ciblée de ce gène réalisée à partir des échantillons métagénomique de *Tara* Oceans.

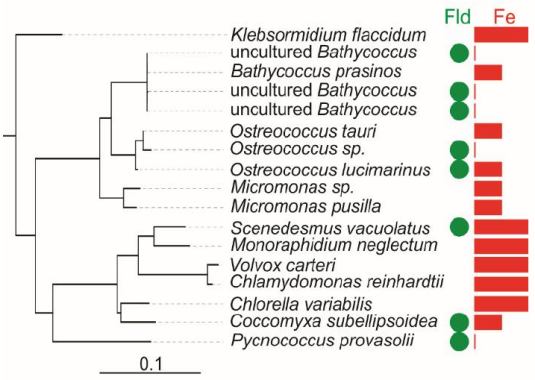


Figure I.2.1 | Distribution phylogénétique des gènes codants pour la flavodoxine (Fld) chez les eucaryotes photosynthétiques. L'arbre phylogénétique des chlorophytes et du charophyte *K. flaccidum* a été construit à partir des ARNr 18S. Pour chaque espèce, la présence (cercles vert) ou l'absence de la Fld ainsi que la concentration en Fer (barre rouge) des sites d'isolation sont indiquées. Pour la majorité des cas, la Fld est présente lorsque la concentration en Fer est faible. La Fld est retrouvée chez l'écotype méditerranéen côtier de *Bathycoccus prasinos* mais n'est pas retrouvée dans les trois métagénomes, ce qui indiquerait que le gène de la Fld est un gène dispensable. Les autres mamiellales *Ostreococcus* et *Micromonas* sont également inclus dans l'analyse. (Figure extraite de Karlusich *et al.*2015[300])

I.2.2 Récupération du gène de la flavodoxine chez TOSAG39-1

La séquence de la flavodoxine du métagénome de *Bathycoccus* de l'*upwelling* du Chili à la profondeur 5m (échantillon T142 disponible dans la base de données de l'EMBL-EBI sous le numéro d'accession CAFX01000001-CAFX01015049) a été utilisée pour savoir si le SAG de *Bathycoccus* (TOSAG39-1) possède ce gène. Suite à l'alignement de la séquence protéique de la flavodoxine sur les scaffolds de TOSAG39-1, deux séquences ont été récupérées sur les gènes 5577 et 5581 (Figure I.2.2). Étant donné que ce métagénome a été décrit comme étant similaire au génome de la souche RCC1105 qui ne possède pas la flavodoxine (Figure S5 de l'annexe I) et qu'il a été suggéré qu'au moins deux génotypes de *Bathycoccus* coexistent dans celui-ci[40], il est possible que ce gène de la flavodoxine fasse partie des gènes dispensables de *Bathycoccus prasinos*.

```
Query= Fld Bathycoccus prasinos Chili 5m

Length=54
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  GSBathycoccus1_scaffold_389                                          112    2e-30
  GSBathycoccus1_scaffold_388                                          112    2e-30


> GSBathycoccus1_scaffold_389
Length=7177

 Score =  112 bits (280),  Expect = 2e-30, Method: Composition-based stats.
 Identities = 53/54 (98%), Positives = 53/54 (98%), Gaps = 0/54 (0%)
 Frame = -1

Query  1     LNSFRDFSALLVGTPTWNTGAEEMRSGTTWDNILEEVRSENLQGKKVAVFGCGD  54
             L SFRDFSALLVGTPTWNTGAEEMRSGTTWDNILEEVRSENLQGKKVAVFGCGD
Sbjct  6784  LXSFRDFSALLVGTPTWNTGAEEMRSGTTWDNILEEVRSENLQGKKVAVFGCGD  6623


> GSBathycoccus1_scaffold_388
Length=7177

 Score =  112 bits (280),  Expect = 2e-30, Method: Composition-based stats.
 Identities = 53/54 (98%), Positives = 53/54 (98%), Gaps = 0/54 (0%)
 Frame = -1

Query  1     LNSFRDFSALLVGTPTWNTGAEEMRSGTTWDNILEEVRSENLQGKKVAVFGCGD  54
             L SFRDFSALLVGTPTWNTGAEEMRSGTTWDNILEEVRSENLQGKKVAVFGCGD
Sbjct  6784  LXSFRDFSALLVGTPTWNTGAEEMRSGTTWDNILEEVRSENLQGKKVAVFGCGD  6623
```

Figure I.2.2 | Alignement du gène de la flavodoxine du métagénome de l'échantillon T142 sur le TOSAG39-1. Un alignement tblastn avec une *e-value* à 10^{-3} a été réalisé à partir de la séquence protéique du gène de la flavodoxine du métagénome T142 contre les scaffolds de TOSAG39-1. Les alignements à 98% d'identité ont permis de récupérer les deux séquences de la flavodoxine chez TOSAG39-1.

I.2.3 Le gène de la flavodoxine est-il un gène dispensable chez *Bathycoccus prasinos* ?

Le gène de la flavodoxine est absent dans le génome de la souche RCC1105. Il est donc impossible de détecter celui-ci dans la liste des gènes dispensables fournie dans l'étude précédente. Cependant, il est possible d'aligner les lectures des métagénomes ne contenant qu'un des deux écotypes de *Bathycoccus* sur ce gène afin d'observer son abondance relative dans les différents échantillons (figure I.2.3). Ainsi, les lectures des échantillons ne contenant que le génome de TOSAG39-1 ont été alignées sur la séquence de la flavodoxine. La corrélation entre l'abondance métagénomique de ce gène et l'abondance de TOSAG39-1 dans ces échantillons démontre que le gène de la flavodoxine est toujours présent dans le génome de TOSAG39-1. Cela a été confirmé par l'absence de ce gène dans la liste des gènes dispensables de TOSAG39-1. De plus, le métagénome de l'océan atlantique tropical qui est similaire à TOSAG39-1 possède effectivement ce gène lui aussi. Lorsque l'on aligne les lectures des échantillons ne contenant que le génome de RCC1105 sur le gène de la flavodoxine, la corrélation de l'abondance métagénomique de ce gène et l'abondance de RCC1105 ne s'observent que sur certains échantillons. Cela indiquerait l'existence d'un autre génotype de *bathycoccus prasinos* qui possède le gène de la flavodoxine.

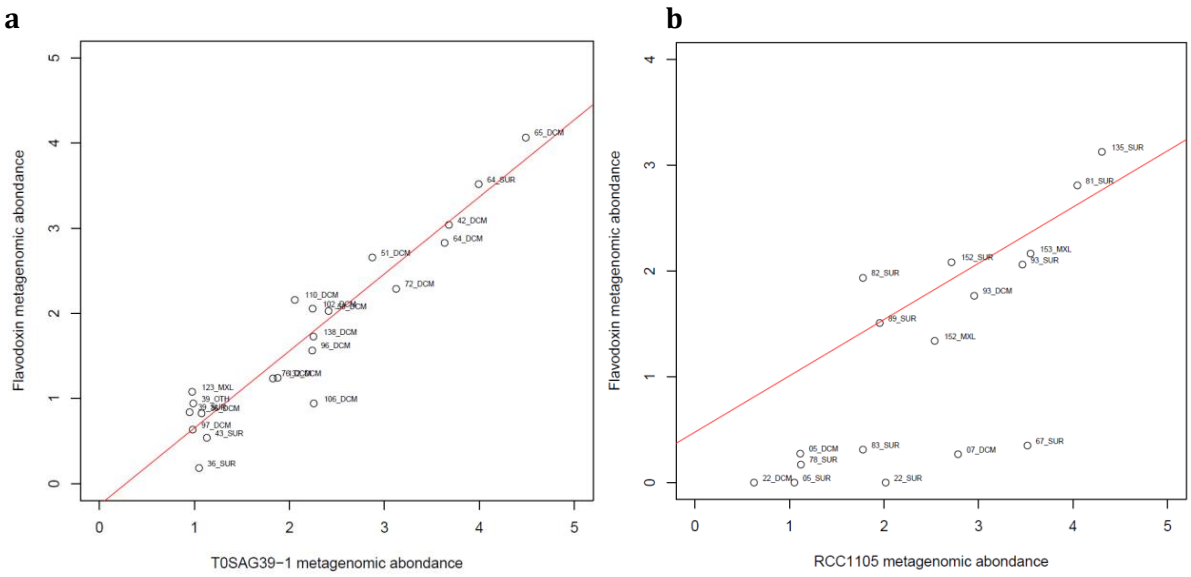


Figure I.2.3 | Recherche d'un autre génotype de RCC1105 possédant le gène de la flavodoxine. **a**, Scaterplot représentant l'abondance relative de TOSAG39-1 en fonction de l'abondance relative de la flavodoxine dans les échantillons ne possédant que le génome de TOSAG39-1. **b**, Scaterplot représentant l'abondance relative de RCC1105 en fonction de l'abondance relative de la flavodoxine dans les échantillons ne possédant que le génome de RCC1105.

I.2.4 Tentative de récupération du gène de la flavodoxine de *Bathycoccus prasinos* dans les échantillons métagénomiques

Pour pouvoir valider l'existence d'un deuxième génotype de *Bathycoccus prasinos* possédant le gène de la flavodoxine, on a cherché à reconstruire la région génomique de *Bathycoccus prasinos* correspondante. Pour cela, les assemblages des échantillons métagénomiques ne contenant que la souche RCC1105, qui seraient susceptible de contenir le génotype possédant la flavodoxine, ont été réalisés. Les contigs issus des assemblages de ces échantillons ont été alignés sur la séquence de la flavodoxine du métagénome T142. Des alignements avec un fort pourcentage d'identité avec la séquence de référence ont été obtenus mais la taille des contigs assemblés étant trop courte, il a été impossible de recouvrir l'ensemble de la séquence du gène de la flavodoxine.

Ainsi, les observations obtenues avec la méthode de métagénomique ciblée sur les échantillons *Tara* Oceans suggèrent l'existence d'un deuxième génotype de *Bathycoccus prasinos* qui possède la flavodoxine. Le SAG de *Bathycoccus* (TOSAG39-1) a été décrit dans la partie précédente comme étant présent dans les échantillons pauvre en fer et dans des eaux plus chaude que l'écotype côtier RCC1105. Il est possible que ce dernier ait subi une pression de sélection en faveur de la ferrédoxine et que d'autres génotypes aient conservé le gène de la flavodoxine. Des analyses plus approfondies sont nécessaires pour conclure sur ce scénario.

I.3. Caractérisation de « gènes inconnus » chez *Bathycoccus prasinos* par une méthode de phylostratigraphie

I.3.1 La phylostratigraphie

Les gènes inconnus correspondent aux gènes n'ayants aucune similarité détectable avec une séquence connue d'une autre espèce et qui n'ont donc pas de fonction déductible par similarité. Ces gènes, également appelés gènes orphelins[302] par l'absence d'homologues avec d'autres lignées, ont une origine évolutive peu connue. Pourtant, ils peuvent représenter jusqu'à un tiers des gènes chez les bactéries, les archées ou encore les phages[303]. Ces gènes auraient une importance dans le développement de stratégie d'adaptation et d'interaction avec l'environnement spécifique à un taxon[303–306]. Il a notamment été montré chez le microcrustacé *Daphnia pulex* que beaucoup de ses gènes spécifiques ont des fonctions inconnues mais liés à une réponse environnementale[305]. Il est possible d'identifier les gènes orphelins dans un contexte phylogénétique afin d'observer l'émergence ou la perte de gènes au sein d'une lignée taxonomique[307]. Cette procédure, appelée phylostratigraphie, représente donc l'origine des gènes dans un contexte de comparaison de génomes séquencés et annotés sur plusieurs niveaux de hiérarchie phylogénétique. Pour construire la phylostratigraphie des gènes d'un organisme, on procède à un alignement blast de ses gènes sur les gènes d'autres organismes plus ou moins proches phylogénétiquement. Chacun des gènes qui ont un ou plusieurs alignements sera assigné sur le nœud correspondant au plus lointain ancêtre commun (LCA). Chaque nœud représente un *phylostratum* avec la correspondance du nombre de gènes apparus à ce niveau taxonomique. Cela implique que chaque lignée taxonomique possède une fraction de gènes restreinte à ce groupe. Ces gènes sont nommés *Taxonomicaly restricted genes* (TRGs)[303,308]. On a donc pour chaque phylostratum, le nombre de gènes restreint à ce groupe taxonomique. Si un gène ne s'aligne sur aucune référence, celui-ci sera placé au dernier nœud de l'arbre et correspondra à un gène orphelin (Figure I.3.1).

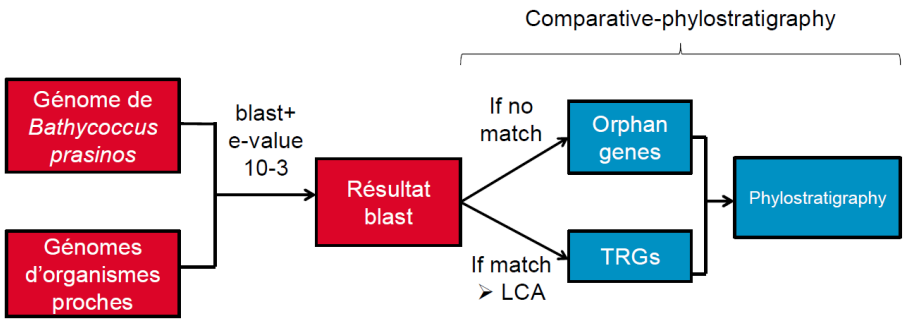


Figure I.3.1 | Méthode pour l'étude phylostratigraphique des gènes d'un organisme. L'organisme étudié est *Bathycoccus prasinos.* Un alignement blast+ avec un seuil de *e-value* à 10^{-3} est réalisé contre les gènes d'organismes proches afin d'étudier l'évolution d'apparition de ces gènes dans les différentes lignées.

I.3.2 Détection des gènes orphelins de *Bathycoccus prasinos*

La plupart des gènes inconnus détectés dans le métagénome planctonique est d'origine eucaryote[309]. Il y a donc un large répertoire de fonction de gènes non exploré dans l'océan. L'étude de ce catalogue de gènes dans différents environnements pourrait permettre de mieux comprendre les mécanismes d'adaptation et d'évolution des micro-organismes planctoniques. Une analyse de ces gènes a d'abord été effectuée sur le génome de la souche RCC1105 de *Bathycoccus prasinos*. L'arbre phylostratigraphique a été réalisé à partir des génomes d'eucaryotes photosynthétiques disponibles dans la base de données pico-PLAZA[310] (le génome de *Bathycoccus prasinos* est alors enlevé de la base). Le choix de cette base de données s'explique par le fait qu'elle contient des génomes séquencés et annotés fiables, contrairement à la banque de données UniRef[311] qui peut contenir des erreurs d'assignations et d'annotation. De plus, pour une analyse sur l'évolution des gènes orphelins en rapport avec l'adaptation des micro-organismes planctoniques à un environnement donné, les gènes d'intérêt seront plutôt ceux apparus dans les dernières lignées. D'autant plus que l'utilisation des deux bases de données UniRef et pico-PLAZA présente des résultats similaires en terme de nombre de TRGs dans les phylostratum phylogénétiquement proches de *Bathycoccus prasinos* (Figure I.3.2.a). Ainsi, les séquences codantes des gènes de RCC1105 ont été alignées sur les gènes des génomes présents dans la base pico-PLAZA. Plus la *e-value* est faible et plus l'alignement sera stringent ce qui augmentera le nombre de gènes orphelins (Figure I.3.2.b). Un seuil de la *e-value* entre 10^{-3} et 10^{-4} a été évalué comme étant optimum pour maximiser la sensibilité et la spécificité des alignements chez la drosophile[302]. Il est utilisé comme seuil de référence dans d'autres analyses phylostratigraphiques[312]. Dans les analyses suivantes, un seuil de 10^{-3} a donc été utilisé. L'arbre phylostratigraphique de *Bathycoccus prasinos* a été construit et on dénombre un total de 742 gènes orphelins (≈ 10% du génome) ou TRGs au niveau du phylostratum *Bathycoccus prasinos* (figure I.3.3).

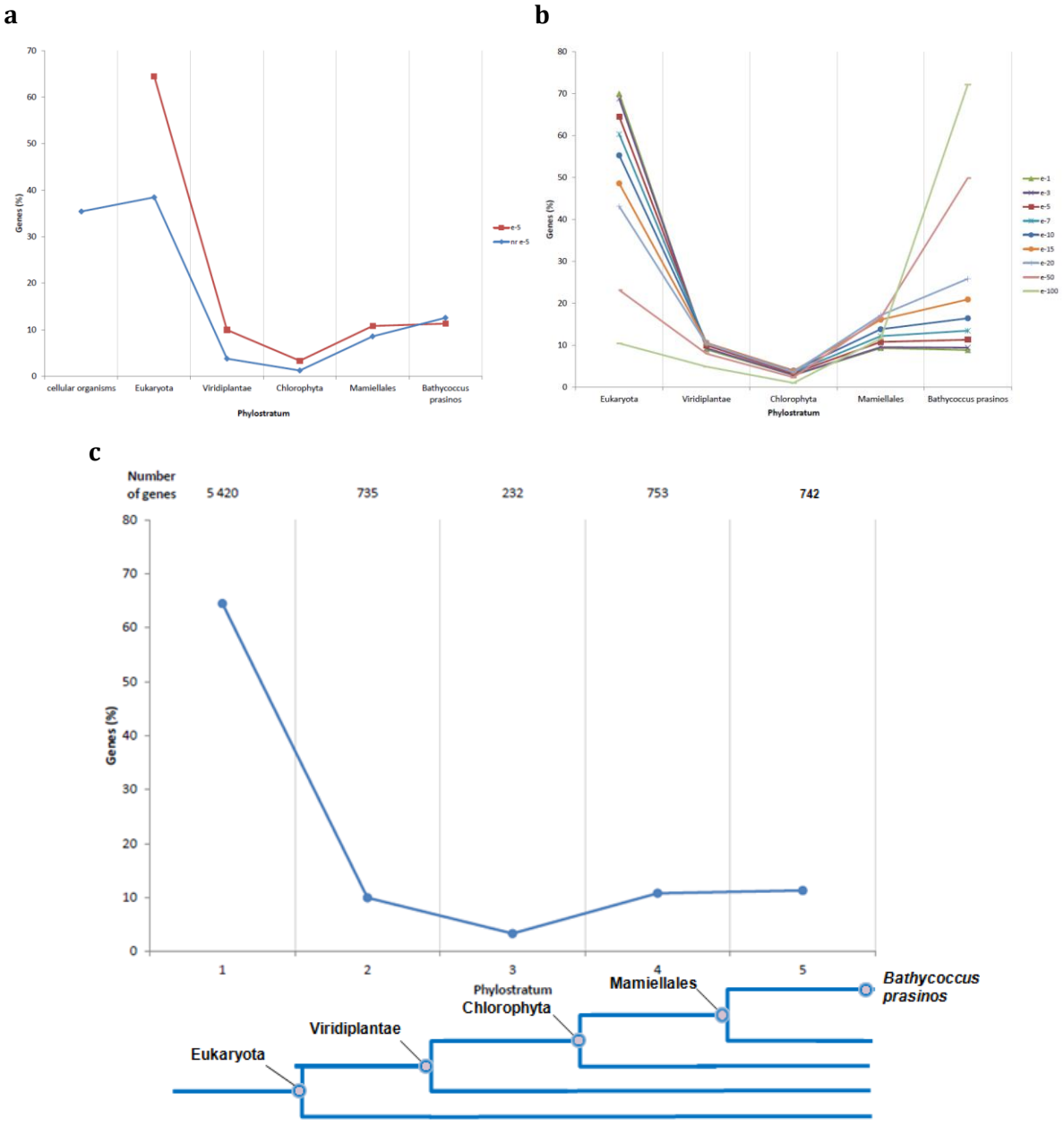


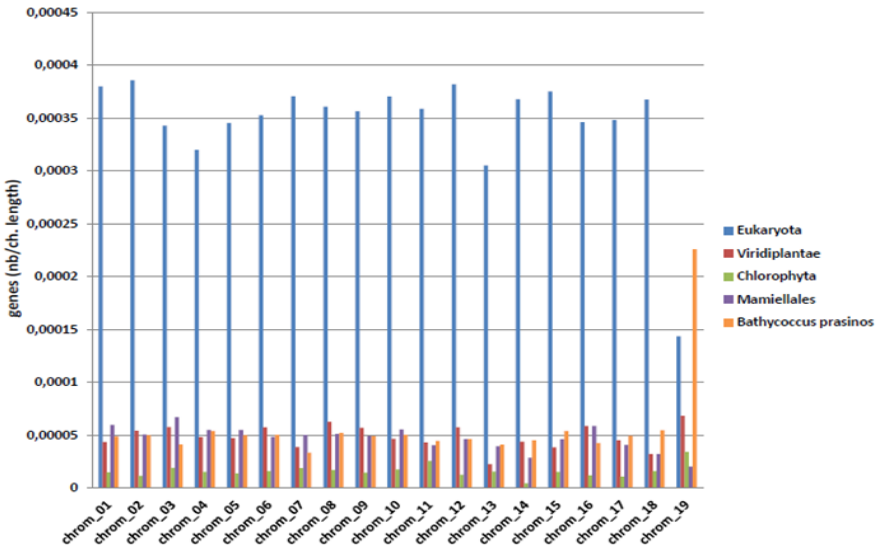
Figure I.3.2 | Évaluation de la banque de données et de la *e-value* à utiliser pour l'analyse phylostratigraphique de *Bathycoccus prasinos* RCC1105. (a), Phylostratigraphie de *Bathycoccus prasinos* en utilisant la banque de référence *nr* (courbe rouge) ou la base de données pico-PLAZA (courbe bleue). Le pourcentage de gène pour chaque phylostratum est représenté en abscisse. La base de données nr ne possédant pas uniquement les eucaryotes, le phylostratum *cellular organisms* est ajouté. (b), Phylostratigraphie de *Bathycoccus prasinos* avec différents seuils d'*e-value*. (c), Phylostratigraphie de *Bathycoccus prasinos* avec un seuil d'*e-value* inférieur à 10^{-3}.

I.3.3 Caractérisation structurale des gènes orphelins de *Bathycoccus prasinos*

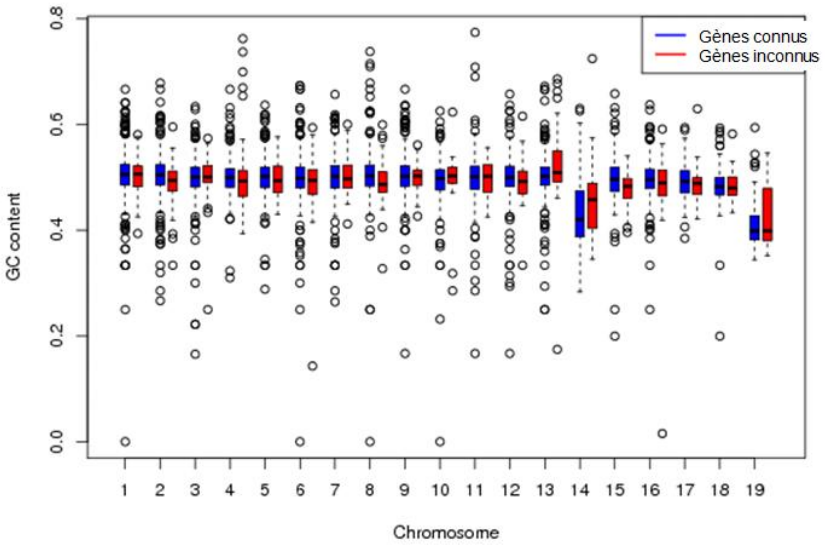
Après avoir sélectionné les gènes orphelins de la souche RCC1105, il est possible de réaliser la caractérisation structurale de ces derniers. Ainsi, la majorité des gènes inconnus ou orphelins est présente sur le chromosome 19 de *Bathycoccus* (figure I.3.3.a). Ce chromosome a été décrit comme étant un *outlier* du fait de ses différences structurales et fonctionnelles en comparaison aux autres chromosomes[68]. Le contenu en GC est légèrement plus faible pour les gènes inconnus (Mann-Whitney test p = 0.000885). De plus, ce contenu en GC est également plus faible pour les deux chromosomes *outlier* (figure I.3.3.b). Enfin, les gènes inconnus sont plus petits et présentent moins d'introns que les gènes connus (figure I.3.3.c). Ces distinctions sur la structure des gènes inconnus avaient déjà été présentées auparavant[302,312–315]. Pour vérifier que ces gènes inconnus ne sont pas de fausses prédictions, leur niveau d'expression a été calculé dans les échantillons où la souche RCC1105 est présente. Pour cela, la normalisation RPKM (*Reads Per Kilobase per Million*) a été réalisée. Celle-ci combine une normalisation inter et intra-échantillons afin de corriger les comptages pour prendre en compte la taille de la librairie et la longueur des gènes. Ici, le calcul correspond donc au nombre de lectures qui s'alignent sur le gène étudié, divisé par le nombre total de lectures séquencées dans l'échantillon multiplié par la taille du gène. Dans ces échantillons, la valeur moyenne du RPKM métatranscriptomique des gènes connus (0.56 avec un écart moyen de 0.75) est équivalente à celle des gènes inconnus (0.51 avec un écart moyen de 0.68), ce qui valide leur existence. Enfin, il y a plus de gènes dispensables dans les gènes orphelins (4.4% des gènes orphelins) que dans les gènes connus (0.4% des gènes connus) de *Bathycoccus prasinos*. Il a été présenté dans l'article 1 (chapitre I.1) que l'expression des gènes dispensables dans les stations où ils sont présents est plus importante que les autres gènes. Il y a donc un gain ou une perte de ses gènes orphelins qui ont une fonction en lien direct avec l'environnement. Cela pourrait supposer une contrainte évolutive plus importante de ces gènes.

Il a été montré auparavant que les gènes orphelins sont perdus à un taux plus important que les gènes non orphelins. L'hypothèse d'un turnover rapide de ces gènes avait été posée[307,314] et cela serait le reflet d'une exigence fonctionnelle spécifique à une lignée[312]. Il est donc possible que certains gènes, à la fois orphelins et dispensables chez *Bathycoccus prasinos*, aient une fonction nécessaire à sa présence dans un environnement donné.

a



b



c

	Gènes inconnus	Gènes connus
Nombre de gènes	742	7128
Nombre de gène sans intron	687 (93%)	6015 (84%)
Taille des gènes en nt. (moy. : med.)	1125.75 : 906	1657.42 : 1317
Nombre d'exons /gene (moy. : med.)	1.08 : 1	1.20 : 1
Taille des CDS en nt. (moy. : med.)	1115.58 : 885	1624.02 : 1287
Nombre d'introns	58	1456
Taille des introns en nt. (moy. : med.)	130.05 : 75	163.49 : 133
GC content (moy. : med.)	0.490 : 0.493	0.493 : 0.498
GC1 content (moy. : med.)	0.501 : 0.505	0.496 : 0.500
GC2 content (moy. : med.)	0.412 : 0.406	0.418 : 0.412
GC3 content (moy. : med.)	0.557 : 0.564	0.565 : 0.578

Figure I.3.3 | Comparaison de la position chromosomique et de la structure des gènes inconnus (ou orphelins) et des gènes connus de *Bathycoccus prasinos* RCC1105. a, Provenance chromosomique des gènes présents dans chaque phylostratum. Le nombre de gènes a été normalisé par la taille du chromosome. **b**, Pourcentage du contenu en GC des gènes connus (boite bleue) et inconnus (boite rouge) sur chaque chromosome. **c**, Tableau récapitulatif de la caractérisation structurale des gènes inconnus et connus de RCC1105.

I.4. Étude de génomique comparative des mamiellales

L'étude comparative de la biogéographie et du profil phylostratigraphique d'organismes phylogénétiquement proches permet d'analyser les différences et les similitudes de localisations géographiques ainsi que l'origine phylogénétique de leurs gènes. Pour cela les trois genres de mamiellales, *Ostreococcus sp.* , *Bathycoccus sp.* et *Micromonas sp.* se trouvent être de bons organismes d'étude car leurs génomes sont disponibles dans les bases de données et ils sont décrits comme étant abondants dans les océans. Leurs génomes et annotations structurales ont été récupérés sur le site de l'institut Ghent (http://bioinformatics.psb.ugent.be/). Ainsi, la phylostratigraphie et la biogéographie ont été réalisées à partir des souches *Micromonas pusilla* CCMP1545[62], *Micromonas sp.* RCC299, *Ostreococcus lucimarinus* CCE9901[59], *Ostreococcus* RCC809, *Ostreococcus tauri*[61] et *Bathycoccus prasinos* RCC1105[68].

I.4.1 Phylostratigraphie des mamiellales

Le profil phylostratigraphique des différents mamiellales présente des similitudes et des différences sur le pourcentage de gènes présents dans un phylostratum (figure I.4.1). Ainsi, pour chaque souche, le pourcentage de gènes présent dans les phylostratum viridiplantae et chlorophyta est à peu près identique. Ces gènes apparus au même moment dans l'évolution des mamiellales ne semblent pas être soumis à pression de sélection et sont conservés chez l'ensemble des mamiellales. Il y a des variations dans le nombre de gènes présents dans le phylostratum mamiellales et celui de l'espèce étudié. Cette variabilité dans l'apparition ou dans la perte de gènes dans ces génomes illustre la différence de la dynamique d'évolution du contenu en gène des mamiellales. Certains de ces gènes sont sans doute nécessaires au développement d'une espèce dans son environnement comme cela est le cas de la flavodoxine. Dans le cadre de l'étude des mamiellales, la famille de gènes des glycosyltransferases illustre bien cette variabilité dans l'évolution du contenu en gènes. En effet, *Bathycoccus prasinos* présente un *scale* qui entoure sa cellule. Ce *scale* a été perdu au cours de l'évolution chez *Micromonas* et *Ostreococcus*[68]. Les gènes codants pour la sialyltransferase sont impliqués dans la formation de ce *scale*. Ces gènes qui correspondent à la famille 29 des glycosyltransferases sont au nombre de 41 chez *Bathycoccus prasinos*. 40 de ces gènes sont apparus au phylostratum *Viridiplantae* et 1 est spécifique au phylostratum *Bathycoccus prasinos*. Ces gènes nécessaires à la formation du *scale* chez *Bathycoccus* ont été perdus chez *Micromonas* et *Ostreococcus*. L'environnement biotique ou physico-chimique de ces organismes est peut être favorable, levant la pression évolutive sur le maintien de cette structure et sur ses gènes. La comparaison de la biogéographie de ces différentes espèces de mamiellales est possible.

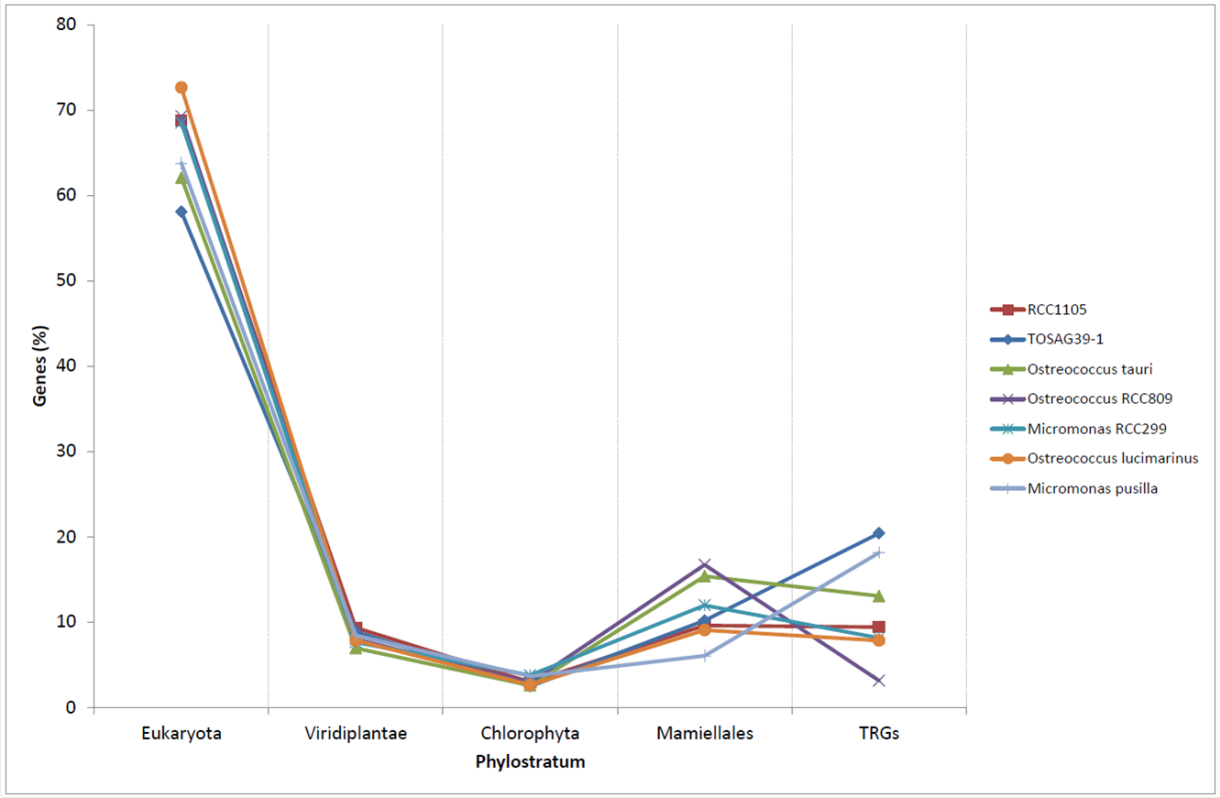


Figure I.4.1 | Phylostratigraphie des trois genres de mamiellales. Les gènes des souches *Micromonas pusilla* CCMP1545, *Micromonas sp.* RCC299, *Ostreococcus lucimarinus* CCE9901, *Ostreococcus* RCC809, *Ostreococcus tauri, Bathycoccus prasinos* RCC1105 ont été alignés avec blast+ sur les gènes des organismes présents sur la base de données pico-PLAZA.

I.4.2 Biogéographie des mamiellales.

Ces différentes espèces de mamiellales présentent des localisations distinctes dans les océans (figure I.4.2). *Micromonas pusilla* est retrouvé dans quelques échantillons de surface (*upwelling* du Chili, début du gyre de l'Atlantique Nord et dans le détroit de Gibraltar) alors que *Micromonas* RCC299 est retrouvé en surface et en DCM dans des stations similaires au SAG de *Bathycoccus* (TOSAG39-1). *Ostreococcus* RCC809 est abondant dans les eaux froides (les trois *upwelling*s et station au large du Chili) alors qu'*Ostreococcus lucimarinus* est présent dans des eaux plus tempérées (océan Indien, Pacifique équatorial et début du gyre de l'Atlantique Nord). Ces deux espèces *d'Ostreococcus,* comme les deux espèces de *Bathycoccus*, sont dans des environnements différents. *Ostreococcus tauri* n'est pas retrouvé dans les océans, ce qui est attendu puisque celui-ci provient de l'étang de Thau. Ainsi, ces différences de localisation entre ces espèces de mamiellales peuvent être une piste pour expliquer l'évolution divergente du contenu en gène observée dans les différents phylostratum. Une comparaison plus fine sur ces différences biogéographiques et des préférences physico-chimiques des mamiellales est prévue pour de futures analyses.

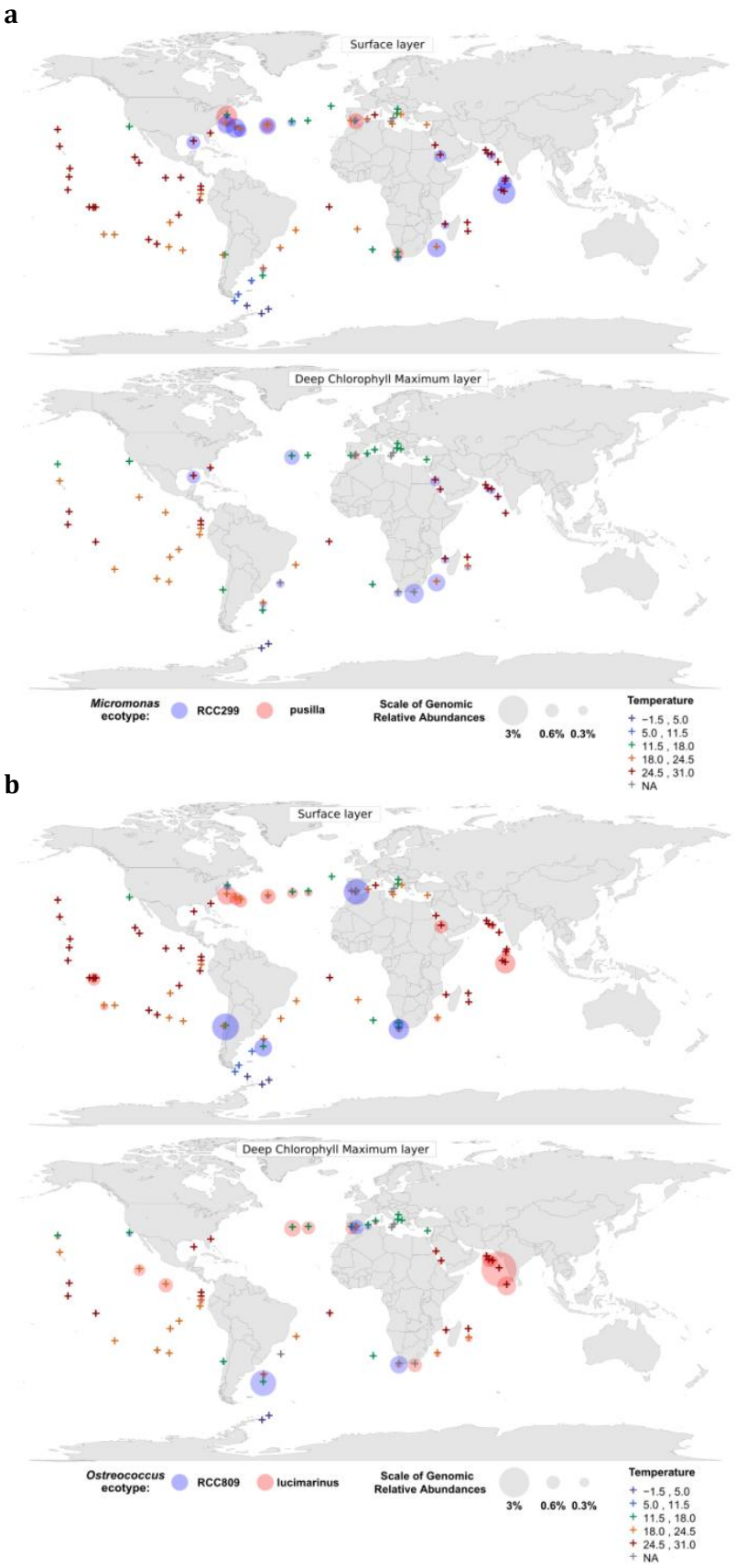


Figure I.4.2 | Biogéographie des mamiellales à partir des données métagénomiques. a, *Micromonas pusilla* CCMP1545 (cercles rouge) et *Micromonas sp.* RCC299 (cercles bleu). **b,** *Ostreococcus lucimarinus* CCE9901 (cercles rouge) et *Ostreococcus* RCC809 (cercles bleu).

Conclusion du chapitre I

La caractérisation de la biogéographie et de la plasticité génétique de micro-organismes planctoniques a permis de montrer la capacité des données métagénomiques générées par le projet *Tara* Oceans à étudier la diversité d'une espèce ainsi que de ses gènes. De plus, les conditions physico-chimiques semblent avoir un impact sur cette dynamique de la structure des génomes et de leur biogéographie dans l'océan. Cependant, une étude de métagénomique comparative est nécessaire pour passer de l'échelle d'un génome à un grand ensemble représentatif des génomes qui compose une communauté planctonique afin d'étudier l'organisation de cet écosystème planctonique dans les océans.

Chapitre II

Évaluation et amélioration d'outils permettant d'étudier l'organisation génomique des communautés virales, bactériennes et eucaryotes à l'échelle de la planète à partir des données métagénomiques

Il a été présenté dans la partie III.5 que la métagénomique comparative permettait de comparer la totalité du matériel génétique généré par le séquençage d'échantillons métagénomiques. L'outil *Compareads* capable de réaliser une comparaison *de novo* de métagénomes a été évalué sur une partie des échantillons *Tara* Oceans. Son amélioration a permis de fournir un outil qui rend possible l'étude de l'organisation génomique des communautés microplantoniques sur l'ensemble des échantillons du projet *Tara* Oceans.

II.1 Article 2: Environmental characteristics of Agulhas rings affect interocean plankton transport

Compareads[285] est un outil conçu pour la métagénomique comparative *de novo*. Il permet de détecter les lectures similaires entre deux échantillons métagénomiques en utilisant l'information des *k-mers*. Cet outil a été évalué sur un petit nombre d'échantillons du projet *Tara* Oceans. Dans le cadre d'un projet collaboratif, 11 stations provenant de trois océans (Indien, Atlantique Sud et Austral) sur quatre fractions de taille (0.22 à 3 µm, 0.8 à 5 µm, 20 à 180 µm et 180 à 2000 µm) aux profondeurs surface et DCM ont été utilisées. La comparaison du matériel génétique de l'ensemble des fractions de taille d'organisme partagé par ces trois bassins a pu être réalisée.

Ainsi, la similarité des échantillons à l'intérieur d'un même océan (intra-océan) est plus forte que celle des échantillons issus de deux océans (inter-océans). Les échantillons de l'océan Austral ont peu de lectures partagées avec les échantillons des océans Atlantique Sud et Indien. Il est probable que la différence des conditions environnementales de l'océan Austral comparée aux deux autres océans fait qu'une grande partie des micro-organismes planctoniques présents dans cet océan ne sont pas les mêmes que ceux présents dans les océans Atlantique Sud et Indien. De plus, le pourcentage de lectures similaires entre les échantillons de l'océan Indien et

de l'océan Atlantique Sud est proche du pourcentage de lectures partagées entre les échantillons intra-océan (dans l'Atlantique Sud et dans l'Indien). Ces deux océans partageraient des communautés microplanctoniques similaires. En effet, en analysant la dissymétrie des matrices de similarité de chaque taille de filtre, on observe qu'une partie des séquences de l'océan Atlantique est similaire à des séquences de l'océan Indien, mais une moindre partie de séquences de l'océan Indien est similaire à des séquences de l'océan Atlantique Sud. Cela implique qu'une partie des séquences génomiques des micro-organismes planctoniques est retrouvée dans l'océan Atlantique Sud, l'inverse est moins fréquent. Cela peut s'expliquer par la connexion de ces deux océans avec les anneaux d'Agulhas. Ces anneaux se formant dans l'océan Indien piègeraient le plancton de cet océan et le transporteraient dans l'océan Atlantique Sud. Des échantillonnages au niveau d'un anneau au début du gyre de l'Atlantique Sud 9 mois après sa formation, puis d'un autre anneau au niveau des côtes brésiliennes 3 ans après sa formation ont été réalisés. Le pourcentage de lectures que partage l'échantillon prélevé dans le « jeune » anneau est plus important avec l'océan Indien qu'avec l'océan Atlantique. Il en est de même pour l'échantillon prélevé au niveau de l'anneau formé il y a trois ans. Ces similarités de séquences entre ces échantillons métagénomiques sont en adéquation avec l'explication du captage et du transport des micro-organismes planctoniques par les anneaux d'Agulhas. Il a également été observé qu'un échantillon dans l'océan Atlantique Sud présente très peu de similarité de lecture avec l'ensemble des autres échantillons. En effet, celui-ci a été prélevé dans une zone d'upwelling qui contient une communauté en microplanctons différente des autres échantillons. Enfin, les mêmes analyses ont été effectuées avec les données V9 de l'ARNr et des résultats identiques ont été observés via l'étude de la similarité de barcodes partagés entre les échantillons de ces trois bassins. Ces analyses valident la capacité de cette méthode de métagénomique comparative *de novo* à réaliser des analyses sur l'organisation génomique des communautés planctoniques.

Ces résultats ont été intégrés dans une publication parue le 22 Mai 2015 dans la revue *Science* présentée ci-après. Les informations supplémentaires concernant l'analyse sur ces connexions génétiques entre les bassins se trouvent en annexe II.

Environmental characteristics of Agulhas rings affect interocean plankton transport

Emilie Villar, Gregory K. Farrant, Michael Follows, Laurence Garczarek, Sabrina Speich, Stéphane Audic, Lucie Bittner, Bruno Blanke, Jennifer R. Brum, Christophe Brunet, Raffaella Casotti, Alison Chase, John R. Dolan, Fabrizio d'Ortenzio, Jean-Pierre Gattuso, Nicolas Grima, Lionel Guidi, Christopher N. Hill, Oliver Jahn, Jean-Louis Jamet, Hervé Le Goff, Cyrille Lepoivre, Shruti Malviya, Eric Pelletier, Jean-Baptiste Romagnan, Simon Roux, Sébastien Santini, Eleonora Scalco, Sarah M. Schwenck, Atsuko Tanaka, Pierre Testor, Thomas Vannier, Flora Vincent, Adriana Zingone, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans Coordinators, Silvia G. Acinas, Peer Bork, Emmanuel Boss, Colomban de Vargas, Gabriel Gorsky, Hiroyuki Ogata, Stéphane Pesant, Matthew B. Sullivan, Shinichi Sunagawa, Patrick Wincker, Eric Karsenti, Chris Bowler, Fabrice Not, Pascal Hingamp and Daniele Iudicone (May 21, 2015)
Science **348** (6237), . [doi: 10.1126/science.1261447]

Editor's Summary

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

II.2 Article 3: *COMMET*: comparing and combining multiple metagenomic datasets.

Compareads ne permettant pas de réaliser les comparaisons sur l'ensemble des échantillons du projet *Tara* Oceans, des améliorations de cet outil ont été nécessaires. Un travail en collaboration avec l'équipe GenScale à l'INRIA de Rennes dans le cadre d'un projet ANR, a permis le développement de ces améliorations.

COMMET (COmpare Multiple METagenomes) est implémenté en C++ contrairement à *Compareads* qui est codé en C. L'optimisation de l'utilisation de la mémoire vive permet d'accélérer le temps de calcul de la similarité génomique entre les échantillons métagénomiques (au moins 2 fois plus rapide) tout en gardant le même algorithme et donc la même sensibilité que *Compareads*. L'utilisation de vecteurs binaires pour indiquer si une lecture est présente ou non dans deux jeux de données permet de générer des fichiers de sortie peu volumineux ce qui divise par 100 l'impact sur le disque dur. L'utilisation d'opérations booléennes entre les résultats des intersections d'échantillons permet également de retrouver facilement et rapidement les séquences d'un échantillon qui sont présentes ou non dans les intersections de plusieurs autres échantillons. Différents modules ont été créés pour manipuler les vecteurs de bits représentant les lectures similaires à deux échantillons. Un premier module (*filter_reads*) permet de *réaliser* l'étape de filtrage et de sous échantillonnage des lectures des jeux métagénomiques. Un second module (*index_and_search*) permet de réaliser la comparaison de similarité génomique entre plusieurs jeux de données métagénomiques comme le fait *Compareads* mais en utilisant la représentation en vecteurs de binaires. Un troisième module (*bvop*) permet de réaliser les opérations pour récupérer les lectures dans, ou en dehors des intersections des différents échantillons comparés, toujours via l'utilisation des vecteurs binaires. Le dernier module (*extract_reads*) permet de récupérer les séquences brutes des lectures en sortie des trois premiers modules. Un script python qui implémente ces quatre modules a été développé pour permettre une utilisation simple de *COMMET* et permet de réaliser l'ensemble des comparaisons en une seule exécution. De plus, ce script génère plusieurs sorties graphiques dont les heatmap ainsi que des dendrogrammes représentant la similarité génomique entre les échantillons comparés.

L'outil *COMMET* a fait l'objet d'une publication dans un article de conférence (IEEE BIBM 2014 à Belfast) présenté ci-après.

COMMET: comparing and combining multiple metagenomic datasets

Nicolas Maillet, Guillaume Collet, Thomas Vannier, Dominique Lavenier, Pierre Peterlongo

COMMET: comparing and combining multiple metagenomic datasets

Nicolas Maillet*, Guillaume Collet†, Thomas Vannier‡, Dominique Lavenier* and Pierre Peterlongo*⁕
*INRIA / IRISA-UMR CNRS 6074, EPI GenScale, Rennes, France
†INRIA / IRISA-UMR CNRS 6074, EPI Dyliss, Rennes, France
‡CEA Genoscope / CNRS UMR 8030 / Université d'Évry, Evry, France
⁕ Corresponding author: Pierre Peterlongo pierre.peterlongo@inria.fr

***Abstract*—Metagenomics offers a way to analyze biotopes at the genomic level and to reach functional and taxonomical conclusions. The bio-analyzes of large metagenomic projects face critical limitations: complex metagenomes cannot be assembled and the taxonomical or functional annotations are much smaller than the real biological diversity. This motivated the development of *de novo* metagenomic read comparison approaches to extract information contained in metagenomic datasets.**

However, these new approaches do not scale up large metagenomic projects, or generate an important number of large intermediate and result files. We introduce COMMET ("COmpare Multiple METagenomes"), a method that provides similarity overview between all datasets of large metagenomic projects.

Directly from non-assembled reads, all against all comparisons are performed through an efficient indexing strategy. Then, results are stored as bit vectors, a compressed representation of read files, that can be used to further combine read subsets by common logical operations. Finally, COMMET computes a clusterization of metagenomic datasets, which is visualized by dendrogram and heatmaps.

Availability: http://github.com/pierrepeterlongo/commet

I. INTRODUCTION

NGS revolution enabled the emergence of the metagenomic field where an environment is sequenced instead of an individual or a species, opening the way to a comprehensive understanding of environmental microbial communities. Large metagenomic projects such as MetaSoil [1], MetaHit [2] or Tara Oceans [3] witness this evolution. Analyzing of metagenomic data is a major bottleneck. For instance, assembly tests over "simple" simulated metagenomes showed that N50 is only slightly larger than read sizes [4]. This situation becomes even worth on complex datasets, such as seawater, where millions of distinct species coexist. In this case, biodiversity can be estimated by using statistical approaches [5] or by mapping reads on reference banks [6], [7]. Nevertheless, statistical approaches are limited to a few dozens of species with limited differences in their relative abundance. In addition, the mapping approaches are limited to current knowledge contained in reference banks that suffer from their incompleteness and their inherent errors [8].

A key point of substantial metagenomic projects stands in the number of metagenomes they produce. Then, similarities and differences between metagenomes can be exploited as a source of information, measuring external effects like pollution sources, geographic locations, and patient microbial gut environment [2], [9]. A few methods were proposed to compare metagenomes using external information sources such as taxonomic diversity [10] or functional content [11]. However, these methods are biased because as they are based on partial knowledge.

Methods were proposed to compare metagenomes without using any *a priori* knowledge. These *de novo* methods use global features like GC content [12], genome size [13] or sequence signatures [14]. Theses methods face limitations as they are based on rough imprecise criteria and as they only compute a similarity distance: they do not extract similar elements between samples. We believe that it is possible to go further by comparing metagenomic samples at the read sequence level. This provides a higher precision distance and, importantly, it provides reads that are similar between datasets or that are specific to a unique dataset, enabling their latter analysis: assembly with better coverage or comparison with other metagenomic samples. Such comparisons may be performed using Blast [15] or Blat [16] like tools. Unfortunately, these methods do not scale up on large comparative metagenomic studies in which hundreds of millions of reads have to be compared to other hundreds of millions of reads. For instance, one can estimate that comparing a hundred of metagenomes each composed by a hundred of millions of reads of size 100 would require centuries of CPU computation. The crAss approach [17] constructs a reference metagenome by cross assembling reads of all samples. Then, it maps the initial reads on the so obtained contigs and several measures are derived, based on the repartition of mapped reads. This method provides results of high quality. However, due to its assembly and mapping approach, it does not scale up to large metagenomic datasets. Simpler methods such as TriageTools [18] or Compareads [19] measure the sequence similarity of a read with a databank by counting the number of k-mers (words of length k) shared with the databank. Due to memory consumption, TriageTools cannot use k values larger than 15 and is thus limited to small datasets (a few hundred of thousands reads of length 100). The Compareads tool scales up to large datasets with a small memory footprint and acceptable running time. However, applied on large metagenomic projects, this tools generates an important number of large intermediate result files. In practice, applying Compareads to N datasets generates N^2 resulting new datasets, each of the size of the original ones at worse. Additionally, Compareads leads to highly redundant computation raising up the execution time. These drawbacks are serious bottlenecks limiting the practical usage of Compareads.

In this paper, we introduce COMMET ("COmpare Multiple METagenomes"), a fast software that provides a global similarity overview between all datasets of a metagenomic project. COMMET is based on the Compareads philosophy that consists in determining similarity between two metagenomic datasets by extracting common reads using k-mer approach: two reads are considered similar if they share t non-overlapping k-mers (t and k are parameters). A metagenomic project involving N datasets will thus requires the computation of N^2 intersections which is both time- and storage-consuming. To keep computation time as low as possible, the computation of the N^2 intersections has been strongly improved compared to the Compareads approach through an efficient indexing strategy in which each file is fully indexed only once. In addition, to save storage space, intersections between metagenomic datasets are represented as bit vectors. This compact representation reduces the storage space by two orders of magnitude. Moreover, it provides an easy way to filter and sub-sample reads, or to combine various results by applying logical operations. Finally, COMMET computes a clusterization of metagenomic datasets, which is visualized by dendrogram and heatmaps.

II. METHOD

A. Comparing two sets of reads

The COMMET algorithm to compare two sets of read is based on the Compareads [19] methodology. It consists in finding reads from a set A that are similar to at least one read from a set B. The similarity between two reads is based on a minimal number t of non-overlapping identical k-mers. This core operation is directed : it provides reads from A similar to reads from B but it does not provides reads from B similar to reads from A. Note that, as explained below, this operation is based on a heuristic. Thus we denote this operation by $A \overrightarrow{\cap} B$.

Computing $A \overrightarrow{\cap} B$ consists in two steps. Firstly, k-mers from B are indexed in a Bloom filter like data-structure [20]. Secondly, non-overlapping k-mers of reads from A are searched in the Bloom filter. A read r from A sharing t non-overlapping k-mers with the Bloom filter is considered similar to at least one read from B. However, the algorithm does not check that these k-mers co-occur on a single read from B, which is a source of false positives. Readers are invited to refer to [19] for having more details on precision results.

We recall that the following strategy is applied in order to limit the second source of false positives. First $A \overrightarrow{\cap} B$ is computed. Then, instead of navely computing $B \overrightarrow{\cap} A$, $B \overrightarrow{\cap} (A \overrightarrow{\cap} B)$ is computed. This limits the indexed reads of A to those already detected as similar to at least one read from B. Finally, the symmetrical operation is performed: $A \overrightarrow{\cap} (B \overrightarrow{\cap} (A \overrightarrow{\cap} B))$.

The previously exposed strategy to fully compare sets A and B within three consecutive $\overrightarrow{\cap}$ operations has also the advantage to limit the indexation effort. Indeed, only the first $A \overrightarrow{\cap} B$ operation indexes the full set B. The two other operations only index subsets of A and B.

While comparing read samples A and B, the final results of interest are the reads of A similar to reads of B computed by $A \overrightarrow{\cap} (B \overrightarrow{\cap} (A \overrightarrow{\cap} B))$ and reads of B similar to reads of A computed by $B \overrightarrow{\cap} (A \overrightarrow{\cap} B)$. For sake of simplicity, we denotes these two sets as, respectively, $A \overrightarrow{\Cap} B$ and $B \overrightarrow{\Cap} A$.

In the following sections we present the COMMET novelties: represent read subsets with a limited disk space impact, new read filtering and read subsets manipulation features, compare multiple sets of reads, visualize dataset's similarities as heatmaps and dendrogram.

B. Read subsets representation

In COMMET we propose a simple yet compact data structure to represent a read subset: a vector of bits where each bit represents a read of the original read set. This is what we call the "bit vector representation". As shown below, this representation enables to filter and to subsample read files, to represent $\overrightarrow{\cap}$ (and thus $\overrightarrow{\Cap}$) results and to easily perform logical operation between read subsets.

Note that with such a representation, a bit vector needs hundreds to thousand times less disk space than a classical uncompressed fastq file. Note also that this way of coding read subsets is not limited to the COMMET framework. It may be applied to any other programs that manipulate read subsets. Thus, the COMMET tool includes a C++ library of reusable components to manipulate read subsets.

In the COMMET framework, the bit vector representation is used as inputs and/or outputs of all tools. In particular they are used in the following operations:

1) Read subsampling and filtering: With huge datasets, it may appear necessary to subsample, for instance limiting each read file to a same number m of a few millions reads. This is immediate by creating a bit vector in which only the first m bits are set to 1, while others are set to 0.

Raw NGS reads also usually need to be filtered on several practical characteristics (read size, read complexity, ...). Thus, a bit vector is a direct representation of a filtered result: bit values associated to selected reads are set to 1, the others to 0. A combination of subsampling and filtering allows to select only the m first reads that fulfill the filtration criteria.

2) Representing the similar reads: Results of any $\overrightarrow{\cap}$ operation is represented by a bit vector. Bit values of reads from the query set detected as similar to at least one read from the reference set are set to 1 and the others are set to 0.

3) Compute logical operations on read subsets: The bit vector representation is ideally suited to perform fundamental logical operations. COMMET provides a module to perform the AND, OR and NOT operations between distinct subsets of a single initial set of reads.

As presented in the simple case study (Section II-F), these operations, although simple, are powerful while dealing with read subsets. They allow to combine comparison results and so to focus on read subsets intersections or exclusions.

These logical operations perform very efficiently, both in terms of execution time and memory footprint. Moreover, it is worth to notice that they do not generate large result files, as results of these logical operations are also represented as bit vectors. This allows to intensively manipulate read subsets with no technical limitations.

C. Dealing with more than two datasets

We recall that the computation of the $A \overrightarrow{\cap} B$ core operation involves indexation and search. Once the k-mers of the reads from B are indexed, then the k-mers of the reads from A are sequentially search in the index. If more than a threshold number t of such k-mers are find in the index, then the given read from A is considered as similar to a read from B, which means that the associated value in the bit vector is set to 1.

Consider $S = \{R_1, \ldots, R_N\}$ a set of $N \geq 2$ read sets. Applying COMMET on the whole S implies that $\forall (i,j) \in [1,N]^2, i < j$, three ordered operations are performed:

1) $R_i \overrightarrow{\cap} R_j$
2) $R_j \overrightarrow{\overrightarrow{\cap}} R_i = R_j \overrightarrow{\cap} (R_i \overrightarrow{\cap} R_j)$
3) $R_i \overrightarrow{\overrightarrow{\cap}} R_j = R_i \overrightarrow{\cap} (R_j \overrightarrow{\cap} (R_i \overrightarrow{\cap} R_j))$

Note that for each couple (i,j), the order (i,j) or (j,i) only slightly changes the overall results of the three operations. To avoid redundancies, we limit these operations to $i < j$.

1) Factorizing the indexation: In practice, applying COMMET on S implies to perform the $R_i \overrightarrow{\cap} R_j$ operations for all $i < j$. In particular, $R_1 \overrightarrow{\cap} R_N \ldots R_{N-1} \overrightarrow{\cap} R_N$ have to be computed. For these $N-1$ computations, the k-mer index of R_N is the same. To avoid redundancies, the R_N index is computed only once and the $N-1$ remaining sets are compared to R_N using this single index. In general, while R_{ref} ($ref \in [2,N]$) is indexed, the index is conserved in RAM memory during the computation of the $ref - 1$ comparisons $R_{query} \overrightarrow{\cap} R_{ref}$, with $query < ref$.

2) Results visualization: Comparisons of $N \geq 2$ read sets $\{R_1, \ldots, R_N\}$ provide useful metrics that give an overview of the genomic diversity of the studied samples. Those metrics are summarized in three matrices M_1, M_2, and M_3 with values calculated as follows:

- $M_1(i,j) = |R_i \overrightarrow{\overrightarrow{\cap}} R_j|$
- $M_2(i,j) = 100 \times \frac{|R_i \overrightarrow{\overrightarrow{\cap}} R_j|}{|R_i|}$
- $M_3(i,j) = 100 \times \frac{|R_i \overrightarrow{\overrightarrow{\cap}} R_j| + |R_j \overrightarrow{\overrightarrow{\cap}} R_i|}{|R_i| + |R_j|}$.

$M_1(i,j)$ with $(i,j) \in [1,N]^2$, is the raw number of reads from R_i that are similar to at least one read from R_j. As read sets may be of different sizes, $M_2(i,j)$ is the percentage of reads from R_i similar to at least one read from R_j. Those two first matrices are asymmetrical. M_3 is a symmetrical matrix. $M_3(i,j)$ is the percentage of similar reads between the two sets with respect to the total number of reads in R_i and R_j.

For each matrix, a heatmap is generated . Additionally, M_3 is used to construct a dendrogram representation by hierarchical clustering (see Fig 2 for an example of a heatmap and a dendrogram generated by COMMET).

D. The COMMET modules

COMMET integrates four independent modules written in C++, all manipulating, as inputs and outputs, the bit vector representation of read subsets. Additionally, COMMET provides a python script (Commet.py) that takes $N \geq 2$ read sets, filters them, compares them and generates explicit representations of comparative results, see Section II-E.

1) Filtering and subsampling reads: Thanks to the first module, $filter\_reads$, each read of each dataset (fasta or fastq format, gzipped or not) is filtered out according to user-defined criteria: minimal read length, number of undefined bases, and Shannon complexity [21], used to remove low complexity sequences. The result is a bit vector for each input read file. $Filter\_reads$ can also subsample each read set by limiting the number of selected reads to a user defined parameter m. The m first reads that passed the filters are selected.

2) Performing the $\overrightarrow{\cap}$ core operation: The second module, $index\_and\_search$, performs the $\overrightarrow{\cap}$ core operation, representing results using the bit vector representation. It inputs a set of read sets (the queries) to be searched in an indexed read set (the bank). A read set may be composed of several read files. Each file could be associated to a bit vector. In this latter case, $index\_and\_search$ only considers reads whose associated bit values are set to 1.

3) Manipulating read subsets: The third module, $bvop$ (bit vector operations) inputs one or two bit vectors. In this second case, the two bit vectors should represent subsets of the same initial set. This module performs the NOT operation on a single bit vector, and the AND, OR, and $AND\ NOT$ operations on two bit vectors.

4) From bit vectors to read files: Given an original read file and its bit vector, the last module, $extract\_reads$, generates an explicit representation of any read subset.

E. Automatization for $N \geq 2$ read sets

COMMET includes a python script (Commet.py) which inputs $N \geq 2$ read sets. This pipeline i) filters reads, given user-defined parameters, ii) compares all-against-all read sets, and iii) outputs a user-friendly visualization of results. The outputs consist in the three matrices in csv format, their heatmaps and a dendrogram as described in Section II-C2. The dendrogram is realized using the *hclust R* function, computing a hierarchical complete clustering.

F. Combining read subsets use case

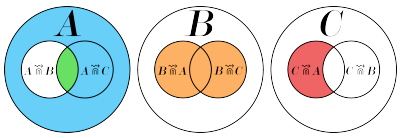


Fig. 1. Logical operations on intersections between A, B and C extract read subsets of interest. The blue subset corresponds to the $A\ AND\ NOT(A \overrightarrow{\overrightarrow{\cap}} B)$ operation. The green subset corresponds to the $(A \overrightarrow{\overrightarrow{\cap}} B)\ AND\ (A \overrightarrow{\overrightarrow{\cap}} C)$ operation. The orange subset corresponds to the $(B \overrightarrow{\overrightarrow{\cap}} A)\ OR\ (B \overrightarrow{\overrightarrow{\cap}} C)$ operation. The red subset corresponds to the $(C \overrightarrow{\overrightarrow{\cap}} A)\ AND\ NOT(C \overrightarrow{\overrightarrow{\cap}} B)$ operation.

By using the $bvop$ module, logical operations can be performed between inputs/outputs of the COMMET pipeline output. For instance, reads from A not similar to any read from set B (blue subset of Fig 1) are obtained by first applying $NOT(A \overrightarrow{\overrightarrow{\cap}} B)$ operation. Reads from A similar to at least one read from B and one read from C (green subset

TABLE I. 28 METAGENOMES FROM THE IMG/M DATABASE

Identifiers	Description
SWITGRA	Rhizosphere soil from *Panicum virgatum*
SUBGIN	Oral TM7 microbial community of Human
TERMITE2, TERMITE1	Gut microbiome of divers termites
SOILM, SOILD, SOILL	Soil microbiome from divers locations
OMIN, MESO, EUPHO	Divers marine planktonic communities
BEETLE	*Dendroctonus ponderosae*
ACOFUNT, ACOFUNB, CLOFUN, ACEFUN, TRAFUN, FUNCOMB	Fungus garden of divers ants
FUNTER	Fungus-growing termite worker
WALLABY	Forestomach microbiome of tammar wallaby
RICE	Endophytic microbiome from rice
SNAIL	*Achatina fulica*
SNOCT	*Sirex noctilio* microbiome
XALARV, XAAD	*Xyleborus affinis* microbiome (larvae, adult)
HGUT7, HGUT8	Human gut community
PANDA2, PANDA5	Wild panda gut microbiome

of Fig 1), are identified by computing the AND operation: $(A \overset{\rightsquigarrow}{\cap} B)\ AND\ (A \overset{\rightsquigarrow}{\cap} C)$. In the same spirit, reads from B similar to at least one read from A or one read from C (orange subset of Fig 1), are found by computing the OR operation: $(B \overset{\rightsquigarrow}{\cap} A)\ OR\ (B \overset{\rightsquigarrow}{\cap} C)$. Operations may be combined to obtain more complex results as, for instance, the red subset of Fig 1, representing reads from C similar to at least one read from A, but not similar to any read from B. This would be done by applying the $(C \overset{\rightsquigarrow}{\cap} A)\ AND\ NOT(C \overset{\rightsquigarrow}{\cap} B)$ operation.

III. RESULTS

A. COMMET *efficiently compares multiple metagenomes*

We tested COMMET on a set of 28 metagenomes from the IMG/M database [7] (see Table I). These 28 metagenomes were compared with options $k = 33$, $t = 2$ and $m = 10000$. Computations were done using COMMET (Commet.py) and Compareads (v1.3.1) on a 2.9 GHz Intel Core i7 processor with 8GB of RAM and a Solid-State Drive. COMMET calculated the 756 intersections in 35 minutes while Compareads took 81 minutes. In this experiment, COMMET is 2x faster than Compareads thanks to its indexing strategy (each file is fully indexed only once). The obtained dendrograms, shown in Figure 2, are biologically coherent. The different fungus samples are grouped together as well as soil samples, marine planktonic communities and insects. The two human gut microbiome samples are far from other species, as well as the two panda gut microbiome samples.

B. Metasoil study

The MetaSoil study focuses on untreated soils of Park Grass Experiment, Rothamsted Research, Hertfordshire, UK. One of the goals of this study is to assess the influence of depth, seasons and extraction procedure on the sequencing [22]. To achieve this, the 13 metagenomes from MetaSoil, two other soil metagenomes and a sea water metagenome, were compared at the functional level using MG-RAST [23]. This approach identified 835 functional subsystems present in at least one of those metagenomes. On Figure 3.a, samples were clustered using the relative number of reads associated with the 835 functions. This figure shows that the extraction procedure correlates with sample clusters: two metagenomic samples processed with the same extraction procedure share more similarities at the functional level than two samples processed with different extraction procedures [1].

This study was reproduced with COMMET on all available metagenomes. The generated bit vectors weigh 68MB while

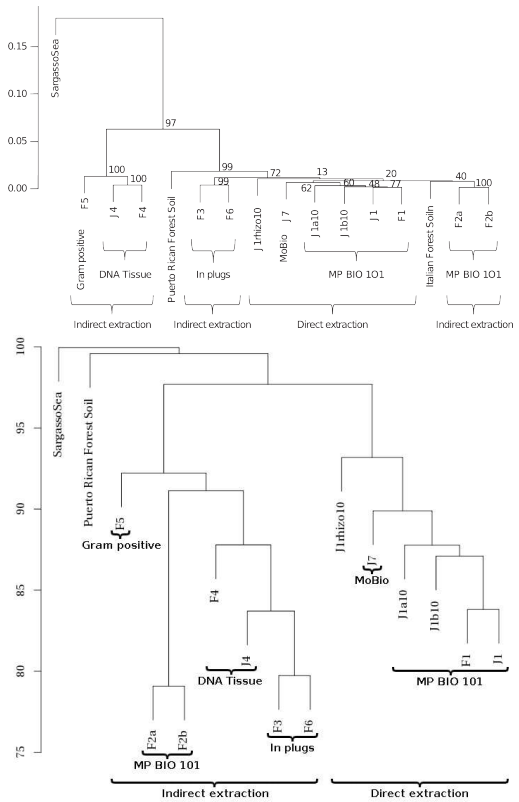


Fig. 3. Dendrograms from MetaSoil study (top, figure from [1]) and COMMET analysis (bottom), comparing the 13 MetaSoil samples, an other soil metagenome and a seawater metagenome (Sargasso Sea).

the explicit representation of the fasta results requires 6.4GB. The storage footprint is thus divided by a factor 100. This ratio is even higher if using fastq format or if dealing with larger read files. The COMMET computation time was 828 minutes (the same set treated by Compareads took 2981 minutes).

Although COMMET uses another metric, the produced dendrogram is highly similar to the MetaSoil one (see Fig 3). On both dendrograms, samples coming from direct extraction are clustered together and external metagenomes are far from the MetaSoil's. Moreover, on the COMMET dendrogram, all samples coming from indirect extraction are clustered together, which is not the case in the MetaSoil study. Even if the two comparing methods are different, they lead to the same conclusion: extraction procedures have a critical impact on sequencing.

IV. CONCLUSION

COMMET gives a global similarity overview of all datasets of a large metagenomic project. It performs all-against-all comparisons of N datasets by factorizing indexation phases. Disk I/Os and storage footprint are highly limited thanks to a new read subset representation which reduces the storage space by at least two orders of magnitude compare to explicit fasta or fastq format. Interestingly, this read subset representation is a powerful way to compute extremely fast boolean operations between read subsets without copying large read files. This enables to focus on reads that fulfill several distinct constraints of interest. The advantages of this representation and of

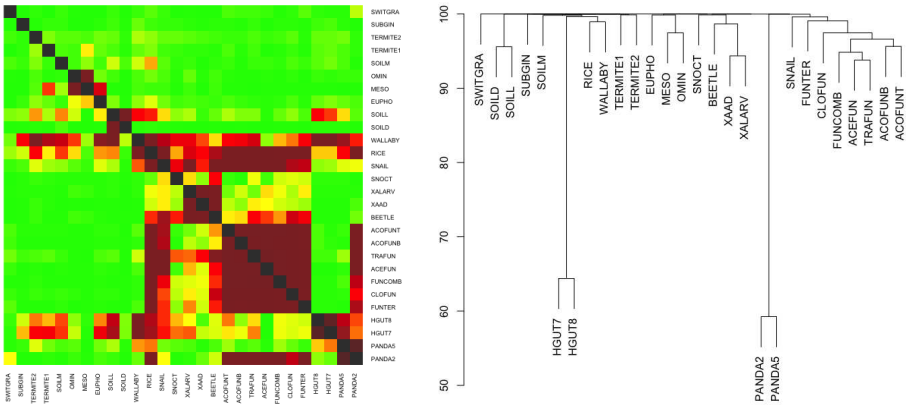


Fig. 2. Heatmap (left) and dendrogram (right) representation of the results of the comparison of 28 datasets from the IMG/M database. Results are given with $t = 2$, $m = 10000$ and $k = 33$. The heatmap is constructed from the matrix M_2 and is thus asymmetrical. The dendrogram is constructed from the matrix M_3 by the hierarchical clustering procedure available in *R* (method "complete").

the boolean manipulation are not limited to the COMMET framework. Thus, COMMET includes a C++ library of reusable components to manipulate read subsets.

COMMET produces graphical outputs that sum up all-against-all comparisons results and open the way for further statistical analysis, thanks to the provided similarity matrices.

A future work will consists in quickly identify significant clusters of read sets by applying rougher comparative metrics (such as the GC content) or a statistical framework based on Principal Component Analysis (PCA). Then, COMMET should be used to go further by precisely compute the shared reads between read sets inside clusters, or between clusters.

COMMET is available under the A-GPL license: http://github.com/pierrepeterlongo/commet.

ACKNOWLEDGMENT

Authors warmly thank Claire Lemaitre for her precious advices and her help designing the *R* functions. This work was supported by the french ANR-2010-COSI-004 *MAPPI* and by the ANR-12-BS02-0008 *Colibread* projects. Guillaume Collet's work is funded by the investment expenditure program IDEALG 1192 ANR-10-BTBR-02-04-11.

REFERENCES

[1] T. O. Delmont *et al.*, "Structure, fluctuation and magnitude of a natural grassland soil metagenome," pp. 1677–1687, 2012.

[2] J. Qin *et al.*, "A human gut microbial gene catalogue established by metagenomic sequencing," *Nature*, vol. 464, no. 7285, pp. 59–65, Mar. 2010. [Online]. Available: http://dx.doi.org/10.1038/nature08821

[3] E. Karsenti *et al.*, "A holistic approach to marine Eco-systems biology," *PLoS Biology*, vol. 9, 2011.

[4] M. Pignatelli and A. Moya, "Evaluating the fidelity of De Novo short read metagenomic assembly using simulated data," *PLoS ONE*, vol. 6, 2011.

[5] Y. Wang *et al.*, "Metacluster 5.0: A two-round binning approach for metagenomic data for low-abundance species in a noisy sample," *Bioinformatics*, vol. 28, 2012.

[6] D. H. Huson *et al.*, "MEGAN analysis of metagenomic data." *Genome research*, vol. 17, pp. 377–386, 2007.

[7] V. M. Markowitz *et al.*, "IMG/M: The integrated metagenome data management and comparative analysis system," *Nucleic Acids Research*, vol. 40, 2012.

[8] A. M. Schnoes *et al.*, "Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies," *PLoS Comput Biol*, vol. 5, no. 12, p. e1000605, Dec 2009.

[9] T. O. Delmont, P. Simonet, and T. M. Vogel, "Describing microbial communities and performing global comparisons in the omic era," pp. 1625–1628, 2012.

[10] S. Jaenicke *et al.*, "Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing," *PLoS ONE*, vol. 6, 2011.

[11] M. O. A. Sommer, G. Dantas, and G. M. Church, "Functional characterization of the antibiotic resistance reservoir in the human microflora." *Science (New York, N.Y.)*, vol. 325, pp. 1128–1131, 2009.

[12] K. U. Foerstner *et al.*, "Environments shape the nucleotide composition of genomes." *EMBO reports*, vol. 6, pp. 1208–1213, 2005.

[13] J. Raes *et al.*, "Prediction of effective genome size in metagenomic samples." *Genome biology*, vol. 8, p. R10, 2007.

[14] B. Jiang *et al.*, "Comparison of metagenomic samples using sequence signatures." *BMC genomics*, vol. 13, p. 730, Jan. 2012.

[15] S. F. Altschul *et al.*, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022283605803602

[16] W. J. Kent, "BLAT—The BLAST-Like Alignment Tool," *Genome Research*, vol. 12, no. 4, pp. 656–664, Mar. 2002. [Online]. Available: http://www.genome.org/cgi/doi/10.1101/gr.229202

[17] B. E. Dutilh *et al.*, "Reference-independent comparative metagenomics using cross-assembly: crAss." *Bioinformatics (Oxford, England)*, vol. 28, no. 24, pp. 3225–31, Dec. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23074261

[18] D. Fimereli, V. Detours, and T. Konopka, "TriageTools : tools for partitioning and prioritizing analysis of high-throughput sequencing data," pp. 1–8, 2013.

[19] N. Maillet *et al.*, "Compareads: comparing huge metagenomic experiments." *BMC bioinformatics*, vol. 13 Suppl 1, no. Suppl 19, p. S10, Jan. 2012.

[20] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," vol. 13, pp. 422–426, 1970.

[21] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9230594

[22] T. O. Delmont *et al.*, "Accessing the soil metagenome for studies of microbial diversity," *Applied and Environmental Microbiology*, vol. 77, no. 4, pp. 1315–1324, Feb 2011.

[23] F. Meyer *et al.*, "The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC bioinformatics*, vol. 9, no. 1, p. 386, Jan 2008.

Conclusion du chapitre II

Ainsi, la méthode *Compareads* utilisée en métagénomique comparative *de novo* pour récupérer l'information sur la similarité génomique entre plusieurs échantillons peut permettre d'étudier l'organisation génomique des communautés planctoniques. De plus, le développement de l'outil *COMMET* permet de réaliser cette étude sur l'ensemble des échantillons du projet *Tara* Oceans. Il est nécessaire pour mieux comprendre l'impact des facteurs environnementaux sur l'organisation génomique et spatiale des communautés micro-planctoniques d'intégrer des informations environnementales aux analyses des distances génomiques.

Chapitre III

Étude de l'impact des courants océaniques et des paramètres physico-chimiques sur l'organisation génomique des communautés de microplanctons

La biogéographie, présenté en partie I.3.5, s'intéresse à l'étude de la distribution des organismes vivants sur Terre et cherche à expliquer les raisons de leur répartition géographique. Les chapitres précédents ont montré qu'il existait une variabilité dans la composition en micro-organismes planctoniques et en gènes dans des régions océaniques présentant des caractéristiques physico-chimiques différentes. Les courants océaniques en transportant le plancton dans des régions présentant des contrastes environnementaux important contribuent directement à cette variabilité. Il est possible avec les données métagénomiques du projet *Tara* Oceans de comparer les métagénomes afin de réaliser une biogéographie des communautés de plancton viral, bactérien et eucaryote dans les eaux océaniques de surface. L'évaluation de l'impact des courants océaniques et des paramètres physico-chimiques sur l'organisation du contenu génomique des communautés de micro-organismes planctoniques peut également être effectuée.

III. Article 4 : Global plankton biogeography is shaped via ocean circulation dynamics

Le calcul de la similarité des lectures des 644 échantillons métagénomiques du projet *Tara* Oceans a permis de proposer la première biogéographie des communautés virales, bactériennes et eucaryotes dans les eaux océaniques de surface. Pour cela, les lectures provenant des échantillons de six fractions de tailles d'organismes (0 à 0.2 µm, 0.22 à 3 µm, 0.8 à 5 µm, 5 à 20 µm, 20 à 180 µm et 180 à 2000 µm) aux deux profondeurs, surface et DCM, ont été comparées avec l'outil *COMMET* ainsi qu'avec l'outil *Simka*, présenté dans la partie III.5.7. Ce dernier permet d'obtenir différents estimateurs de distance écologique ou de dissimilarité génétique entre deux échantillons. Cette distance a permis de regrouper par une méthode de *clustering*, les échantillons similaires d'un point de vue génétique. À partir de ces regroupements d'échantillons, la réalisation de la biogéographie des communautés microplanctoniques dans chaque taille de filtre a été effectuée. Ces régions sont appelées *genocenoses* comme référence

aux *biocenoses* qui correspondent à l'ensemble des êtres vivants coexistants dans un espace écologique donné. Cette biogéographie est différente selon les fractions de tailles d'organismes. L'utilisation d'un set de gènes construits et annotés taxonomiquement dans le cadre de ce projet à permis de caractériser les organismes qui forment ces regroupements d'échantillons. Ces micro-organismes planctoniques ne sont pas les mêmes selon les *genocenoses*. La caractérisation de l'environnement physico-chimique de chaque genocenose a été réalisée avec les données environnementales mesurées lors de l'expédition. L'environnement physico-chimique varie également entre les *genocenoses*. La couverture géographique des *genocenoses* a été comparée aux régions et aux provinces de Longhurst. Les *genocenoses* des fractions de petite taille (virus, bactéries et protistes) concordent mieux avec les régions décrites par Longhurst que les fractions de grande taille (zooplancton).

L'impact des courants océaniques sur la dynamique d'évolution des communautés microplanctoniques a ensuite été mesuré. Pour cela, le temps minimum nécessaire à une particule pour se déplacer entre deux stations a été mis en relation avec leur dissimilarité métagénomique. Cette étude a été réalisée sur les stations de surface le long des courants de l'océan Atlantique Nord. Le *turnover* (roulement) des communautés micro-planctoniques a pu être estimé au regard de leur composition génomique. Pour cela, un *decay time* (temps de roulement) a été calculé en prenant en compte l'évolution du modèle exponentiel de la beta-diversité spatio-temporelle avec le temps lagrangien existant entre ces stations. Ce *decay time* est de près d'un an pour les communautés bactériennes et protistes des petites fractions alors qu'il est de plus de deux ans pour les grandes fractions eucaryotes. Les courants océaniques affectent différemment le *turnover* des communautés planctoniques selon les tailles des organismes. Cela se répercute donc sur l'organisation des *genocenoses* dans l'ensemble des océans.

Ces résultats ont fait l'objet d'un article soumis le 13 Décembre 2016 pour publication dans la revue *Nature*. Cet article est présenté ci-après. Les informations supplémentaires se trouvent en annexe III.

Global plankton biogeography is shaped via ocean circulation dynamics

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

Tara Oceans Coordinators:

Silvia G. Acinas[29], Peer Bork[18,24,25], Emmanuel Boss[27], Chris Bowler[17], Colomban de Vargas[1], Mick Follows[12], Gabriel Gorsky[14], Nigel Grimsley[30,31], Pascal Hingamp[32], Daniele Iudicone[2], Olivier Jaillon[3,4,5], Stefanie Kandels-Lewis[18,19], Lee Karp-Boss[27], Eric Karsenti[17,19], Uros Krzic[33], Fabrice Not[34], Hiroyuki Ogata[35], Stéphane Pesant[22,23], Jeroen Raes[36,37], Emmanuel G. Reynaud[38], Christian Sardet[14,39], Mike Sieracki[40,41], Sabrina Speich[42,43], Lars Stemmann[14], Matthew B. Sullivan[16,28], Shinichi Sunagawa[18,26], Didier Velayoudon[44], Jean Weissenbach[3,4,5], Patrick Wincker[3,4,5]

** Co-first authors
* Co-corresponding authors

1 CNRS, UMR 7144, EPEP & Sorbonne Universités, UPMC Université Paris 06, Station Biologique de Roscoff, 29680 Roscoff, France.
2 Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.
3 CEA - Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, Evry France.
4 CNRS, UMR 8030, 2 rue Gaston Crémieux, Evry France.
5 Université d'Evry, UMR 8030, CP5706, Evry France.
6 Nippon Foundation-Nereus Program and Changing Ocean Research Unit, Institute for the Oceans and Fisheries, University of British Columbia. Aquatic Ecosystems Research Lab. 2202 Main Mall. Vancouver, BC V6T 1Z4. Canada.
7 Institut Pasteur - Bioinformatics and Biostatistics Hub - C3BI, USR 3756 IP CNRS - Paris, France.
8 INRIA/IRISA, Genscale team, UMR6074 IRISA CNRS/INRIA/Université de Rennes 1, Campus de Beaulieu, 35042, Rennes, France.
9 Jacobs University Bremen gGmbH, Campus Ring 1, D-28759 Bremen, Germany.
10 Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany.
11 Oxford e-Research Centre, University of Oxford, 7 Keble Road, OX1 3QG Oxford, Oxfordshire, UK.
12 Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA.
13 Dipartimento di Fisica e Astronomia 'G. Galilei' & CNISM, INFN, Università di Padova, Via Marzolo 8, 35131 Padova, Italy.
14 Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France.
15 Department of Oceanography, University of Hawaii, Honolulu, Hawaii 96822, USA.
16 Department of Microbiology, The Ohio State University, Columbus, OH 43214, USA.
17 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.
18 Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany.

19 Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg Germany.
20 Sorbonne Universités, UPMC Univ Paris 06, UMR 7093 LOV, F-75005, Paris, France.
21 CNRS, UMR 7093 LOV, F-75005, Paris, France.
22 MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany.
23 PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany.
24 Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany.
25 Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany.
26 Institute of Microbiology, Department of Biology, ETH Zurich, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.
27 School of Marine Sciences, University of Maine, Orono, Maine 04469, USA.
28 Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus OH 43214 USA.
29 Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC), Barcelona, Catalonia, Spain.
30 CNRS, UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France.
31 Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France.
32 Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, France.
33 Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
34 CNRS, UMR 7144, Sorbonne Universités, UPMC Université Paris 06, Station Biologique de Roscoff, 29680 Roscoff, France.
35 Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-001, Japan.
36 Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.
37 VIB Center for Microbiology, Herestraat 49, 3000 Leuven, Belgium.
38 Earth Institute, University College Dublin, Dublin, Ireland.
39 CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-mer, France.
40 National Science Foundation, Arlington, VA 22230, USA.
41 Bigelow Laboratory for Ocean Sciences East Boothbay, ME, USA.
42 Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, 29820 Plouzané, France.
43 Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France.
44 DVIP Consulting, Sèvres, France.

Conclusion du chapitre III

La métagénomique comparative de l'ensemble des échantillons du projet *Tara* Oceans a permis de réaliser la première biogéographie des communautés de plancton viral, bactérien et eucaryote dans les eaux océaniques de surface. Ces distances génomiques associées aux temps lagrangien et aux variations physico-chimiques entre stations ont révélé l'impact des courants océanique et de l'environnement sur l'organisation génomique des communautés planctoniques.

Conclusion générale

Au cours de ce travail de thèse, j'ai étudié la dynamique de la structure des génomes et de leur biogéographie dans l'océan. J'ai également analysé dans quelle mesure les facteurs océaniques influencent cette dynamique. Cela a été possible via des analyses à plusieurs niveaux de l'écosystème planctonique des eaux de surface océanique et en intégrant les données génomiques, métagénomiques ainsi que les paramètres environnementaux correspondant aux échantillons récoltés lors de l'expédition *Tara* Oceans.

I. Dynamique de la structure des génomes et de leur biogéographie à l'échelle d'une espèce

L'étude de la diversité génomique et spatiale du phytoplancton *Bathycoccus prasinos* a été effectuée à partir du génome de référence (RCC1105) et d'une partie d'un second génome (TOSAG39-1) obtenu lors de l'expédition *Tara* Oceans par une méthode d'amplification à cellule unique (SAG). L'analyse de génomique comparative de ces deux génomes partageant le même ARNr 18S a révélé qu'il s'agit vraisemblablement de deux espèces distinctes de *Bathycoccus*. De plus, une analyse de métagénomique ciblée a permis de décrire une biogéographie différente pour ces deux écotypes. Ces deux espèces de *Bathycoccus* sont présentes dans des environnements différents et pourraient ainsi avoir développé des stratégies écologiques divergentes. Ces résultats n'auraient pas pu être observés avec l'utilisation de métabarcodes génétiques comme les séquences V9 qui sont ici moins résolutives que les données métagénomiques. Dans les échantillons *Tara* Oceans, il n'existe pas de troisième écotype de *Bathycoccus*. Cependant, il n'est pas exclu qu'un autre écotype soit présent dans d'autres environnements tels que les zones polaires. Pour cela, l'expédition *Tara Polar Circle* a réalisé des échantillonnages dans l'océan Arctique. Il est prévu d'effectuer les mêmes analyses dans ces stations Arctiques.

L'analyse de métagénomique ciblée a également révélé une variabilité du contenu en gènes de la souche RCC1105 dans les différents échantillons. Ces gènes dispensables ont été caractérisés structuralement et validés. Ces gènes, souvent en cassettes, sans fonction connue et majoritairement localisés sur le chromosome 19, sont restreints géographiquement. Cela implique l'existence de plusieurs types génomiques au sein d'une même espèce qui varient en fonction du milieu écologique. Ces gènes pourraient participer à un processus évolutif pour l'adaptation d'un génotype dans un environnement donné. Ces analyses sur *Bathycoccus* ont fait

l'objet d'une publication dans la revue *Scientific Report* et sont présentées dans le chapitre I.1 de cette thèse.

L'analyse de la présence ou de l'absence du gène de la flavodoxine chez les deux génomes de *Bathycoccus* dans les échantillons *Tara* Oceans a permis de montrer que ce gène est semble-t-il toujours présent dans le génome de TOSAG39-1 et qu'il serait un gène dispensable chez RCC1105. Un autre génotype de *Bathycoccus prasinos* qui possède le gène de la flavodoxine serait présent dans les échantillons *Tara* Oceans. Il est possible que *Bathycoccus* présente différentes stratégies d'évolution sur ce gène en fonction de l'environnement où il se trouve. Il devrait être possible en améliorant les méthodes d'assemblage de métagénomes d'obtenir la portion de séquence complète englobant le gène de la flavodoxine de *Bathycoccus prasinos.* Cela permettrait de valider la présence ou non de ce gène dans les échantillons. De plus, une corrélation de cette information avec les paramètres environnementaux aiderait à la compréhension des mécanismes d'apparition de ce gène chez *Bathycoccus*.

Une analyse phylostratigraphique a permis de détecter des gènes inconnus dans le génome de RCC1105. L'analyse structurale de ces gènes a montré que la majorité d'entre eux est présente sur le chromosome 19. De plus, ces gènes ont un contenu en GC plus faible, sont plus petits et présentent moins d'introns que les gènes connus. L'utilisation des données métatranscriptomiques a validé la prédiction de ces gènes inconnus. La majorité de ces gènes inconnus est dispensable. Les informations sur leurs caractéristiques structurales, leurs localisations sur les chromosomes ainsi que leurs absences dans certaines stations laissent supposer l'existence d'une contrainte évolutive plus importante sur ces gènes en lien avec leur environnement. Une analyse ciblée de certaines de ces cassettes de gènes dispensables et inconnus pourrait être envisagée. Par exemple, la mise en culture d'un génotype de *Bathycoccus* possédant ou non, les cassettes de gènes dans différents environnements physico-chimiques ou biotiques pourrait permettre d'étudier l'influence de ces gènes sur la biologie de *Bathycoccus*, voire même de pouvoir nous éclairer sur une fonction potentielle de ces gènes.

La comparaison des profils phylostratigraphique des génomes des autres mamiellales disponibles (*Ostreococcus sp*. et *Micromonas sp.*) a été réalisée et a permis de montrer des variations plus ou moins importantes du pourcentage de gènes présents dans différents phylostratum. Cette variabilité dans l'apparition ou dans la perte de gènes dans ces génomes illustre la différence de la dynamique d'évolution du contenu en gènes des mamiellales. Les gènes codants pour la glycosyltransferases présents chez *Bathycoccus* et absents chez *Micromonas* et *Ostreococcus* sont un exemple dans cette variabilité génique. La biogéographie des mamiellales a permis de montrer que certaines espèces étaient retrouvées dans des

localisations et des environnements différents. Ces différences peuvent être une piste pour expliquer l'évolution divergente du contenu en gènes observée dans les différents phylostratum. Cela est peut être le cas pour les gènes de la glycosyltransferases impliqués dans la formation du *scale* chez *Bathycoccus* qui ont été perdus chez *Micromonas* et *Ostreococcus*. Une comparaison plus fine des différences biogéographiques et des préférences physico-chimiques des mamiellales est prévue dans le cadre de futures analyses.

II. Dynamique de l'organisation biogéographique des communautés planctoniques globales

Le passage à l'échelle pour l'étude de la biogéographie des communautés d'organismes planctoniques a été réalisé via l'utilisation d'un outil de métagénomique comparative. Pour cela, l'évaluation de l'outil *Compareads* permettant de connaître la similarité en lectures entre deux jeux de données métagénomiques a été réalisée sur une partie des échantillons du projet *Tara* Oceans. La méthode de métagénomique comparative *de novo* utilisée a permis d'analyser l'organisation génomique des communautés planctoniques dans ces échantillons. Les similarités de séquences entre différents bassins océaniques (Indien et Atlantique Sud) sont en adéquation avec l'explication du captage et du transport des micro-organismes planctoniques par les anneaux d'Agulhas. Ces analyses ont validé la capacité de cette méthode de métagénomique comparative *de novo* à réaliser des analyses sur l'organisation génomique des communautés planctoniques. Suite à un travail collaboratif avec d'autres laboratoires du consortium *Tara,* ces résultats ont été intégrés dans une publication parue dans la revue *Science* et sont présentés dans le chapitre II.1 de cette thèse. Cependant, pour des raisons de temps de calculs et de stockages des données générées trop importants, l'outil *Compareads* ne permet pas de réaliser les comparaisons sur l'ensemble des échantillons du projet *Tara* Oceans. Le développement de l'outil *COMMET* (COmpare Multiple METagenomes) a été réalisé lors d'un travail en collaboration avec l'équipe GenScale à l'INRIA de Rennes. Cet outil associé à des supercalculateurs a permis de réaliser les comparaisons de l'ensemble des échantillons *Tara* Oceans. *COMMET* a fait l'objet d'une publication dans un article de conférence (IEEE BIBM 2014 à Belfast) et est présenté dans le chapitre II.2 de cette thèse.

Le calcul de la similarité des lectures des 644 échantillons métagénomiques du projet *Tara* Oceans a permis de proposer la première biogéographie établie à partir de données métagénomiques des communautés virales, bactériennes et eucaryotes dans les eaux océaniques de surface. Le terme *genocenose,* en référence aux *biocenoses* qui correspondent à l'ensemble des êtres vivants coexistant dans un espace écologique donné, a été proposé pour décrire ces

regroupements d'échantillons. Ces *genocenoses* sont différentes selon les fractions de tailles d'organismes. La comparaison de ces *genocenoses* avec les provinces établies par Longhurst a montré que les *genocenoses* des organismes de petite taille (virus, bactéries et petits eucaryotes) concordent mieux avec les régions décrites par Longhurst que les *genocenoses* des organismes de grande taille (zooplancton). Cela est dû au fait que le découpage de Longhurst reflète surtout la distribution spatiale des organismes phytoplanctoniques car elle intègre la mesure de données chimiques telles que la quantité en chlorophylle *a*. Une analyse taxonomique a permis de montrer que ce ne sont pas les mêmes organismes qui participent aux regroupements d'échantillons dans les *genocenoses*. L'environnement physico-chimique (température, nitrate, phosphate, fer et les autres nutriments) varie également entre les *genocenoses*. Cette observation est bien illustrée par les échantillons issus de zones d'upwelling géographiquement distantes qui se trouvent regroupées au sein de la même *genocenose*. Comme cela a été décrit lors de l'évaluation de *Compareads* dans le chapitre II.2, ces échantillons ont peu de similarités génomiques avec les autres (sauf pour les échantillons subpolaires avec qui elles partagent plus de similarités de séquences). Leur environnement physico-chimique très contrasté par rapport aux autres échantillons fait que les communautés planctoniques s'y trouvant sont très peu similaires par rapport à celles qui constituent les autres genocenoses. De plus, il a été montré que les communautés de planctons présentes dans les fractions de grande taille sont plus affectées par les variations de température que par la limitation en nutriments contrairement aux organismes des petites fractions de taille qui sont plus rapidement impactés par les variations en nutriments. Ainsi, selon la taille des organismes, l'environnement n'a pas le même impact sur la composition des communautés planctoniques correspondantes ce qui explique en partie les différences de *genocenoses* entre les fractions de taille d'organisme. Cette sensibilité à l'environnement physico-chimique varie en fonction du temps lagrangien séparant les stations.

Les distances génomiques associées aux temps lagrangiens entre stations ont révélé qu'à une échelle de temps inférieure à 1 an et demi, la circulation océanique est le premier facteur qui a une influence sur la dynamique de l'organisation génomique des communautés planctoniques. Le *turnover* (roulement) de la génomique des communautés micro-planctoniques a pu être estimé sur les stations le long des courants de l'océan Atlantique Nord. Ces courants, en transportant et en mélangeant les organismes ainsi que les nutriments, affectent différemment ce *turnover* des communautés planctoniques selon les tailles des organismes. Les grandes fractions eucaryotes voient leur diversité moins impactée par les courants océaniques que les communautés bactériennes et protistes des petites fractions de taille. En effet, les grandes fractions de taille sont surtout composées d'eucaryotes multicellulaires qui ont une plus grande espérance de vie que les bactéries ou petits eucaryotes. Ils sont donc transportés sur de plus

longues distances. Cela explique que ce turnover plus long se répercute sur un étalement géographique plus important des *genocenoses* des grandes fractions de taille d'organismes qui traversent donc de multiples régions biogéochimiques. Ces résultats obtenus suite à un travail collaboratif avec d'autres laboratoires du consortium ont fait l'objet d'un article qui est actuellement en cours de révision pour la revue *Nature* et sont présentés dans le chapitre III de cette thèse.

III. Discussions et perspectives

Les résultats obtenus pendant ma thèse ont contribué à montrer l'apport de la connaissance des génomes et de la génomique en général à la compréhension des écosystèmes marins. De plus, ces résultats ont permis de mieux comprendre dans quelles mesures les facteurs océaniques impactent la dynamique de la structure des génomes planctoniques et de leur biogéographie dans l'océan.

III.1. Impact de l'environnement sur la variabilité de la structure en gènes et de la biogéographie chez la micro-algue *Bathycoccus*.

Des analyses antérieures ont déjà présenté l'existence de la variation du répertoire de gènes chez des bactéries[316–320]. Il a notamment été montré par la comparaison de deux écotypes de *Prochlorococcus* que la variation du contenu en gènes serait liée à l'adaptation à différentes intensités lumineuses[321]. Des hypothèses sur ce type d'adaptation avaient été proposées chez *Ostreococcus*[60] mais sont encore sources de débat[47]. Les deux écotypes de *Bathycoccus* sont présents à des niveaux de profondeurs et d'intensités lumineuses différentes. Cependant, ces génomes peuvent parfois partager une même localisation. En effet, TOSAG39-1 qui est surtout retrouvé en DCM est également observé dans des échantillons de surfaces où ont lieu des courants verticaux dans lesquels les deux espèces se mélangent. C'est le cas au niveau des courants d'Aghulas ainsi que dans le Gulf Stream. Il y a dans les échantillons présents le long des courants allant de l'océan Indien vers l'océan Atlantique Sud un remplacement progressif d'Est en Ouest de TOSAG39-1 par le génome de RCC1105. On observe la même distribution des deux écotypes d'Ouest en Est au niveau du gyre de l'Atlantique Nord. L'évolution le long de ce gyre du gradient de température et de la concentration en nutriments joue une influence sur la distribution de ces deux écotypes. Ainsi, les mouvements verticaux semblent transporter TOSAG39-1 de la profondeur DCM à la surface dans certains sites, puis celui-ci serait emporté par les courants horizontaux. Le long de ces courants horizontaux, la distribution de *Bathycoccus* évolue avec le changement des paramètres environnementaux. Chez les eucaryotes comme

MAST-4[322,323] ainsi que l'écotype arctique de *Micromonas*[46], la température influencerait également la distribution géographique de ces micro-organismes. La structure des *genocenoses* montre que pour le filtre de taille bactérien, deux *genocenoses* différentes (*Genocenoses* 5 et 7) sont présentes à l'Est et à l'Ouest du Cap de Bonne Espérance ainsi qu'au niveau du gyre de l'Atlantique Nord. La dynamique de la distribution géographique des deux écotypes de *Bathycoccus* décrite précédemment serait également valable pour un certain nombre de bactéries. Pour les *genocenoses* de la fraction de taille 0.8 à 5 µm dans laquelle *Bathycoccus* est retrouvé, une seule *genocenose* est observée au niveau de ces deux courants. Cependant, l'arbre de similarité des échantillons métagénomiques présente une séparation entre les échantillons situés à l'Est et à l'Ouest des ces deux régions avec une distance de dissimilarité moins élevée que pour les bactéries. La dynamique de la distribution des deux écotypes de *Bathycoccus* le long de ces courants ne serait alors pas un cas isolé chez les petits eucaryotes. En effet, la biogéographie des autres mamiellales a montré qu'il semblerait y avoir une distribution géographique similaire chez *Ostreococcus lucimarinus* et *Ostreococcus* RCC809. Le séquençage futur d'autres génomes via la méthode d'amplification par cellule unique pourrait permettre de découvrir chez d'autres espèces micro-planctoniques cette dynamique dans leur biogéographie. Cependant, dans ces analyses, les interactions biotiques ne sont pas prises en compte. Les facteurs biotiques représentent l'ensemble des interactions du vivant sur le vivant dans un écosystème. Une étude récente chez *Ostreococcus tauri* mis en culture avec son virus Otv5, a montré que des gènes présents sur le chromosome 19 sont surexprimés et que la taille de ce chromosome varie dans les lignées cellulaires résistantes à ce virus[324]. Les résultats sur les gènes dispensables de *Bathycoccus prasinos* très représentés sur ce chromosome *outlier* n'invalident pas l'hypothèse du rôle de ce chromosome sur la fonction immunitaire. Certains de ces évènements d'apparitions ou de pertes de gènes pourraient être le résultat d'un mécanisme de défense immunitaire contre les virus. Des virus du genre *Prasinovirus* infectant les *Mamiellales* ont été séquencés et leurs génomes sont disponibles sur les bases de données[325–328]. Une étude comparative sur la biogéographie de *Mamiellales* et de leurs *Prasinovirus* correspondants pourrait être envisagée pour étudier l'impact des interactions biotiques sur la biogéographie de ces micro-algues. De plus, des analyses ciblées sur les cassettes de gènes dispensables pourraient permettre de mieux comprendre les mécanismes impliqués dans cette résistance virale. Une hypothèse avait été émise sur le fait que des bactéries cosmopolites possèdent des gènes spécifiques, ou bien des variations génétiques qui ont une relation avec les propriétés écologiques de leur environnement[329]. L'évolution de certaines classes de gènes comme les gènes dispensables dont fait partie la flavodoxine, ou encore de gènes restreints taxonomiquement comme cela est le cas pour les gènes codant pour la glycosyltransferase, dans certaines conditions environnementales pourrait être un avantage évolutif qui permettrait une

acclimatation d'un organisme dans différents environnements et le rendrait ainsi cosmopolite, comme cela est le cas pour *Bathycoccus*.

III.2. Impact de l'environnement sur la biogéographie des communautés micro-planctoniques globales

Les différences observées sur le nombre et les recouvrements géographiques des *genocenoses* entre les petits et les gros organismes planctoniques impliquent une séparation écologique entre les producteurs primaires (ainsi que les autres organismes inférieurs à 20 µm) et les plus gros prédateurs dans l'écosystème planctonique. Le plancton présent dans les grandes fractions de taille est transporté le long de différents gradients de nutriments locaux. Les plus petits organismes ne peuvent pas traverser ces gradients. Cela suggère une grande plasticité dans la composition des communautés planctoniques. La diversité des réseaux d'interactions biotiques qui en découle doit, comme les paramètres physico-chimiques, jouer un rôle dans la dynamique d'organisation des communautés planctoniques globales. Par exemple, il est connu depuis plusieurs années que les virus à ADN infectent les algues[330,331] et ont un impact sur le *turnover* des populations de planctons[332–334]. Une prise en compte des interactions entre les différents organismes planctoniques doit être intégrée dans les analyses futures. De plus, les différences de *genocenoses* observées entre les tailles d'organismes montrent l'utilité d'analyser les métagénomes par fractions de tailles. Il a également été nécessaire d'utiliser les données métagénomiques pour distinguer les génomes d'organismes planctoniques à un niveau taxonomique résolutif. Les marqueurs génétiques moins résolutifs tels que les ARN16S ou 18S présentent en effet des temps de *turnover* différents de ceux calculés avec les données métagénomiques. Le *turnover* des communautés bactériennes et des petits eucaryotes calculé avec les métabarcodes dans le gyre de l'Atlantique Nord est plus long que celui calculé avec les données métagénomiques. Cela rejoint les observations décrites dans la partie précédente où la dynamique de distribution des deux écotypes de *Bathycoccus* le long de ces courants ne peut pas être observée avec les données 18S. On observe également moins de contraste entre les similarités des échantillons à l'Ouest et à l'Est de ce courant avec les métabarcodes qu'avec les données métagénomiques. Un résultat similaire est obtenu au niveau du Cap de Bonne Espérance (voir dans l'annexe 3 figure supplémentaire 3 et 9). Ainsi, les données métagénomiques se trouvent être plus adéquates que les métabarcodes pour de futures descriptions sur la dynamique des génomes et des communautés planctoniques.

La circulation de l'eau de mer dans l'ensemble des océans est régie par les effets combinés du vent, des différences de températures, de densités, de salinités mais également des interactions au sein des courants marins. Le plancton, transporté par ces mouvements d'eau, est

intimement lié à ces différents facteurs. Le changement climatique a un impact sur les grands courants marins[335–337] ainsi que sur la distribution des organismes planctoniques[338]. Ces courants qui dispersent les nutriments et transportent le plancton ont une influence primordiale sur le *turnover* de la composition des communautés planctoniques et donc sur les *genocenoses* correspondantes. Cela suggère que les *genocenoses* sont des entités dynamiques qui reflètent un équilibre local entre l'effet de l'environnement et les interactions biotiques. Ainsi, des modifications futures de la circulation océanique ou de la variation des conditions environnementales à large échelle peut potentiellement redistribuer la composition des communautés micro-planctoniques. Cela induirait une nouvelle biogéographie du plancton global dans les océans. L'utilisation des modèles de prédictions sur l'évolution des courants océaniques et des paramètres physico-chimiques au cours du 21ème siècle[339] peut éventuellement permettre de proposer une organisation des futures *genocenoses*.

Enfin, il a été observé une forte similarité génomique entre des échantillons très éloignés géographiquement (exemples des upwellings). Des résultats préliminaires avec l'intégration des données issues du séquençage *Tara Polar Circle* dans les comparaisons métagénomiques, montrent également une forte similarité génomique entre les échantillons de l'océan Austral et les échantillons arctiques comparés aux échantillons des autres océans. De plus, une étude récente chez l'espèce *Oïtona similis* qui est un copépode cosmopolite, a présenté qu'une lignée arctique était plus proche génétiquement des lignées de l'océan Austral que de celles de l'hémisphère Nord[340]. Ainsi, la dynamique rapide de la structure et de la biogéographie des génomes en réponse à l'environnement physico-chimique (et biotique) lors du transport du plancton par les courants océaniques peut supposer que le plancton est capable de traverser l'ensemble des océans via la circulation thermohaline. Cela expliquerait l'existence d'organismes cosmopolites, comme *Bathycoccus* ainsi que les fortes similarités génomiques observées entre des communautés planctoniques éloignées géographiquement. L'intégration des données métagénomiques issues du séquençage du projet *Tara Polar Circle* (mais également d'autres expéditions) ainsi que des données provenant des échantillons récoltés à la profondeur mésopélagique des océans permettrait d'approfondir l'étude de cette dynamique du plancton océanique global.

Références

1. Hays, G. C., Richardson, A. J. & Robinson, C. Climate change and marine plankton. *Trends Ecol. Evol.* **20,** 337–344 (2005).
2. Longhurst, A. Ecological Geography of the Sea, Academic Press, London. (2007).
3. Oliver, M. J. & Irwin, A. J. Objective global ocean biogeographic provinces. *Geophys. Res. Lett.* **35,** L15601 (2008).
4. Reygondeau, G. *et al.* Dynamic biogeochemical provinces in the global ocean. *Glob. Biogeochem. Cycles* **27,** 1046–1058 (2013).
5. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (1995).
6. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546, 563–567 (1996).
7. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
8. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
9. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev. MMBR* **68,** 669–685 (2004).
10. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
11. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 14250–14255 (2002).
12. Breitbart, M. *et al.* Metagenomic Analyses of an Uncultured Viral Community from Human Feces. *J. Bacteriol.* **185,** 6220–6223 (2003).
13. Venter, J. C. *et al.* Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
14. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLOS Biol* **5,** e77 (2007).
15. Hensen, V. *Uber die Bestimmung des Planktons oder des im Meere treibenden Materials an Pflanzen und Tieren*. (1887).
16. Trégouboff, G. & Rose, M. in *Manuel de planctonologie méditerranéenne, Tome 1* (1957).
17. Ochoa, J., Maske, H., Sheinbaum, J. & Candela, J. Diel and lunar cycles of vertical migration extending to below 1000 m in the ocean and the vertical connectivity of depth-tiered populations. *Limnol. Oceanogr.* **58,** 1207–1214 (2013).
18. Brock, T. D. Sea microbes. *Limnol. Oceanogr.* **25,** 199–200 (1980).
19. Scola, B. L. *et al.* A Giant Virus in Amoebae. *Science* **299,** 2033–2033 (2003).
20. Philippe, N. *et al.* Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* **341,** 281–286 (2013).
21. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438,** 86–89 (2005).
22. Sun, S. *et al.* Structural Studies of the Sputnik Virophage. *J. Virol.* **84,** 894–897 (2010).
23. Sardet, C. *Plancton, Aux origines du vivant.* (Ulmer, 2013).
24. Fenchel, T. Marine Plankton Food Chains. *Annu. Rev. Ecol. Syst.* **19,** 19–38 (1988).

25. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* **281,** 237–240 (1998).
26. Wilson, S. E., Ruhl, H. A. & Smith, K. L. Zooplankton fecal pellet flux in the abyssal northeast Pacific: A 15 year time-series study. *Limnol. Oceanogr.* **58,** 881–892 (2013).
27. Darwin, C. *L'origine des espèces au moyen de la sélection naturelle ou la lutte pour l'existence dans la nature*. (Alfred Costes, 1859).
28. Scanlan, D. J. *et al.* Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev. MMBR* **73,** 249–299 (2009).
29. Delwiche, null. Tracing the Thread of Plastid Diversity through the Tapestry of Life. *Am. Nat.* **154,** S164–S177 (1999).
30. Burki, F. The Eukaryotic Tree of Life from a Global Phylogenomic Perspective. *Cold Spring Harb. Perspect. Biol.* **6,** a016147 (2014).
31. Palmer, J. D. A single birth of all plastids? *Nature* **405,** 32–33 (2000).
32. Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. & Bhattacharya, D. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.* **21,** 809–818 (2004).
33. Marin, B. Nested in the Chlorellales or Independent Class? Phylogeny and Classification of the Pedinophyceae (Viridiplantae) Revealed by Molecular Phylogenetic Analyses of Complete Nuclear and Plastid-encoded rRNA Operons. *Protist* **163,** 778–805 (2012).
34. De reviers, B. in *La classification des algues* **10,** 59–111 (Biosystema, 1993).
35. Fawley, M. W., Yun, Y. & Qin, M. Phylogenetic analyses of 18s rdna sequences reveal a new coccoid lineage of the prasinophyceae (Chlorophyta). *J. Phycol.* **36,** 387–393 (2000).
36. Lewin, R. A., Krienitz, L., Goericke, R., Takeda, H. & Hepperle, D. Picocystis salinarum gen. et sp. nov. (Chlorophyta) – a new picoplanktonic green alga. *Phycologia* **39,** 560–565 (2000).
37. Guillou, L. *et al.* Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. *Protist* **155,** 193–214 (2004).
38. Viprey, M., Guillou, L., Ferréol, M. & Vaulot, D. Wide genetic diversity of picoplanktonic green algae (Chloroplastida) in the Mediterranean Sea uncovered by a phylum-biased PCR approach. *Environ. Microbiol.* **10,** 1804–1822 (2008).
39. Marin, B. & Melkonian, M. Molecular Phylogeny and Classification of the Mamiellophyceae class. nov. (Chlorophyta) based on Sequence Comparisons of the Nuclear- and Plastid-encoded rRNA Operons. *Protist* **161,** 304–336 (2010).
40. Vaulot, D. *et al.* Metagenomes of the Picoalga Bathycoccus from the Chile Coastal Upwelling. *PLoS ONE* **7,** e39648 (2012).
41. Simmons, M. P. *et al.* Abundance and biogeography of picoprasinophyte ecotypes and other phytoplankton in the Eastern North Pacific Ocean. *Appl. Environ. Microbiol.* AEM.02730-15 (2016). doi:10.1128/AEM.02730-15
42. van Baren, M. J. *et al.* Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants. *BMC Genomics* **17,** 267 (2016).
43. Díez, B., Pedrós-Alió, C. & Massana, R. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl. Environ. Microbiol.* **67,** 2932–2941 (2001).
44. Worden, A. Z. & Not, F. in *Microbial Ecology of the Oceans* (ed. Kirchman, D. L.) 159–205 (John Wiley & Sons, Inc., 2008).

45. Šlapeta, J., López-García, P. & Moreira, D. Global Dispersal and Ancient Cryptic Species in the Smallest Marine Eukaryotes. *Mol. Biol. Evol.* **23,** 23–29 (2006).
46. Lovejoy, C. *et al.* Distribution, Phylogeny, and Growth of Cold-Adapted Picoprasinophytes in Arctic Seas1. *J. Phycol.* **43,** 78–89 (2007).
47. Demir-Hilton, E. *et al.* Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *ISME J.* **5,** 1095–1107 (2011).
48. Monier, A., Sudek, S., Fast, N. M. & Worden, A. Z. Gene invasion in distant eukaryotic lineages: discovery of mutually exclusive genetic elements reveals marine biodiversity. *ISME J.* **7,** 1764–1774 (2013).
49. Worden, A. Z., Nolan, J. K. & Palenik, B. Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnol. Oceanogr.* **49,** 168–179 (2004).
50. Monier, A. *et al.* Oceanographic structure drives the assembly processes of microbial eukaryotic communities. *ISME J.* **9,** 990–1002 (2015).
51. Li, W. K. W., McLaughlin, F. A., Lovejoy, C. & Carmack, E. C. Smallest Algae Thrive As the Arctic Ocean Freshens. *Science* **326,** 539–539 (2009).
52. Monier, A., Worden, A. Z. & Richards, T. A. Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ. Microbiol. Rep.* (2016).
53. Piganeau, G., Eyre-Walker, A., Jancek, S., Grimsley, N. & Moreau, H. How and why DNA barcodes underestimate the diversity of microbial eukaryotes. *PloS One* **6,** e16342 (2011).
54. Chrétiennot-Dinet, M.-J. *et al.* A new marine picoeucaryote: Ostreococcus tauri gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia* **34,** 285–292 (1995).
55. Eikrem, W. & Throndsen, J. The ultrastructure of Bathycoccus gen. nov. and B. prasinos sp. nov., a non-motile picoplanktonic alga (Chlorophyta, Prasinophyceae) from the Mediterranean and Atlantic. *Phycologia* **29,** 344–350 (1990).
56. Butcher, R. W. Contributions to our knowledge of the smaller marine algae. *J. Mar. Biol. Assoc. U. K.* **31,** 175–191 (1952).
57. Courties, C. *et al.* Smallest eukaryotic organism. *Nature* **370,** 255–255 (1994).
58. Subirana, L. *et al.* Morphology, genome plasticity, and phylogeny in the genus ostreococcus reveal a cryptic species, O. mediterraneus sp. nov. (Mamiellales, Mamiellophyceae). *Protist* **164,** 643–659 (2013).
59. Palenik, B. *et al.* The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 7705–7710 (2007).
60. Rodríguez, F. *et al.* Ecotype diversity in the marine picoeukaryote Ostreococcus (Chlorophyta, Prasinophyceae). *Environ. Microbiol.* **7,** 853–859 (2005).
61. Blanc-Mathieu, R. *et al.* An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies. *BMC Genomics* **15,** 1103 (2014).
62. Worden, A. Z. *et al.* Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. *Science* **324,** 268–272 (2009).
63. Foulon, E. *et al.* Ecological niche partitioning in the picoplanktonic green alga Micromonas pusilla: evidence from environmental surveys using phylogenetic probes. *Environ. Microbiol.* **10,** 2433–2443 (2008).
64. Simmons, M. P. *et al.* Intron invasions trace algal speciation and reveal nearly identical Arctic and Antarctic Micromonas populations. *Mol. Biol. Evol.* (2015).

65. Johnson, P. W. & Sieburth, J. M. In-Situ Morphology and Occurrence of Eucaryotic Phototrophs of Bacterial Size in the Picoplankton of Estuarine and Oceanic Waters1. *J. Phycol.* **18,** 318–327 (1982).
66. Collado-Fabbri, S., Vaulot, D. & Ulloa, O. Structure and seasonal dynamics of the eukaryotic picophytoplankton community in a wind-driven coastal upwelling ecosystem. *Limnol. Oceanogr.* **56,** 2334–2346 (2011).
67. Not, F. *et al.* A single species, Micromonas pusilla (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. *Appl. Environ. Microbiol.* **70,** 4064–4072 (2004).
68. Moreau, H. *et al.* Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biol.* **13,** R74 (2012).
69. Tansley, A. G. *The use and abuse of vegetational concepts and terms*. **16,** (Stor, 1935).
70. ASTOLFI J.-P. *Procédures d'apprentissage en sciences expérimentales*. (Paris INRP, 1985).
71. Möbius, K. The Oyster and the Oyster-Culture. *Rep. US Comm. Fish Fish.* (1883).
72. Mantoura, R. F. C. & Llewellyn, C. A. The rapid determination of algal chlorophyll and carotenoid pigments and their breakdown products in natural waters by reverse-phase high-performance liquid chromatography. *Anal. Chim. Acta* **151,** 297–314 (1983).
73. Murray, A. P., Gibbs, C. F., Longmore, A. R. & Flett, D. J. Determination of chlorophyll in marine waters: intercomparison of a rapid HPLC method with full HPLC, spectrophotometric and fluorometric methods. *Mar. Chem.* **19,** 211–227 (1986).
74. Claustre, H. The trophic status of various oceanic provinces as revealed by phytoplankton pigment signatures. *Limnol. Oceanogr.* **39,** 1206–1210 (1994).
75. Latasa, M., Bidigare, R. R., Ondrusek, M. E. & Kennicutt, M. C. HPLC analysis of algal pigments: a comparison exercise among laboratories and recommendations for improved analytical performance. *Mar. Chem.* **51,** 315–324 (1996).
76. Cullen, J. J., Ciotti, Á. M., Davis, R. F. & Lewis, M. R. Optical detection and assessment of algal blooms. *Limnol. Oceanogr.* **42,** 1223–1239 (1997).
77. Trees, C. C., Clark, D. K., Bidigare, R. R., Ondrusek, M. E. & Mueller, J. L. Accessory pigments versus chlorophyll a concentrations within the euphotic zone: A ubiquitous relationship. *Limnol. Oceanogr.* **45,** 1130–1143 (2000).
78. Kirkpatrick, G. J., Millie, D. F., Moline, M. A. & Schofield, O. Optical discrimination of a phytoplankton species in natural mixed populations. *Limnol. Oceanogr.* **45,** 467–471 (2000).
79. Havskum, H., Schlter, L., Scharek, R., Berdalet, E. & Jacquet, S. Routine quantification of phytoplankton groups‹microscopy or pigment analyses? *Mar. Ecol. Prog. Ser.* **273,** 31–42 (2004).
80. Kemp, P. . Handbook of Methods in Aquatic Microbial Ecology. *Phycologia* **33,** 308–308 (1994).
81. Wetzel, R. G. & Likens, G. E. *Limnological Analyses*. (Springer New York, 2000).
82. Jacquet, S., Lennon, J. & Vaulot, D. Application of a compact automatic sea water sampler to high frequency picoplankton studies. *Aquat. Microb. Ecol.* **14,** 309–314 (1998).
83. Jacquet, S., Lennon, J.-F., Marie, D. & Vaulot, D. Picoplankton population dynamics in coastal waters of the northwestern Mediterranean Sea. *Limnol. Oceanogr.* **43,** 1916–1931 (1998).
84. Vaulot, D. & Marie, D. Diel variability of photosynthetic picoplankton in the equatorial Pacific. *J. Geophys. Res. Oceans* **104,** 3297–3310 (1999).
85. Peperzak, L., Vrieling, E. G., Sandee, B. & Rutten, T. Immuno flow cytometry in marine phytoplankton research. *Sci. Mar.* **64,** 165–181 (2000).

86. Veldhuis, M. J. W. & Kraay, G. W. Application of flow cytometry in marine phytoplankton research: current applications and future perspectives. *Sci. Mar.* **64,** 121–134 (2000).
87. Li, W. K. & Dickie, P. M. Monitoring phytoplankton, bacterioplankton, and virioplankton in a coastal inlet (Bedford Basin) by flow cytometry. *Cytometry* **44,** 236–246 (2001).
88. Chisholm, S. W. *et al.* A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334,** 340–343 (1988).
89. Olson, R. J., Frankel, S. L., Chisholm, S. W. & Shapiro, H. M. An inexpensive flow cytometer for the analysis of fluorescence signals in phytoplankton: Chlorophyll and DNA distributions. *J. Exp. Mar. Biol. Ecol.* **68,** 129–144 (1983).
90. Yentsch, C. M. & Horan, P. K. Cytometry in the Aquatic Sciences. *Cytometry* **10,** 497–499 (1989).
91. Olson, R. J., Zettler, E. R., Chisholm, S. W. & Dusenberry, J. A. in *Particle Analysis in Oceanography* (ed. Demers, S.) 351–399 (Springer Berlin Heidelberg, 1991).
92. Wiebe, P. H. & Benfield, M. C. in *Encyclopedia of Ocean Sciences* 3237–3253 (Elsevier, 2001).
93. Sieracki, C. K., Sieracki, M. E. & Yentsch, C. S. An imaging-in-flow system for automated analysis of marine microplankton. *Mar. Ecol. Prog. Ser.* **168,** 285–296 (1998).
94. Grosjean, P. Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. *ICES Journal of Marine Science: Journal du Conseil* 518–525 (2004).
95. Bell, J. L. & Hopcroft, R. R. Assessment of ZooImage as a tool for the classification of zooplankton. *J. Plankton Res.* **30,** 1351–1367 (2008).
96. Hebert, P. D. N., Ratnasingham, S. & Waard, J. R. de. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. B Biol. Sci.* **270,** S96–S99 (2003).
97. Valentini, A., Pompanon, F. & Taberlet, P. DNA barcoding for ecologists. *Trends Ecol. Evol.* **24,** 110–117 (2009).
98. Biastoch, A., Böning, C. W. & Lutjeharms, J. R. E. Agulhas leakage dynamics affects decadal variability in Atlantic overturning circulation. *Nature* **456,** 489–492 (2008).
99. Gordon, A. L. Oceanography: The brawniest retroflection. *Nature* **421,** 904–905 (2003).
100. Biastoch, A., Böning, C. W., Schwarzkopf, F. U. & Lutjeharms, J. R. E. Increase in Agulhas leakage due to poleward shift of Southern Hemisphere westerlies. *Nature* **462,** 495–498 (2009).
101. Beal, L. M., De Ruijter, W. P. M., Biastoch, A., Zahn, R. & 136, S. W. G. On the role of the Agulhas system in ocean circulation and climate. *Nature* **472,** 429–436 (2011).
102. Cermeño, P. & Falkowski, P. G. Controls on Diatom Biogeography in the Ocean. *Science* **325,** 1539–1541 (2009).
103. Schmidt, M. W., Spero, H. J. & Lea, D. W. Links between salinity variation in the Caribbean and North Atlantic thermohaline circulation. *Nature* **428,** 160–163 (2004).
104. Kelling, G. & Stanley, D. J. Sedimentary evidence of bottom current activity, Strait of Gibraltar region. *Mar. Geol.* **13,** 51–60 (1972).
105. Fuglister, F. G. Annual variations in current speeds in the Gulf Stream System. *Journal of Marine Research* 119–127 (1951).
106. Bolhuis, H. & Cretoiu, M. S. in *The Marine Microbiome: An Untapped Source of Biodiversity and Biotechnological Potential* 5–6 (Springer, 2016).
107. Durand, M.-H. *et al. Global versus local changes in upwelling systems*. (ORSTOM, 1998).
108. Capone, D. G. & Hutchins, D. A. Microbial biogeochemistry of coastal upwelling regimes in a changing ocean. *Nat. Geosci.* **6,** 711–717 (2013).

109. Aminot, A., Chaussepied, M. & Oceans, B. (France) C. N. pour l'Exploitation des. Manuel des analyses chimiques en milieu marin. *CNEXO* (1983).
110. Alzieu, C. in *Aquaculture* 16–43 (Tec et Doc, 1989).
111. Le Menn, M. *Instrumentation et métrologie en océanographie physique*. (Hermes-Lavoisier, 2007).
112. Fofonoff, N. P. Physical properties of seawater: A new salinity scale and equation of state for seawater. *J. Geophys. Res.* **90,** 3332 (1985).
113. Bartz, R., Ronald, J., Zaneveld, V. & Pak, H. A Transmissometer For Profiling And Moored Observations In Water. in (eds. White, M. B. & Stevenson, R.) **160,** 102–109 (Ocean Optics, 1978).
114. Sutherland, T. F., Lane, P. M., Amos, C. L. & Downing, J. The calibration of optical backscatter sensors for suspended sediment of varying darkness levels. *Mar. Geol.* **162,** 587–597 (2000).
115. Bricaud, A., Morel, A., Babin, M., Allali, K. & Claustre, H. Variations of light absorption by suspended particles with chlorophyll a concentration in oceanic (case 1) waters: Analysis and implications for bio-optical models. *J. Geophys. Res. Oceans* **103,** 31033–31044 (1998).
116. Gould, W. J. in *From Swallow floats to Argo—the development of neutrally buoyant floats* **52,** 529–543 (Elsevier, 2005).
117. Renner, S. S., Lomolino, M. V., Riddle, B. R. & Brown, J. H. Biogeography, third edition. *Syst. Biol.* **55,** 845 (2006).
118. Fasham, M. J. R. Ocean biogeochemistry: the role of the ocean carbon cycle in global change. *Glob. Change - IGBP Ser.* (2003).
119. Somerville, M. *Physical geography*. (Blanchard and Lea, 1862).
120. Beklemishev, C. W. On the spatial structure of plankton communities in dependence of oceanic circulation. Boundaries of ranges of oceanic plankton animals in the North Pacific. *Okeanologia* 1059–1072 (1961).
121. McGowan, J. A. Ocean biogeography of the Pacific. In : The micropaleontology of the oceans,. 3–74 (1971).
122. Emery, W. & Meincke, J. Global water masses - summary and review. *Oceanol. Acta* **9,** 383–391 (1986).
123. Yentsch, C. S. & Garside, J. C. Patterns of phytoplankton abundance and biogeography. in **Pelagic Biogeography,** 278–284 (1986).
124. Platt, T., Caverhill, C. & Sathyendranath, S. Basin-scale estimates of oceanic primary production by remote sensing: The North Atlantic. *J. Geophys. Res.* **96,** 15147 (1991).
125. Sathyendranath, S., Longhurst, A., Caverhill, C. M. & Platt, T. Regionally and seasonally differentiated primary production in the North Atlantic. *Deep Sea Res. Part Oceanogr. Res. Pap.* **42,** 1773–1802 (1995).
126. Reygondeau, G. *et al.* Biogeography of tuna and billfish communities: Biogeography of tuna and billfish communities. *J. Biogeogr.* **39,** 114–129 (2012).
127. Wiebe, P. H. & Benfield, M. C. From the Hensen net toward four-dimensional biological oceanography. *Progree Oceanogr.* 7–136 (2003).
128. Frazer, J. H. The history of plankton sampling. in **UNESCO,** 11–18 (Zooplankton sampling, 1968).
129. Cleve, P. T. Plankton collected by the Swedish expedition to Spitzbergen in 1898. *PA Norstedt Söner* (1899).

130. Lohmann, H. Über das Nannoplankton und die Zentrifugierung kleinster Wasserproben zur Gewinnung desselben in lebendem Zustande. *Int. Rev. Gesamten Hydrobiol. Hydrogr.* **4,** 1–38 (1911).
131. Heackel, E. *Report on the scientific results of the voyage of H.M.S. Challenger during the years 1873-1876*. **XVIII,** (Neill, 1880).
132. Fuhrman, J. A. Microbial community structure and its functional implications. *Nature* **459,** 193–199 (2009).
133. Martiny, J. B. H. *et al.* Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* **4,** 102–112 (2006).
134. Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5,** e16 (2007).
135. Barberán, A., Fernández-Guerra, A., Bohannan, B. J. M. & Casamayor, E. O. Exploration of community traits as ecological markers in microbial metagenomes. *Mol. Ecol.* **21,** 1909–1917 (2012).
136. Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl. Acad. Sci.* **106,** 1374–1379 (2009).
137. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
138. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467 (1977).
139. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 560–564 (1977).
140. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).
141. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51 Pt 1,** 263–273 (1986).
142. Sanger, F. *et al.* Nucleotide sequence of bacteriophage |[phi]|X174 DNA. *Nature* **265,** 687–695 (1977).
143. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14 : Article : Nature. *Nature* **421,** 601–607 (2003).
144. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
145. Housby, J. N. & Southern, E. M. Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. *Nucleic Acids Res.* **26,** 4259–4266 (1998).
146. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309,** 1728–1732 (2005).
147. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9,** 387–402 (2008).
148. Hosomichi, K., Shiina, T., Tajima, A. & Inoue, I. The impact of next-generation sequencing technologies on HLA research. *J. Hum. Genet.* **60,** 665–673 (2015).
149. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).
150. Medini, D. *et al.* Microbiology in the post-genomic era. *Nat. Rev. Microbiol.* **6,** 419–430 (2008).
151. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci.* **100,** 3960–3964 (2003).

152. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323,** 133–138 (2009).
153. Korlach, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol.* **472,** 431–455 (2010).
154. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26,** 1146–1153 (2008).
155. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4,** 265–270 (2009).
156. Ding, F. *et al.* Single-molecule mechanical identification and sequencing. *Nat. Methods* **9,** 367–372 (2012).
157. Beer, M. & Zobel, C. R. Electron stains. II: Electron microscopic studies on the visibility of stained DNA molecules. *J. Mol. Biol.* **3,** 717–726 (1961).
158. Madoui, M.-A. *et al.* Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics* **16,** 327 (2015).
159. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13,** 341 (2012).
160. Liu, L. *et al.* Comparison of Next-Generation Sequencing Systems. *BioMed Res. Int.* **2012,** e251364 (2012).
161. Wooley, J. C., Godzik, A. & Friedberg, I. A Primer on Metagenomics. *PLOS Comput Biol* **6,** e1000667 (2010).
162. Pignatelli, M. & Moya, A. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PloS One* **6,** e19984 (2011).
163. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
164. Bik, E. M. *et al.* Molecular analysis of the bacterial microbiota in the human stomach. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 732–737 (2006).
165. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308,** 1635–1638 (2005).
166. Balzer, S., Malde, K. & Jonassen, I. Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinforma. Oxf. Engl.* **27,** i304-309 (2011).
167. Bernal, A., Ear, U. & Kyrpides, N. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* **29,** 126–127 (2001).
168. Reddy, T. B. K. *et al.* The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.* **43,** D1099-1106 (2015).
169. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
170. Chikhi, R. & Rizk, G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms Mol. Biol.* **8,** 22 (2013).
171. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinforma. Oxf. Engl.* **27,** 578–579 (2011).
172. Dayarian, A., Michael, T. P. & Sengupta, A. M. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics* **11,** 345 (2010).

173. Gao, S., Sung, W.-K. & Nagarajan, N. Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **18,** 1681–1691 (2011).
174. Marcy, Y. *et al.* Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 11889–11894 (2007).
175. Eberwine, J., Sul, J.-Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nat. Methods* **11,** 25–27 (2014).
176. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Res.* **37,** D793–D796 (2009).
177. Tringe, S. G. *et al.* Comparative Metagenomics of Microbial Communities. *Science* **308,** 554–557 (2005).
178. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17,** 175–188 (2016).
179. Baker, M. Method offers DNA blueprint of a single human cell. *Nature* (2012). doi:10.1038/nature.2012.12088
180. Arneson, N., Hughes, S., Houlston, R. & Done, S. Whole-Genome Amplification by Degenerate Oligonucleotide Primed PCR (DOP-PCR). *Cold Spring Harb. Protoc.* **2008,** pdb.prot4919 (2008).
181. Spits, C. *et al.* Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.* **1,** 1965–1970 (2006).
182. Lasken, R. S. Single-cell sequencing in its prime. *Nat. Biotechnol.* **31,** 211–212 (2013).
183. Seeleuthner, Y. Single-cell genomics reveals hidden functional complexity in the marine microbial loop. *submited*
184. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* (2015). doi:10.1101/pdb.top084970
185. Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.* **5,** 516–535 (2010).
186. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proc. R. Soc. B Biol. Sci.* **270,** 313–321 (2003).
187. Moon-van der Staay, S. Y., De Wachter, R. & Vaulot, D. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409,** 607–610 (2001).
188. Massana, R., Balagué, V., Guillou, L. & Pedrós-Alió, C. Picoeukaryotic diversity in an oligotrophic coastal site studied by molecular and culturing approaches. *FEMS Microbiol. Ecol.* **50,** 231–243 (2004).
189. Derelle, E. *et al.* Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 11647–11652 (2006).
190. Stoeck, T., Hayward, B., Taylor, G. T., Varela, R. & Epstein, S. S. A multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. *Protist* **157,** 31–43 (2006).
191. Worden, A. Z. Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquat. Microb. Ecol.* **43,** 165–175 (2006).
192. Countway, P. D. *et al.* Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). *Environ. Microbiol.* **9,** 1219–1232 (2007).

193. Shalchian-Tabrizi, K., Kauserud, H., Massana, R., Klaveness, D. & Jakobsen, K. S. Analysis of environmental 18S ribosomal RNA sequences reveals unknown diversity of the cosmopolitan phylum Telonemia. *Protist* **158,** 173–180 (2007).
194. Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W. & Huse, S. M. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PloS One* **4,** e6372 (2009).
195. Brown, M. V. *et al.* Microbial community structure in the North Pacific ocean. *ISME J.* **3,** 1374–1386 (2009).
196. Vigil, P. *et al.* Rapid shifts in dominant taxa among microbial eukaryotes in estuarine ecosystems. *Aquat. Microb. Ecol.* **54,** 83–100 (2009).
197. Le Bescot, N. *et al.* Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. *Environ. Microbiol.* **18,** 609–626 (2016).
198. Bucklin, A., Lindeque, P. K., Rodriguez-Ezpeleta, N., Albaina, A. & Lehtiniemi, M. Metabarcoding of marine zooplankton: prospects, progress and pitfalls. *J. Plankton Res.* **38,** 393–400 (2016).
199. Cai, L., Ye, L., Tong, A. H. Y., Lok, S. & Zhang, T. Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets. *PLOS ONE* **8,** e53649 (2013).
200. Prescott, L. M., Harley, J. P. & Klein, D. A. *Microbiology*. (McGraw-Hill, 2002).
201. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5,** R245-249 (1998).
202. Gilbert, J. A. & Dupont, C. L. Microbial Metagenomics: Beyond the Genome. *Annu. Rev. Mar. Sci.* **3,** 347–371 (2011).
203. Maillet, N. Comparaison de novo de données de séquençage issues de très grands échantillons métagénomiques : application sur le projet Tara Oceans. (Rennes 1, 2013).
204. Lindner, M. S. & Renard, B. Y. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Res.* **41,** e10 (2013).
205. de Vargas, C. *et al.* Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348,** 1261605 (2015).
206. Suenaga, H. Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environ. Microbiol.* **14,** 13–22 (2012).
207. Cuvelier, M. L. *et al.* Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl. Acad. Sci.* **107,** 14679–14684 (2010).
208. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40,** D71-75 (2012).
209. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–D230 (2014).
210. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* **31,** 6633–6639 (2003).
211. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32,** D277-280 (2004).
212. Kennedy, J., Marchesi, J. R. & Dobson, A. D. Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microb. Cell Factories* **7,** 27 (2008).
213. Kennedy, J. *et al.* Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems. *J. Appl. Microbiol.* **111,** 787–799 (2011).

214. Sommer, M. O. A., Dantas, G. & Church, G. M. Functional Characterization of the Antibiotic Resistance Reservoir in the Human Microflora. *Science* **325,** 1128–1131 (2009).
215. Bouhajja, E., Agathos, S. N. & George, I. F. Metagenomics: Probing pollutant fate in natural and engineered ecosystems. *Biotechnol. Adv.* **34,** 1413–1426 (2016).
216. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
217. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
218. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21,** 487–493 (2011).
219. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinforma. Oxf. Engl.* **26,** 2460–2461 (2010).
220. Drezen, E., Durand, P. & Lavenier, D. KLAST, a Blast-like tool for fast sequence similarity searches. in (2014).
221. John, D. E., Zielinski, B. L. & Paul, J. H. Creation of a pilot metatranscriptome library from eukaryotic plankton of a eutrophic bay (Tampa Bay, Florida). *Limnol. Oceanogr. Methods* **7,** 249–259 (2009).
222. Warnecke, F. & Hess, M. A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. *J. Biotechnol.* **142,** 91–95 (2009).
223. Poretsky, R. S. *et al.* Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* **71,** 4121–4126 (2005).
224. Gilbert, J. A. *et al.* Detection of Large Numbers of Novel Sequences in the Metatranscriptomes of Complex Marine Microbial Communities. *PLOS ONE* **3,** e3042 (2008).
225. Bailly, J. *et al.* Soil eukaryotic functional diversity, a metatranscriptomic approach. *ISME J.* **1,** 632–642 (2007).
226. Grant, S. *et al.* Identification of Eukaryotic Open Reading Frames in Metagenomic cDNA Libraries Made from Environmental Samples. *Appl. Environ. Microbiol.* **72,** 135–143 (2006).
227. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E2329-2338 (2014).
228. Poretsky, R. S. *et al.* Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ. Microbiol.* **11,** 1358–1375 (2009).
229. Liang, P. & Pardee, A. B. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257,** 967–971 (1992).
230. Shrestha, P. M., Kube, M., Reinhardt, R. & Liesack, W. Transcriptional activity of paddy soil bacterial communities. *Environ. Microbiol.* **11,** 960–970 (2009).
231. Pace, N. R., Stahl, D. A., Lane, D. J. & Olsen, G. J. in *Advances in Microbial Ecology* (ed. Marshall, K. C.) 1–55 (Springer US, 1986).
232. Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.* **40,** 337–365 (1986).
233. Dusko Ehrlich, S. & MetaHIT consortium. [Metagenomics of the intestinal microbiota: potential applications]. *Gastroenterol. Clin. Biol.* **34 Suppl 1,** S23-28 (2010).
234. Delmont, T. O., Robe, P., Clark, I., Simonet, P. & Vogel, T. M. Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J. Microbiol. Methods* **86,** 397–400 (2011).
235. Monier, J.-M. *et al.* Metagenomic exploration of antibiotic resistance in soil. *Curr. Opin. Microbiol.* **14,** 229–235 (2011).

236. Alberti, A., Poulain, J. & Engelen, S. Marine plankton from viruses to metazoans: nucleotide sequences. *submited*
237. Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348,** 1261498 (2015).
238. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537,** 689–693 (2016).
239. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348,** 1261359 (2015).
240. Afshinnekoo, E. *et al.* Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Syst.* **1,** 72–87 (2015).
241. Poinar, H. N. *et al.* Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA. *Science* **311,** 392–394 (2006).
242. Inagaki, F. *et al.* Exploring deep microbial life in coal-bearing sediment down to ~2.5 km below the ocean floor. *Science* **349,** 420–424 (2015).
243. Haeckel, E. H. in *Oecologie und Chorologie* **II,** 286–289 (1866).
244. Möbius, K. *Die Auster und die Austernwirtschaft*. (Hempel & Parey, 1877).
245. DeLong, E. F. The microbial ocean from genomes to biomes. *Nature* **459,** 200–206 (2009).
246. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17,** 377–386 (2007).
247. Huson, D. H. *et al.* MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Comput. Biol.* **12,** e1004957 (2016).
248. Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9,** 386 (2008).
249. Markowitz, V. M. *et al.* IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* **36,** D534-538 (2008).
250. Davenport, C. F. *et al.* Genometa - A Fast and Accurate Classifier for Short Metagenomic Shotgun Reads. *PLOS ONE* **7,** e41224 (2012).
251. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
252. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25,** 1754–1760 (2009).
253. Freitas, T. A. K., Li, P.-E., Scholz, M. B. & Chain, P. S. G. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* (2015).
254. Dröge, J., Gregor, I. & McHardy, A. C. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinforma. Oxf. Engl.* **31,** 817–824 (2015).
255. Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11,** 80 (2010).
256. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).
257. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16,** 236 (2015).
258. Ames, S. K. *et al.* Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics* (2013).

259. Greenfield, N. & Minot, S. One Codex. (2014). Available at: https://www.onecodex.com/.
260. Lindgreen, S., Adair, K. L. & Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* **6,** 19233 (2016).
261. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9,** 811–814 (2012).
262. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
263. Liu, B., Gibbons, T., Ghodsi, M. & Pop, M. MetaPhyler: Taxonomic profiling for metagenomic sequences. in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 95–100 (2010). doi:10.1109/BIBM.2010.5706544
264. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10,** 1196–1199 (2013).
265. Vaulot, D., Eikrem, W., Viprey, M. & Moreau, H. The diversity of small eukaryotic phytoplankton (< or =3 microm) in marine ecosystems. *FEMS Microbiol. Rev.* **32,** 795–820 (2008).
266. Wilkins, D. *et al.* Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics. *Environ. Microbiol.* **15,** 1318–1333 (2013).
267. Hajibabaei, M. The golden age of DNA metasystematics. *Trends Genet. TIG* **28,** 535–537 (2012).
268. Yu, D. W. *et al.* Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol. Evol.* **3,** 613–623 (2012).
269. Whittaker, R. H. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol. Monogr.* **30,** 279–338 (1960).
270. Moreno, C. E. & Rodríguez, P. A consistent terminology for quantifying species diversity? *Oecologia* **163,** 279–282 (2010).
271. Anderson, M. J. *et al.* Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecol. Lett.* **14,** 19–28 (2011).
272. Vellend, M. Do commonly used indices of β-diversity measure species turnover? *J. Veg. Sci.* **12,** 545–552 (2001).
273. Whittaker, R. H. Evolution and Measurement of Species Diversity. *Taxon* **21,** 213–251 (1972).
274. Sørensen, T. J. *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons.* (I kommission hos E. Munksgaard, 1948).
275. Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* **27,** 326–349 (1957).
276. Obura, D. The Diversity and Biogeography of Western Indian Ocean Reef-Building Corals. *PLOS ONE* **7,** e45013 (2012).
277. Primo, C. & Vázquez, E. Zoogeography of the southern African ascidian fauna. *J. Biogeogr.* **31,** 1987–2009 (2004).
278. Cermeño, P., Vargas, C. de, Abrantes, F. & Falkowski, P. G. Phytoplankton Biogeography and Community Stability in the Ocean. *PLOS ONE* **5,** e10037 (2010).
279. De Klerk, H. M., Crowe, T. M., Fjeldså, J. & Burgess, N. D. Biogeographical patterns of endemic terrestrial Afrotropical birds. *Divers. Distrib.* **8,** 147–162 (2002).
280. Kreft, H. & Jetz, W. A framework for delineating biogeographical regions based on species distributions. *J. Biogeogr.* **37,** 2029–2053 (2010).
281. Milici, M. *et al.* Bacterioplankton Biogeography of the Atlantic Ocean: A Case Study of the Distance-Decay Relationship. *Aquat. Microbiol.* 590 (2016).

282. Bendif, E. M. *et al.* Genetic delineation between and within the widespread coccolithophore morpho-species Emiliania huxleyi and Gephyrocapsa oceanica (Haptophyta). *J. Phycol.* **50,** 140–148 (2014).
283. Dutilh, B. E. *et al.* Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinforma. Oxf. Engl.* **28,** 3225–3231 (2012).
284. Fimereli, D., Detours, V. & Konopka, T. TriageTools: tools for partitioning and prioritizing analysis of high-throughput sequencing data. *Nucleic Acids Res.* **41,** e86 (2013).
285. Maillet, N., Lemaitre, C., Chikhi, R., Lavenier, D. & Peterlongo, P. Compareads: comparing huge metagenomic experiments. *BMC Bioinformatics* **13,** S10 (2012).
286. Bloom, B. H. Space/Time Trade-offs in Hash Coding with Allowable Errors. *Commun ACM* **13,** 422–426 (1970).
287. Jaccard, P. The Distribution of the Flora in the Alpine Zone.1. *New Phytol.* **11,** 37–50 (1912).
288. Seth, S., Välimäki, N., Kaski, S. & Honkela, A. Exploration and retrieval of whole-metagenome sequencing samples. *Bioinformatics* **30,** 2471–2479 (2014).
289. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinforma. Oxf. Engl.* **27,** 764–770 (2011).
290. Rizk, G., Lavenier, D. & Chikhi, R. DSK: k-mer counting with very low memory usage. *Bioinforma. Oxf. Engl.* **29,** 652–653 (2013).
291. Välimäki, N. & Puglisi, S. J. Distributed String Mining for High-Throughput Sequencing Data. in *Algorithms in Bioinformatics* (eds. Raphael, B. & Tang, J.) 441–452 (Springer Berlin Heidelberg, 2012).
292. Ulyantsev, V. I., Kazakov, S. V., Dubinkina, V. B., Tyakht, A. V. & Alexeev, D. G. MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data. *Bioinformatics* btw312 (2016). doi:10.1093/bioinformatics/btw312
293. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17,** 132 (2016).
294. Broder, A. On the Resemblance and Containment of Documents. in *Proceedings of the Compression and Complexity of Sequences 1997* 21– (IEEE Computer Society, 1997).
295. Narayanan, M. & Karp, R. M. Gapped Local Similarity Search with Provable Guarantees. in *Algorithms in Bioinformatics* (eds. Jonassen, I. & Kim, J.) 74–86 (Springer Berlin Heidelberg, 2004).
296. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33,** 623–630 (2015).
297. Drew, J. & Hahsler, M. Strand: Fast Sequence Comparison Using Mapreduce and Locality Sensitive Hashing. in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* 506–513 (ACM, 2014). doi:10.1145/2649387.2649436
298. Rasheed, Z. & Rangwala, H. A Map-Reduce Framework for Clustering Metagenomes. in *2013 IEEE International Symposium on Parallel Distributed Processing, Workshops and Phd Forum* 549–558 (2013). doi:10.1109/IPDPSW.2013.100
299. Benoit, G. *et al.* Multiple Comparative Metagenomics using Multiset k-mer Counting. *peerj* (2016).
300. Karlusich, J. J. P., Ceccoli, R. D., Graña, M., Romero, H. & Carrillo, N. Environmental selection pressures related to iron utilization are involved in the loss of the flavodoxin gene from the plant genome. *Genome Biol. Evol.* evv031 (2015). doi:10.1093/gbe/evv031

301. Monier, A. *et al.* Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. *Environ. Microbiol.* **14,** 162–176 (2012).
302. Domazet-Loso, T. & Tautz, D. An evolutionary analysis of orphan genes in Drosophila. *Genome Res.* **13,** 2213–2219 (2003).
303. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. G. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet. TIG* **25,** 404–413 (2009).
304. Long, M., Betrán, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4,** 865–875 (2003).
305. Colbourne, J. K. *et al.* The ecoresponsive genome of Daphnia pulex. *Science* **331,** 555–561 (2011).
306. Khalturin, K. *et al.* A Novel Gene Family Controls Species-Specific Morphological Traits in Hydra. *PLOS Biol.* **6,** e278 (2008).
307. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12,** 692–702 (2011).
308. Wilson, G. A. *et al.* Orphans as taxonomically restricted and ecologically important genes. *Microbiol. Read. Engl.* **151,** 2499–2501 (2005).
309. Jaillon, O. Marine plankton reveals a large unknown side of the eukaryotic gene repertoire. *(submited)*
310. Vandepoele, K. *et al.* pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environ. Microbiol.* **15,** 2147–2153 (2013).
311. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31,** 926–932 (2015).
312. Palmieri, N., Kosiol, C. & Schlötterer, C. The life cycle of Drosophila orphan genes. *eLife* **3,** (2014).
313. Toll-Riera, M., Castelo, R., Bellora, N. & Albà, M. M. Evolution of primate orphan proteins. *Biochem. Soc. Trans.* **37,** 778–782 (2009).
314. Cai, J. J. & Petrov, D. A. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol. Evol.* **2,** 393–409 (2010).
315. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* **487,** 370–374 (2012).
316. Acuña, L. G. *et al.* Architecture and Gene Repertoire of the Flexible Genome of the Extreme Acidophile Acidithiobacillus caldus. *PLOS ONE* **8,** e78237 (2013).
317. Coleman, M. L. *et al.* Genomic islands and the ecology and evolution of Prochlorococcus. *Science* **311,** 1768–1770 (2006).
318. Fernández-Gómez, B. *et al.* Patterns and architecture of genomic islands in marine bacteria. *BMC Genomics* **13,** 347 (2012).
319. Gonzaga, A. *et al.* Polyclonality of concurrent natural populations of Alteromonas macleodii. *Genome Biol. Evol.* **4,** 1360–1374 (2012).
320. Kashtan, N. *et al.* Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. *Science* **344,** 416–420 (2014).
321. Rocap, G. *et al.* Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424,** 1042–1047 (2003).
322. Rodríguez-Martínez, R., Rocap, G., Salazar, G. & Massana, R. Biogeography of the uncultured marine picoeukaryote MAST-4: temperature-driven distribution patterns. *ISME J.* **7,** 1531–1543 (2013).

323. Lin, Y.-C. *et al.* Distribution patterns and phylogeny of marine stramenopiles in the north pacific ocean. *Appl. Environ. Microbiol.* **78,** 3387–3399 (2012).
324. Yau, S. *et al.* A Viral Immunity Chromosome in the Marine Picoeukaryote, Ostreococcus tauri. *PLOS Pathog* **12,** e1005965 (2016).
325. Derelle, E. *et al.* Life-Cycle and Genome of OtV5, a Large DNA Virus of the Pelagic Marine Unicellular Green Alga Ostreococcus tauri. *PLOS ONE* **3,** e2250 (2008).
326. Derelle, E. *et al.* Diversity of Viruses Infecting the Green Microalga Ostreococcus lucimarinus. *J. Virol.* **89,** 5812–5821 (2015).
327. Moreau, H. *et al.* Marine Prasinovirus Genomes Show Low Evolutionary Divergence and Acquisition of Protein Metabolism Genes by Horizontal Gene Transfer. *J. Virol.* **84,** 12555–12563 (2010).
328. Weynberg, K. D., Allen, M. J., Gilg, I. C., Scanlan, D. J. & Wilson, W. H. Genome Sequence of Ostreococcus tauri Virus OtV-2 Throws Light on the Role of Picoeukaryote Niche Separation in the Ocean. *J. Virol.* **85,** 4520–4529 (2011).
329. Ramette, A. & Tiedje, J. M. Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microb. Ecol.* **53,** 197–207 (2007).
330. Van Etten, J. L., Lane, L. C. & Dunigan, D. D. DNA viruses: the really big ones (giruses). *Annu. Rev. Microbiol.* **64,** 83–99 (2010).
331. Dodds, J. A. Viruses of marine algae. *Experientia* **35,** 440–442 (1979).
332. Bratbak, G., Thingstad, F. & Heldal, M. Viruses and the microbial loop. *Microb. Ecol.* **28,** 209–221 (1994).
333. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nat. Rev. Microbiol.* **5,** 801–812 (2007).
334. Mojica, K. D. A. & Brussaard, C. P. D. Factors affecting virus dynamics and microbial host–virus interactions in marine environments. *FEMS Microbiol. Ecol.* **89,** 495–515 (2014).
335. Parker, A. Atlantic Meridional Overturning Circulation is stable under global warming. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E2760–E2761 (2016).
336. Lozier, M. S. Overturning in the North Atlantic. *Annu. Rev. Mar. Sci.* **4,** 291–315 (2012).
337. Rahmstorf, S. *et al.* Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. *Nat. Clim. Change* **5,** 475–480 (2015).
338. Beaugrand, G., Reid, P. C., Ibañez, F., Lindley, J. A. & Edwards, M. Reorganization of North Atlantic Marine Copepod Biodiversity and Climate. *Science* **296,** 1692–1694 (2002).
339. Bopp, L. *et al.* Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models. *Biogeosciences* **10,** 6225–6245 (2013).
340. Cornils, A., Wend-Heckmann, B. & Held, C. Global phylogeography of Oithona similis s.l. (Crustacea, Copepoda, Oithonidae) – A cosmopolitan plankton species or a complex of cryptic lineages? *Mol. Phylogenet. Evol.* **107,** 473–485 (2017).

Références des ouvrages qui m'ont inspiré pendant la rédaction de cette thèse :

Grégoire Trégouboff. Manuel de planctologie méditerranéenne. Centre National de la Recherche Scientifique. Paris. Tome 1, 1957.

Alan F. Chalmers. Qu'est-ce que la science? Le Livre de Poche. Edition 14, 2015.

Éric Karsenti et Dino Di Meo. Tara Oceans, Chroniques d'une expédition scientifique. Actes SUD. 2012.

Noan Le Bescot. Patrons de biodiversité à l'échelle globale chez les dinoflagellés planctoniques marins. Biodiversité. Université Pierre et Marie Curie - Paris VI, 2014.

Tristan Biard. Diversité, biogéographie et écologie des Collodaires (Radiolaires) dans l'océan mondial. Océanographie. Université Pierre et Marie Curie - Paris VI, 2015.

Jennifer Sengenès. Développement de méthodes de séquençage de seconde génération pour l'analyse des profils de méthylation de l'ADN. Biologie moléculaire. Université Pierre et Marie Curie - Paris VI, 2012.

Nicolas Maillet. Comparaison *de novo* de données de séquençage issues de très grands échantillons métagénomiques. Informatique. Université de Rennes 1, 2013.

Pierre Peterlongo. Lire les lectures : analyse de données de séquençage. INRIA Rennes Bretagne Atlantique. Equipe projet GenScale, 2016.

Artem Kourlaiev. Évolution du système de gestion des chaines de traitements du Genoscope. Université Aix Marseille, 2014.

Annexes

Annexe I. Informations supplémentaires de l'article : Survey of the green picoalga *Bathycoccus* genomes in the global ocean.

Supplementary Information for

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

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

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

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

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

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

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

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

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

Genomic data

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

Single-cell isolation and amplification

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

DNA sequencing and assembly

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

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

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

Gene prediction on the TOSAG39-1 assembly

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

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

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

TOSAG39-1 and RCC1105 genomic comparison

Best reciprocal hits (BRH)

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

Synteny and collinear genes analysis

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

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

Comparison between *Bathycoccus* genomes and MMETSP transcriptome

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

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

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

Metagenomic fragment recruitment

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

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

Gene set filtering

Mitochondrial and plastid genes

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

Foreign sequences in TOSAG39-1 assembly

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

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

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

Estimation of cross-species mapped genes

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

Abundance counts

Relative genomic abundance

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

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

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

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

Metabarcoding

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

Analyses of dispensable genes

Identification and characterization

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

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

Validation of dispensable cassette genes in metagenomes

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

Analysis of environmental parameters

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

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

$$x = \log(Chl)$$

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

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

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

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

rRNA operon comparison

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

Functional analysis of dispensable genes

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

Supplementary Figures

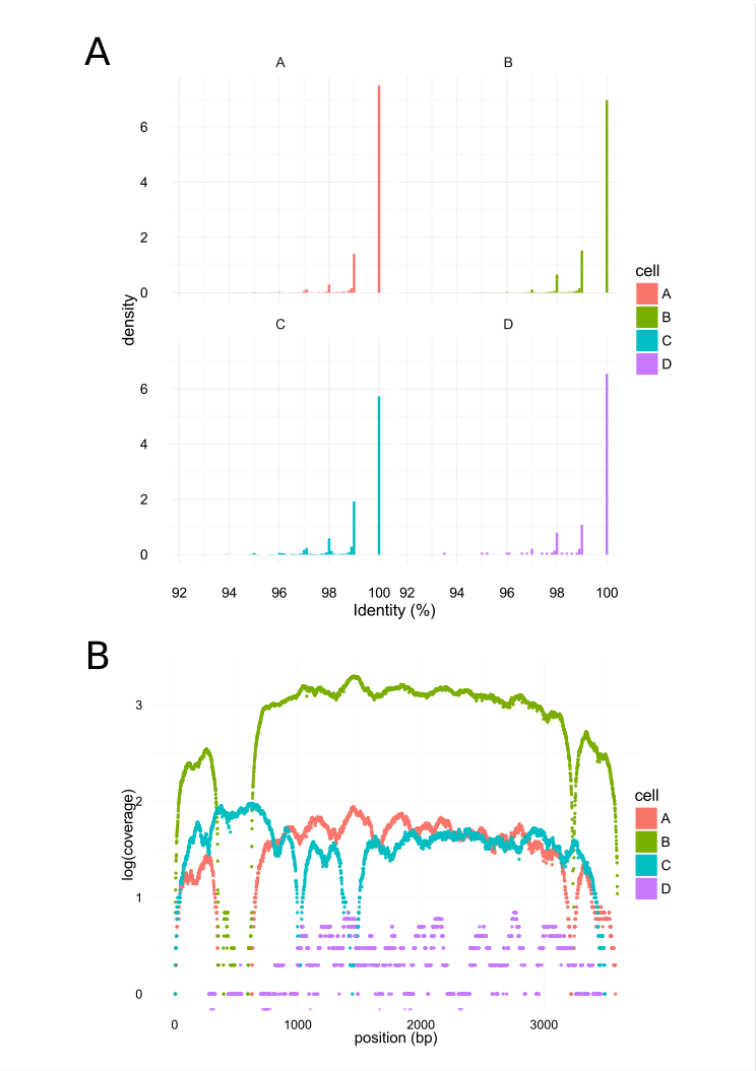


Figure S1. A. Distribution of identity percent of reads from each individual cell A (red), B (green), C (blue) and D (purple) once mapped onto the final combined SAG assembly. B. Example of the contributions of reads of each cell A (red), B (green), C (blue) and D (purple) along one contig of the final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

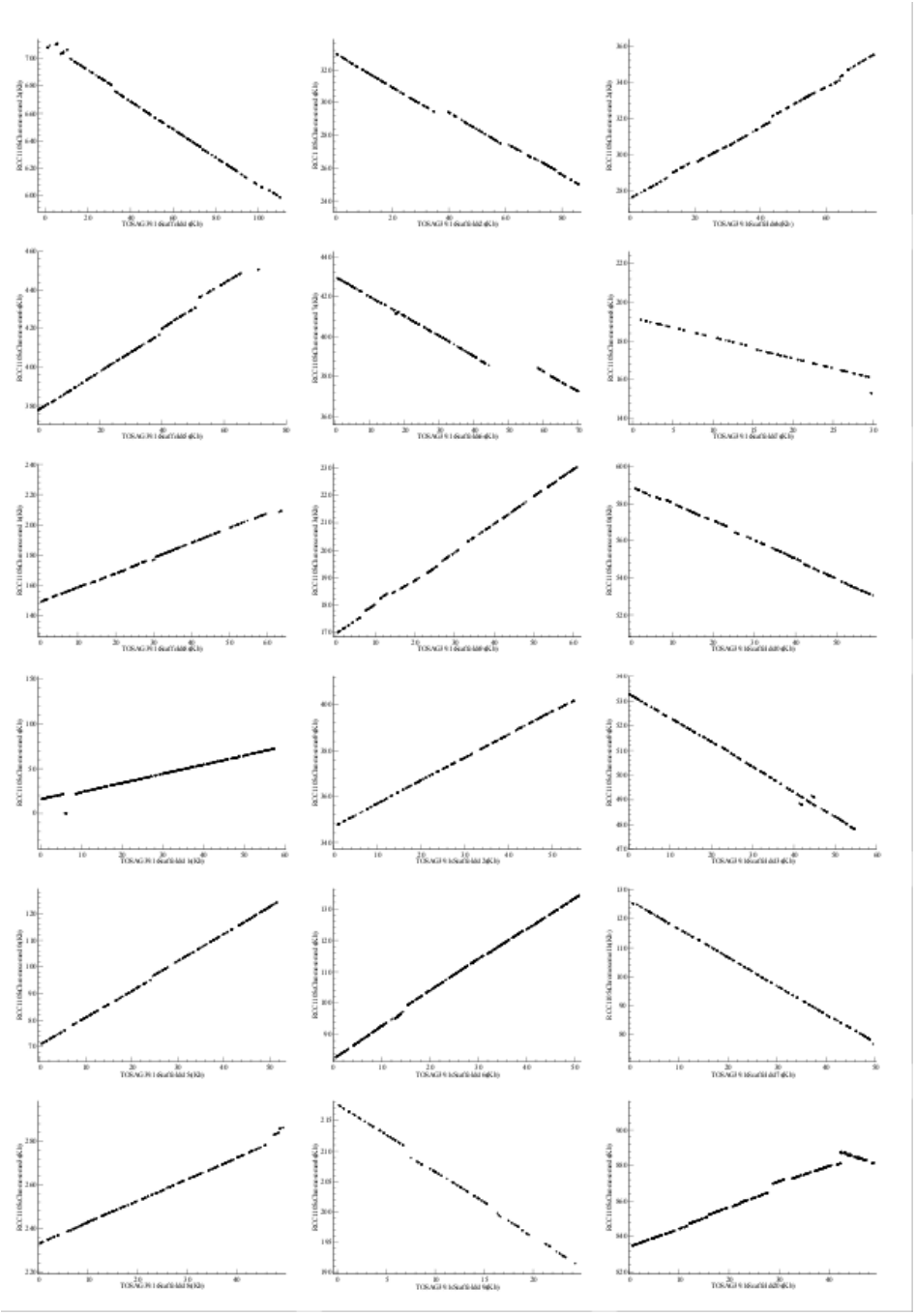


Figure S2. Synteny conservation between the two *B*. *prasinos* genomes. The RCC and TOSAG39-1 genomes are displayed on the X- and Y-axis, respectively. Dots correspond to regions conserved at the protein level (tblastx hits). Only the 18 longest scaffolds of TOSAG39-1 are represented. The two genomes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

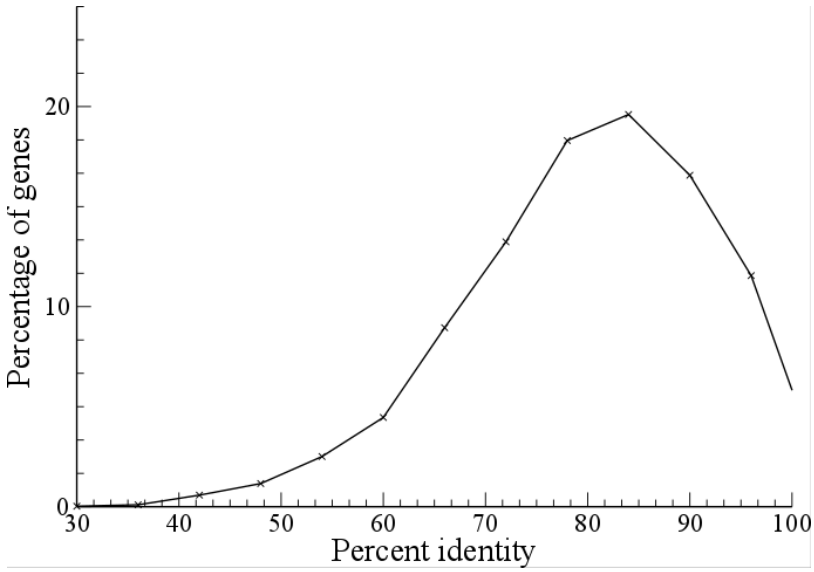


Figure S3. Distribution of orthologous gene divergence at the protein level between *Bathycoccus* RCC1105 and TOSAG39-1.

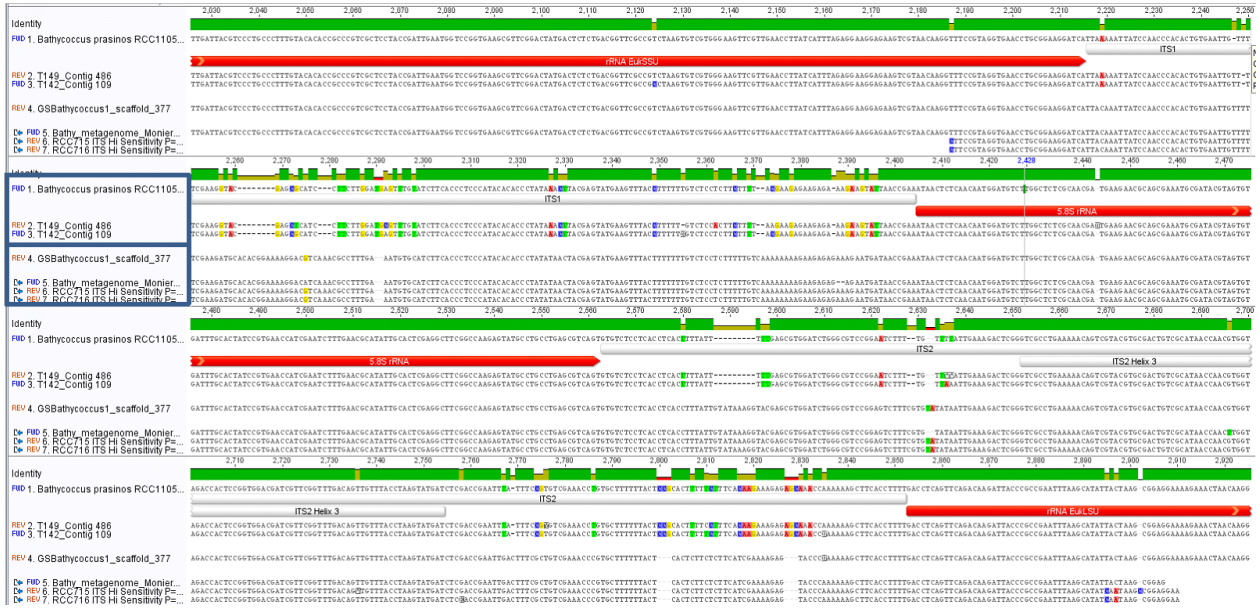


Figure S4. rRNA operon. Comparison of the rRNA ITS region between the two *Bathycoccus* genomes. RCC1105 and two metagenomes from the Chile upwelling[3] share identical ITS1 and ITS2, while TOSAG39-1 ITSs are identical to those of a metagenome from the Atlantic Ocean DCM[4] and to those from strains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

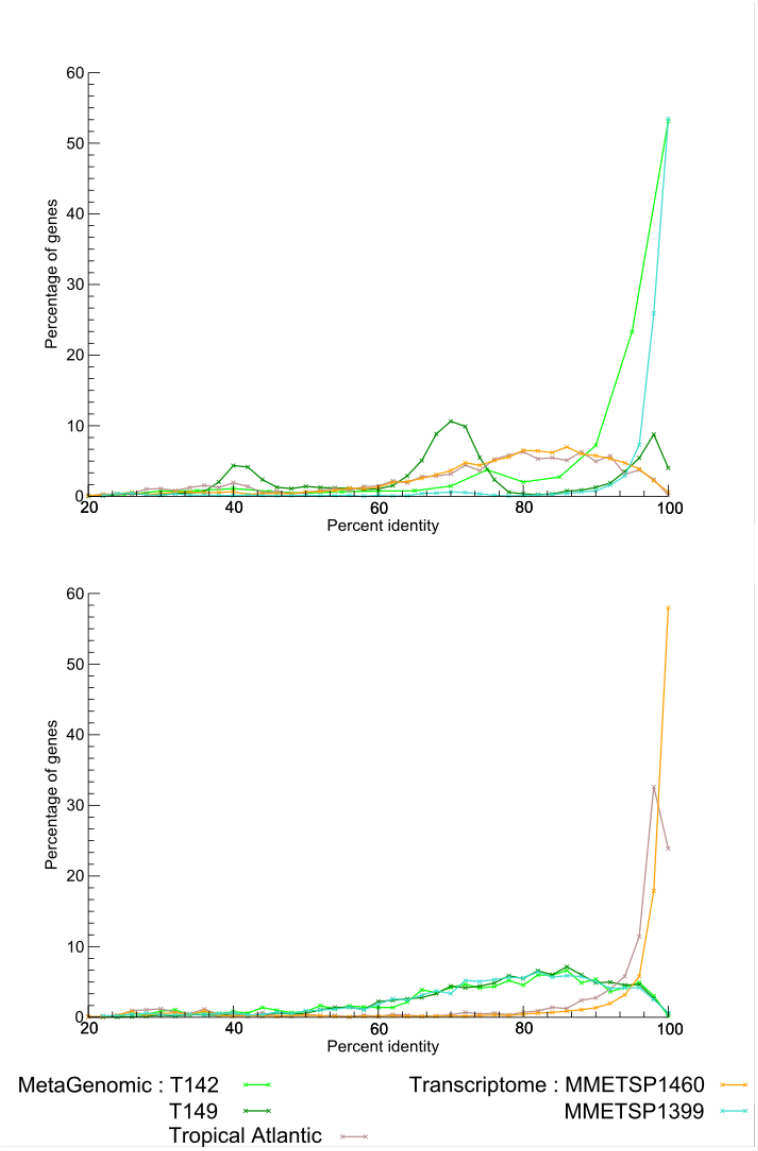


Figure S5. Affiliations of three metagenomes containing *Bathycoccus* and two *Bathycoccus* transcriptomes of the MMETSP database to the two genome assemblies. Distributions correspond to similarities at the amino acid level for one *Bathycoccus* genome assembly (top: RCC1105, bottom: TOSAG39-1) with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

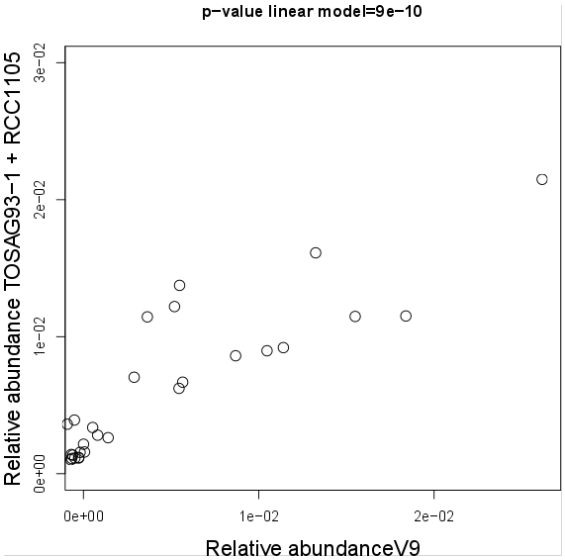


Figure S6. Correlation between the abundance of *Bathycoccus* estimated from whole metagenomes (two genomes summed) and V9 amplicons abundances.

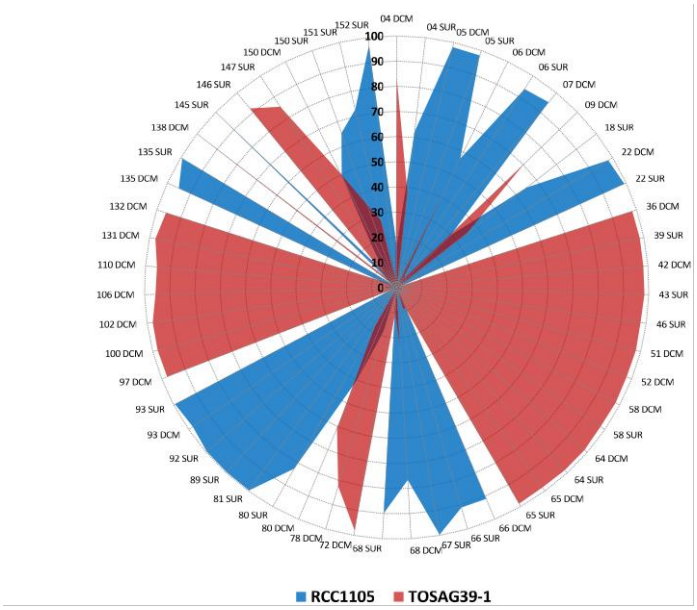


Figure S7. Relative contribution of each genome at *Bathycoccus*-rich stations. Within the 58 DCM and surface samples where *Bathycoccus* metagenomic abundance represents more than 0.01%, one of the two *Bathycoccus* genome was dominant (>70% of the *Bathycoccus* metagenomic reads) in 91% of the cases.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

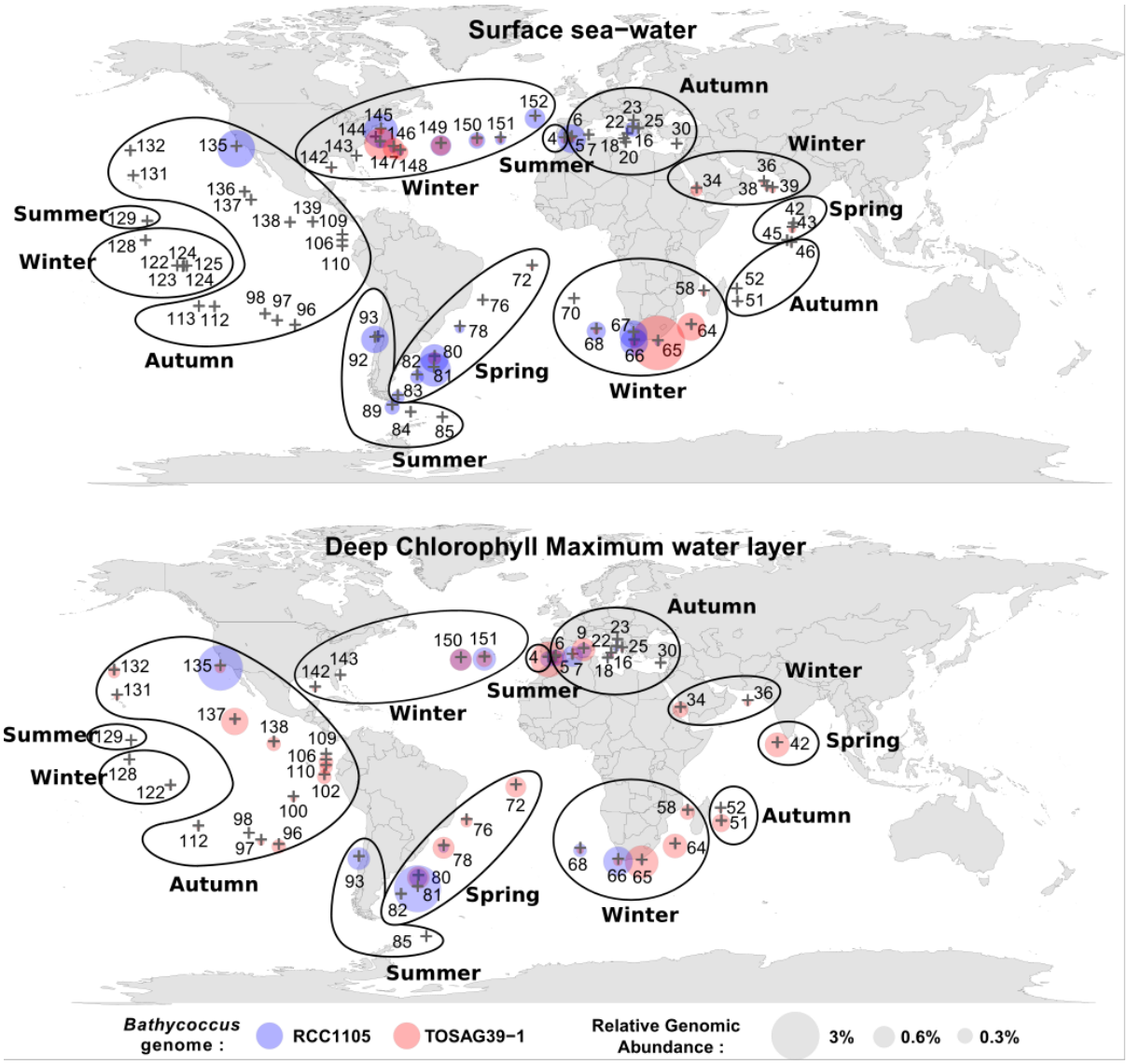


Figure S8. Map of relative metagenomic abundances of the two *Bathycoccus* in *Tara* Oceans stations with sampling season. This map was created using R-package maps_2.1-6, mapproj_1.1-8.3, gplots_2.8.0 and mapplots_1.4 (version R-2.13, https://cran.r-project.org/web/packages/maps/index.html).

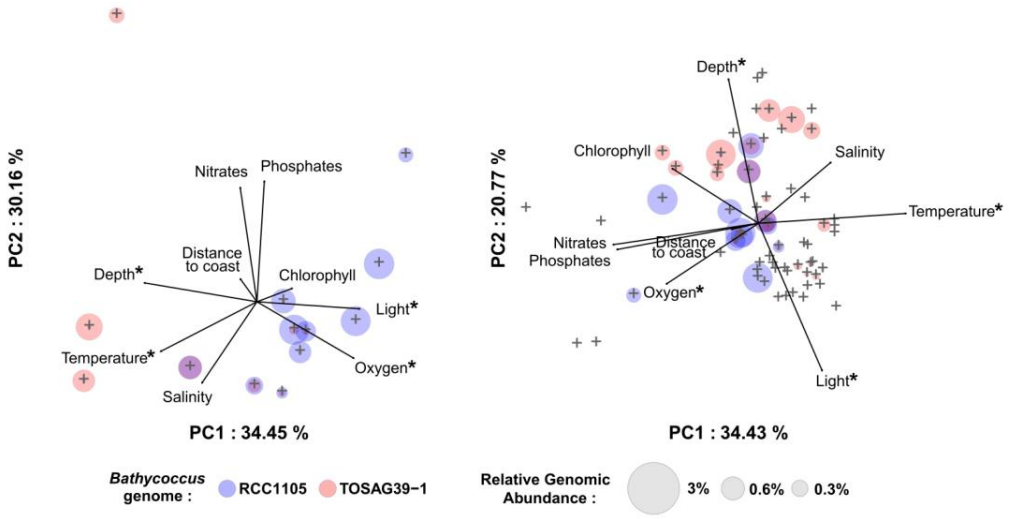


Figure S9. Principal Component Analysis including Photosynthetically Active Radiation (PAR). Left: Using only 13 samples for which we measured a large relative genomic abundance of *Bathycoccus* that have available PAR (indicated as light). Right: Idem but with all *Tara* Oceans samples that have available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

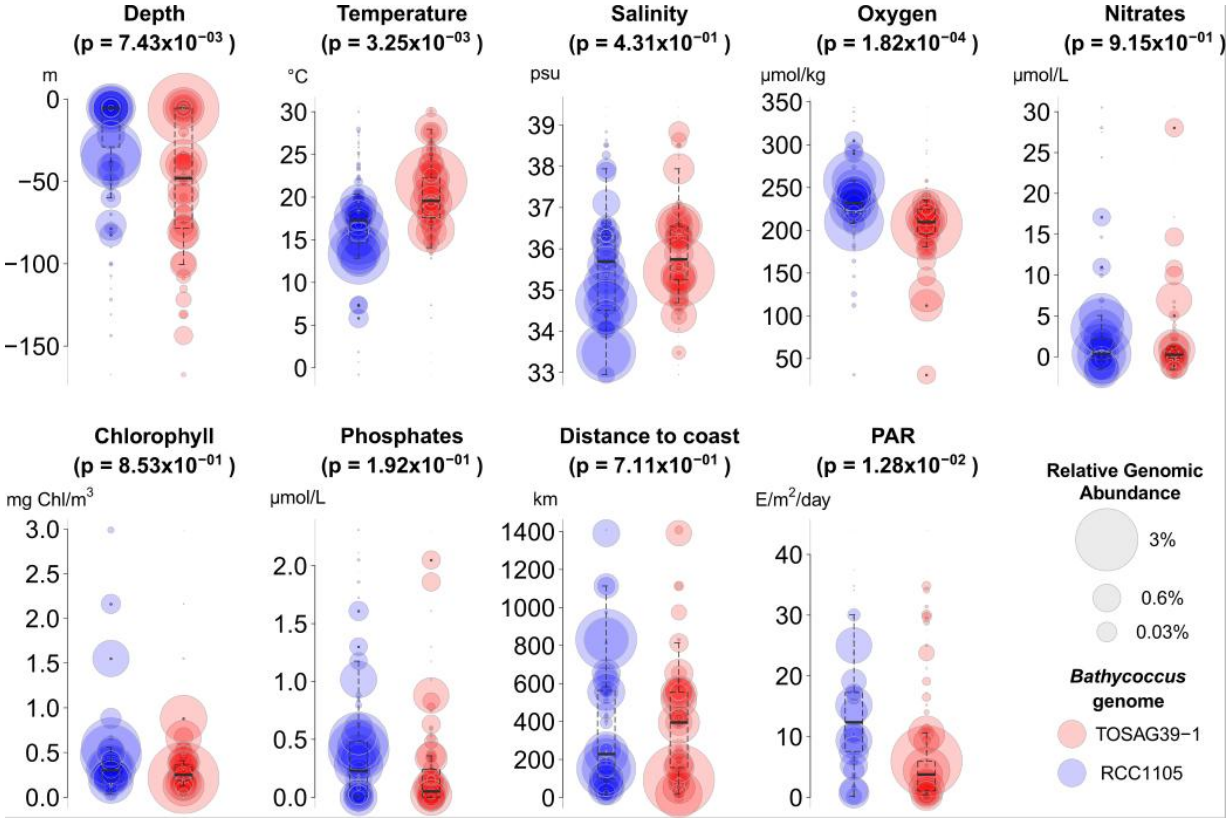


Figure S10. Environmental parameters and genomic abundances of *Bathycoccus*. PAR (Photosynthetically Active Radiation) corresponds to AMODIS satellite data for surface samples and to computed estimations for DCM samples. Temperature, oxygen, depth and light are parameters that gave significantly different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

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

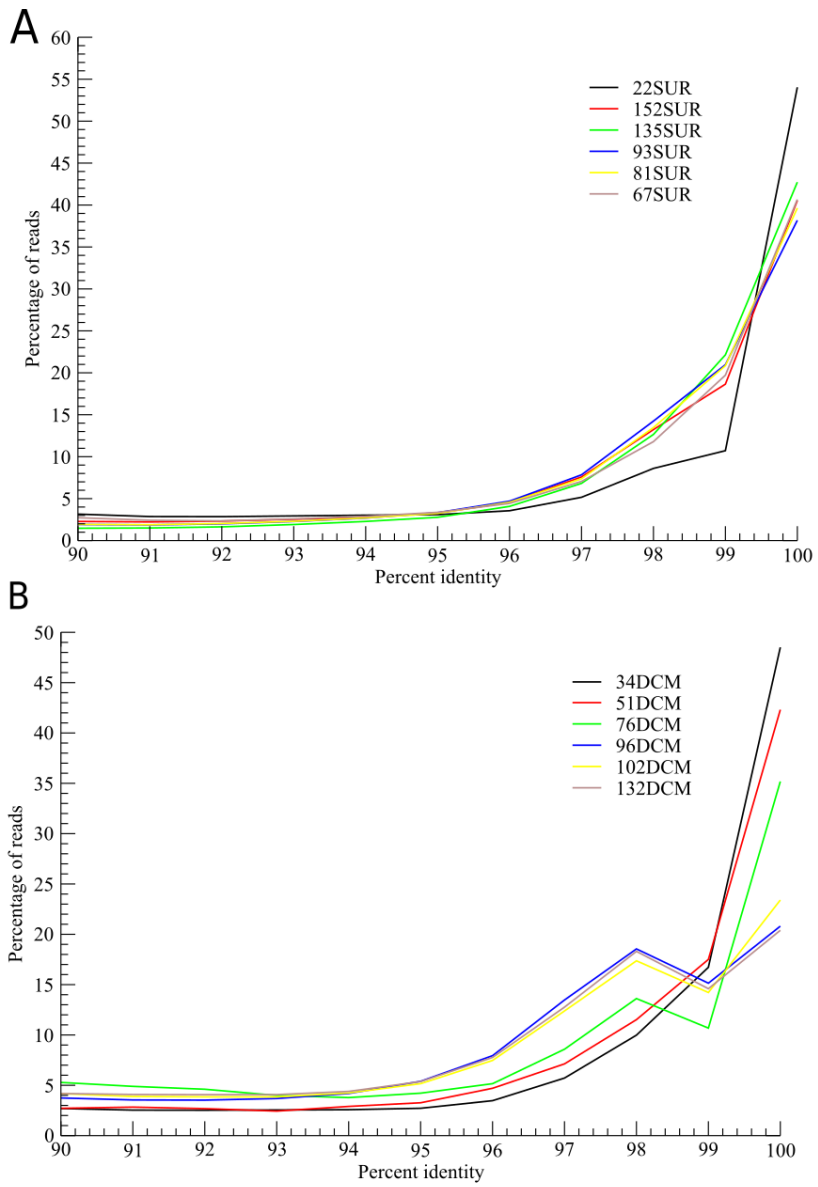


Figure S12. Distribution of identity percent of *Tara* Oceans metagenomic reads mapped onto RCC1105 genome (A) and TOSAG39-1 assembly (B). We only used *Tara* Oeans samples where the presence of only one *Bathycoccus* genome was detected.

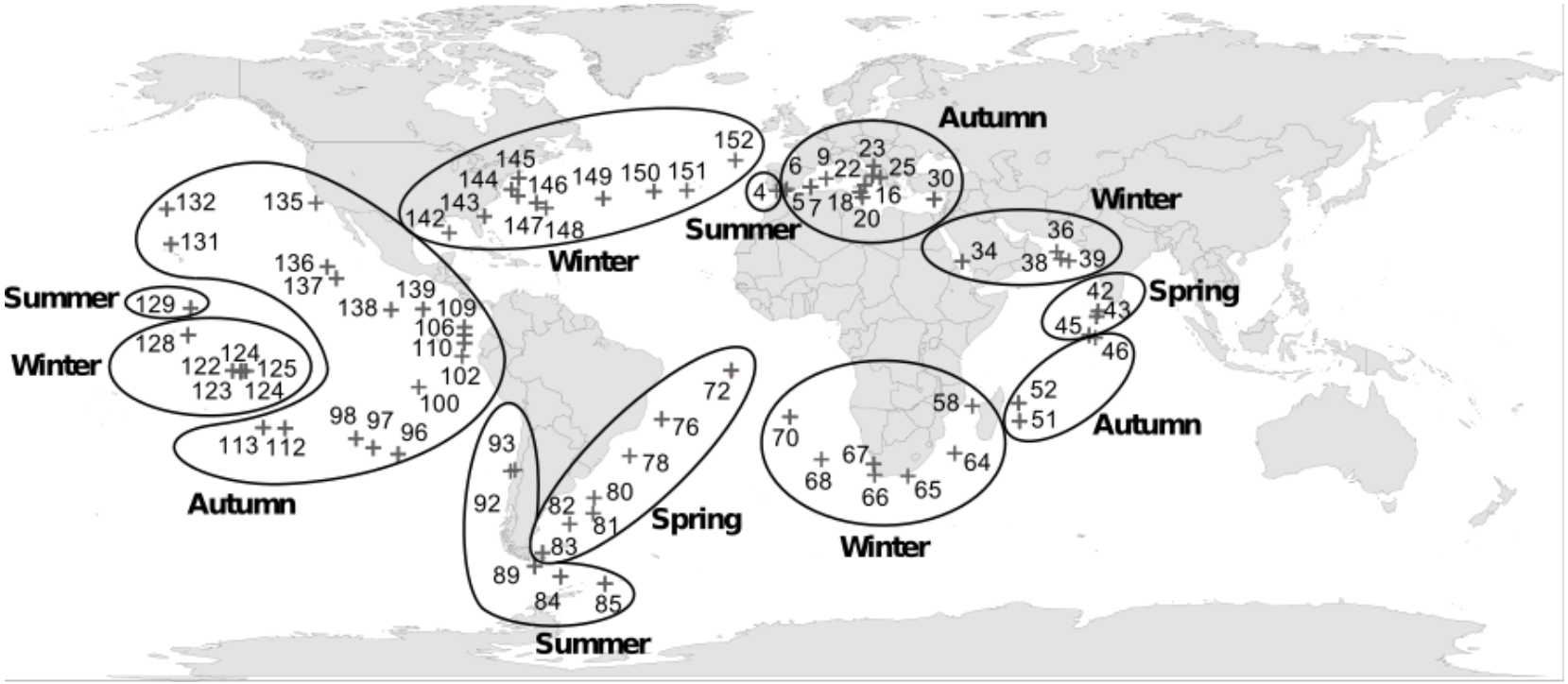


Figure S13. Map of the stations of the *Tara* Oceans expedition with seasons when sampled. This map was created using R-package maps_2.1-6, mapproj_1.1-8.3, gplots_2.8.0 and mapplots_1.4 (version R-2.13, https://cran.r-project.org/web/packages/maps/index.html).

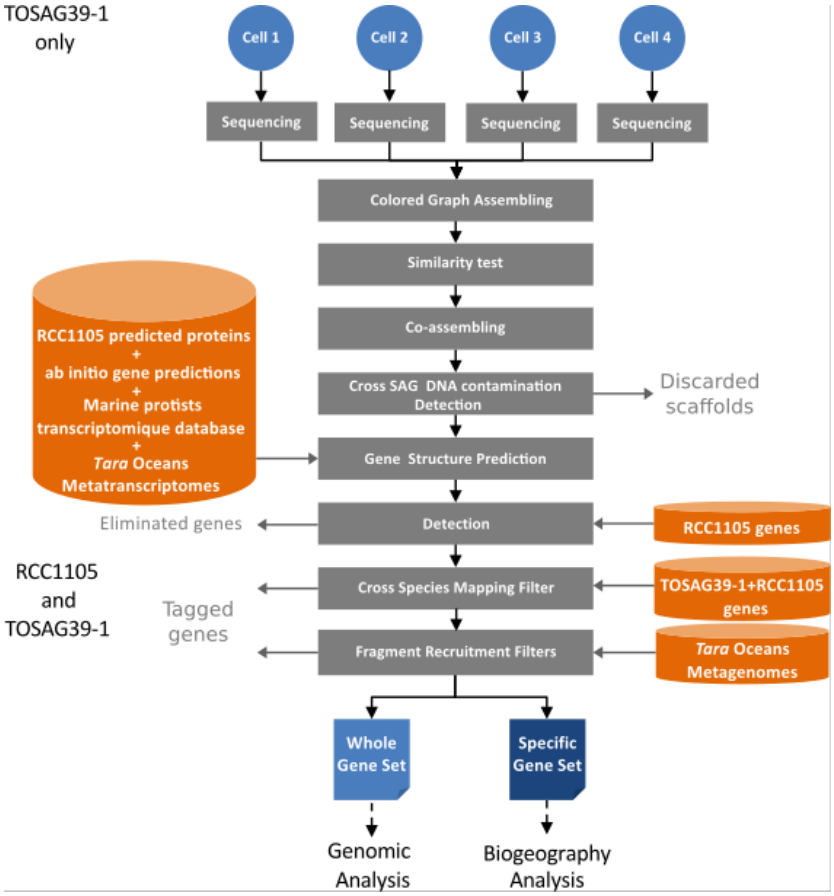


Figure S14. Pipeline for data acquisition and quality control.

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

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

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

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

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

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

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

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

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

Table S4. Dispensable genes of RCC1105.

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

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

Table S5. Dispensable genes of TOSAG39-1.

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

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

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

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

References

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

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

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

Annexe II. Informations supplémentaires de l'article : Environmental characteristics of Agulhas rings affect interocean plankton transport.

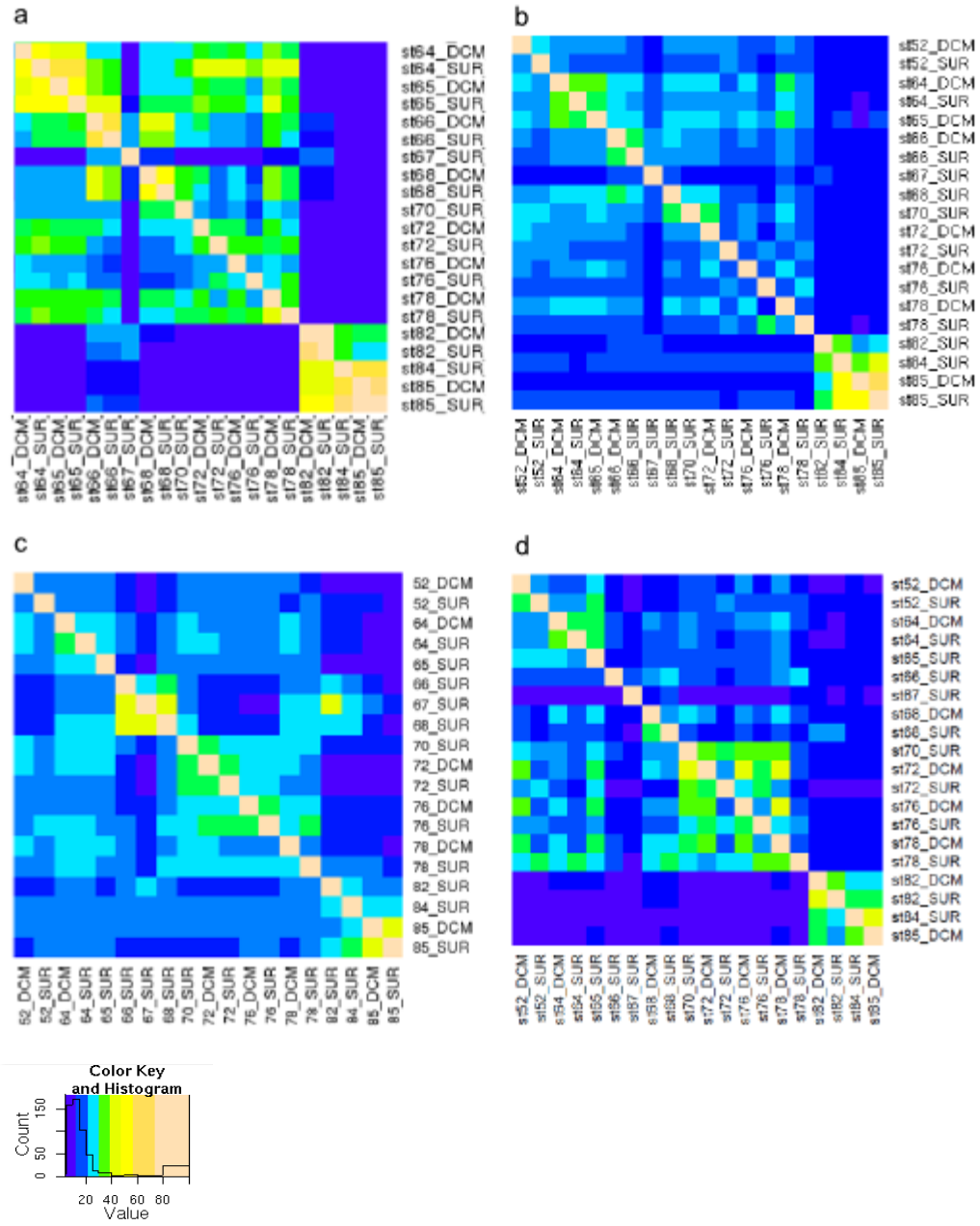


Figure annexe II | Similarité génomique entre les échantillons récoltés dans l'océan Indien, Atlantique Sud et Austral. Le pourcentage de similarité a été calculé avec l'outil *Compareads* sur les échantillons de surface et DCM pour les filtres :
a, 0.22 à 3 µm ; **b,** 0.8 à 5 µm ; **c,** 20 à 180 µm et **d,** 180 à 2000 µm.

Annexe III. Informations supplémentaires de l'article : Global plankton biogeography is shaped via ocean circulation dynamics.

Titre: Dynamique de la structure des génomes et de leur biogéographie dans l'océan : analyses comparatives des données métagénomiques du projet *Tara* Oceans pour l'étude de la microalgue *Bathycoccus* et des communautés planctoniques globales.

Mots clés: métagénomique, génomique comparative, *Tara* Oceans, *Bathycoccus prasinos*, biogéographie, plancton

Résumé: Le plancton représente l'ensemble des organismes qui dérivent le long des courants marins. Par sa partie phytoplancton, il produit autant d'oxygène que toutes les plantes terrestres et est, à travers le cycle du carbone, un important régulateur de la machine climatique ainsi que de l'acidité des océans. De plus, il est à la base de la chaîne alimentaire. L'écosystème planctonique joue donc un rôle important dans les équilibres nécessaires à la vie sur Terre. Pourtant, celui-ci reste peu connu. Avec le développement du séquençage haut débit et de la métagénomique, il est maintenant possible d'étudier les séquences d'ADN des micro-organismes présents dans des échantillons issus de l'océan. Le projet *Tara* Oceans (2009-2012) est la première expédition à avoir réalisé une collecte et un séquençage des micro-organismes planctoniques présents dans les eaux de surface à l'échelle de la planète tout en intégrant des mesures environnementales.

Cette thèse consiste à étudier l'organisation génomique de l'écosystème planctonique dans les eaux océaniques de surface. Pour cela, il a été utilisé les séquences d'ADN de 644 échantillons métagénomiques correspondant à 6 fractions de taille d'organisme planctonique, allant des virus aux petits métazoaires, ainsi que les données environnementales des 113 stations *Tara* Oceans correspondantes. L'objectif étant de mieux comprendre dans quelle mesure les facteurs océaniques impactent la dynamique de la structure des génomes et de leur biogéographie dans l'océan.

L'étude de la diversité génomique et spatiale du phytoplancton *Bathycoccus prasinos* a été réalisée à partir des séquences du génome de référence et d'une partie d'un second génome obtenu lors de l'expédition par une méthode d'amplification à cellule unique (SAG). La comparaison de ces deux génomes partageant la même séquence de l'ARNr 18S a révélé qu'il s'agissait de deux espèces distinctes de *Bathycoccus*. Une analyse de métagénomique ciblée a permis de décrire la biogéographie de ces deux écotypes qui sont présents dans des environnements différents. Enfin, cette analyse a révélé une variabilité du contenu en gènes dans les différents échantillons ce qui induit une grande plasticité génomique au sein d'une même espèce.

Il est nécessaire de passer a un niveau global pour étudier l'organisation des communautés planctoniques dans les océans. La métagénomique comparative sur l'ensemble des échantillons *Tara* Oceans et sur les différentes fractions de tailles d'organismes permet ce passage à large échelle. L'évaluation de l'outil *Compareads* permettant de connaître la similarité en lectures entre deux jeux de données métagénomiques a été réalisée sur une partie des échantillons du projet *Tara* Oceans. Cette analyse a montré qu'il était possible avec les données métagénomiques d'étudier les modifications de la diversité génomique des communautés micro-planctoniques dans différents océans. L'analyse de la variabilité génomique le long de grands courants océaniques est alors envisageable. Un travail collaboratif pour l'amélioration de *Compareads* a permis le développement de l'outil *COMMET* qui, associé à des super calculateurs permet de réaliser les comparaisons de l'ensemble des échantillons *Tara* Oceans. Avec une heuristique similaire, le calcul de distances de diversité beta entre les échantillons métagénomiques a permis de proposer la première biogéographie des communautés virales, bactériennes et eucaryotes. Il a été démontré que les courants océaniques et les variations physico-chimiques ont un impact différent sur l'organisation génomique des communautés micro-planctoniques qui serait plus ou moins important selon l'échelle de temps et la taille des micro-organismes.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

Title: Dynamic of the structure of the genomes and of their biogeography: comparative analysis of *Tara* Oceans metagenomic data to study the *Bathycoccus* microalgeae and global planktonic communities.

Key words: metagenomic, comparative genomic, *Tara* Oceans, *Bathycoccus prasinos*, biogeography, plankton

Abstract: Plankton is composed of all organisms that drift along the currents. Through its phytoplankton part, it produces as much oxygen as all terrestrial plants and it is an important regulator of the climatic system as well as the acidity of the oceans. It is also at the base of the food web. The planktonic ecosystem plays an important role in the necessary balances for life on Earth, however it remains poorly know. With the progress of high-throughput sequencing in the last few years and with the metagenomic approach it is possible to study the DNA sequences of micro-organisms directly sampled from the ocean. The *Tara* Oceans expedition (2009-2012) performed a vast sampling expedition of plankton in all oceans that collected *in situ* environmental parameters and sequenced planktonic organisms from surface water.

This thesis consists in studying the genomic organization of the planktonic ecosystem in the surface ocean waters. For this purpose, the DNA sequences of 644 metagenomic samples corresponding to 6 plankton size fractions, ranging from viruses to the small metazoan, as well as environmental data from the 113 corresponding *Tara* Oceans stations are used. The aim is to understand at the scale of individual organism and of global micro-planktonic communities how the genomic structure and the geographic distribution is influenced by ocean circulation and environmental variations.

The study of the genomic and geographic diversity of the phytoplankton *Bathycoccus prasinos* was done using the sequences of a reference genome and a part of a second genome obtained during the expedition with the Single Amplification Genome method (SAG). The comparison of these two genomes sharing the same 18S rRNA revealed that they were two distinct species of *Bathycoccus*. A targeted metagenomic analysis helped to describe the biogeography of these two ecotypes that are present in different environments. Finally, this analysis showed variability in the gene content in samples which shows the existence of a strong genomic plasticity within species.

It is necessary to scale up to global scale to study the plankton community's organization in the oceans. The comparison of metagenomes of all *Tara* Oceans samples and at different size fractions of organisms allows this passage to a large scale. The evaluation of *Compareads*, a tool allowing us to know the similarity between two metagenomic dataset was done on a part of the *Tara* Oceans samples. This analysis showed the possibility to study the changes of genomic diversity of micro-planktonic communities in different oceans. Analysis of genomic variability along major ocean currents is then possible. The tool *COMMET* is an upgrade of *Compareads* which, combined with supercomputer makes it possible to carry out the comparisons of all the *Tara* Oceans samples. With a similar approach, the beta diversity distances between metagenomic samples made it possible to propose the first biogeography of the viral, bacterial and eukaryotic communities. It has been shown that ocean currents and physico-chemical variations have a different impact on the genomic organization of the microplanktonic communities, which would be more or less important depending on the time scale and the micro-organisms size fraction.